

EXHIBIT W

Spring 2022

**Washington Comprehensive
Assessment Program**

Technical Report

NOVEMBER 21, 2022

TABLE OF CONTENTS

1. Introduction.....	1
1.1 Overview.....	1
1.2 Background.....	1
1.3 Elements of the Washington Comprehensive Assessment Program, 2017–2022	3
1.3.1 State-Level Assessments in English Language Arts, Mathematics, and Science	3
1.3.2 Alternate Assessments.....	4
1.3.3 Other Washington State Assessments	4
1.4 Criterion-Referenced Tests	4
1.5 Appropriate Use of Test Scores	5
Summary	6
Glossary of Abbreviations and Acronyms	6
2. Test Development	9
2.1 Content Standards	9
2.2 Test Specifications.....	9
2.2.1 Test Specifications—Smarter Balanced Tests.....	9
2.2.2 Test Specifications—Washington Comprehensive Assessment of Science (WCAS)	10
2.3 Item Types	11
2.3.1 Item Types—Smarter Balanced Tests	11
2.3.2 Item Types—WCAS.....	12
2.4 Test Design	13
2.4.1 Test Design—Smarter Balanced	13
2.4.2 Test Design—WCAS.....	14
2.5 Test Construction.....	14
2.5.1 Test Construction—Smarter Balanced	14
2.5.2 Test Construction—WCAS.....	15
2.6 Spring 2022 Tests	15
2.6.1 Smarter Balanced Assessments.....	15
2.6.2 WCAS.....	18
Summary	20
3. Item Development.....	21
3.1 Item Development.....	21
3.1.1 Item Development—Smarter Balanced	21
3.1.2 Item Development— Washington Comprehensive Assessment of Science (WCAS)	22
3.2 Content Reviews and Bias and Sensitivity Reviews	23
3.2.1 Smarter Balanced Assessments.....	23
3.2.2 WCAS.....	23
3.3 Item Piloting.....	24
3.3.1 Item Piloting—Smarter Balanced	24
3.4 Field-Test Items Analysis.....	24

3.4.1	Field-Test Items Analysis—Smarter Balanced.....	24
3.4.2	Field-Test Items Analysis—WCAS.....	24
3.4.3	Classical Item Analysis Statistics.....	25
3.4.4	IRT Analysis.....	26
3.4.5	Differential Item Functioning (DIF).....	26
3.5	Item Data Review	30
3.5.1	Smarter Balanced Assessments.....	30
3.5.2	WCAS.....	30
	Summary	30
4.	Calibration and Equating	32
4.1	Item Response Theory (IRT)	32
4.2	Item Calibration	33
4.3	Smarter Balanced.....	33
4.4	WCAS.....	33
4.4.1	Post-Equating Procedure.....	34
4.4.2	Post-Equating: WCAS.....	35
4.5	Equating Result.....	35
	Summary	36
5.	Test Administration	37
5.1	Testing Windows	37
5.2	Test Administration	37
5.2.1	Administrative Roles.....	38
5.2.2	Online Administration.....	39
5.2.3	Paper-Pencil Test and Accommodated Paper Administration.....	41
5.2.4	Online Braille Test Administration.....	42
5.3	Training and Information for Test Coordinators and Administrators.....	43
5.3.1	Online Training.....	43
5.4	Test Security	45
5.4.1	Student-Level Testing Confidentiality.....	45
5.4.2	System Security	46
5.4.3	Security of the Testing Environment.....	47
5.4.4	Test Security Violations	49
5.5	Student Participation.....	49
5.5.1	Homeschooled Students	49
5.5.2	Exempt Students.....	49
5.6	Universal Tools, Designated Supports, and Accommodations.....	50
5.6.1	Universal Tools for All Students.....	51
5.6.2	Designated Supports.....	54
5.6.3	Accommodations.....	57
5.6.4	Spring 2022 Summary.....	60
5.7	Data Forensics Program.....	71

5.7.1 Changes in Student Performance.....	71
5.7.2 Test-Taking Time	72
5.7.3 Inconsistent Item Response Pattern (Person Fit)	72
5.7.4 Item-Response Change.....	73
5.7.5 Observed Online Test-Taking Time	73
5.7.6 Prevention and Recovery of Disruptions in Test Delivery System	75
5.7.7 High-Level System Architecture	76
5.7.8 Automated Backup and Recovery	78
5.7.9 Other Disruption Prevention and Recovery	78
Summary	79
6. Achievement-Level Setting	80
6.1 Overview.....	80
6.2 Smarter Balanced Assessments.....	80
6.3 WCAS	80
6.4 Cut Scores	81
Summary	83
7. Scoring.....	84
7.1 Estimating Student Ability Using Maximum Likelihood Estimation.....	84
7.2 Theta to Scale Score Transformation.....	85
7.3 Conversion Tables for WCAS	86
7.4 Lowest/Highest Obtainable Scores	87
7.5 Scoring All Correct and All Incorrect Cases	88
7.6 Rules for Calculating Strengths and Weaknesses for Reporting Categories	88
7.6.1 Claim Scores for Smarter Balanced Assessments.....	88
7.6.2 Reporting Area Proficiency Range for the WCAS.....	88
7.7 Attemptedness Rule	89
7.8 Target Scores for Smarter Balanced Assessments.....	90
7.8.1 Target Scores Relative to Student’s Overall Estimated Ability	90
7.8.2 Target Scores Relative to Proficiency Standard (Level 3 Cut).....	91
7.9 Handscoring	92
7.9.1 Rangefinding	93
7.9.2 Handscoring for Smarter Balanced Assessments	94
7.9.3 Handscoring for WCAS.....	98
7.9.4 Rater Agreements.....	99
7.10 Test Results.....	101
Summary	102
8. Reliability.....	103
8.1 Smarter Balanced Assessments.....	103

8.1.1 Marginal Reliability.....	103
8.1.2 Conditional Standard Error of Measurement.....	106
8.1.3 Classification Accuracy and Consistency.....	108
8.2 Washington Comprehensive Assessment of Science (WCAS)	112
8.2.1 Internal Consistency.....	112
8.2.2 Standard Error of Measurement	116
8.2.3 Conditional Standard Error of Measurement.....	117
8.2.4 Classification Accuracy and Consistency	119
Summary	120
9. Validity	121
9.1 Smarter Balanced Assessments.....	121
9.1.1 Evidence on Test Content.....	121
9.1.2 Evidence on Relations to Other Variables.....	127
9.1.3 Student Abilities vs. Test Difficulties	129
9.2 Washington Comprehensive Assessment of Science (WCAS)	132
9.2.1 Correlations Among Reporting Areas	132
9.2.2 Dimensionality Analysis.....	133
9.2.3 Evidence on Relations to Other Variables.....	136
9.2.4 Student Abilities vs. Test Difficulties	138
Summary	140
10. Reporting.....	141
10.1 Smarter Reporting System	141
10.1.1 Types of Score Reports In SRS.....	142
10.1.2 Group Reporting	143
10.1.3 Paper Report.....	143
10.2 SRS Report Pages	144
10.2.1 Custom Aggregate Reports	146
10.2.2 Assigned Student Groups Reports.....	152
10.3 Electronic Family Report.....	155
10.4 Interpretation of Reported Scores	160
10.4.1 Scale Score.....	160
10.4.2 Standard Error of Measurement.....	160
10.4.3 Achievement Level	161
10.4.4 Achievement Category for Claims/Reporting Areas.....	161
10.4.5 Achievement Category for Targets	162
10.4.6 Aggregated Score.....	163
10.5 Appropriate Uses for Scores and Reports.....	163
Summary	164
11. Quality Control	165
11.1 Quality Control in Test Configuration.....	165

11.1.1	Platform Review.....	166
11.1.2	User Acceptance Testing and Final Review	166
11.2	Quality Assurance in Document Processing.....	167
11.3	Quality Assurance in Data Preparation.....	167
11.3.1	Quality Assurance in Handscoring.....	167
11.3.2	Handscoring QA Monitoring Reports.....	168
11.3.3	Monitoring by OSPI.....	168
11.3.4	Identifying, Evaluating, and Informing the State on Alert Responses	168
11.4	Quality Assurance in Scoring	169
11.5	Quality Assurance in Reporting.....	170
11.5.1	Student Data Files Quality Assurance.....	170
11.5.2	Data Reports in SRS Quality Assurance	171
11.5.3	Family Report Quality Assurance.....	171
	Summary	172
	References.....	173

LIST OF TABLES

Table 2.1: Differences Between Spring 2019 and Spring 2022 ELA Test Blueprints	16
Table 2.2: Differences Between Spring 2019 and Spring 2022 Mathematics Test Blueprints	16
Table 2.3: Changes in Average Number of Unique Targets Assessed by Each Claim in ELA CAT Component	17
Table 2.4: Changes in Average Number of Unique Targets Assessed by Each Claim in Mathematics CAT Component.....	17
Table 2.5: Changes in Average Testing Times: ELA	18
Table 2.6: Changes in Average Testing Times: Mathematics	18
Table 2.7: Grade 5 WCAS Test Specification	19
Table 2.8: Grade 8 WCAS Test Specification	19
Table 2.9: Grade 11 WCAS Test Specification	19
Table 3.1: Classical Item Analyses Flagging Criteria, Pilot Items	26
Table 3.2: DIF Categories for 1-Point Items	29
Table 3.3: DIF Categories for Multiple Points Items	29
Table 3.4: WCAS Content Review with Data, Spring 2022 Field Test Administration Results ..	30
Table 4.1: WCAS Grades 5, 8, and 11 Post-Equating Sample Size and Percentage of Tested Student Population, 2022 Spring Administration.....	35
Table 4.2: Model Fit, WCAS, 2022 Administration.....	35
Table 5.1: Spring 2022 Testing Windows	37
Table 5.2: Summary of Tests and Testing Options in Spring 2022.....	38
Table 5.3: Responsibilities of Key Personnel 2021–22	38
Table 5.4: Number of Students Who Took Paper-Pencil ELA and Math Tests in Spring 2022 Administration.....	41
Table 5.5: Number of Students Who Took the Accommodated Paper WCAS in Spring 2022 Administration.....	42
Table 5.6: Summary of Smarter Balanced and WCAS Tools, Supports, and Accommodations ..	51
Table 5.7: Total Students with Allowed Embedded Designated Supports–ELA	60
Table 5.8: Total Students with Allowed Non-Embedded Designated Supports–ELA.....	61
Table 5.9: Total Students with Allowed Embedded Accommodations–ELA	62
Table 5.10: Total Students with Allowed Non-Embedded Accommodations–ELA.....	63
Table 5.11 Total Students with Allowed Embedded Designated Supports–Mathematics	64
Table 5.12: Total Students with Allowed Non-Embedded Designated Supports–Mathematics ...	65
Table 5.13: Total Students with Allowed Embedded Accommodations–Mathematics	66

Table 5.14: Total Students with Allowed Non-Embedded Accommodations—Mathematics	67
Table 5.15: Total Students with Allowed Embedded Designated Supports—WCAS	68
Table 5.16: Total Students with Allowed Non-Embedded Designated Supports—WCAS	69
Table 5.17: Total Students with Allowed Embedded Accommodations—WCAS	70
Table 5.18: Total Students with Allowed Non-Embedded Accommodations—WCAS	70
Table 5.19: Smarter Balanced ELA Test-Taking Time, Spring 2022 Administration	74
Table 5.20: Smarter Balanced Mathematics Test-Taking Time, Spring 2022 Administration	75
Table 5.21: WCAS Test-Taking Time, Spring 2022 Administration	75
Table 6.1: WCAP Cut Scores—Smarter Balanced Assessments	82
Table 6.2: WCAP Cut Scores—WCAS	83
Table 7.1: Scaling Constants on the Reporting Metric	86
Table 7.2: Scale Score Cuts—Smarter Balanced	86
Table 7.3: Scale Score Cuts—WCAS	86
Table 7.4: Lowest and Highest Obtainable Scores—Smarter Balanced	87
Table 7.5: Lowest and Highest Obtainable Scores—WCAS	87
Table 7.6: Reporting Area Level Summary for WCAS, Form A	89
Table 7.7: Number of Hand-Scored Items in 2021–22 Smarter Balanced Summative Item Pool, by Grade and Subject	94
Table 7.8: Interrater Agreement—ELA Smarter Balanced for Full-Write Items	100
Table 7.9: Interrater Agreement—ELA Smarter Balanced for Short-Answer Items	100
Table 7.10: Interrater Agreement—Mathematics Smarter Balanced	101
Table 7.11: Interrater Agreement—WCAS	101
Table 8.1: Marginal Reliability for Smarter Balanced ELA and Mathematics	104
Table 8.2: Marginal Reliability Coefficients for Overall and by Student Group: ELA	105
Table 8.3: Marginal Reliability Coefficients for Overall and by Student Group: Mathematics	105
Table 8.6: Smarter Balanced Classification Accuracy and Consistency	111
Table 8.7: Grade 5 WCAS Form A Test Reliability Estimates	113
Table 8.8: Grade 8 WCAS Form A Test Reliability Estimates	114
Table 8.9: Grade 11 WCAS Form A Test Reliability Estimates	115
Table 8.10: Reporting Area Reliabilities by Test, WCAS, 2022 Administration	116
Table 8.11: Classification Consistency and Accuracy	119
Table 9.1: Percentage of ELA Delivered Tests Meeting Blueprint Requirements for Each Claim and Number of Passages Administered (Grades 3–5)	122

Table 9.2: ELA Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Number of Passages Administered (Grades 6–8, HS).....	123
Table 9.3: Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 3–5).....	124
Table 9.4: Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 6–8).....	125
Table 9.5: Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets (HS)	126
Table 9.6: Average and Range of the Number of Unique Targets Assessed within Each Claim, Across All Delivered Tests.....	126
Table 9.7: Percentage of Students in Each Smarter Balanced Achievement Level by Course Grade in 2015, ELA	128
Table 9.8: Percentage of Students in Each Smarter Balanced Achievement Level by Course Grade in 2015, Mathematics	129
Table 9.9: Intercorrelations, WCAS 2022 Administration	132
Table 9.10: Percentage of Students in Each WCAS Achievement Level by Science Course Grade in 2018	137
Table 10.1: Permissions and Reports Available to TIDE Users by Role	142
Table 10.2: Types of Student Groups	143
Table 11.1: Overview of QA Reports	170

LIST OF FIGURES AND EXHIBITS

Figure 8.1: CSEM for Smarter Balanced ELA	107
Figure 8.2: CSEM for Smarter Balanced Mathematics	108
Figure 8.3: TIF and CSEM for the WCAS	117
Figure 9.1: Student Ability–Item Difficulty Distribution for ELA.....	130
Figure 9.2: Student Ability–Item Difficulty Distribution for Mathematics.....	131
Figure 9.3: Scree Plots for WCAS.....	134
Figure 9.4: WCAS Student Ability—Item Difficulty Distributions.....	138
Exhibit 10.1: Home Page—TA-user Level.....	145
Exhibit 10.2: Home Page—District Level	146
Exhibit 10.3: Custom Aggregate Reports	147
Exhibit 10.4: Subject Detail Page for ELA by Gender—District Level.....	148
Exhibit 10.5: Grade 8 WCAS Reporting Area Report.....	149
Exhibit 10.6: Target Report for Math Grade 5—Custom Aggregate Level	150
Exhibit 10.7: Target Report for Grade 3 ELA—Teacher Level	151
Exhibit 10.8: Student Group Results for Grade 11 WCAS	152
Exhibit 10.9: Student Results for Grade 11 WCAS.....	153
Exhibit 10.10: Student Test History Report.....	154
Exhibit 10.11: Student ISR for Grade 11 WCAS	155

Exhibit 10.12: Smarter Balanced ELA Sample Electronic Family Score Report.....	157
Exhibit 10.13: WCAS Sample Electronic Family Report	158

LIST OF APPENDICES

Appendix A: Classical Item Analysis Results and DIF Results for State-Specific Tests
Appendix B: IRT Results for State-Specific Tests
Appendix C: Conversion Tables for State-Specific Tests
Appendix D: Scale Score Summary for Accountability
Appendix E: Percentage of Students by Achievement Level for Accountability
Appendix F: Scale Score Summary for Graduation
Appendix G: Percentage of Students by Achievement Level for Graduation
Appendix H: Historical Data

1. INTRODUCTION

1.1 OVERVIEW

The Washington Comprehensive Assessment Program (WCAP) consists of multiple assessments spanning different grades and content areas. The 2021–22 assessments included Smarter Balanced English language arts (ELA), Smarter Balanced mathematics, and the state-specific Washington Comprehensive Assessment of Science (WCAS). The scope and subject of this report are limited to the technical characteristics of the regular state-level assessments, administered to the majority of students at specified grade levels. This technical report documents the planning, development, delivery, and analyses of the summative spring 2022 WCAP tests.

Chapter 1 provides an overview of the tests. Chapters 2 and 3 describe the test and item development. Psychometric analyses are provided in Chapters 3, 4, 8, and 9, including item analyses, calibration and equating, test reliability, and test validity. An overview of the 2022 test administration is described in Chapter 5. Chapter 6 describes the performance standards and how these standards were established. Test score summaries are provided in Chapter 7. Score reporting is documented in Chapter 10. Chapter 11 describes the quality control procedures used.

Washington’s assessment system is designed to fulfill all federal census-testing requirements. In addition, the high school ELA and mathematics assessments can also be used, per state legislation, as one of multiple graduation pathways. Meeting a graduation pathway is just one of many requirements of a student to earn a high school diploma.

Student performance on assessments is summarized in tables throughout this report, and the student population included in those tables varies based on the purpose of the assessment. Accountability tests (Smarter Balanced ELA and mathematics grades 3–8 and high school, and WCAS grades 5, 8, and 11) are summarized by grade level of the test by including every student who receives a student data file (SDF). Several tables in the appendices summarize the data for high school students who took the ELA or mathematics tests for Graduation Pathway purposes.

1.2 BACKGROUND

In 1993, Washington embarked on the development of a comprehensive change effort with the primary goal to improve teaching and learning in Washington schools. Created by the state legislature in 1993, the Commission on Student Learning was charged with three important tasks to support this effort:

1. Establish Essential Academic Learning Requirements (EALRs) as the basis for state Learning Standards that describe what all students should know and be able to do in five content areas: reading, writing, communication, mathematics, and science. Technology was added by the state Legislature in 2011.
2. Develop an assessment system to measure student progress toward achieving the EALRs at three grade levels.
3. Recommend an accountability system that recognizes and rewards successful schools and provides support and assistance to less successful schools.

The EALRs and state Learning Standards in reading, writing, communications, and mathematics were adopted in 1995 and revised in 1997, while those for science were adopted in 1996 and revised in 1997. The mathematics and science standards were revised and adopted again in 2008 and 2009, respectively. In 2011, the state-developed reading, writing, communications, and mathematics standards were replaced by adoption of the *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects* and for *Mathematics* (CCSS). In 2013, the state-developed science standards were replaced by adoption of the *Next Generation Science Standards* (NGSS). Upon adoption, the CCSS and NGSS were rebranded as the “Washington State K–12 [content area] Learning Standards” and are referred to as “the standards.” (See <https://www.k12.wa.us/student-success/learning-standards-instructional-materials> for links to current Learning Standards in all subject areas.) In this document the phrase “Learning Standards” will be used to refer to these academic content standards.

The assessments for reading, writing, and mathematics at grade 4 were operational in 1997, with those for grade 7 operational in spring 1998. The grade 10 assessments in these content areas were pilot-tested in spring 1998 and operational in spring 1999. Participation in the grade 4 assessment became mandatory for all public schools in spring 1998. Participation in the grade 7 and 10 assessments was voluntary until spring 2000. Participation in the grades 3, 5, 6, and 8 reading and mathematics assessments was voluntary in 2004 and 2005 and became mandatory in spring 2006.

Science was implemented as a voluntary operational administration for grades 8 and 10 in spring 2003 and became mandatory in 2004. Grade 5 science was a voluntary operational administration in spring 2004 with mandatory implementation in spring 2005.

In 2011, new mathematics End-of-Course (EOC) tests in Algebra 1/Integrated Mathematics 1 (EOC 1) and Geometry/Integrated Mathematics 2 (EOC 2) were introduced to replace the mathematics High School Proficiency Exam (HSPE). In spring 2012, a new Biology EOC test was introduced, replacing the science HSPE test. These EOC tests were taken by students enrolled in the course regardless of their enrolled grade level.

Following the adoption of the CCSS as the Learning Standards in 2011, the WCAP system adopted the Smarter Balanced ELA and mathematics assessments in 2015, and has given those tests since then, as described in the following paragraph. In spring 2016, the last reading and writing HSPE tests were administered for the Class of 2016 and earlier. In spring 2018, the last mathematics EOC exams were administered for the Class of 2018 and earlier.

The Smarter Balanced assessments in ELA and mathematics were administered for the first time in spring 2015 to students in grades 3–8 and 11 in all Washington public elementary and secondary schools. In July 2017, the Washington legislature moved the high school testing grade for ELA and mathematics from grade 11 to grade 10 starting with the 2018 administration. Smarter Balanced then established cut scores for students testing in grade 10, which Washington adopted and will be used in this report. For ELA, the test blueprints developed by Smarter Balanced in 2015 were used through 2018. In the 2019 test administration, Smarter Balanced updated the ELA summative blueprint, shortening the overall test length by three to four items. There was no change in the mathematics test blueprints.

Following the adoption of the NGSS as the Learning Standards in 2013, Washington developed a new test based on those Learning Standards and first administered the Washington Comprehensive

Assessment of Science (WCAS) in 2018 to students in grades 5, 8, and 11, and has given those tests since then. In spring 2017, the last Biology EOC exam was administered.

Due to the disruptions caused by the COVID-19 pandemic, there was no accountability testing during the 2019–20 school year. For the 2020–21 school year, testing was delayed from spring 2021 to fall 2021 as Washington submitted and was granted an accountability, school identification, and related reporting requirements waiver from the U.S. Department of Education for the 2020–21 school year. Spring testing resumed during the 2021–22 school year.

For this spring 2022 administration, Washington adopted the Smarter Balanced adjusted blueprint for both math and ELA. The adjusted blueprints are provided in Section 2.6. The WCAS blueprint used in spring 2022 was the same as used since 2018.

1.3 ELEMENTS OF THE WASHINGTON COMPREHENSIVE ASSESSMENT PROGRAM, 2017–2022

Washington’s assessment program has several major components, including state-level summative assessments in ELA, mathematics, and science (including alternate assessments in these content areas); English language proficiency assessments (general and alternate); the Washington Kindergarten Inventory of Developing Skills (WaKIDS); the Smarter Balanced interim assessments in ELA and mathematics; and classroom-based assessments in subjects like the arts, social studies, and technology.

1.3.1 State-Level Assessments in English Language Arts, Mathematics, and Science

Washington’s statewide accountability assessments require students to select and construct responses to demonstrate their knowledge, skills, and understanding in each of the Learning Standards—from multiple-choice, technology-enhanced (e.g., table match, drag-and-drop, and hot-text items), and short-answer items to essays and problem-solving tasks. Student-, school-, district-, and state-level scores are reported for the operational assessments. The WCAS operational test forms in science are fixed-form, meaning that all students taking each assessment are expected to respond to the same items, under the same conditions, and during the same testing window during the school year. In Smarter Balanced ELA and mathematics, a portion of the tests use computer-adaptive testing (CAT), so students see different items depending on their answers to previous items on the test. The other portion of the ELA and mathematics test is a Performance Task (PT) which are distributed to students at random from a pool of available PTs.

All of the WCAP assessments are untimed; that is, students may have as much time as they reasonably need to complete their work. Guidelines for providing accommodations to students with special needs have been developed to encourage the inclusion of as many students as possible in the general assessments. Special needs students include those in special education programs, multilingual learners (ML), migrant students, and highly capable students. A broad range of accommodations allows nearly all students access to some or all parts of the assessment. Details can be found in the *Guidelines on Tools, Supports, & Accommodations for State Assessments* (<https://wa.portal.cambiumast.com/resources/wa-guidelines/guidelines-on-tools-supports-and-accommodations-for-state-assessments>).

Classroom teachers and curriculum specialists throughout Washington assisted with the development of items for all assessments. For the WCAS, content work groups were created at each grade level. Working with content and assessment specialists, these work groups helped to define the test and item specifications consistent with the science learning standards, participated in item writing, reviewed all items prior to field testing, and provided final review and recommendations to approve selected items after field testing. A separate Bias and Sensitivity committee, composed of individuals who reflect Washington's diversity, also conducted a sensitivity review of all items for words or content that might be potentially offensive to students or parents, or might disadvantage some students for reasons unrelated to the assessed skill or concept. Teachers from around the state also participated in various activities related to the development of the ELA and mathematics Smarter Balanced assessments used in grades 3–8 and high school. Chapter 2 of this report provides further details about the test development processes.

1.3.2 Alternate Assessments

Students with disabilities are expected to take the regular WCAP tests, with or without necessary accommodations, unless the Individualized Education Program (IEP) team determines a student is unable to participate in one or more content areas, even with accommodations. In these instances, the IEP team may elect to administer the Washington Access to Instruction and Measurement (WA-AIM) assessment. The WA-AIM was designed for students with significant cognitive disabilities, a very small percentage of the total school population. Information on WA-AIM can be found at <https://www.k12.wa.us/student-success/testing/state-testing-overview/assessment-students-cognitive-disabilities-wa-aim>.

1.3.3 Other Washington State Assessments

This report does not include information about the other assessments used in Washington. Visit the OSPI website at <https://www.k12.wa.us/> to learn more about the following: English language proficiency assessments (general and alternate); the Washington Kindergarten Inventory of Developing Skills (WaKIDS); the Smarter Balanced interim assessments in ELA and mathematics; and classroom-based assessments in other subjects like the arts, social studies, and technology.

1.4 CRITERION-REFERENCED TESTS

The purpose of an achievement test (or standards-based test) is to determine how well a student has learned important concepts and skills and how schools and districts are performing over time. Test scores are used to make inferences in terms of the domain of behavior that students exhibit (Crocker & Algina, 1986, p. 192). When a student's achievement is compared to a targeted level of performance (e.g., the cut score for proficient), this is considered to be a criterion-referenced (or standards-based) interpretation.

The state-level assessments are criterion-referenced tests. Student performance should be interpreted in terms of how well students have achieved the Learning Standards as measured by the test.

Criterion-referenced tests can measure the degree to which students have achieved a desired set of learning targets, conceptual understandings, and skills that are at grade level or developmentally

appropriate. They can also be helpful in make decisions about the success or the usefulness of an instructional or administrative program. Much care and attention ensure that the items on the test represent only the desired content and that there are sufficient numbers of items for each learning target to make reliable statements about students' degree of achievement/behavior related to that content domain. When a standard is defined on a criterion-referenced test, examinee scores are compared to the standard to make inferences about whether students have attained the desired level of achievement (i.e., has the student mastered the material taught?).

To assess all of the desired concepts and skills in a domain would require inordinate testing time. Well-designed state or national achievement tests always include samples from the domain of desired concepts and skills. Therefore, when state or national achievement tests are used, a student's performance on the sample of items in the test is an estimate of how the student would perform in the domain if it were more broadly defined. To obtain a broader measure of student achievement in a specific domain, it is necessary to use more than results from state testing. Results of state assessments should be used in conjunction with additional, local measures to inform state and local policies, practices, and decisions. District and classroom assessments, teacher observations, projects, and other educational activities that inform teachers' day-to-day instructional decisions are all necessary to include in conversations about interpreting and using state achievement test data.

1.5 APPROPRIATE USE OF TEST SCORES

The primary purpose of WCAP results are calculating school and district accountability, to meet the requirements of the federal Every Student Succeeds Act (ESSA) of 2015. WCAP tests also give local and state policy makers information to support schools. State and federal accountability for 2021–22 was based on Smarter Balanced and WA-AIM ELA and mathematics participation and scores in grades 3–8 and high school and on WCAS and WA-AIM science participation in grades 5, 8, and 11. The percentage of students meeting standard and the percentage of students participating in the tests are factored into these calculations.

Once tests are administered, scale scores (total test) are generated for each content area test as well as reporting area scores for the WCAS. Because of the use of the Smarter Balanced adjusted blueprint, claim results were not calculated or reported for spring 2022. The performance data are reported at the individual student, school, district, and state levels. The total test scale score is used to classify students into achievement levels in terms of their level of knowledge and skill in the subject area. Additionally, reporting area scores provide more specificity about a student's achievement in each of several specific knowledge or skill areas covered by the WCAS tests. For the WCAS, the percentages of raw score points earned by the student on each reporting area are reported to provide teachers, parents, and students more detailed information about students' learning and performance on those areas of the test.

The information in these reports (scale score, achievement levels, and reporting area score indicators) can be used with local information and evidence about student learning to help with school, district, and state curriculum planning and instructional decisions.

While school and district scores may be useful in curriculum and instructional planning, it is important to exercise extreme caution when interpreting individual reports. The items included on WCAP tests are samples from a larger content domain. Scores from one test given on a single

occasion should never be used in isolation to make important decisions about students' course or program placement, the type of instruction they receive, or retention at a given grade level in school. It is important that multiple sources of information be used when making decisions about individuals, and individual scores on WCAP tests can be included along with classroom-based and other local evidence of student learning (e.g., scores from district testing programs) to inform those decisions. Multiple individuals who are familiar with the student's progress and achievement—including parents, teachers, school counselors, school psychologists, specialist teachers, and the students themselves—should be brought together to make such decisions collaboratively.

Additionally, when comparing results for the WCAP tests, one is limited to comparing results only within the same content area and grade level. A person may compare results for the same content area and grade, within a school, between schools, between a school and its district or the state, or between years. For example, results can be compared for grade 5 science WCAS in 2018 and grade 5 science WCAS in 2019. Additionally, results from the 2019 WCAS are not comparable to those from the previous science test, last administered in 2017. In 2015, a new test was used for both ELA and mathematics in grades 3–8 and high school. Therefore, the 2019 results are not comparable to the 2014 results in mathematics or in reading and writing, but they are comparable to the 2015 results. There are no 2020 scores in mathematics, ELA, or WCAS due to the COVID-19 pandemic. Results from the fall 2021 shortened ELA, mathematics, WCAS tests should not be compared to any previous or future results due the differences in timing of testing, which students took which tests, and design of the tests used.

SUMMARY

Washington's assessment program has several components, but only the summative accountability tests are examined in this report. This report focuses on the spring 2022 administration of the Smarter Balanced assessments and the state-specific WCAS. Washington is a member of the Smarter Balanced Assessment Consortium and offers Smarter Balanced tests in its state-level ELA and mathematics assessments for grades 3–8 and high school. The WCAS grades 5, 8, and 11 assessments are referred to in this document as state-specific exams, separate from Smarter Balanced assessments. Further details about Smarter Balanced assessments are available at <http://www.smarterbalanced.org/>.

The Office of Superintendent of Public Instruction (OSPI) is committed to developing an instructionally relevant, accessible, evidence-based assessment system. Smarter Balanced and the WCAS are criterion-based, developed from the Learning Standards. Teachers and other professionals who provide pre-service and in-service training to teachers should be thoroughly familiar with the Learning Standards and the assessments that measure them.

GLOSSARY OF ABBREVIATIONS AND ACRONYMS

A glossary of abbreviations and acronyms commonly used in this technical report is given below for reference.

Abbreviation or Term	Meaning
ASL	American Sign Language

Abbreviation or Term	Meaning
CAI	Cambium Assessment, Inc.
CAT	computer-adaptive test(ing)
CBT	computer-based test(ing)
CCC	Crosscutting Concept in NGSS
CSEM	conditional standard error of measurement
DCI	Disciplinary Core Idea in NGSS
DIF	differential item functioning
DOR	Database of Record
ELA	English language arts
ESSA	Every Student Succeeds Act
Form	A compilation of test items and/or tasks that comprise the full test.
GPCM	generalized partial credit model
HOSS	highest obtainable scale score
HOT	highest obtainable theta (score)
IEP	Individualized Education Program
IRT	item response theory
JAWS	Job Access with Speech
LOSS	lowest obtainable scale score
LOT	lowest obtainable theta (score)
MC	multiple-choice item, worth 1 point
MI	Measurement Incorporated
ML	Multilingual learner
MLE	maximum likelihood estimate
MS	multiple select item
NGSS	Next Generation Science Standards
OSPI	Office of Superintendent of Public Instruction
ORG	Organization
PCM	partial credit model
PE	performance expectation
PPT	paper-pencil testing
PT	performance task
Purp	Purpose
QA	quality assurance
SA	short-answer item
SBAC	Smarter Balanced Assessment Consortium

Abbreviation or Term	Meaning
SC	School Coordinator
SD	standard deviation
SE	standard error
SEM	standard error of measurement
SEP	Science and Engineering Practice from NGSS
TA	Test Administrator
TAM	<i>Test Administration Manual</i>
TDS	Test Delivery System
TEI	technology-enhanced item
TEST	Questions or tasks designed to measure students' performance on specific academic content standards.
TIDE	Test Information Distribution Engine
TIF	test information function
ITS	Item Tracking System
UAT	user acceptance testing
VIPP	Variable-Data Intelligent PostScript Printware
WA-AIM	Washington Access to Instruction and Measurement
WCAP	Washington Comprehensive Assessment Program
WCAS	Washington Comprehensive Assessment of Science

2. TEST DEVELOPMENT

2.1 CONTENT STANDARDS

The content of Washington Comprehensive Assessment Program (WCAP) tests is derived from the Washington State Learning Standards (see <https://www.k12.wa.us/student-success/learning-standards-instructional-materials> for links to the Learning Standards in all subject areas). These Learning Standards define what Washington students should know and be able to do by the end of grades 3–8 and 10 in English language arts (ELA) and mathematics, and by the end of grades 5, 8, and 11 in science. WCAP tests measure the Learning Standards for ELA and mathematics in grades 3–8 and high school, and science in grades 5, 8, and 11. Mathematics and ELA tests in grades 3–8 and 10 measure the Learning Standards (Common Core State Standards) for English Language Arts and Mathematics adopted in 2011; science tests in grades 5, 8, and 11 measure the *Washington State 2013 K–12 Science Learning Standards* which are the *Next Generation Science Standards (NGSS)* adopted in 2013. In this document the phrase “Learning Standards” will be used to refer to these academic content standards.

2.2 TEST SPECIFICATIONS

For any new tests, specifications must be developed, describing common agreement on the meaning and interpretation of the Learning Standards and identifying which Learning Standards could be assessed on a statewide test. It is important that the vendor, educator work groups, and staff at the Office of Superintendent of Public Instruction (OSPI) are in agreement not only on what students are expected to know and be able to do but also on how these skills and knowledge will be assessed. Washington educators and OSPI content staff participate in this process for all summative tests in Washington.

2.2.1 Test Specifications—Smarter Balanced Tests

Washington educators and OSPI content staff participated in the test specification development process through activities of the Smarter Balanced Assessment Consortium (SBAC) for the Learning Standards in ELA and mathematics.

Among the guiding principles for the Consortium’s work were the following ideals, as described in the SBAC End of Grant Report (<https://portal.smarterbalanced.org/library/en/v1.0/end-of-grant-report.pdf>):

- Assessments are grounded in a thoughtful, standards-based curriculum and are managed as part of an integrated system of standards, curriculum, assessment, instruction, and teacher development;
- Assessments produce evidence of student performance on challenging tasks that evaluate student achievement on the Common Core State Standards;
- Educators are integrally involved in the development and scoring of assessments;
- The development and implementation of the assessment system is a state-led effort with a transparent and inclusive governance structure;

- Assessment, reporting, and accountability systems provide useful information on multiple measures that is educative for all stakeholders; and
- Design and implementation strategies adhere to established professional standards.

These ideals provide the foundation for a comprehensive assessment system that is developed with attention to technical rigor. The SBAC technical report provides a detailed description of all test development procedures (<https://validity.smarterbalanced.org/reports-and-specifications/>).

Test specifications define the kinds and numbers of items on the assessment, the blueprint and physical layout of the assessment, the amount of time to be devoted to each content area, and the scores to be generated once the test is administered. It is important at this stage to define the goals of the assessment and the ways in which the results will be used to ensure that the structure of the test will support the intended uses. The test specifications are the building blocks to developing equivalent test forms in subsequent years and to creating new items to supplement the item pool. The final test specifications document contains some or all of the following topics:

- Purpose of the assessment
- Claims or strands
- Item types
- General considerations of testing time and style
- Test scoring
- Distribution of test items by item type

Smarter Balanced test blueprints and item and task specification documents are available on the Smarter Balanced Development and Design website (<https://contentexplorer.smarterbalanced.org/test-development/>). In spring 2022, Washington used the adjusted blueprints for both math and ELA.

2.2.2 Test Specifications—Washington Comprehensive Assessment of Science (WCAS)

OSPI content staff led Washington educators through the process of developing the science test design and item specifications based on the NGSS beginning in 2015 to guide development of the WCAS.

Among the guiding principles for this process were the following objectives:

- Design an assessment that reflects how science content is taught and tested in the classroom.
- Use Washington educators in assessment development.
- Develop high-quality item clusters and stand-alone items that achieve alignment with the Science and Engineering Practices (SEPs), Disciplinary Core Ideas (DCIs), and Crosscutting Concepts (CCCs) represented in a performance expectation (PE) or PE

bundle and attend to the three-dimensional nature of the standards.

- Design an assessment that allows for valid and reliable inferences to be drawn from the results.
- Design an assessment that ensures fair and accurate assessment of students in special populations.

The most recent Test Design and Item Specifications documents for the WCAS were published in August 2019 and updated as recently as January 2021 on the OSPI webpage (<https://www.k12.wa.us/student-success/testing/state-testing/washington-comprehensive-assessment-science/wcas-educator-resources>). The specifications contain the following topics:

- Purpose of the assessment
- Structure of the test
- Item types
- Test design (including testing times and test blueprint)
- Overview of the learning standards
- Item specifications

2.3 ITEM TYPES

2.3.1 Item Types—Smarter Balanced Tests

The Smarter Balanced math and ELA tests are comprised of a variety of item types and items at different depths of knowledge.

Smarter Balanced tests use multiple-choice, multiple select, equation/numeric, table input, matching, hot text, short text, essay, and technology-enhanced items. Technology-enhanced items (TEIs) are present in both the ELA and mathematics assessments. All TEIs included on Smarter Balanced summative assessments, whether they are part of the CAT portion of the assessment or embedded within a performance task, were developed in accordance with an established TEI template. These templates, which are applicable across grade levels and content areas, describe a single interaction, response data collected as a result of that interaction, and the logic applied to score the response data. Across all of these item/task types, technology-enhanced items take advantage of technological innovations to allow students to demonstrate their knowledge and skills in ways that are not possible with traditional item types.

Smarter Balanced established cognitive complexity as a specific consideration in item development by adopting a Cognitive Rigor Matrix that integrates Bloom's (revised) Taxonomy of Educational Objectives and Webb's Depth of Knowledge Levels. The Smarter Balanced General Item Specifications document is accompanied by an extensive set of accompanying grade-level and content-specific documents that provide detailed requirements for writing five types of items and tasks designed to measure the full range of cognitive complexity of the standards: selected-response items, constructed-response items, extended-response items, technology-enhanced items, and performance tasks. More detailed information can be found in the

Mathematics Item Specifications for grades 3–5, 6–8, and high school, respectively; Mathematics Performance Task Specifications; Sample ELA Item Specifications for specific grades, claims, and targets; Sample ELA Performance Task Specifications; ELA Stimulus Specifications; Technology-Enhanced Item Guidelines; and the Smarter Balanced General Item Specifications. These documents can be found on the Smarter Balanced Test Development and Design website (<https://contentexplorer.smarterbalanced.org/test-development>).

All item types, when carefully constructed, allow for inclusion of challenging content and have the capability to measure higher-order thinking skills. Selected-response items allow students to demonstrate complex thinking skills such as formulating comparisons or contrasts or identifying causes and effects. Constructed-response and extended-response items often allow for greater complexity by requiring students to supply a response rather than selecting from a list of possible responses. Performance tasks provide a measure of the student’s performance in integrating knowledge and skills across multiple content standards and better assess capacities such as depth of understanding, research skills, and complex analysis than stand-alone items found on the CAT.

2.3.2 Item Types–WCAS

The WCAS contain multiple item types.

- In edit-task-inline-choice (ETC) items, students select words, numbers, or phrases from drop-down lists to complete a statement. The number of drop-down lists in an item will typically be between two and four. Students must answer all parts correctly for a maximum score of 1 point (scored 0 or 1). ETC items are machine-scored.
- In grid or graphic gap match items, students place arrows, symbols, labels, or other graphical elements onto a background graphic, or interact with and construct simple graphs. Grid items are worth a maximum score of 1 point (scored 0 or 1) and are machine-scored.
- In multiple-choice items, students select the one best answer from among at least four choices. In multiple-select items, students choose a specified number of correct responses from a list of choices. Both multiple-choice and multiple-select items are worth a maximum score of 1 point (scored 0 or 1) and are machine-scored.
- In short-answer items, students produce their own response based on a specific task statement. Short-answer items are worth a maximum score of 1 point (scored 0 or 1) or 2 points (scored 0, 1, or 2) and are hand-scored by well-trained professional scorers using a detailed rubric and training set.
- In simulations, students use a simulation to control an investigation and/or generate data. The data can be scored directly or used to answer related questions, or both. Simulations vary in their interaction, design, and scoring. Some simulations are not scored and are used by students to generate information to use to answer other items. Simulations that are scored are worth a maximum score of 1 point (scored 0 or 1) or 2 points (scored 0, 1, or 2) and are either machine-scored or hand-scored by well-trained professional scorers using a detailed rubric.

- In table input items, students complete a table by typing numeric responses into the cells of the table using the keyboard. Table input items are worth a maximum score of 1 point (scored 0 or 1) and are machine-scored.
- In table match items, students check boxes within the cells of a table to make identifications, classifications, or predictions. Students must answer all parts correctly for a maximum score of 1 point (scored 0 or 1). Table match items are machine-scored.
- In hot text items, student move statements into the cells of a table to describe an ordered sequence. Students must answer all parts correctly for a maximum score of 1 point (scored 0 or 1). Hot text items are machine-scored.

The WCAS also includes multipart items. See pages 5-7 of the Test Design and Item Specifications documents on the OSPI webpage (<https://www.k12.wa.us/student-success/testing/state-testing/washington-comprehensive-assessment-science/wcas-educator-resources>) for details.

Chapter 7 provides further detail about the handscoring process and results for the different subject area tests.

2.4 TEST DESIGN

2.4.1 Test Design—Smarter Balanced

Smarter Balanced summative assessments are technology-based and include a computer-adaptive test (CAT) component along with a performance task component. The final blueprints for the Smarter Balanced summative assessments, available at <https://contentexplorer.smarterbalanced.org/test-development/>, leverage technology both to provide innovative ways for students to access test content and to measure student performance more reliably and precisely through the use of a CAT component. Use of a CAT component necessitates an exceptionally large and robust item pool. Therefore, Smarter Balanced paid particular attention to item characteristics, beginning with pilot testing of items within the pool. Summary statistics for the CAT portion of the assessments from the pilot test are presented in Smarter Balanced's technical reports (<https://validity.smarterbalanced.org/reports-and-specifications/>). Through use of quantitative and qualitative information from item development workshops, and information from the pilot test, available items were inventoried and a field-test plan was created to yield an adequate item pool. Field-test data were analyzed using both classical and item response theory (IRT) statistics, as well as content and scoring decisions, to create the final item pool.

This quality item pool, along with the test blueprint, provides the basis for the Smarter Balanced CAT algorithm to provide a precise and efficient measure of student performance. For each student's test, the blueprints specify the proportions of items in each area, but not the order in which the student will encounter them. The Smarter Balanced blueprints specify a range of items to be administered in each claim for each assessment, with a collection of constraint sets. For each student's test, the CAT adaptive algorithm optimizes item selection in order to meet blueprint specifications, while also targeting test information to student ability to improve the precision of the estimate of student achievement.

2.4.2 Test Design—WCAS

The WCAS is a technology-based fixed-form test, meaning that items were developed to be delivered in an online test and all students in the grade level receive the same test items. The WCAS is composed of item clusters and stand-alone items aligned with the performance expectations (PEs) in the Learning Standards. Advisory groups composed of national education experts, science assessment experts, and science educators recommend the item cluster structure for large-scale assessment of the standards because item clusters involve significant interaction with stimulus materials leading to a demonstration of the students' application of knowledge and skills. Stand-alone items increase PE coverage that can be achieved in a single test administration.

Item clusters that assess a PE bundle make up the core of the WCAS. A PE bundle is generally two or three related PEs that are used to explain or make sense of a scientific phenomenon or a design problem. A phenomenon gives an item cluster conceptual coherence. The items within an item cluster are interconnected and focused on the given phenomenon. Items are also structured to support a student's progression through the cluster.

Students must make sense of the phenomenon or a design problem for an item cluster by using the Science and Engineering Practices (SEPs), Disciplinary Core Ideas (DCI), and Crosscutting Concepts (CCCs) represented in the PE bundle. PE bundles are often within a single domain but may include PEs from different domains. PE bundles sometimes share a similar practice or crosscutting concept or may include multiple practices or crosscutting concepts. Each item within the cluster will align with two or three dimensions (2-D, 3-D) from one or more of the PEs in the bundle. Achieving as full coverage as possible requires developing items that target a variety of the dimensions represented in the PE bundle. In all cases, item clusters achieve full coverage of the dimensions of each PE within a PE bundle.

The final blueprints for the WCAS are available in each grade level Test Design and Item Specifications document on the OSPI webpage (<https://www.k12.wa.us/student-success/testing/state-testing/washington-comprehensive-assessment-science/wcas-educator-resources>).

2.5 TEST CONSTRUCTION

2.5.1 Test Construction—Smarter Balanced

The Smarter Balanced adaptive test algorithm selects items until a defined percentage of the test has been administered, sampling items to meet item selection criteria. Item selection occurs in two discrete stages: 1) blueprint satisfaction, and 2) match to ability. A decision point is reached with a substantial portion of content covered. At the decision point, the distance of the estimated score from the college content readiness cut score (Level 3) is evaluated. From the pool, the algorithm selects subsequent items with the best content and measurement characteristics. The algorithm delivers the remainder of the blueprint until termination of the test once all test constraints have been met. If the following conditions occurs, the item pool will expand to include items from adjacent grades that address content in the target test grade: 1) on-grade content coverage requirements have been met, such that over two-thirds of the CAT session has been administered; 2) the estimate of performance is clearly far below or far above the proficiency score; and 3) items in the expanded pool will better satisfy content and measurement requirements. Additional

information is available in the Smarter Balanced technical report available on the Smarter Balanced Reports and Specifications website (<https://validity.smarterbalanced.org/reports-and-specifications/>).

2.5.2 Test Construction—WCAS

Unlike the Smarter Balanced tests, the WCAS assessments are fixed-form tests. Operational forms are created for each test administration, typically in the fall after data review of the field-test items. OSPI assessment content specialists and vendor psychometricians jointly select items according to test build specifications and test blueprints. There are a number of factors that must be considered during the test construction process. Items are selected to: 1) satisfy the test map, 2) meet target test difficulty, and 3) result in an overall test with balanced content (a variety of SEPs and CCCs). A test development checklist is used to review the initial test assembled during the test build. Test build is an iterative process to balance test content and statistical properties.

Test specifications guide the item selection process to ensure that all relevant standards and reporting areas are represented in each operational form. Representation of all gender and ethnic groups—in aspects including topics of science stimuli and item contexts—is reviewed to ensure that scenarios in science and stimulus materials used include balanced representations of groups. Items are selected to cover a range of difficulty levels on each of the science scales.

When a new operational form is created for each test administration, test scores must be equated to the baseline scale to maintain score interpretability over time. The baseline scale was determined following achievement level setting in 2018, following the first operational test administration; the scale is maintained until performance-level standards are revisited or redefined. The test developer's primary objective is to construct a new, parallel operational test form for each administration with target statistical characteristics and criteria to allow for comparability across test administrations. The better the match to these criteria, the better the equating accuracy of test scores among different test administrations.

Operational test forms are constructed such that test forms across administrations have difficulties that are as similar as possible. The weighted mean item response theory (IRT) difficulty is used as a statistical target for evaluating the test form's difficulty. The IRT item difficulty of each operational item is multiplied by the item's maximum raw score to obtain the item's weighted IRT difficulty. The sum of weighted item IRT difficulties is divided by the maximum total raw test score to compute the overall weighted mean IRT difficulty for the test. The weighted mean IRT difficulty for an operational form should closely approximate historical weighted mean IRT difficulties.

2.6 SPRING 2022 TESTS

2.6.1 Smarter Balanced Assessments

Smarter Balanced tests are administered in the format of a computer-adaptive test (CAT) and a performance task (PT). That is, in the CAT, the item(s) selected for a student at the time depends on the student ability estimate based on all items administered to the students at the time. The Smarter Balanced tests are delivered via CAI's CAT delivery system that takes both content requirements and the adaptive nature of the CAT into account simultaneously. Details about the

CAT algorithm used for Smarter Balanced tests can be found in the Smarter Balanced technical reports (<http://www.smarterapp.org/documents/AdaptiveAlgorithm.pdf>).

2.6.1.1 CHANGES IN TEST BLUEPRINT OF SUMMATIVE ASSESSMENTS

Starting with the 2020–21 Smarter Balanced summative assessments, Smarter Balanced has offered member states the option to administer the summative assessments either with the full blueprint from 2018–19 or with an adjusted blueprint for ELA and mathematics. In the adjusted blueprint, the CAT portion of the blueprint is reduced by approximately 50 percent in each claim. Given that PTs are designed to be integrated tasks, the blueprints associated with the PTs have not been adjusted.

Because the CAT was approximately half as long as tests based on the full blueprint, testing times were expected to be significantly shorter; test reliability was expected to be lower, but still sufficiently high, for the ELA and mathematics assessments. Because the number of items per claim was too small, claim scores were not generated for the adjusted blueprint.

Washington chose to administer the Smarter Balanced adjusted blueprints for grades 3–8 and high school in the spring 2022 summative assessment administration. Tables 2.1 and 2.2 present the differences in the blueprint requirements for each claim between the full and adjusted blueprints for ELA and mathematics.

Table 2.1: Differences Between Spring 2019 and Spring 2022 ELA Test Blueprints

Component	Claim	Items in 2019 Full Blueprint	Items in Spring 2022 Blueprint	Changes in Spring 2022 Blueprint
CAT	Total Items	36–42	20–22	
	Claim 1 Reading	14–19	8–10	Grades 3–5: a total of 8 items with one passage from each of 1-LT and 1-IT. Grades 6–8 and HS: a total of 10 items with one 1-LT* passage and two 1-IT* passages.
	Claim 2 Writing	6	4	All grades: removed 2 items from target 9
	Claim 3 Listening	8–9	4	All grades: administered two passages
	Claim 4 Research	8	4	All grades: removed 4 items
PT	Total Items	2	2	
	Claim 4 Research	1	1	All grades: no change
	Claim 2 Full Write	1	1	

* 1-LT: Literary Text; 1-IT: Informational Text

Table 2.2: Differences Between Spring 2019 and Spring 2022 Mathematics Test Blueprints

Component	Claim	Items in 2019 Full Blueprint	Items in Spring 2022 Blueprint	Changes in Spring 2022 Blueprint
CAT	Total Items	30–36	16–18	
	Claim 1	16–22	9–11	
	Claim 2	3	1	All grades: reduced 50% of total test length
	Claim 3	8	4	
	Claim 4	3	2	

PT	Claims 2, 3, & 4	4–6	4–6	All grades: no change
-----------	------------------	-----	-----	-----------------------

2.6.1.2 IMPACT OF CHANGES IN SUMMATIVE TEST BLUEPRINTS

As expected, the shortened CAT length had an impact on the target coverage, the overall testing time, and the test-score reliability.

Target Coverage

The average number of unique content targets covered in the CAT component for the full blueprints and the adjusted blueprints is listed by claim and grade in Tables 2.3–2.4. The average number of unique targets was decreased in claim 1 only for ELA and in all claims for mathematics in the adjusted blueprint. The Smarter Balanced blueprints do not require all targets to be administered to each individual test, but all targets are covered at the aggregate level, across all tests, in the adjusted blueprint.

Table 2.3: Changes in Average Number of Unique Targets Assessed by Each Claim in ELA CAT Component

Grade	2019 Full Blueprint				Spring 2022 Adjusted Blueprint				Decrease in Average Number of Unique Targets			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
3	10.1	4.0	1.0	3.0	7.5	4.0	1.0	3.0	2.6	0	0	0
4	10.7	4.0	1.0	3.0	7.6	4.0	1.0	3.0	3.1	0	0	0
5	11.4	4.0	1.0	3.0	7.4	4.0	1.0	3.0	4.0	0	0	0
6	10.3	4.0	1.0	3.0	9.1	4.0	1.0	3.0	1.2	0	0	0
7	10.7	4.0	1.0	3.0	9.2	4.0	1.0	3.0	1.5	0	0	0
8	10.9	4.0	1.0	3.0	9.0	4.0	1.0	3.0	1.9	0	0	0
HS	10.0	4.0	1.0	3.0	8.3	4.0	1.0	3.0	1.7	0	0	0

Table 2.4: Changes in Average Number of Unique Targets Assessed by Each Claim in Mathematics CAT Component

Grade	2019 Full Blueprint				Spring 2022 Adjusted Blueprint				Decrease in Average Number of Unique Targets			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
3	10.9	2.0	5.7	3.0	9.0	1.0	3.6	2.0	1.9	1.0	2.1	1.0
4	10.0	2.0	5.4	3.0	9.0	1.0	3.6	2.0	1.0	1.0	1.8	1.0
5	9.0	2.0	5.3	3.0	8.0	1.0	3.4	2.0	1.0	1.0	1.9	1.0
6	10.0	2.0	4.6	3.0	8.6	1.0	3.0	2.0	1.4	1.0	1.6	1.0
7	8.0	2.0	4.6	3.0	6.3	1.0	3.4	2.0	1.7	1.0	1.2	1.0
8	10.0	2.0	4.8	3.0	9.0	1.0	3.4	2.0	1.0	1.0	1.4	1.0
HS	14.8	2.0	5.0	3.0	9.8	1.0	3.3	2.0	5.0	1.0	1.7	1.0

Testing Time

The overall testing time was greatly reduced in all grades. The average testing time decreased 64–134 minutes for ELA. The average testing time decreased 51–140 minutes for mathematics. The reduction in the overall testing times are primarily caused by the reduced time for the CAT due to changes in the test blueprints. There were also reductions in times for the PT components across both ELA and math that are unexplained by the use of the adjusted blueprint as the PT portions were the same design in 2019 and 2022. The changes in average testing times are presented in Tables 2.5–2.6.

Table 2.5: Changes in Average Testing Times: ELA

Grade	2019 Full Blueprint			Spring 2022 Adjusted Blueprint			Decrease in Testing Time		
	Overall	CAT	PT	Overall	CAT	PT	Overall	CAT	PT
3	4:40	2:00	2:40	2:44	0:58	1:46	1:56	1:02	0:54
4	5:04	2:07	2:57	2:52	0:58	1:55	2:12	1:09	1:02
5	5:05	2:09	2:56	2:51	0:58	1:53	2:14	1:11	1:03
6	4:41	2:14	2:27	2:31	1:08	1:23	2:10	1:06	1:04
7	4:19	1:59	2:20	2:30	1:06	1:25	1:49	0:53	0:55
8	4:05	1:56	2:09	2:31	1:06	1:25	1:34	0:50	0:44
HS	3:35	1:50	1:45	2:31	1:11	1:20	1:04	0:39	0:25

Table 2.6: Changes in Average Testing Times: Mathematics

Grade	2019 Full Blueprint			Spring 2022 Adjusted Blueprint			Decrease in Testing Time		
	Overall	CAT	PT	Overall	CAT	PT	Overall	CAT	PT
3	2:37	1:43	0:54	1:27	0:50	0:36	1:10	0:53	0:18
4	2:47	1:54	0:53	1:26	0:52	0:34	1:21	1:02	0:19
5	3:17	1:58	1:19	1:37	0:52	0:45	1:40	1:06	0:34
6	2:54	1:53	1:00	1:20	0:45	0:35	1:34	1:08	0:25
7	2:18	1:42	0:37	1:08	0:44	0:24	1:10	0:58	0:13
8	2:32	1:51	0:41	1:15	0:47	0:28	1:17	1:04	0:13
HS	2:11	1:31	0:41	1:20	0:48	0:32	0:51	0:43	0:09

Reliability of Total Scores

As expected, the reliability of total scores decreased in all grades due to the reduction in the number of items on the test. Although the reliability decreased, it was still high enough to report total scores and achievement levels. In the 2019 administration, the reliability for total scores was 0.92 for all grades in ELA and ranged from 0.92 to 0.95 in mathematics. In the spring 2022 administration, the reliability for total scores ranged from 0.87 to 0.88 for ELA and from 0.84 to 0.91 for mathematics.

2.6.2 WCAS

The spring 2022 administration consisted of one version of operational items per grade for the online assessments, noted as Test Form A. Field-test items developed for the NGSS were

embedded in the online versions. The 2022 accommodated forms (designated as Form 2 at each grade level) were administered to students unable to test online.

The left panels of Tables 2.7–2.9 show the test blueprints, and the right panels show the 2022 forms. The comparisons show a match in range between the 2022 forms and their associated blueprints on the WCAS at all grades.

Table 2.7: Grade 5 WCAS Test Specification

Reporting Areas	Test Blueprint	Spring 2022 Test Form A	Spring 2022 Test Form 2
	Points per reporting area	Points per reporting area	Points per reporting area
Practices and Crosscutting Concepts in Physical Sciences	10–18	14	15
Practices and Crosscutting Concepts in Life Sciences	7–14	12	12
Practices and Crosscutting Concepts in Earth and Space Sciences	7–15	12	11
Total Number of Points	35	38	38

Table 2.8: Grade 8 WCAS Test Specification

Reporting Areas	Test Blueprint	Spring 2022 Test Form A	Spring 2022 Test Form 2
	Points per reporting area	Points per reporting area	Points per reporting area
Practices and Crosscutting Concepts in Physical Sciences	10–18	14	14
Practices and Crosscutting Concepts in Life Sciences	11–19	16	16
Practices and Crosscutting Concepts in Earth and Space Sciences	7–14	12	12
Total Number of Points	40	40	40

Table 2.9: Grade 11 WCAS Test Specification

Reporting Areas	Test Blueprint	Spring 2022 Test Form A	Spring 2022 Test Form 2
	Points per reporting area	Points per reporting area	Points per reporting area
Practices and Crosscutting Concepts in Physical Sciences	12–20	18	18
Practices and Crosscutting Concepts in Life Sciences	12–20	15	17
Practices and Crosscutting Concepts in Earth and Space Sciences	9–17	12	10
Total Number of Points	45	45	45

SUMMARY

Content of the WCAP tests is derived from the Washington State Learning Standards and measures what students should know and be able to do in the tested grades. The types of items that appear in Smarter Balanced assessments and WCAS are diverse—varying from conventional multiple-choice items to writing equations and performing tasks—allowing these tests to assess student skills at various levels of complexity. The Smarter Balanced assessment consists of a performance task and a CAT, which uses an algorithm that selects items with the best content and ability measurement characteristics. The WCAS is a fixed-form test, constructed such that test forms across administrations have difficulties that are as similar as possible.

3. ITEM DEVELOPMENT

3.1 ITEM DEVELOPMENT

3.1.1 Item Development—Smarter Balanced

Item development for the accountability tests in English language arts (ELA) and mathematics was conducted by Smarter Balanced.

Smarter Balanced involved hundreds of educators from member states in the process. All K–12 participants

- were certified/licensed to teach in the applicable content area in a K–12 public school;
- were currently teaching in a public school within a Smarter Balanced governing state;
- had taught ELA and/or mathematics in grades 3 through 8 and/or high school within the past three years;
- had previously reviewed the Common Core State Standards for the content area for which they were writing items and/or performance tasks;
- submitted a statement of interest that described their interest in developing Smarter Balanced items and/or performance tasks as well as their qualifications for doing so; and
- completed training and achieved qualifications through the Smarter Balanced certification process.

All higher-education faculty

- were currently employed, or recently retired from, a college or university located within a Smarter Balanced governing state;
- had taught development and/or entry-level courses in English, English composition, mathematics, statistics, or a related discipline within the last three years;
- had previously reviewed the Common Core State Standards for the content area for which they were writing items and/or performance tasks; and
- completed training and achieved qualifications through the Smarter Balanced certification process.

Selected educators were required to participate in a series of online training activities. Training modules covered general item-writing guidelines and specifications, as well as specifications for writing specific types of tasks aligned to individual claims and targets. Item writers also received training regarding Smarter Balanced Bias and Sensitivity guidelines, the item-authoring systems, and item-tagging requirements. See the Smarter Balanced technical reports at

<https://validity.smarterbalanced.org/reports-and-specifications/> for a description of the item development process.

3.1.2 Item Development— Washington Comprehensive Assessment of Science (WCAS)

The Washington Comprehensive Assessment of Science (WCAS) for grades 5, 8, and 11 were developed by the Office of Superintendent of Public Instruction (OSPI) science assessment team staff with support from the development vendor. Item development for the WCAS began in winter of 2015.

The first step in the test development process is to select groups of educators to work with staff from OSPI and the vendor to develop the test items. Each work group includes 10 to 12 persons from throughout the state, most of whom are classroom teachers and curriculum specialists with teaching experience at or near the grades and in the content areas that are to be assessed. Participants are invited to apply for the Item Writing work group. They are chosen to represent Washington's student demographics.

In addition, all participants:

- are certified/licensed to teach in Washington State;
- are currently teaching or recently retired from teaching in a public or charter school in Washington State;
- have content and grade-level expertise;
- have knowledge of the state science standards; and
- are willing to disseminate information about the assessment development process.

Participation in the Item Writing work groups is a professional development opportunity for selected educators. Participants include novice and experienced item writers. In 2015 and 2016, all training was done during the face-to face work groups. In 2017, 2018, and 2019 educators were required to participate in a six-hour online training course prior to the face-to face work group. Training modules covered general assessment development information, required item-writing activities, information for aligning items and item clusters to NGSS performance expectations and associated dimensions (Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts). Participants also received training regarding sample item clusters.

Using the state science standards (NGSS performance expectations and Appendices), item writers prepare new items and scoring rubrics. Raw items are initially produced during these workshops and later refined by OSPI assessment content specialists in collaboration with the vendor's content specialists.

Item writers develop items and stimuli that

- align with two or three dimensions of a performance expectation or performance expectation bundle;

- fulfill the test map specifications;
- display content accurately and clearly;
- are within the grade-level reading range;
- are free of bias; and
- are accessible to students with special needs.

3.2 CONTENT REVIEWS AND BIAS AND SENSITIVITY REVIEWS

3.2.1 Smarter Balanced Assessments

Before any item is field tested, each item developed for a Smarter Balanced summative assessment was subject to reviews for content, bias, and accessibility. As with item development, groups of educators from member states were integrally involved in the item review process. The application process and qualifications for review groups mirrored the requirements for item writers. Like item writers, participants in the accessibility, bias, and sensitivity reviews participated in online training opportunities prior to reviewing items and/or stimuli. Checklists provided additional guidance during the review. Item content was evaluated according to the Item Quality Criteria for ELA and mathematics. Cognitive laboratories provided additional qualitative evidence that items were eliciting the types of response processes intended by the item writers. See the Smarter Balanced technical reports at <https://validity.smarterbalanced.org/reports-and-specifications/> for a description of the item review process.

3.2.2 WCAS

Before any item developed for the WCAS is field tested, it must be reviewed and approved by the Content Review work group and the Bias and Sensitivity committee. Like the Item Writing work groups, the Content Review work group includes Washington educators, curriculum specialists, and educational administrators with grade-level and subject-matter expertise relevant to the specific grade-level content. All participants are selected by OSPI from a pool of Washington educators who complete an application to participate in OSPI professional development activities. The participants engage in the online pre-meeting training course as well as training during the face-to-face meeting. This is another professional development opportunity for Washington educators. A Content Review work group's task is to review the item content and scoring rubric to ensure that each item

- is an appropriate measure of the intended content (learning standards);
- is aligned with two or three dimensions of the intended content (learning standard);
- is appropriate in difficulty for the grade level of the examinees;
- has only one correct or best answer for each multiple-choice and multiple select item; and
- has an appropriate and complete scoring guideline for constructed-response machine-scored and constructed-response hand-scored items.

Items may be revised on the basis of content reviews. Each test item is coded by Performance Expectation (Learning Standards), by at least two dimensions (Science and Engineering Practice, Disciplinary Core Idea, and /or Crosscutting Concept), and by item type (ETC, grid, hot text, multiple-choice, multiple-select, short-answer, simulation, table input, or table match). Items are then presented to the OSPI assessment content specialist for final review and approval before field testing. The final review includes a review of graphics, artwork, and layout.

The Bias and Sensitivity committee is composed of community members who represent the demographics of the students in Washington. The committee reviews each item to identify language or content that might be inappropriate or offensive to students, parents, or community members, or items that might contain stereotypic or biased references to gender, ethnicity, or culture. The Bias and Sensitivity committee reviews each item and accepts it as is, accepts it with suggested edits, or rejects it for use in item pilots.

3.3 ITEM PILOTING

3.3.1 Item Piloting—Smarter Balanced

The Smarter Balanced pilot test administration in spring 2013 deployed the key elements of the program so that the spring 2014 field test could be adjusted in accordance with the data collected in 2013 on the statistical quality of items and tasks. The pilot test also familiarized states, schools, teachers, and students with the item types and tasks that would be part of the Smarter Balanced summative assessments introduced two years later. Whereas the summative assessment includes a computer-adaptive test (CAT) component, the pilot tests were not adaptive. They were based on linear (i.e., fixed-form) assessments delivered by a computer. Pilot test forms were intentionally designed to resemble the future operational tests so that students and teachers would have an additional opportunity to become familiar with the assessment and the types of tasks associated with the Common Core State Standards. For details on the pilot test and the field test, see the Smarter Balanced 2013–14 technical report (<https://validity.smarterbalanced.org/reports-and-specifications/>).

3.4 FIELD-TEST ITEMS ANALYSIS

3.4.1 Field-Test Items Analysis—Smarter Balanced

Field-test items in 2022 were embedded in the CAT and PT versions of the Smarter Balanced ELA and mathematics summative assessments. The various analyses conducted by Smarter Balanced on these items are detailed in the Smarter Balanced technical report (<https://validity.smarterbalanced.org/reports-and-specifications/>).

3.4.2 Field-Test Items Analysis—WCAS

After each field-test administration, student responses for constructed response items (hand-scored and machine-scored) are scored based on scoring rubrics approved by OSPI and the Field Test Rangefinding work groups. The work group members include Washington educators, curriculum specialists, and educational administrators with grade-level and subject-matter expertise. All participants are selected by OSPI from a pool of Washington educators who complete an

application to participate in OSPI professional development activities. The Rangefinding work group looks at a range of student responses to each short answer item and decides how to score each response. This educator work group refines scoring rubrics and produces the materials that will be used by professional hand-scorers to score the field-test items. Rubric Validation is completed by the vendor and OSPI assessment specialists who review rubrics and student responses for machine-scored constructed-response field-test items. Decisions about how every student answer on these items will be machine-scored are finalized.

Item analyses based on classical test theory, item response theory (IRT), and differential item functioning (DIF) are conducted to examine item qualities. The analysis procedures are explained below.

3.4.3 Classical Item Analysis Statistics

Smarter Balanced performs item analyses on embedded field test math and ELA items. Information on these analyses is available in the Smarter Balanced technical reports online at <https://validity.smarterbalanced.org/reports-and-specifications/>. Cambium conducted the analyses on WCAS field test items.

Classical item analyses involve computing a set of statistics for each item by test form. The set of statistics provides key information about the quality of each item. It includes item means, item-test correlations, percentage of students at each response option or score level, and percentage of students omitting the item.

For 1-point items, the item mean, or p -value, is the proportion of examinees that selected the correct answer choice. Item-test correlation (point-biserial) is computed as the item-total correlation. For 2-point items, the item mean is the sum of score points (0, 1, and 2) weighted (multiplied) by the proportion of students scored at that score and then divided by the maximum possible score 2, that is, the item mean is expressed as the average of weighted score points. Adjusted-polyserial correlation is computed as the item-total correlation. In IRT, these statistics of classical test theory are equivalent to item difficulty and item discrimination.

The item mean ranges from 0.0 to 1.0. p -values that exceed 90% or are lower than 30% are considered too high or too low. The point-biserial/adjusted-polyserial correlations are indexes of the relationship between performance on an item and overall performance on the test. They range from -1.00 to $+1.00$. A large positive value indicates a tendency for students with high scores on the overall test to earn higher item scores and students with low scores on the overall test tend to earn lower item scores. A low point-biserial/adjusted-polyserial index (close to 0) indicates no relationship between the performance on the item and the performance on the whole test. However, a large negative point-biserial (an extreme case) value implies that students who earn higher item scores tend to earn lower overall test scores, and students who earn lower item scores tend to earn higher test scores. This contradiction is an indication of a faulty test item. Point-biserial/adjusted-polyserial correlations are usually expected to be greater than 0.25, but these values can be deflated when item content is unfamiliar to students, regardless of student performance on the entire test, or when the item cannot well distinguish between students with different abilities.

Point-biserials/adjusted-polyserials for each incorrect answer option are correlations between each incorrect answer choice and the overall test, and are expected to have negative values, indicating

that high-scoring students tend not to do well on the item, whereas low-scoring students tend to do well on the item. A positive point-biserial/adjusted-polyserial between an incorrect answer option and the overall test score may indicate an incorrect item key.

Table 3.1 shows flagging criteria for field-test items. Note that, in WCAS Data Review meetings, all field-test items, both flagged and unflagged, are reviewed by the committee members.

Table 3.1: Classical Item Analyses Flagging Criteria, Pilot Items

Statistics	Value
Low Item Mean	<0.3
High Item Mean	>0.9
Point Biserial/Adjusted Polyserial	<0.25

3.4.4 IRT Analysis

The Smarter Balanced assessment used generalized partial credit model (GPCM) and the WCAS used partial credit model (PCM) (Masters, 1982) for item calibration and scoring. Please refer to Section 4.1 for a more detailed discussion of the GPCM and PCM models. For Smarter Balanced assessments, please refer to the latest Smarter Balanced technical report for details about the IRT analysis for items (<https://validity.smarterbalanced.org/reports-and-specifications/>).

The goodness of fit of items indicates how well items fit the model. For the WCAS, the mean square Infit and Outfit are used as the fit indices. Both are chi-square-based. They indicate one aspect of item quality.

- The Infit is weighted by the model information and more sensitive to the discrepant observations close to middle range of a scale.
- The Outfit statistic is not weighted and more sensitive to the unexpected observations at locations toward the two ends of a scale.

Both statistics have an expected value of 1 that indicates perfect item-model fit. Values greater than 1.0 indicate noise and unmodeled variance in the data. Values less than 1.0 indicate that the data fit the measurement model better than expected, which could indicate some degree of local dependence among items. For both statistics, a range of [0.7, 1.3] is adopted as the productive range. When either statistic is greater than 1.3, the item is suspected as reduced productivity to unproductive or even distorted measurement.

3.4.5 Differential Item Functioning (DIF)

DIF analyses are also performed on the pilot items. DIF is observed when examinees from different demographic groups with the same ability (students matched on operational total test score) perform differently on the same item. DIF analyses were conducted for the purpose of further content review to flag items that might assess different constructs for different student groups. For the WCAS, the following DIF groups are included: male vs. female, White vs. African American, White vs. Hispanic, and White vs. Native American.

- In DIF analysis, test-takers in each student group are ranked relative to their total test score (conditioning on ability). Examinees in the focal group (e.g., females) are compared to examinees in the reference group (e.g., males) relative to their performance on individual items. In the 2022 administration, DIF analyses were conducted on both base form operational and field-test items for gender groups (male/female) and ethnicity groups (White/Asian, White/African American, White/Hispanic, and White/Native American), as well as multilingual learners (ML). The groups of male and White are referenced as the reference group and the other groups as the focal group.
- If the item is more difficult for the reference or the focal group, when conditioning on ability, the item may be biased or may be measuring something different from the intended construct. However, it may be also related to actual differences in relevant knowledge or skills (item impact) or statistical Type I error. As a result, DIF statistics are used only to identify items that are potentially functioning differentially. Subsequent review by content experts and Bias and Sensitivity committees is required to determine whether there is an identifiable source and meaning of performance differences.
- A generalized Mantel–Haenszel ($MH\chi^2$) procedure was applied to calculate DIF. The generalizations include: (1) adaptation to polytomous items, and (2) improved variance estimators to render the test statistics valid under complex sample designs.
- This procedure uses each student's raw score on the operational items on a given test to divide into 10 intervals for Smarter Balanced and 5 for WCAS to compute the $MH\chi^2$ DIF statistics. The analysis program computes the $MH\chi^2$ value, the conditional odds ratio, and the MH-delta for dichotomous items; the $GMH\chi^2$ and the standardized mean difference (SMD) are computed for polytomous items.

The MH chi-square statistic (Holland & Thayer, 1988) is calculated as

$$MH\chi^2 = \frac{(|\sum_k n_{R1k} - \sum_k E(n_{R1k})| - 0.5)^2}{\sum_k var(n_{R1k})}$$

where $k = \{1, 2, \dots, K\}$ for the strata, n_{R1k} is the number of correct responses for the reference group in stratum k , and 0.5 is a continuity correction. The expected value is calculated as

$$E(n_{R1k}) = \frac{n_{+1k}n_{R+k}}{n_{++k}}$$

where n_{+1k} is the total number of correct responses, n_{R+k} is the number of students in the reference group, and n_{++k} is the number of students, in stratum k , and the variance is calculated as

$$var(n_{R1k}) = \frac{n_{R+k}n_{F+k}n_{+1k}n_{+0k}}{n_{++k}^2(n_{++k}-1)}$$

where n_{F+k} is the number of students in the focal group, n_{+1k} is the number of students with correct responses, and n_{+0k} is the number of students with incorrect responses, in stratum k .

The MH conditional odds ratio is calculated as

$$\alpha_{MH} = \frac{\sum_k n_{R1k} n_{F0k} / n_{++k}}{\sum_k n_{R0k} n_{F1k} / n_{++k}}.$$

The MH-delta (Δ_{MH} ; Holland & Thayer, 1988) is then defined as

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

The GMH statistic generalizes the MH statistic to polytomous items (Somes, 1986), and is defined as

$$GMH\chi^2 = (\sum_k \mathbf{a}_k - \sum_k E(\mathbf{a}_k))' (\sum_k \text{var}(\mathbf{a}_k))^{-1} (\sum_k \mathbf{a}_k - \sum_k E(\mathbf{a}_k)),$$

where $\mathbf{a}_k$ is a $(T-1) \times 1$ vector of item response scores, corresponding to the T response categories of a polytomous item (excluding one response). $E(\mathbf{a}_k)$ and $\text{var}(\mathbf{a}_k)$, a $(T-1) \times (T-1)$, the variance matrix, are calculated analogously to the corresponding elements in $MH\chi^2$, in stratum k .

The standardized mean difference (SMD; Dorans & Schmitt, 1991) is defined as

$$SMD = \sum_k p_{FK} m_{FK} - \sum_k p_{RK} m_{RK},$$

where $p_{FK} = \frac{n_{F+k}}{n_{F++}}$ is the proportion of the focal group students in stratum k ,

$m_{FK} = \frac{1}{n_{F+k}} (\sum_t a_t n_{Ftk})$ is the mean item score for the focal group in stratum k , and

$m_{RK} = \frac{1}{n_{R+k}} (\sum_t a_t n_{Rtk})$ is the mean item score for the reference group in stratum k .

Standardized mean difference is defined by

$$SMD = \sum_k p_{FK} m_{FK} - \sum_k p_{RK} m_{RK}$$

where $p_{FK} = n_{F+k} / n_{F++}$ is the proportion of the focal group students in stratum k ,

$m_{FK} = 1/n_{F+k} \left(\sum_t a_t n_{Ftk} \right)$ is the mean item score for the focal group in stratum k , and

$m_{RK} = 1/n_{R+k} \left(\sum_t a_t n_{Rtk} \right)$ is the mean item score for the reference group in stratum k .

The classification logic used for flagging items is based on a combination of significance testing and absolute differences. Items that are not statistically significantly different between the focal and reference groups based on the $MH\chi^2$ $P \geq 0.05$ are considered to have similar performance

between the two studied groups; these items are considered to be functioning comparably in both groups. For items with $MH\chi^2 p < 0.05$, the effect size is used to determine the direction and severity of item DIF. For 1-point items, $|\Delta_{MH}|$ is the effect size. Negative Δ_{MH} DIF statistics favor the reference group and positive values favor the focal group. For multiple point items, $|SMD/SD|$ is the effect size where SD is the total group standard deviation of the item scores on logit metric. A negative SMD/SD value indicates that the item is more difficult for the focal group, whereas a positive value indicates that the item is more difficult for the reference group. Tables 3.2 and 3.3 show the rule to classify DIF into one of three categories, A, B, and C. A category A DIF is minor DIF, a category B DIF is mild DIF, and a category C DIF is severe DIF. Items with C DIF are intended to be reviewed again at the item data review meetings.

DIF analyses were not conducted if the sample size for either the reference group or the focal group was less than 100 or if the sample size for the two groups combined was less than 400.

Table 3.2: DIF Categories for 1-Point Items

DIF Category	Definition
A	$MH\chi^2$ is not significant.
B	$MH\chi^2$ is significant and $ \hat{\Delta}_{MH} \geq 1.5$.
C	$MH\chi^2$ is significant and $ \hat{\Delta}_{MH} \geq 1.5$.

Table 3.3: DIF Categories for Multiple Points Items

DIF Category	Definition
A (negligible)	Mantel Chi-square p -value > 0.05 OR $ SMD/SD \leq 0.17$
B (slight to moderate)	Mantel Chi-square p -value < 0.05 and $0.17 < SMD/SD \leq 0.25$
C (moderate to large)	Mantel Chi-square p -value < 0.05 and $ SMD/SD > 0.25$

For WCAS field-test items, the classical item analysis results and DIF result can be found in Appendix A. The IRT analysis can be found in Appendix B. Data for spring 2022 operational items is also included in both appendices; that data comes from when those operational items were field tested in previous years' administrations.

3.5 ITEM DATA REVIEW

3.5.1 Smarter Balanced Assessments

Field-test data were analyzed by Smarter Balanced using both classical and IRT statistics, as well as content and scoring decisions, to create the final item pool. See the Smarter Balanced technical reports for more information (<https://validity.smarterbalanced.org/reports-and-specifications/>).

3.5.2 WCAS

For the WCAS, Content Review with Data work groups were held to evaluate the quality of field-test items. The work group members included Washington educators, curriculum specialists, and educational administrators with grade-level and subject-matter expertise. All work group members are selected by OSPI from a pool of Washington educators who complete an application to participate in OSPI professional development activities. OSPI content specialists and psychometricians from the vendor facilitated the Content Review with Data meeting.

For the meeting, item review cards are provided, which include item means, item-test correlations, IRT item difficulties, item fit statistics, and DIF categories. In addition, item content alignment is provided and verified. During the meeting, the committee identifies items that function poorly (too easy, too difficult, or have low or negative item-test correlations, distractors that are drawing very few students) and makes recommendations to either revise or reject such items. Finally, items that are flagged for C DIF are examined closely to see if there is a content reason that could explain the C DIF. If there is a content explanation of the C DIF for an item, the item will be rejected for operational use. See Tables 3.1–3.3 for the flagging rules and the cuts for DIF categories.

During these reviews, the educators make recommendations to accept, revise, or reject each item. If items are accepted, they are added to the operational pool for future administrations. Table 3.4 shows the final outcome for items field tested in spring 2022. At the 2022 Content Review with Data, 160 items were accepted, 3 items were revised and will be re-field tested, and 9 items were rejected. The reasons for rejecting items included poor Classical Item Analysis Statistics and DIF statistics.

Table 3.4: WCAS Content Review with Data, Spring 2022 Field Test Administration Results

Grade	Reviewed			Revised			Rejected		
	1-Point Items	2-Point Items	Total Items	1-Point Items	2-Point Items	Total Items	1-Point Items	2-Point Items	Total Items
5	34	13	47	-	-	-	5	1	6
8	35	11	46	-	-	-	1	-	1
11	58	21	79	2	1	3	2	-	2

SUMMARY

Developing bias-free items that accurately reflect a student’s skill set in a content area is critical to the integrity of assessments. Smarter Balanced Consortium and OSPI staff have adopted rigorous standards and criteria governing the screening, recruitment, and training of educators who are part of the teams that develop and review test items.

For the WCAS, OSPI recruited classroom teachers and curriculum specialists throughout Washington to participate in various committees created to develop, assess, and review items used in the tests—the Item Writing work group, the Content Review work group, the Bias and Sensitivity committee, and the Content Review with Data work group. In the Content Review with Data meeting, classical item analysis statistics of each field-test item, as well as the flagging criteria and rules for approving, rejecting, or updating the content of field-test items, are provided and clearly communicated to work group members.

4. CALIBRATION AND EQUATING

Calibration is the statistical process used to obtain item response theory (IRT) parameters. Equating is the process used to put the calibrated parameters onto the existing scale.

For both Smarter Balanced assessments and the Washington Comprehensive Assessment of Science (WCAS), if there are field-test items, calibration and equating are necessary for item data review purposes. For the WCAS, if there are first-year operational items, calibration and equating are conducted for scoring purposes.

4.1 ITEM RESPONSE THEORY (IRT)

For item calibration and scoring, Smarter Balanced assessments use a generalized partial credit model (GPCM), while the WCAS uses a partial credit model (PCM). The GPCM used by Smarter Balanced assessments takes the form of a two-parameter logistic model where the probability $p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})$ of earning a specific score point at a specific theta point relates to both item discrimination a_i and item difficulty b_{i,m_i} for item i . In PCM, the probability $p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})$ of earning a specific score point at a specific theta point takes the form of the one-parameter logistic model, where an item is described by item difficulty parameter only.

In the case of items with one score point, $m_i = 1$,

$$p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = \begin{cases} \frac{\exp(Da_i(\theta_j - b_{i,1}))}{1 + \exp(Da_i(\theta_j - b_{i,1}))} = p_{ij}, \text{ if } z_{ij} = 1 \\ \frac{1}{1 + \exp(Da_i(\theta_j - b_{i,1}))} = 1 - p_{ij}, \text{ if } z_{ij} = 0 \end{cases};$$

in the case of items with two or more points,

$$p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = \begin{cases} \frac{\exp(\sum_{k=1}^{z_{ij}} Da_i(\theta_j - b_{i,k}))}{s_{ij}(\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})}, \text{ if } z_{ij} > 0 \\ \frac{1}{s_{ij}(\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})}, \text{ if } z_{ij} = 0 \end{cases},$$

where $s_{ij}(\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = 1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l Da_i(\theta_j - b_{i,k}))$, and $D = 1.7$.

In the formula, a_i is the item discrimination, and $b_{i,k}$ is the item difficulty for item i at score point k . For PCM, $a_i = 1$.

4.2 ITEM CALIBRATION

The item parameters were estimated by maximizing the joint likelihood function of PCM:

$$\arg \max \delta L(\delta) = \prod_{i=1}^R \prod_{j=1}^N \frac{\exp \sum_{k=1}^{x_i} Da_i(\theta_j - b_{i,k})}{1 + \sum_{j=1}^{m_i} \exp \sum_{k=1}^j Da_i(\theta_j - b_{i,k})}$$

where R indexes the total number of items, N indexes the total number of students, and $b_{i,k}$ is the step parameter for step k on item i . Each step parameter is located at the point where the likelihood function for that step is maximized along the ability scale.

4.3 SMARTER BALANCED

The adoption of the Smarter Balanced English language arts and mathematics assessments in 2015 offered an alternative model of calibrating student ability. Essentially, Smarter Balanced assessments differ from the WCAS in the following ways:

- Smarter Balanced assessments use a two-parameter GPCM.
- For scoring purposes, Smarter Balanced assessments applied item parameters that were pre-equated from field-test responses.
- Smarter Balanced assessments are computer-adaptive tests. Depending on the items presented in a test, two students receiving the same raw score will often receive two different scale scores.

For detailed descriptions of the calibration of item parameters and student ability of Smarter Balanced tests, please refer to the Smarter Balanced technical report (<https://validity.smarterbalanced.org/reports-and-specifications/>).

4.4 WCAS

The purpose of calibration and equating is to calibrate first-time operational items and put them on the existing scale for scoring purposes. Therefore, only operational and first-time operational items are involved. Field-test items are not included.

- WCAS uses the one-parameter PCM.
- If there are first-time operational items in the base WCAS form, post-equated parameters are used in scoring.
- WCAS is an online, fixed-form test. The WCAS are number-correct scored, which means that two students having the same raw score will receive the same scale score.

4.4.1 Post-Equating Procedure

In 2022, post-equating was done for WCAS tests. Post-equating refers to the calibration and equating process that occurs after a test administration. For the WCAS, first-time operational items are concurrently calibrated with the “anchor” items—items that have been used operationally in earlier administrations and are on the existing scale. Anchoring on these items calibrates the newly operational items to the existing scale.

The 2022 post-equating was conducted using all testing records that met the following conditions:

- Students attempted the test by responding to two or more items.
- Students took the English version.
- Students did not take the Braille version.

Post-equating included the following general steps:

1. Identify the anchor items in the test form and their associated bank value item difficulties.
2. Calibrate the 2022 base form items without anchoring—free calibration of the 2022 base form items.
3. Calculate the mean item difficulty of the anchor items using the item bank values (from Step 1).
4. Calculate the mean item difficulty of the anchor items from the 2022 free calibration (from Step 2).
5. Compute the equating constant as the difference of Step 3 minus Step 4 results.
6. Add the equating constant to each of the anchor item parameters from 2022 free calibration (2022 “adjusted difficulty”) so that the mean equals that of the mean of the banked values.
7. Subtract, by item, the 2022 “adjusted difficulties” from the bank anchor difficulties.
8. Flag the items with an absolute difference greater than 0.3.
9. Review the flagged items. Anchor items are retained in the anchor set unless additional information about the flagged item(s) suggests that the item(s) should not be used as an anchor item. The item difficulty of the stable anchor items will remain unchanged, or fixed, in calibrating the item difficulty of the first-time operational items. Anchor items that are considered as unstable will be treated as first-time operational items and re-calibrated.
10. The bank value of the item difficulty of the anchor items and the newly calibrated parameter of the first-time operational items are then used to estimate student ability. The newly operational items and their calibrated parameters from 2022 will take on the role of anchor items in future administrations.

4.4.2 Post-Equating: WCAS

The 2022 WCAS assessments had two sets of operational items: (a) items used for calibrating student performance and (b) items used in the accommodation forms such as paper and Braille in each of the tested grades. For all grades, the different sets of operational items were offered for online and accommodated forms. The accommodated forms were not used in equating because they were subject to pre-equating.

For WCAS, online tests were used for item calibration, and the parameters derived were applied for ability estimation for both online and paper accommodated forms. There is no need to compare online performance with paper performance because almost all students took the test online.

The following describes post-equating conducted on the online tests:

The WCAS equating sample for grades 5, 8, and 11 was based on the scored online tests that CAI received from the scoring vendor on July 2, 2022. Table 4.1 shows the size of the post-equating sample and as a percentage of all students who took the WCAS tests. In the spring 2022 administration, due to item parameter stability check, one item in grade 5 was flagged, two items in grade 8 and two items in grade 11 were flagged as well. After OSPI's review of these items, the decision was made to keep all items in the anchor set. Both first-time operational and anchor items were concurrently calibrated by fixing their parameters to the bank values.

Table 4.1: WCAS Grades 5, 8, and 11 Post-Equating Sample Size and Percentage of Tested Student Population, 2022 Spring Administration

	Online Tests		
	Grade 5	Grade 8	Grade 11
Sample Size	75,796	78,077	55,727
Percentage of the Tested Population	98.7	98	96.8

4.5 EQUATING RESULT

The item parameters and the fit statistics can be found in Appendix B. The model fit statistics are summarized in Table 4.2. The results show that the data fit the model well.

Table 4.2: Model Fit, WCAS, 2022 Administration

Test	Role	Percentage of Items Between 0.7 and 1.3		Mean (SD)	
		INFIT MNSQ	OUTFIT MNSQ	INFIT MNSQ	OUTFIT MNSQ
Grade 5 WCAS	OP	90%	79%	1.02 (0.166)	1.06 (0.4014)
Grade 8 WCAS	OP	91%	88%	1.02 (0.1667)	1.03 (0.2325)
Grade 11 WCAS	OP	100%	83%	1.00 (0.137)	1.05 (0.2952)

SUMMARY

Smarter Balanced is an online adaptive assessment that uses an algorithm to select the next item that would meet the test blueprint and best match with student ability at that point. Smarter Balanced uses a two-parameter GPCM to calibrate student achievement. The WCAS is fixed-form and uses a one-parameter partial credit model to calibrate student performance. A raw-to-scale score conversion table is used to place a student's scale score within the state-approved achievement levels.

The WCAS adopts a post-equating process to tie the difficulty parameter of first-time operational items to the existing scale. To accommodate the schedule for handscoring items, post-equating is often conducted on tests returned by a certain date, not on total tests returned. Post-equating was conducted for grades 5, 8, and 11. All post-equating samples were found to be fair representations of eligible testers.

The process of post-equating begins with a review of any changes in the item difficulty parameter between the current administration and when the item was first-time operational in an earlier administration. OSPI content staff and the vendor's psychometric team jointly reviewed these items, and the anchor items flagged to the item parameter stability check were evaluated individually and determined to remain in the anchor set.

5. TEST ADMINISTRATION

5.1 TESTING WINDOWS

The Spring 2022 Washington Comprehensive Assessment Program (WCAP) testing windows are shown in Table 5.1. For Smarter Balanced assessments, the testing window spans approximately 12 weeks for the online summative assessments and approximately 6 weeks for the paper-pencil summative assessment. The testing windows were approximately one month for the Washington Comprehensive Assessment of Science (WCAS).

Table 5.1: Spring 2022 Testing Windows

Tests	Grade of Test	Start Date	End Date	Mode
ELA and Mathematics Smarter Balanced	3–8, HS	3/07/2022	6/10/2022	Online Adaptive
	3–8, HS	4/11/2022	5/20/2022	Paper Fixed-Form
WCAS	5, 8, 11	4/11/2022	6/03/2022	Online Fixed-Form
	5, 8, 11	4/11/2022	5/20/2022	Paper Fixed-Form

5.2 TEST ADMINISTRATION

The Smarter Balanced assessments are administered online for most students in Washington or on paper for a small population of students who either have an Individualized Education Program (IEP), 504 Plan, or similar learning plan that specifies paper for large print, braille, Spanish mathematics, or standard print forms and those who do not have access to technology. To ensure that all eligible students in the tested grades were given the opportunity to take the Smarter Balanced assessments, a number of assessment options were available for the spring 2022 administration to accommodate students' needs.

The WCAS is administered online for most students in Washington. The accommodated paper form is available for the small number of students who either have an IEP, 504 plan, or similar learning plan that specifies paper for large print, braille, Spanish, or standard print forms and those who do not have access to technology. To ensure that all eligible students in the tested grades were given the opportunity to take the WCAS, a number of assessment options were available for the spring 2022 administration to accommodate students' needs.

Table 5.2 lists the testing options that were offered in 2021–22. A testing option is selected for each content area. Once the testing option is selected, it applies to all tests within that content area, whether in online or paper-pencil format.

Table 5.2: Summary of Tests and Testing Options in Spring 2022

Assessments	Test Options	Test Mode
ELA and Mathematics Smarter Balanced	English	Online, Accommodated Paper
	Spanish (Mathematics only)	Online, Accommodated Paper
	Braille	Online, Hybrid Adaptive Test (Mathematics only), and Accommodated Paper
	Large Print	Accommodated Paper
WCAS	English Fixed-Form	Online, Accommodated Paper
	Spanish Fixed-Form	Online, Accommodated Paper
	Braille Fixed-Form	Accommodated Paper
	Large Print Fixed-Form	Accommodated Paper

To ensure standardized administration conditions, Test Administrators (TAs) follow procedures outlined in the *Test Administration Manual* (TAM) for each specific test. TAs must review the TAM before testing, ensure that the testing room is prepared for testing (e.g., removing certain classroom posters, arranging desks), and follow make-up procedures for any students who are absent on the day(s) of testing. TAs follow required administration procedures and TA scripts of student directions. TAs read the boxed directions verbatim to students, ensuring standardized administration conditions for all assessments. The latest TAM and *TA Scripts of Student Directions* are available on the WCAP portal (<https://wa.portal.cambiumast.com/>). Contact the Office of Superintendent of Public Instruction (OSPI) for the 2021–22 versions.

5.2.1 Administrative Roles

The key personnel involved with the test administration are District Test Coordinators (DCs), District Administrators (DAs), School Test Coordinators (SCs), Technology Coordinators, and TAs. The main responsibilities of these key personnel are described below.

Table 5.3: Responsibilities of Key Personnel 2021–22

User	Description
District Assessment Coordinator (DC)	DCs are responsible for the following: ▪ general oversight of all test administration activities; ▪ review and approve each school's Test Security and Building Plan (TSBP) and test schedules; ▪ add users, order paper test booklets, set testing windows, and enter appeals in the Test Information Distribution Engine (TIDE); ▪ ensure that all staff are appropriately trained regarding the WCAP assessments administration, security policies, and procedures; ▪ monitor testing progress and ensure that all students participate, as appropriate; ▪ report all required information to the State via the Assessment Reporting Management System (ARMS); and ▪ notify the OSPI State Test Coordinator directly in instances involving test irregularities and breaches.

User	Description
District Administrator (DA)	DAs are responsible for the following: ▪ support the DC by providing general oversight and responsibilities for all test administration activities in their district and schools; ▪ support the DC in adding users in TIDE; ensuring staff are appropriately trained test administration and security policies and procedures; and ▪ assist in the review of school Test Security and Building Plans and testing schedules;
School Test Coordinator (SC) Note: An SC can be a principal, vice principal, Technology Coordinator, counselor, or other LEA member. If possible, an SC should be a person with non-instructional or limited instructional duties so that they can coordinate and monitor testing activity in the school.	SCs are responsible for the following: ▪ general oversight of all TAs and administration activities in their school; ▪ coordinate with technology coordinators to ensure that necessary secure browsers are installed, and any other technical issues are resolved; ▪ create school Test Security and Building Plan and submit for approval by the DC; ▪ ensure TAs are properly trained and have access to the secure test delivery system; ▪ enter and/or verifying test settings and accessibility features, monitor school testing progress, and ensure that all students participate in testing with the appropriate supports; ▪ report all test security incidents to the DC; and ▪ submit appropriate reporting documents to the DC.
Technology Coordinator	Technology Coordinators are responsible for the following: ▪ General oversight of technology needed for all online testing activities; ▪ Configure the devices, software, and networks used for online testing; ▪ Ensure that all non-approved features and software are blocked; and ▪ Assist in troubleshooting technical or infrastructure issues.
Test Administrator (TA)	TAs are responsible for the following: ▪ review all training and administration documents prior to administering any assessments; ▪ review student information prior to testing to ensure that each student receives the proper test with the appropriate supports, and report potential errors to SCs and DCs as appropriate; ▪ administer the appropriate assessments; and ▪ report all potential test incidents to the SC and DC in a manner consistent with state and district policies.

5.2.2 Online Administration

The online assessments allow schools to choose testing dates and to test students in intervals rather than in one long testing period. With online testing, schools have required protocols, but do not need to handle test booklets and address the test booklet security and storage protocols required in district-wide paper-based assessments.

SCs oversee all aspects of testing at their schools and serve as the main points of contact, while TAs administer the online assessments. TAs are trained in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the administration are provided online. All school personnel who serve as SCs and TAs are required to attend the school district's training workshop and sign verification documentation. In addition, a strongly

recommended online TA Certification Course is available before testing begins. Staff members who complete this online course receive a certificate of completion and appear in the online testing system. The TA Certification Course is available on the WCAP portal (<https://wa.portal.cambiumast.com/>).

To start a test session, the TA must first enter the Test Administrator Interface (TA Interface) of the online testing system using their computer. A session ID is generated when the test session is created. Students who are taking the assessment with the TA need to enter their State Student Identifier (SSID) number, their first name, and the session ID into the Student Secure Browser using computers provided by the school. The TA then verifies that the students are taking the appropriate content-area assessment(s), using the correct test opportunity, and being provided with the appropriate assessment accommodations, such as testing in a small group (see Section 5.6 for a list of accommodations). Students can begin testing only after the TA confirms that they are taking the appropriate assessment(s) and approves them to be tested. The TA must read the *Online TA Script of Student Directions* aloud to the students and walk them through the login process.

The Smarter Balanced English language arts (ELA) and mathematics assessments can be started in one test session and completed in a different session. However, the Smarter Balanced CAT must be completed within 45 calendar days of the start date or the assessment opportunity will expire. For the performance tasks (PT), the assessment must be completed within 30 calendar days of the start date or the assessment opportunity will expire.

The WCAS tests can be started in one test session and completed in a different session. WCAS tests expire at the end of the test window.

Once an assessment is started, the student must answer all test questions presented on a page before proceeding to the next page; students are not allowed to skip questions. For the online CAT, a student is allowed to scroll back to review and edit answers, as long as they are in the same test session and their test has not been paused for more than 20 minutes. For the WCAS, a student is allowed to scroll back to review and edit answers, as long as their test has not been paused for more than 20 minutes. Students can only edit answers to WCAS questions that are not locked.

During a test session TAs can pause a single student's assessment, or all of the assessments (for example, to give students a break) from within the TA Interface. It is up to the TA to determine an appropriate stopping point. Students can also pause their tests from within the Student Secure Browser. If a test is paused for more than 20 minutes, when the student logs back in to resume their test they will see the next test page with unanswered questions. Students will only be able to move forward in their test. Students will not be able to return to any previous pages or questions they answered before their test was paused, even if they marked questions for review. This is to ensure the integrity of the assessments.

The TA must remain in the room at all times during a test session to monitor student testing. Once the test session ends, the TA must ensure students have successfully logged out of the system, collect all scratch paper and ancillary materials that students used during the assessment, and clear calculator memories.

5.2.3 Paper-Pencil Test and Accommodated Paper Administration

For Smarter Balanced assessments, paper-pencil versions of the assessments are provided as an accommodation for students who cannot take the assessments online as stated in their IEP, 504 Plan, or similar learning plan or who do not have access to technology. Paper-pencil TAMs were available on the Washington portal (<https://wa.portal.cambiumast.com/>). Contact OSPI for the 2022 version. Braille, Spanish (math only), standard print, and large print versions were available for Smarter Balanced tests.

For the paper-pencil version of the Smarter Balanced ELA and mathematics assessments, each content area has two separate booklets, a test booklet and an answer booklet. The Smarter Balanced CAT and the performance task are combined into one test booklet. In both content areas, three sessions (two for the CAT and one for the performance task) are included in each test booklet so that the TA can break up the assessment into separate sessions. After the students complete the assessments, the DC securely returns the test booklets and the answer booklets to the testing vendor for scoring. The testing vendor scans the answer document and hand-scores the hand-scorable items. Once all of the items have been hand-scored, the testing vendor scores the overall test.

The total number of students who took paper-pencil Smarter Balanced tests is shown in Table 5.4.

Table 5.4: Number of Students Who Took Paper-Pencil ELA and Math Tests in Spring 2022 Administration

Test Form	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	HS	Total
ELA Braille	1	2	5	1	4	4	2	19
ELA Standard	7	14	17	20	23	20	32	133
ELA Large Print	1	2	3	4	1	4	3	18
ELA Total	9	18	25	25	28	28	37	170
Mathematics Braille	2	2	4	2	3	5	7	25
Mathematics Spanish	-	-	-	1	2	1	-	4
Mathematics Standard	8	14	20	24	21	22	32	141
Mathematics Large Print	1	2	3	4	2	5	2	19
Mathematics Total	11	18	27	31	28	33	41	189

For the WCAS, accommodated paper versions of the assessments are provided as an accommodation for students who cannot take the assessments online as stated in their IEP, 504 Plan, or similar learning plan or who do not have access to technology. Paper-pencil TAMs were available on the Washington portal (<https://wa.portal.cambiumast.com/>). Contact OSPI for the 2022 version. Secure SAY scripts were also provided. The tests are administered in a student to proctor ratio of no more than 3 to 1. Braille, Spanish, standard print, and large print versions were available for WCAS testing.

For the paper-pencil version of the WCAS, the student enters their answers into a single, scorable test booklet. Student responses for braille and large print booklets are transcribed by local staff into standard print booklets at the end of testing. Then the DC securely returns the test booklets

and the answer booklets to the testing vendor for scanning, handscoring of the hand-scorable items, and overall test scoring. The total number of students who took the accommodated paper WCAS is shown in Table 5.5.

Table 5.5: Number of Students Who Took the Accommodated Paper WCAS in Spring 2022 Administration

Test Form	Grade 5	Grade 8	Grade 11	Total
Science Braille	5	5	5	15
Science Spanish	-	-	-	-
Science Standard Print	58	46	10	114
Science Large Print	4	2	2	8
Science Total	67	53	17	137

5.2.4 Online Braille Test Administration

In Washington, the WCAP assessments are made available to students who use braille as a mode of instruction via a paper test booklet (see section 5.2.3). Washington also offers the Smarter Balanced assessments online to students who use braille. In 2018–19, Smarter Balanced added the Braille Hybrid Adaptive Test (Braille HAT) for mathematics. The Braille HAT consists of a fixed-form segment, a computer-adaptive segment, and a fixed-form PT. The fixed-form segment includes items with tactile graphics, which can be embossed at the testing location or received as a package of pre-embossed materials. All items on the Braille HAT can be presented to the students using a Refreshable Braille Display (RBD). The WCAS is not currently available in an online braille version.

The braille interface is described below in several formats:

- The braille interface includes a text-to-speech component for mathematics consistent with the read-aloud assessment accommodation. The Job Access with Speech (JAWS) screen-reading software provided by Freedom Scientific is an essential component that students use with the braille interface.
- Mathematics items are presented to students in Nemeth Code or UEB Code via a braille embosser through the adaptive online summative test and a fixed-form PT.
- Students taking the summative ELA assessment can emboss both reading passages and items as they progress through the assessment. If a student has an RBD, a 40-cell RBD is recommended. The summative ELA assessment is presented to the student with items in either contracted or uncontracted Literary Braille (for items containing only text) and via a braille embosser (for items with tactile or spatial components that cannot be read by an RBD).

Before administering the online summative assessments using the braille interface, Technology Coordinators must ensure that the technical requirements are met. These requirements apply to the student's computer, the TA's computer, and any supporting braille technologies used in

conjunction with the braille interface. A *Braille Requirements Manual* was also available for Technology Coordinators on the WCAP Portal (<https://wa.portal.cambiumast.com/>).

5.3 TRAINING AND INFORMATION FOR TEST COORDINATORS AND ADMINISTRATORS

DCs and SCs oversee all aspects of testing at their schools and serve as the main points of contact, while TAs administer the assessments. An online TA Certification Course, webinars, user guides, manuals, PowerPoint presentations, and training sites are used to train TAs on the testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the administration are posted on the WCAP Portal (<https://wa.portal.cambiumast.com/>).

5.3.1 Online Training

Multiple training opportunities were available to the key district staff through the WCAP Portal (<https://wa.portal.cambiumast.com/>). Contact OSPI Assessment Operations office for the 2021–22 versions.

TA Certification Course

All school personnel who serve as TAs are required to attend district-developed training sessions. SCs maintain documentation, including TA signatures, of each individual who has completed the training. In addition to this mandatory training, TAs are strongly encouraged to complete an online TA Certification Course. This web-based course is about 30–45 minutes long and covers information on testing policies and the steps for administering a test session in the online system. The course is interactive, requiring participants to actually start test sessions under different scenarios. At the end of the course, participants need to answer multiple-choice questions about the information provided. Completion of the TA Certification Course is tracked online in TIDE.

Practice and Training Test Site

Separate training sites are available for TAs and students through the WCAP Portal (<https://wa.portal.cambiumast.com/>). TAs practice administering assessments and starting and ending test sessions on the TA Practice Interface site. The site can also be used to administer the practice tests or training tests to students. Students can practice taking an online assessment with a TA-generated test session ID in the Student Secure Browser or on the Practice and Training Test site using a browser like Chrome, Edge, or Firefox.

The practice tests mirror the full blueprint Smarter Balanced summative assessments for ELA and mathematics. Each test provides students with a grade-specific testing experience, including a variety of question types (approximately 30 items each in ELA and mathematics), and a performance task.

The training tests are designed to provide students and teachers with opportunities to quickly familiarize themselves with the online platform and navigational tools they will use for the online tests. Training tests include almost all item types that are included in the operational item pool for that content area. Training tests are available for mathematics, ELA, and WCAS and are organized

by grade bands for ELA (grades 3–5, 6–8, and high school) and mathematics (grades 3–5, 6, 7–8, and high school) and grades 5, 8, and 11 for WCAS.

A student, parent, or member of the public can log in directly to the Practice and Training Test site as a “Guest” using a browser like Chrome, Edge, or Firefox. These “guest” sessions do not use a TA-generated test session ID, nor do they use an SSID. “Guests” can take any of the practice or training tests for ELA, mathematics, and WCAS to learn about the student testing experience.

Manuals and User Guides

The *Test Administration Manual (TAM)* and test-specific *TA Script of Student Test Directions* provide information for TAs administering the Smarter Balanced summative assessments in ELA and mathematics and the WCAS. The *TA Scripts* for online tests include screenshots of both parts of the Test Delivery System (TDS): the Student Secure Browser and the Test Administrator Interface.

The *Quick Guide for Setting Up Online Testing Technology* outlines the basic technology requirements for administering an online assessment, including operating system requirements and supported web browsers.

The *Configurations, Troubleshooting, and Advanced Secure Browser Installation Guide* provides instructions for downloading and installing the secure browser on supported operating systems used for online assessments. Guides are available for Chrome OS, iOS/iPadOS, MAC, and Windows.

The *Technical Requirements for Online Testing* provides technology staff with the technical specifications for online testing, including information on Internet and network requirements, general hardware and software requirements, and the text-to-speech function.

The *Test Information Distribution Engine User Guide* is designed to help users navigate TIDE. Users can find information on managing user account information, student account information, student test settings and accommodations, and appeals.

The *Test Administrator Interface User Guide* is designed to help users navigate the TDS, including the Student Secure Browser and the Test Administrator Interface, and help TAs manage and administer online testing for students.

The *Operating System Support Plan for Test Delivery System* describes CAI’s plan for supporting operating systems during the upcoming test administration and following years. This plan helps districts and schools manage operating system deployments based on the support timelines.

The *Braille Requirements Document for Online Systems* provides information about supported hardware and software requirements for braille testing and instructions for configuring JAWS. Information about navigating a test with JAWS is also included.

The *Guidelines on Tools, Supports, and Accommodations for State Assessments* helps to guide decisions associated with accessibility features available to students during state testing.

The *Test Coordinators Manual* assists in the administration and security of the assessments. This manual provides DCs with information on the security, coding, logistical, and paper-handling/online requirements at the district and school levels.

Training Modules

The following training modules help users understand how each CAI system works. The modules were provided as PowerPoint presentations. All modules are posted on the WCAP Portal (<https://wa.portal.cambiumast.com/>).

The *Student Secure Browser for Online Testing Module* explains the onscreen layout of the test; the functionality of the test tools; and how students log in to the testing system, select a test, and navigate through it.

The *Technology Requirements for Online Testing Module* provides current information about technology requirements, site readiness, supported devices, and secure browser installation.

The *Test Administrator Interface for Online Testing Module* presents an overview of how to navigate the Test Administrator Interface.

The *Test Information Distribution Engine Module* provides an overview of TIDE. It includes information on logging in to TIDE; managing user accounts; and managing student information, rosters, and appeals.

Accommodated Test Administration Training provides an overview of ordering, receiving, administering, processing and returning accommodated assessment materials.

The *Braille Training Module for Test Administrators* provides support on the process for administering online tests to students using braille, braille types, and emboss requests.

The *Braille Training Module for Technology Coordinators* provides support on the process for device and software configuration, embossing requests, and using the braille sign-in to the Student Secure Browser.

5.4 TEST SECURITY

All test content, test materials, and student-level testing information is secure for both online and paper-pencil assessments. The importance of maintaining test security and the integrity of test items is communicated and documented through the webinar trainings and in user guides, modules, and manuals. Features in the online system are developed to maintain test security. This section describes system security, student confidentiality, and policies on testing impropriety.

5.4.1 Student-Level Testing Confidentiality

All of the testing contractor's secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the Family Educational Rights and Privacy Act (FERPA) and other federal laws. Secure transmission and password-protected access are basic features of the current system and ensure authorized data

access. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins. The testing contractor's systems use role-based security models that ensure that users may access only the data to which they are entitled and may edit data only in accordance with their user rights.

There are three dimensions related to identifying that the correct students are accessing appropriate test content:

1. *Test eligibility* refers to the assignment of a test to a particular student.
2. *Test tools, supports, or accommodations* refers to the assignment of a test setting to specific students based on needs.
3. *Test session* refers to the authentication process of a TA creating and managing a test session, the TA reviewing and approving a test (and its settings) for every student, and the student signing on to take the test.

FERPA prohibits the public disclosure of student information or test results. The following are examples of prohibited practices:

- Providing login information (username and password) to other authorized TIDE users or to unauthorized individuals
- Sending a student's name and SSID number together in an email message. If information must be sent via email or fax, include only the SSID number, not the student's name.
- Having a student log in and test under another student's SSID number

Student test materials and reports should not be exposed so that student names could be identified with student results, except by authorized individuals with an appropriate need-to-know status.

All students, including homeschooled students, must be enrolled or registered at their testing schools in order to take the online, paper-pencil, braille, standard print, or large print assessments. Student enrollment information, including demographic data, is generated using an OSPI file and uploaded nightly via a secured file transfer site to the online test registration system during the testing period.

Students log in to the online assessment using their legal first name, their SSID number, and a test session ID. Only students are permitted to log in to an online test session. TAs, proctors, or other personnel are not permitted to log in to the system on behalf of students, although they are permitted to assist students who need help logging in. For the paper-pencil versions of the assessments, TAs are required to ensure that the student pre-ID label is affixed to the student's answer document (ELA and mathematics) or test booklet (WCAS).

5.4.2 System Security

The objective of system security is to ensure that all data are protected and accessed correctly by the appropriate user groups. It is about protecting data and maintaining data and system integrity

at all times, including ensuring that all personal information is secured, that transferred data (whether sent or received) are not altered in any way, that the data source is known, and that any service can be performed only by a specific, designated user.

A hierarchy of control

As described in Section 5.2 Test Administration, district personnel, SCs, and TAs share defined roles and levels of access to the testing system. When TIDE opens for the school year, CAI rolls over user accounts from the prior school year and resets all passwords. DCs are responsible for selecting and entering new SCs' information into TIDE, and the SC is responsible for entering new TAs' information into TIDE. Throughout the year, DCs are also expected to delete from TIDE information for any staff members who have transferred to other schools, resigned, or no longer serve as SCs or TAs.

Password protection

All access points for different roles—at the state, district, school principal, and school staff levels—require a password to log in to the system. Newly added SCs and TAs receive separate passwords through their personal email addresses assigned by the school.

Secure browser

A role of the Technology Coordinator is to ensure that the CAI Student Secure Browser is properly installed on the computers used for administration of the online assessments. Developed by the testing contractor, the Student Secure Browser prevents students from accessing other computers or Internet applications and from copying test information. The Student Secure Browser suppresses access to commonly used browsers such as Internet Explorer, Edge, Chrome, and Firefox, and prevents students from searching for answers on the Internet or communicating with other students. The summative assessments can be accessed only through the Student Secure Browser and not through other Internet browsers.

5.4.3 Security of the Testing Environment

The SCs and TAs work together to determine appropriate testing schedules based on the number of computers available, the number of students in each tested grade, and the average amount of time needed to complete each assessment.

Testing personnel are instructed in the online training and user manuals that assessments should be administered in testing rooms that do not crowd students. Good lighting, ventilation, and freedom from noise and interruptions are important factors to be considered when selecting testing rooms.

TAs must establish procedures to maintain a quiet environment during each test session, recognizing that some students may finish more quickly than others. If students are allowed to leave the testing room when they finish, TAs are required to explain the procedures for leaving without disrupting others and to tell students where they are expected to report once they leave. If students are expected to remain in the testing room until the end of the session, TAs are encouraged to prepare some quiet work for students to do after they finish the assessment.

If a student needs to leave the room briefly, the TA is required to pause the student's assessment. As described in section 5.2.2 Online Administration, the 20-minute pause rule was implemented to prevent students from using the time to look up answers.

Room Preparation

Instructional materials for math, English language arts (ELA), and science content within the testing location must be removed or covered. This includes, but is not limited to vocabulary lists, definitions, maps, scientific cycles, mathematics formulas, graphic organizers, problem-solving strategies, etc. displayed on wall charts, students' desks, bulletin boards, nametags, chalkboards, dry-erase boards, or on posters as these might assist students in answering questions. These materials may invalidate students test results.

Materials related to social emotional learning do not need to be removed or covered. This includes, but is not limited to, resources related to emotional regulation, management, self-awareness, or coping; multiple intelligences or learning mindset; classroom behavior expectations or social contracts; feelings; etc.

The cell phones of students must be turned off and stored in the testing room out of sight. TAs are encouraged to minimize access to the testing rooms by posting signs in halls and entrances in order to promote optimum testing conditions; they should also post "Testing—Do Not Disturb" signs on the doors of testing rooms.

Seating Arrangements

TAs obtain the student seating chart from the SC. TAs should provide adequate spacing between students' seats. Because the WCAS is a fixed-form test all students will see the same test questions as other students in the same grade level. Students should be seated so that they will not be tempted to look at the answers of others. Because the online Smarter Balanced CAT is adaptive, it is unlikely that students will see the same test questions as another student. For the PTs, different forms are spiraled within a classroom so that students receive different performance tasks. As with the WCAS, students should be discouraged from communicating during the ELA and mathematics tests through appropriate seating arrangements.

After the Test

The TA must collect and account for any ancillary materials (scratch paper, test tickets, printed reading passages, and questions for any content-area assessment [from students using the print on demand accommodation], etc.) that were provided to students prior to the release of students from the testing location. TAs follow the school's *Test Security and Building Plan* for returning materials to the SC to be securely shredded immediately or stored in a locked area if they are to be used again.

Specific instructions for pencil-paper processing are provided in the *Accommodated Test Administration Training* presentation on how to package and secure the test booklets to be returned to the testing contractor's office.

5.4.4 Test Security Violations

Everyone who administers or proctors the assessments is responsible for understanding the security procedures. Prohibited practices, as detailed in the *Professional Standards and Security, Incidents, and Reporting Guidelines*, are categorized into three groups:

1. Impropropriety. This is an unusual circumstance that has a low impact on the individual or group of students who are testing and has a low risk of potentially affecting student performance on the test, test security, or test validity. These circumstances can be corrected at the local level and are not required to be submitted to the SEA if no impact to student performance or test security is noted. (Examples: Student[s] misconduct distracting the test session, fire drill during test session, cell phone rings from secured location.)

2. Irregularity: This is an unusual circumstance that impacts an individual or group of students who are testing and may potentially affect student performance on the test or test validity. These circumstances can likely be corrected at the local level. A Test Incident report is required to be entered into the Assessment Reporting Management System (ARMS) in the Educational Data System (EDS) and submitted to the SEA for review. (Examples: Student[s] accessing or using unauthorized material or electronic equipment during testing, student[s] left unattended during a test session, TA assistance outside of administration protocols, Student tested under another student's login.)

3. Breach: This is any test administration event that poses a threat to the validity of the test. A breach in state testing requires immediate attention. A Test Incident report is required to be entered into ARMS and submitted to the SEA for review. (Examples: Test content left unsecured, test content or student responses being reviewed, retained or shared with other persons or in social media, adults modifying student answers.)

5.5 STUDENT PARTICIPATION

All students (including retained students) currently enrolled in grades 3–8 and high school at public schools in Washington are expected to participate in the Smarter Balanced assessments. All students in grades 5, 8, and 11 are expected to participate in the WCAS assessments.

5.5.1 Homeschooled Students

Students who are homeschooled may participate in the WCAP assessments at the request of their parent or guardian. Districts should have a plan for dealing with these requests and providing these students with testing opportunities.

5.5.2 Exempt Students

The following students are exempt from participating in the Smarter Balanced assessment:

- A student who has a significant medical emergency
- A multilingual learner (ML) student who moved to the country within the year (ELA exemption only)

5.6 UNIVERSAL TOOLS, DESIGNATED SUPPORTS, AND ACCOMMODATIONS

Washington has adopted the Usability, Accessibility, and Accommodations Guidelines (UAAG) created by the Smarter Balanced Assessment Consortium, while enacting state-specific adjustments particular to the WCAS exams (paper-pencil and/or online). The result is the development of the *Guidelines on Tools, Supports, and Accommodations (Guidelines)* for state assessments. The information in the Guidelines is intended for district- and school-level personnel and decision-making teams, including IEP, Section 504 Plan, and multilingual learner (ML) teams, to use in preparing for and implementing the Smarter Balanced assessments and the WCAS.

The *Guidelines* provides information for classroom teachers, multilingual learner/English language development educators, special education teachers, and instructional assistants to use in selecting and administering universal tools, designated supports, and accommodations for students who need them. The *Guidelines* is also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment. The *Guidelines* is available at the WCAP portal at <https://wa.portal.cambiumast.com/resources/wa-guidelines/guidelines-on-tools-supports-and-accommodations-for-state-assessments>.

The *Guidelines* applies to all students. It emphasizes an individualized approach to the implementation of assessment practices for those students who have diverse needs and participate in large-scale content assessments. The *Guidelines* focuses on universal tools, designated supports, and accommodations for Washington’s assessments. At the same time, the *Guidelines* supports important instructional information about the connection between accessibility and accommodations for students who participate in the assessments.

Table 5.6 lists the summary of universal tools, designated supports, and accommodations used for Smarter Balanced assessments and the WCAS. As shown in the table, the embedded resources are part of the Student Secure Browser, whereas non-embedded resources are provided outside of the Student Secure Browser. In addition, some resources are available for paper-pencil tests only.

Universal tools are provided to all students who choose to use them based on student preference. A DC, DA, or SC can deactivate some of the preselected universal tools in TIDE for a student who may be distracted by the ability to access a specific tool during a test session. Designated supports are features that are available for use by any student for whom the need has been indicated by a team of educators with parent/guardian and student input. Accommodations are available for students for whom there is documentation of the need on an Individualized Education Program (IEP), 504 Plan, or other similar learning plans. State-level users, DCs, DAs, and SCs have the ability to set designated supports and accommodations based on their specific user role. Designated supports and accommodations must be set in TIDE before a student begins testing. Table 5.6 shows a complete list of all accessibility supports available. Accessibility supports vary by test type, grade level, and the content being assessed. See the Guidelines for detailed information about each feature listed in table 5.6.

Table 5.6: Summary of Smarter Balanced and WCAS Tools, Supports, and Accommodations

	Universal Tools	Designated Supports	Accommodations
Embedded	Breaks Calculator Digital Notepad English Dictionary English Glossary Expandable Items Expandable Passages Global Notes Highlighter Keyboard Navigation Line Reader Mark for Review Periodic Table Spell Check Strikethrough Thesaurus Zoom—Student Level Zoom—Test Level	Color Contrast Dual Language Spanish Translations Test (includes translated directions) Hybrid Masking Tool Illustration Glossaries Masking Mouse Pointer Streamline Text-to-Speech (student responses) Text-to-Speech (test content) Translated Test Directions (Spanish only) Translations Glossaries Zoom Test Level with Streamline	American Sign Language Audio Transcriptions Braille Closed Captioning Emboss Permissive Mode Print on Demand Speech-to-Text Text-to-Speech (test content)
Non-Embedded	Breaks English Dictionary Periodic Table Scratch and/or Graph Paper Technological Assistance with Test Navigation Thesaurus	Amplification Bilingual Dictionary Color Contrast Color Overlays Illustration Glossaries Magnification Device Medical Supports Noise Buffers Read Aloud in English Read Aloud in Spanish Read Aloud Student Scribe Separate Setting Simplified Test Directions Translated Test Directions	100s Number Table Abacus Alternate Response Options American Sign Language Braille Test Booklet Calculator Large Print Test Booklet Multiplication Table Read Aloud in English Scribe Spanish Test Booklet Speech-to-Text Standard Print Test Booklet Translation Glossaries for Paper Testing Word Prediction

5.6.1 Universal Tools for All Students

Universal tools are provided to all students by default, and students choose when to use them based on student preference. Universal tools are accessibility features and resources of the assessment that are either provided as digitally delivered embedded components within the Test Delivery System (TDS), or outside of TDS as non-embedded, which can support computer-based or accommodated form (paper) testing. In the spring 2022 test administration, the following features

of universal tools were available for *all* students to access. For specific information on how to access and use these features, refer to the *Guidelines on Tools, Supports, and Accommodations*.

Embedded

Breaks (subject to pause rules): The number of items per session can be flexibly defined based on the student's need. There is no limit on the number of breaks that a student might be given. Available for ELA, mathematics, and science tests.

Pause Rules: Best practice for pausing during the WCAS and CAT portion of the Smarter Balanced tests is for students to finish all questions on the page and then click the pause button; students should not click the next button to move to the next page of questions. The *TA Script of Student Directions* contains specific instructions for the TA to give students when a pause is needed.

- If the WCAS or the CAT portion of a Smarter Balanced Test is paused for less than 20 minutes, the student can return to previous test pages and change the response to any questions the student has already answered within that segment (with the exception of locking items in the WCAS). The student may not return to a previous segment.
- If the WCAS or the CAT portion of a Smarter Balanced Test is paused for more than 20 minutes, the student will log back in and see the next test page with unanswered questions. Students will not be able to return to any previous pages or questions they answered before pausing their test, even if they marked questions for review.

Calculator: This tool is an embedded, on-screen digital calculator can be accessed for calculator-allowed items when students click the calculator tool button. This tool is available for calculator-allowed items in grades 6-8 and high school mathematics and all questions in grades 5, 8, and 11 science tests.

Digital Notepad: This tool is used for making notes about an item. The digital notepad is item-specific and is available through the end of the test segment. Notes are not saved when the student moves on to the next segment or after a break of more than 20 minutes. This tool is available for ELA, mathematics, and science tests.

English Dictionary: An English dictionary is available for the full-write portion of an ELA performance task. A full-write segment is part 2 of a performance task.

English Glossary: Grade- and context-appropriate definitions of specific, construct-irrelevant terms are shown on the screen via a pop-up window. Terms are pre-selected and indicated throughout the tests by a gray dotted outline. This tool is available for ELA, mathematics, and science tests.

Expandable Item and Passages: These allow the student to expand each stimulus or item so that it takes up a larger portion of the screen as the student reads. The student can then retract the screen to its original size. A student has the ability to change the screen display from the default of 40% stimulus and 60% item to 5% stimulus and 95% item or 95% stimulus and 5% item. This tool is available for ELA, mathematics, and science tests.

Global Notes: This is a notepad that is available for ELA performance tasks in which students complete a full write. The student clicks the notepad icon for the notepad to appear. During the ELA performance tasks, the notes are retained from segment to segment so that the student may go back to the notes even though he or she cannot go back to specific items in the previous segment.

Highlighter: This allows the student to highlight passages or sections of passages and test questions. This tool is available for ELA, mathematics, and science tests.

Keyboard Navigation: A student can navigate through the test using a keyboard instead of a mouse or touch screen. This tool is available for ELA, mathematics, and science tests.

Line Reader: This tool assists in reading by highlighting a single line of text in a stimulus or question. This tool is available for ELA, mathematics, and science tests.

Mark for Review: A student can mark a question for review to return to it later. However, markings are not saved when a student moves on to the next segment or after pausing the test for more than 20 minutes. This tool is available for ELA, mathematics, and science tests.

Periodic Table: An embedded onscreen periodic table can be accessed for permitted items when students click on the periodic table tool button. This tool is available for the grades 8 and 11 science tests.

Spell Check: A writing tool for checking the spelling of words in student responses. Spell check only highlights misspelled words; it does not provide the correct spelling. This tool is available for ELA, mathematics, and science tests.

Strikethrough: This allows the student to strike through answer options for selected-response items. This tool is available for ELA, mathematics, and science tests.

Zoom Student Level and Zoom Test Level: These are tools that allow either the student to zoom in on an individual item or the entire test to be enlarged on test questions, text, or graphics. These tools are available for ELA, mathematics, and science tests.

Non-Embedded

Breaks: Breaks may be given at predetermined intervals or after completion of sections of the assessment for students taking a paper-based test. Individual students are sometimes allowed to take breaks to address cognitive fatigue if they are experiencing heavy assessment demands. Available for ELA, mathematics, and science tests.

English Dictionary: An English dictionary can be provided for the ELA performance task, part 2 full write. A full write is the second part of a performance task. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

Periodic Table: For grades 8 and 11. A printable version of the periodic table is delivered with the accommodated paper test materials. This tool is available for science test.

Scratch and/or Graph Paper: Students may use blank scratch paper to make notes, write computations, record responses, or create graphic organizers.

ELA: Plain or lined scratch paper, whiteboards with markers to make notes or plan responses may be made available. Graph paper is not permitted.

Math and science: Plain or lined paper, graph paper, or whiteboard with a marker may be used on all math and science assessments. Graph paper is required for math in grades 6–8 and HS.

Assistive Technology (AT) Devices: If the construct being measured is not impacted, AT devices, including low-tech AT (Math Window) are permitted to make notes, including the use of digital graph paper. The AT device needs to be familiar to the student and/or consistent with the IEP or 504 Plan. Access to internet must be disabled on AT devices. Permissive mode may be required to support AT devices.

ELA/math CAT: If a student needs to take the CAT in more than one session, scratch paper, whiteboards, and/or AT devices must be collected at the end of each session, securely stored, and made available to the student at the start of the next CAT testing session. Once the student completes the CAT, the scratch paper must be collected and securely destroyed, whiteboards should be erased, and notes on AT devices erased to maintain test security.

Science: If a student needs to take the WCAS in more than one session, scratch paper, whiteboards, and/or AT devices must be collected at the end of each session, securely stored, and made available to the student at the start of the next WCAS testing session. Once the student completes the WCAS, the scratch paper must be collected and securely destroyed, whiteboards should be erased, and notes on AT devices erased to maintain test security.

ELA/math Performance Tasks: If a student needs to take the performance task in more than one session, scratch paper, whiteboards, and/or AT devices must be collected at the end of each session, securely stored, and made available to the student at the start of the next performance task testing session. Once the student completes the performance task, the scratch paper must be collected and securely destroyed, whiteboards should be erased, and notes on AT devices erased to maintain test security.

Spanish Periodic Table: For grades 8 and 11. A printable version of the Spanish periodic table is delivered with the Spanish translated paper test materials. This tool is available for science test.

Technological Assistance with Navigation: Students without the necessary computer skills may have a trained TA help with mouse point-and-click and drag-and-drop items, onscreen tool and button navigation (e.g., back, next, submit, start, stop), and keyboarding. TA assistance does not include identifying correct tool buttons. The TA is allowed to assist only with the technology as indicated by the student and must never assist with actual answer responses. Choosing answers for a student is a test incident and will result in an invalid assessment. This tool is available for the science test.

Thesaurus: A thesaurus provides synonyms of terms while a student interacts with text included in the assessment. A thesaurus can be provided for the ELA performance task, part 2 full write. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

5.6.2 Designated Supports

Designated supports for assessments are those features that are available for use by any student for whom the need has been indicated by an educator (or team of educators) with parent/guardian and student input. Approved designated supports do not compromise the learning expectations,

construct, grade-level standard, or intended outcome of the assessments Scores achieved by students using designated supports will be included for federal accountability purposes. It is recommended that a consistent process be used to determine these supports for individual students. All educators making these decisions should be trained in the process and should understand the range of designated supports available. OSPI has identified digitally embedded and non-embedded designated supports for students for whom an adult or team has indicated a need for the support. For specific information on how to access and use these features, refer to the *Guidelines on Tools, Supports, and Accommodations* The following are the designated supports:

Embedded

Color Contrast: This support allows the screen background or font color to be set. This may include reversing the colors for the entire interface or choosing the color of font and background. Black on white, reverse contrast, black on rose, medium gray on light gray, and yellow on blue were offered for the online assessments. This support is available for ELA, mathematics, and science tests.

Dual language Spanish Translations Test: This support provides the full Spanish translation of each test item above the original item in English. Students taking the Spanish math and science tests may respond to items in English, Spanish, or a combination of both. This support also provides Spanish translation of test directions prior to beginning the actual test items. This support is available for mathematics and science tests.

Hybrid Masking Tool: This support assists in reading by showing a single line of text in a stimulus or question, while masking the rest of the content on the screen. When the line reader button is selected, use of the arrow keys will move the visible line up and down through the text. This support is available for ELA, mathematics, and science tests.

Illustration Glossaries: This support is provided for selected construct-irrelevant terms for math items. Illustrations for these terms appear on the computer screen when students select the term. Students can also adjust the size of the illustration and move it around the screen. This support is available for the mathematics test.

Masking: This support allows the student to block off content that is not of immediate need or that may be distracting. Students are able to focus their attention on a specific part of a test item by masking. Masking allows students to hide and reveal individual answer options, as well as all navigational buttons and menus. This support is available for ELA, mathematics, and science tests.

Mouse Pointer: This support allows the mouse pointer to be set to a larger size and also for the color to be changed. This support is available for ELA, mathematics, and science tests.

Streamlined Interface Mode: This support provides a streamlined interface of the entire test in an alternate, simplified format in which items are displayed below the stimuli. This support is available for ELA, mathematics, and science tests.

Text-to-Speech (student responses): This support reads aloud the text the student entered via embedded text-to-speech technology when they select the speaker button at the top of the response box. This support is available for ELA, mathematics, and science tests.

Text-to-Speech (test content): This support reads aloud items and/or stimuli to the student via embedded text-to-speech technology. The student can control the speed and raise or lower the volume of the voice via a volume control. This support is available for ELA, mathematics, and science tests.

Translated Test Directions: Spanish translation of test directions for the online tests is a language support available prior to beginning the actual test items. This support is available for mathematics and science tests.

Translations (Glossaries): This support is a language support, provided for selected construct-irrelevant terms. Translations for these terms appear on the computer screen when the student clicks on the word or term. Students can also select the audio icon next to the glossary term and listen to a recording of the glossary, when available. The following language glossaries were offered: Arabic, Burmese, Cantonese, Filipino, Hmong, Korean, Mandarin, Punjabi, Russian, Somali, Spanish, Ukrainian, and Vietnamese. This support is available for mathematics and science tests.

Zoom Test Level with Streamline: This support allows the test platform to be pre-set to be enlarged more than the 3x level available as a universal tool. Test level zoom increases the text and graphics for the entire test to the setting indicated in TIDE. Use of zoom levels 5x–20x also require the streamlined interface mode which arranges the test content vertically. This support is available for ELA, mathematics, and science tests.

Non-Embedded

Amplification: This support allows students to use amplification assistive technology to adjust the volume control beyond the computer’s built in settings. This support is available for ELA, mathematics, and science tests.

Bilingual Dictionary: A bilingual/dual language word-to-word dictionary can be provided for the full-write portion of an ELA performance task.

Color Contrast: This support allows test content of online items to be printed with different colors using print on demand. This support is available for ELA, mathematics, and science tests.

Color Overlays: This support allows color transparencies to be placed over the paper-based assessment. This support is available for ELA, mathematics, and science tests.

Illustration Glossaries: This support provides students grade- and construct-irrelevant images for terms are provided in a supplement to the paper-pencil test and are identified by item number. This support is available for mathematics only.

Magnification Device: This support allows the size of specific areas of the screen (e.g., text, formulas, tables, graphics, navigation buttons) to be adjusted by the student with an assistive technology device or software. Magnification allows increasing the size to a level not provided for by the zoom universal tool. This support is available for ELA, mathematics, and science tests.

Medical Supports: This support allows students to access medical supports for medical purposes. This support is available for ELA, mathematics, and science tests.

Noise Buffers: This support is used to reduce environmental noise and may include ear mufflers, white noise, and/or other equipment. This support is available for ELA, mathematics, and science tests.

Read Aloud in English: This support allows text to be read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Read Aloud Guidelines for Washington State Assessments*. This support is available for ELA, mathematics, and science tests.

Read Aloud in Spanish: This support allows Spanish text to be read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Read Aloud Guidelines for Washington State Assessments*. All or portions of the content may be read aloud. This support is available for mathematics and science tests.

Read Aloud Student: This support allows the student to read the test content out loud to themselves. All or portions of the content may be read aloud. This support is available for ELA, mathematics, and science tests.

Scribe (all items except ELA full-write items): This support allows students to dictate their responses to a human scribe who records verbatim what they dictate. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Scribing Protocol for Washington State Assessments*. This support is available for ELA, mathematics, and science tests.

Separate Setting: This support allows the test location to be altered so that the student is tested in a setting different from what is available for most students. This support is available for ELA, mathematics, and science tests.

Simplified Test Directions: This support allows The TA to simplify or paraphrases the test directions found in the appropriate *TA Script of Student Directions* following the directions outlined in the *Guidelines for Simplified Test Directions for Washington State Assessments*. This support is available for ELA, mathematics, and science tests.

Translated Student Test Directions: This support allows a bilingual adult to read to student or the directions can be printed and given to students for them to read. The Translated Test Directions for Online Testing are available in fifteen languages. This support is available for ELA, mathematics, and science tests.

5.6.3 Accommodations

Accommodations are changes in procedures or materials that increase equitable access during the assessments. Assessment accommodations generate valid assessment results for students who need them; they allow these students to show what they know and can do. Accommodations are available for students with documented IEPs, 504 Plan, or similar learning plans. Approved accommodations do not compromise the learning expectations, construct, grade-level standard, or intended outcome of the assessments. Scores achieved by students using accommodations will be

included for federal accountability purposes. For specific information on how to access and use these features, refer to the *Guidelines on Tools, Supports, and Accommodations*. The accommodations are listed in this section.

Embedded

American Sign Language (ASL): This accommodation allows test content to be translated into ASL video. An ASL human signer and the signed test content are viewed on the same screen. Students may view portions of the ASL video as often as needed. This accommodation is available for ELA and mathematics tests.

Braille: This accommodation is a raised-dot code that individuals read with their fingertips. Graphic material (e.g., maps, charts, graphs, diagrams, illustrations) is presented in a raised format (paper or thermoform). This accommodation is available for ELA and mathematics tests.

Online Braille tests have additional features available:

Audio Transcriptions: For the ELA listening stimuli, a braille transcript of the audio of the listening passages is available for use with refreshable braille interfaces.

Emboss: Allows braille to be presented via embosser.

Closed Captioning: This accommodation is printed text that appears on the computer screen as audio materials are presented. This accommodation is available for ELA tests.

Permissive Mode: This accommodation allows assistive technology devices and software to be used with the secure browser. This accommodation is available for ELA, mathematics, and science tests.

Print-on-Demand: This accommodation allows the student to use paper copies of individual test items printed from the Test Delivery System (TDS). The student requests the printing from within the secure browser and the TA prints the materials from the TA Interface. The student or a scribe enters student answers to items into the TDS. This accommodation is available for ELA, mathematics, and science tests.

Speech-to-Text: This accommodation supports dictation of student responses to test questions. This accommodation is available for ELA, mathematics, and science tests.

Speech-to-Text Language: This accommodation supports dictation of student responses to test questions in Spanish. This accommodation is available for mathematics and science tests.

Text-to-Speech (test content): This accommodation allows passage text to be read aloud to the student via embedded text-to-speech technology. The student can control the speed and raise or lower the volume of the voice via a volume control. This accommodation is available for the ELA test.

Non-Embedded

100s Number Table: This accommodation allows students to use the paper-based table listing numbers from 1–100 published by Smarter Balanced. This accommodation is available for the mathematics test.

Abacus: This accommodation may be used in place of scratch paper for students who typically use an abacus. This accommodation is available for the mathematics and science tests.

Alternate Response Options: This accommodation may include but is not limited to adapted keyboards, large keyboards, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches. This accommodation is available for the ELA, mathematics, and science tests.

American Sign Language: this accommodation allows test content (online or paper) to be translated by a human signer into ASL. This accommodation is available for the science test.

Braille Graphics: This accommodation allows students access to pre-embossed braille graphics for the online mathematics hybrid adaptive test (HAT). This accommodation is available for the mathematics test.

Braille Test Booklet: This accommodation provides students a test booklet with a raised-dot code they read with their fingertips. Graphic material (e.g., maps, charts, graphs, diagrams, illustrations) is presented in a raised format. This accommodation is available for the ELA, mathematics, and science tests.

Calculator: This accommodation allows a non-embedded, stand-alone calculator for students needing a special calculator, such as a braille calculator or a talking calculator. Administration directions will identify items open to calculator use. In those instances, TAs will make calculators available to students. The calculator used must be on the list of eligible devices; refer to the *Calculator and Electronic Device Policy*, available on the WCAP portal at: <https://wa.portal.cambiumast.com/resources/wa-guidelines/calculator-and-electronic-device-policy>. This accommodation is available for the mathematics and science tests.

Large Print Test Booklet: This accommodation provides students a large print paper form of the test that is provided to the student with a visual impairment. The font size for the large print form is 18 point on paper sized 11 x 17. This accommodation is available for the ELA, mathematics, and science tests.

Multiplication Table: This accommodation allows access to the paper-based multiplication table (containing numbers 1–12) published by Smarter Balanced. This accommodation is available for the mathematics test.

Read Aloud in English: This accommodation allows text to be read aloud to the student by a trained and qualified test reader who follows the *Read Aloud Guidelines for Washington State Assessments*. This accommodation is available for the ELA test.

Scribe (ELA full write items only) : Students dictate their responses to a human who records verbatim what they dictate. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Online Test Administration Manual*. This accommodation is available for the ELA test.

Spanish Print Test Booklet: This accommodation allows students access to Spanish print test materials. For Smarter Balanced mathematics, the full Spanish translation of each item is above the original item in English. For WCAS, the entire test is translated in Spanish. Students taking the Spanish math and science tests may respond to items in English, Spanish, or a combination of both. This accommodation is available for the mathematics and science tests.

Speech-to-Text: This accommodation allows students to use their voices as devices to input information into the computer to dictate responses or give commands (e.g., opening application programs, pulling down menus, saving work). Voice recognition software generally can recognize

speech at up to 160 words per minute. Students may use their own assistive technology devices. This accommodation is available for the ELA, mathematics, and science tests.

Standard Print Test Booklet: This accommodation allows students access to standard print test and answer booklets. This accommodation is available for the ELA, mathematics, and science tests.

Translations Glossaries For Paper Testing: Translated paper glossaries are provided for selected construct-irrelevant terms. Only state approved glossaries posted on the WCAP portal may be provided to students. This accommodation is available for the mathematics and science tests.

Word Prediction: This accommodation allows students to begin writing a word and choose from a list of words that have been predicted from word frequency and syntax rules. Students who have documented motor or orthopedic impairments, which severely impairs their ability to provide written or typed responses without the use of assistive technology, may use word prediction. Students may use their own assistive technology devices. This accommodation is available for the ELA, mathematics, and science tests.

5.6.4 Spring 2022 Summary

Tables 5.7–5.14 provide the number of students who utilized any of the offered designated supports and/or accommodations in the Smarter Balanced assessments. Tables 5.15–5.18 provide frequencies for students who were offered the designated supports and/or accommodations in the WCAS.

Table 5.7: Total Students with Allowed Embedded Designated Supports–ELA

Designated Supports	Student Group	Grade						
		3	4	5	6	7	8	HS
Color Contrast	All	98	112	148	95	90	56	42
	ML	20	18	26	15	8	5	2
	IDEA Eligible	24	25	32	37	40	36	29
Hybrid Masking Tool	All	4	27	42	37	6	-	5
	ML	1	7	15	13	-	-	5
	IDEA Eligible	3	14	10	11	2	-	-
Masking	All	494	597	606	629	592	504	220
	ML	143	166	165	170	132	129	49
	IDEA Eligible	244	331	343	306	294	231	191
Mouse Pointer	All	131	77	95	55	39	26	28
	ML	37	23	28	13	12	6	3
	IDEA Eligible	40	56	43	26	33	20	24
Streamline	All	349	419	418	678	766	741	419
	ML	72	91	100	154	129	139	60
	IDEA Eligible	253	297	326	408	488	490	400
Text-to-Speech (student responses)	All	5,243	5,738	5,503	4,693	4,609	4,301	3,150
	ML	1,688	1,903	1,628	1,449	1,302	1,233	808
	IDEA Eligible	2,201	2,639	2,855	2,598	2,698	2,407	1,894
Text-to-Speech (test content): CAT Items	All	13,639	12,605	11,869	7,586	6,272	5,865	3,342
	ML	5,307	4,671	4,012	2,563	1,947	1,816	1,292
	IDEA Eligible	2,639	2,596	2,701	2,106	2,010	2,008	1,480
	All	986	898	887	509	497	475	494

Designated Supports	Student Group	Grade						
		3	4	5	6	7	8	HS
Text-to-Speech (test content): PT Items	ML	312	297	269	119	119	130	107
	IDEA Eligible	306	299	319	335	354	377	313
Text-to-Speech (test content): PT Stimuli	All	164	150	137	115	112	100	148
	ML	37	37	40	43	22	24	54
	IDEA Eligible	62	77	84	87	76	86	132
Text-to-Speech (test content): PT Stimuli and Items	All	17,124	15,862	15,229	11,054	9,658	9,400	6,940
	ML	6,153	5,146	4,836	3,527	2,926	2,832	2,153
	IDEA Eligible	5,376	5,643	6,005	5,204	5,097	4,940	3,921
Zoom Test Level with Streamline	All	6	3	9	21	11	6	2
	ML	2	2	2	7	2	1	1
	IDEA Eligible	5	2	8	7	11	5	2

Table 5.8: Total Students with Allowed Non-Embedded Designated Supports–ELA

Designated Supports	Student Group	Grade						
		3	4	5	6	7	8	HS
Amplification	All	30	39	46	46	24	29	16
	ML	4	8	9	9	3	6	1
	IDEA Eligible	24	28	27	23	13	14	7
Bilingual Dictionary	All	366	279	266	241	200	206	539
	ML	358	275	252	228	192	193	381
	IDEA Eligible	41	39	31	35	31	35	80
Color Contrast	All	10	32	45	31	13	10	14
	ML	2	5	14	4	1	1	-
	IDEA Eligible	9	20	32	15	7	9	11
Color Overlays	All	7	20	18	25	12	10	13
	ML	-	1	-	1	1	1	2
	IDEA Eligible	7	16	14	11	10	8	10
Magnification Device	All	26	45	43	42	35	36	25
	ML	2	5	8	4	10	12	4
	IDEA Eligible	18	36	32	25	27	17	21
Medical Supports	All	14	31	29	40	34	42	31
	ML	1	1	1	2	1	1	4
	IDEA Eligible	7	13	8	9	7	6	6
Noise Buffers	All	353	566	578	466	420	319	226
	ML	33	91	105	53	59	39	46
	IDEA Eligible	312	443	490	374	365	266	198
Read-Aloud in English: Items	All	1,454	1,626	1,671	1,038	901	857	849
	ML	353	477	516	282	210	198	228
	IDEA Eligible	1,055	1,222	1,321	921	820	742	760
Read-Aloud in English: Passages/Stimuli and Items	All	1,443	1,494	1,545	977	846	750	803
	ML	356	402	431	240	188	174	199
	IDEA Eligible	1,025	1,154	1,252	880	779	658	736
Read-Aloud in English: Stimuli	All	239	169	156	120	122	104	161
	ML	45	36	37	25	31	34	43
	IDEA Eligible	184	140	138	115	117	92	142
Scribe (CAT)	All	613	715	738	466	343	233	145
	ML	95	125	157	76	59	42	13
	IDEA Eligible	561	685	693	433	322	220	135
Scribe (PT Segment 1)	All	645	716	693	462	339	239	144
	ML	104	132	155	76	50	42	15
	IDEA Eligible	588	677	652	424	316	226	135

Designated Supports	Student Group	Grade						
		3	4	5	6	7	8	HS
Separate Setting	All	4,538	5,394	5,933	4,881	4,748	4,709	4,601
	ML	808	1,001	1,174	805	732	730	764
	IDEA Eligible	3,624	4,302	4,703	4,059	4,041	3,858	3,895
Simplified Test Directions	All	1,404	1,738	1,639	1,229	1,151	1,062	951
	ML	384	494	495	347	362	320	319
	IDEA Eligible	990	1,248	1,235	970	918	823	772
Translated Test Directions	All	277	392	249	232	270	305	198
	ML	250	336	236	221	257	288	184
	IDEA Eligible	42	48	37	39	42	23	27

Table 5.9: Total Students with Allowed Embedded Accommodations—ELA

Accommodations	Student Group	Grade						
		3	4	5	6	7	8	HS
American Sign Language	All	18	28	31	30	25	27	39
	ML	6	6	5	9	1	1	6
	IDEA Eligible	16	26	30	29	22	25	36
Audio Transcriptions	All	1	1	5	-	2	3	3
	ML	1	1	5	-	2	3	2
	IDEA Eligible	1	-	-	-	-	-	-
Braille	All	1	-	2	1	1	1	2
	ML	1	-	-	-	-	-	-
	IDEA Eligible	-	-	2	-	1	1	2
Closed Captioning	All	55	64	82	105	117	116	177
	ML	12	12	12	31	19	17	61
	IDEA Eligible	39	46	57	69	87	83	134
Emboss	All	1	-	2	1	1	1	2
	ML	1	-	-	-	-	-	-
	IDEA Eligible	-	-	2	-	1	1	2
Permissive Mode	All	77	159	176	197	183	154	104
	ML	19	14	33	33	33	24	23
	IDEA Eligible	68	141	159	178	168	131	93
Print-on-Demand: Items	All	-	1	-	-	2	1	-
	ML	-	1	-	-	1	-	-
	IDEA Eligible	-	-	-	-	1	1	-
Print-on-Demand: Stimuli	All	-	2	-	4	1	5	-
	ML	-	-	-	-	-	2	-
	IDEA Eligible	-	2	-	4	1	5	-
Print-on-Demand: Passages/Stimuli and Items	All	55	107	110	106	101	115	121
	ML	9	15	11	25	18	11	28
	IDEA Eligible	54	100	94	99	95	106	116
Speech-to-Text	All	1,310	1,718	1,867	1,651	1,423	1,278	1,147
	ML	194	297	327	285	210	192	266
	IDEA Eligible	1,155	1,525	1,656	1,501	1,279	1,126	882
Text-to-Speech (test content): Passages	All	68	74	53	92	110	119	63
	ML	10	14	17	29	44	42	12
	IDEA Eligible	60	57	35	65	90	93	53
Text-to-Speech (test content): Passages and Items	All	4,145	4,742	5,003	4,692	4,761	4,653	4,261
	ML	1,004	1,191	1,253	1,202	1,256	1,228	1,052
	IDEA Eligible	2,985	3,502	3,844	3,665	3,689	3,581	3,041

Table 5.10: Total Students with Allowed Non-Embedded Accommodations–ELA

Accommodations	Student Group	Grade						
		3	4	5	6	7	8	HS
Alternate Response Options	All	45	66	78	56	44	38	55
	ML	5	13	17	11	7	5	8
	IDEA Eligible	45	64	75	51	44	35	54
Braille Test Booklet	All	1	2	5	1	4	4	2
	ML	1	1	1	-	1	-	2
	IDEA Eligible	1	2	5	1	4	4	2
Large Print Test Booklet	All	1	2	3	4	1	4	3
	ML	1	-	1	1	-	1	-
	IDEA Eligible	1	1	3	3	1	2	3
Read Aloud Passages (English)	All	220	251	238	221	199	177	207
	ML	40	41	43	54	48	46	37
	IDEA Eligible	205	240	219	202	188	165	199
Read Aloud Passages and Items (English)	All	791	985	1,080	870	741	685	692
	ML	146	247	276	222	180	167	195
	IDEA Eligible	748	864	938	742	629	570	629
Scribe (PT Segment 2)	All	530	557	541	312	237	160	96
	ML	93	119	117	52	41	22	12
	IDEA Eligible	491	529	516	294	226	150	94
Speech-to-Text	All	646	827	894	696	705	636	533
	ML	80	124	153	100	100	106	114
	IDEA Eligible	591	753	820	632	640	567	508
Standard Print Test Booklet	All	7	14	17	20	23	20	32
	ML	1	-	4	2	1	-	2
	IDEA Eligible	5	10	9	10	8	6	11
Word Prediction	All	284	373	395	323	349	251	243
	ML	42	48	69	63	70	50	36
	IDEA Eligible	278	362	381	298	321	235	237

Table 5.11 Total Students with Allowed Embedded Designated Supports—Mathematics

Designated Supports	Student Group	Grade						
		3	4	5	6	7	8	HS
Color Contrast	All	98	103	135	91	85	54	42
	ML	20	18	24	15	9	5	2
	IDEA Eligible	24	23	31	33	40	34	29
Dual Language Spanish Translations Test	All	905	905	838	525	462	487	521
	ML	759	711	630	387	373	387	502
	IDEA Eligible	74	49	73	25	38	21	16
Hybrid Masking Tool	All	4	28	44	36	5	-	4
	ML	1	8	15	13	-	-	4
	IDEA Eligible	3	15	10	11	1	-	-
Illustration Glossaries	All	2,150	2,016	1,804	1,250	1,256	1,144	629
	ML	1,552	1,434	1,240	872	886	826	482
	IDEA Eligible	257	311	345	242	216	199	144
Masking	All	459	582	560	623	579	498	223
	ML	135	164	156	167	126	124	47
	IDEA Eligible	233	325	329	298	282	232	197
Mouse Pointer	All	134	76	89	52	30	23	25
	ML	39	23	25	9	7	4	2
	IDEA Eligible	43	55	37	25	25	19	23
Streamline	All	353	421	422	692	771	742	480
	ML	75	98	105	158	133	149	69
	IDEA Eligible	250	292	326	416	488	477	455
Text-to-Speech (student responses)	All	5,392	5,868	5,546	4,743	4,674	4,503	3,839
	ML	1,790	2,003	1,662	1,486	1,302	1,303	879
	IDEA Eligible	2,135	2,625	2,798	2,555	2,603	2,371	1,901
Text-to-Speech (test content): Items	All	500	453	446	268	262	242	186
	ML	157	138	121	64	62	76	33
	IDEA Eligible	152	160	186	177	190	177	172
Text-to-Speech (test content): Stimuli	All	23	19	16	27	15	17	15
	ML	9	6	5	13	7	2	4
	IDEA Eligible	7	7	6	14	14	11	11
Text-to-Speech (test content): Stimuli and Items	All	18,982	18,451	17,711	12,493	11,058	10,613	8,465
	ML	6,847	6,323	5,625	3,898	3,260	3,127	2,495
	IDEA Eligible	5,799	6,205	6,581	5,684	5,541	5,467	4,579
Translation (Glossary): Spanish	All	1,764	1,681	1,632	1,349	1,166	1,156	965
	ML	1,567	1,422	1,318	1,144	1,057	1,051	914
	IDEA Eligible	194	201	211	250	222	199	202
Translation (Glossary): Other Languages*	All	420	479	384	331	315	270	249
	ML	372	382	317	299	271	243	237
	IDEA Eligible	16	41	29	26	23	28	19
Zoom Test Level with Streamline	All	6	2	9	21	13	6	4
	ML	2	1	2	7	3	1	2
	IDEA Eligible	5	1	8	5	12	6	4

* The most used language was Russian, followed by (in order) Ukrainian, Mandarin, Vietnamese, and Arabic.

Table 5.12: Total Students with Allowed Non-Embedded Designated Supports—Mathematics

Designated Supports	Student Group	Grade						
		3	4	5	6	7	8	HS
Amplification	All	30	37	45	46	23	25	20
	ML	5	6	9	9	2	6	1
	IDEA Eligible	25	28	26	23	13	12	9
Color Contrast	All	7	28	42	35	13	10	12
	ML	1	3	12	4	-	1	-
	IDEA Eligible	6	17	28	19	7	8	8
Color Overlays	All	5	21	15	28	9	8	14
	ML	-	1	-	2	-	1	2
	IDEA Eligible	5	17	12	15	8	6	11
Illustration Glossaries	All	169	189	194	80	127	108	62
	ML	143	119	118	64	107	98	48
	IDEA Eligible	38	51	56	13	25	22	12
Magnification Device	All	26	40	41	45	34	32	27
	ML	2	4	8	4	11	11	5
	IDEA Eligible	18	31	30	28	25	16	21
Medical Supports	All	13	31	30	37	32	41	37
	ML	1	1	1	1	-	1	1
	IDEA Eligible	6	13	8	7	6	6	6
Noise Buffers	All	344	556	563	467	412	316	233
	ML	36	85	100	54	56	37	41
	IDEA Eligible	305	433	480	376	358	264	203
Read-Aloud Items (English)	All	365	363	332	247	203	233	180
	ML	66	99	81	74	57	64	40
	IDEA Eligible	281	296	288	233	195	207	166
Read-Aloud Items (Spanish)	All	9	40	19	18	12	22	9
	ML	8	33	19	14	10	18	6
	IDEA Eligible	2	3	2	7	3	4	4
Read-Aloud Stimuli (English)	All	183	175	139	100	111	97	120
	ML	36	34	30	24	29	32	28
	IDEA Eligible	129	145	128	92	106	85	112
Read-Aloud Stimuli (Spanish)	All	9	37	14	12	16	16	9
	ML	7	29	13	9	10	14	6
	IDEA Eligible	4	4	3	2	6	2	4
Read-Aloud Stimuli and Items (English)	All	1,492	1,708	1,802	1,120	983	896	953
	ML	379	505	547	308	258	246	240
	IDEA Eligible	1,059	1,263	1,379	924	812	702	837
Read-Aloud Stimuli and Items (Spanish)	All	86	174	128	120	114	123	61
	ML	63	133	100	104	95	110	44
	IDEA Eligible	28	35	39	39	38	19	23
Scribe Items	All	618	709	706	481	341	245	146
	ML	104	133	155	81	48	39	13
	IDEA Eligible	562	669	658	447	319	228	132
Separate Setting	All	4,515	5,398	5,945	4,862	4,774	4,728	5,086
	ML	836	1,038	1,188	806	756	765	800
	IDEA Eligible	3,586	4,264	4,673	4,044	4,027	3,826	4,257

Designated Supports	Student Group	Grade						
		3	4	5	6	7	8	HS
Simplified Test Directions	All	1,419	1,720	1,626	1,251	1,152	1,079	1,002
	ML	402	497	508	369	377	336	331
	IDEA Eligible	993	1,215	1,212	965	885	816	817
Translated Test Directions	All	341	478	324	266	344	367	236
	ML	308	419	309	249	318	344	223
	IDEA Eligible	38	50	45	38	45	29	28

Table 5.13: Total Students with Allowed Embedded Accommodations—Mathematics

Accommodations	Student Group	Grade						
		3	4	5	6	7	8	HS
American Sign Language	All	19	27	31	29	24	25	41
	ML	7	5	5	9	1	1	9
	IDEA Eligible	17	25	30	28	21	24	38
Braille	All	-	-	1	-	1	2	2
	ML	-	-	-	-	-	-	1
	IDEA Eligible	-	-	1	-	1	2	2
Emboss	All	-	-	1	-	1	2	2
	ML	-	-	-	-	-	-	1
	IDEA Eligible	-	-	1	-	1	2	2
Permissive Mode	All	78	158	166	187	185	151	123
	ML	18	14	32	37	37	25	25
	IDEA Eligible	67	139	152	161	163	127	112
Print-on-Demand: Items	All	1	-	-	-	1	-	1
	ML	-	-	-	-	-	-	-
	IDEA Eligible	1	-	-	-	-	-	1
Print-on-Demand: Stimuli	All	-	-	-	2	1	4	1
	ML	-	-	-	-	-	1	-
	IDEA Eligible	-	-	-	2	1	4	1
Print-on-Demand: Stimuli and Items	All	43	96	87	92	90	100	134
	ML	12	17	10	24	17	9	26
	IDEA Eligible	40	86	74	86	83	91	125
Speech-to-Text	All	1,222	1,614	1,732	1,508	1,315	1,193	1,139
	ML	196	296	316	265	205	196	251
	IDEA Eligible	1,076	1,421	1,527	1,372	1,169	1,037	874
Speech-to-Text Language	All	22	45	46	51	52	60	91
	ML	20	43	44	50	50	59	85
	IDEA Eligible	6	7	11	19	5	7	12

Table 5.14: Total Students with Allowed Non-Embedded Accommodations–Mathematics

Accommodations	Student Group	Grade						
		3	4	5	6	7	8	HS
100s Number Table	All	1,113	1,432	1,606	1,094	792	641	399
	ML	250	342	454	316	220	217	128
	IDEA Eligible	1,063	1,383	1,546	1,057	768	624	391
Abacus	All	80	48	51	52	24	14	1
	ML	6	5	13	3	4	5	1
	IDEA Eligible	77	48	50	50	24	14	1
Alternate Response Options	All	45	55	64	52	37	25	43
	ML	4	11	13	9	6	4	6
	IDEA Eligible	45	55	61	47	39	23	42
Braille Graphics	All	1	-	1	2	1	2	1
	ML	1	-	-	-	-	-	-
	IDEA Eligible	1	-	1	2	1	2	1
Braille Test Booklet	All	2	2	4	2	3	5	7
	ML	2	1	1	-	-	-	2
	IDEA Eligible	2	2	4	2	3	5	7
Calculator	All	185	185	384	1,092	1,550	1,853	2,322
	ML	33	39	100	244	354	399	499
	IDEA Eligible	169	174	360	1,062	1,493	1,783	2,261
Large Print Test Booklet	All	1	2	3	4	2	5	2
	ML	1	-	1	1	1	1	-
	IDEA Eligible	1	1	3	3	2	2	2
Multiplication Table	All	1,343	2,436	3,528	3,597	3,721	3,595	2,441
	ML	274	502	810	828	834	800	513
	IDEA Eligible	1,275	2,329	3,396	3,484	3,599	3,468	2,375
Spanish Print Test Booklet	All	-	-	-	1	2	1	-
	ML	-	-	-	1	2	1	-
	IDEA Eligible	-	-	-	-	-	-	-
Speech-to-Text	All	514	678	767	578	610	568	520
	ML	71	98	136	83	87	98	110
	IDEA Eligible	476	625	707	522	553	500	492
Standard Print Test Booklet	All	8	14	20	24	21	22	32
	ML	1	-	4	2	1	-	3
	IDEA Eligible	6	10	10	10	7	6	13
Word Prediction	All	225	310	343	259	280	208	230
	ML	39	40	63	54	63	43	37
	IDEA Eligible	221	303	330	234	254	193	223

Table 5.15: Total Students with Allowed Embedded Designated Supports–WCAS

Designated Supports	Student Group	Grade		
		5	8	11
Color Choices	All	52	50	13
	ML	5	4	1
	IDEA Eligible	13	32	9
Dual Language Spanish Translation Test	All	501	429	244
	ML	396	357	235
	IDEA Eligible	30	16	9
Masking	All	449	450	76
	ML	126	104	20
	IDEA Eligible	245	197	64
Hybrid Masking (Enhanced Line Reader)	All	28	-	1
	ML	8	-	-
	IDEA Eligible	5	-	1
Mouse Pointer	All	51	18	11
	ML	10	3	1
	IDEA Eligible	23	13	8
Streamline	All	361	630	209
	ML	82	118	30
	IDEA Eligible	271	384	192
Text-to-Speech (test content): Items	All	430	206	50
	ML	122	66	12
	IDEA Eligible	192	150	46
Text-to-Speech (test content): Stimuli	All	20	15	8
	ML	5	2	3
	IDEA Eligible	4	7	6
Text-to-Speech (test content): Stimuli & Items	All	15,385	9,320	3,979
	ML	5,033	2,850	1,304
	IDEA Eligible	5,885	4,805	2,053
Text-to-Speech (test content): Student Responses	All	5,028	4,095	1,668
	ML	1,530	1,199	407
	IDEA Eligible	2,484	2,072	850
Zoom Test Level with Streamline (5X, 10X, 20X)	All	6	4	3
	ML	1	1	-
	IDEA Eligible	4	2	2
Translation (Glossary) Spanish	All	1,211	1,061	476
	ML	1,048	965	438
	IDEA Eligible	165	179	98
Translation (Glossary) Other Language	All	302	238	126
	ML	269	215	116
	IDEA Eligible	25	24	12

Table 5.16: Total Students with Allowed Non-Embedded Designated Supports–WCAS

Designated Supports	Student Group	Grade		
		5	8	11
Amplification	All	27	25	7
	ML	5	5	1
	IDEA Eligible	18	13	4
Color Contrast	All	20	5	6
	ML	2	-	1
	IDEA Eligible	14	4	6
Color Overlay	All	10	5	19
	ML	-	-	1
	IDEA Eligible	8	4	8
Magnification Device	All	30	29	11
	ML	6	11	1
	IDEA Eligible	26	13	6
Noise Buffers	All	409	260	94
	ML	67	27	18
	IDEA Eligible	342	215	72
Read Aloud: Items (English)	All	226	167	64
	ML	66	52	13
	IDEA Eligible	196	141	56
Read Aloud: Items (Spanish)	All	15	19	3
	ML	12	17	2
	IDEA Eligible	4	3	-
Read Aloud: Stimuli (English)	All	87	101	34
	ML	14	33	6
	IDEA Eligible	79	91	30
Read Aloud: Stimuli (Spanish)	All	8	16	4
	ML	6	14	4
	IDEA Eligible	4	2	-
Read Aloud: Stimuli & Items (English)	All	1,345	619	460
	ML	389	163	113
	IDEA Eligible	1,071	547	408
Read Aloud: Stimuli & Items (Spanish)	All	188	53	36
	ML	111	42	32
	IDEA Eligible	53	8	6
Scribe Items	All	569	167	64
	ML	130	33	6
	IDEA Eligible	529	151	56
Separate Setting	All	4,655	3,795	2,286
	ML	950	606	414
	IDEA Eligible	3,685	3,089	1,806
Simplified Test Directions	All	1,307	861	394
	ML	420	246	158
	IDEA Eligible	983	679	289
Translated Test Directions	All	354	335	93
	ML	299	320	85
	IDEA Eligible	48	27	14

Table 5.17: Total Students with Allowed Embedded Accommodations–WCAS

Accommodations	Student Group	Grade		
		5	8	11
Permissive Mode	All	125	143	34
	ML	27	25	6
	IDEA Eligible	112	122	30
Print-on-Demand: Stimuli	All	-	4	1
	ML	-	1	-
	IDEA Eligible	-	4	1
Print-on-Demand: Stimuli & Items	All	73	84	70
	ML	9	5	15
	IDEA Eligible	64	78	66
Speech-to-Text	All	1517	1012	489
	ML	258	155	121
	IDEA Eligible	1359	878	297
Speech-to-Text Language	All	41	44	30
	ML	38	42	24
	IDEA Eligible	9	5	4

Table 5.18: Total Students with Allowed Non-Embedded Accommodations–WCAS

Accommodations	Student Group	Grade		
		5	8	11
Abacus	All	22	8	1
	ML	4	3	-
	IDEA Eligible	22	8	1
Alternate Response Options	All	52	22	30
	ML	14	4	3
	IDEA Eligible	51	19	30
American Sign Language	All	5	5	15
	ML	1	-	3
	IDEA Eligible	5	5	13
Calculator	All	674	477	234
	ML	126	90	52
	IDEA Eligible	625	411	215
Speech-to-Text	All	270	178	94
	ML	48	45	30
	IDEA Eligible	265	162	88
Word Prediction	All	22	8	1
	ML	4	3	-
	IDEA Eligible	22	8	1

5.7 DATA FORENSICS PROGRAM

The validity of test scores critically depends on the integrity of test administration. Any irregularities in test administration could cast doubt on the validity of the inferences based on those test scores. Multiple factors ensure that tests are administered properly, such as clear test administration policies, effective TA training, and tools to identify possible irregularities in test administrations.

For online administrations, a set of quality assurance (QA) reports is generated during and after the testing window. One of the QA reports focuses on flagging possible testing anomalies. Testing anomalies are analyzed by examining changes in student performance from year to year, test-taking time, item response patterns using a person-fit index, and item response change analyses.

Analyses are performed at the student level and summarized for each aggregate unit, including the testing session, TA, and school. The flagging criteria used for these analyses are described in the next section and are configurable by an authorized user. When the aggregate unit size is small, the aggregate unit is flagged if the percentage of flagged students is greater than 50% in the analysis. The default small aggregate unit size is five or fewer students, but this value is configurable. For each aggregate unit, small groups are identified based on the number of tests included in the aggregate unit from that analysis. Thus, a small unit identified in one analysis may not be a small unit in another analysis. The QA reports are provided to state clients to monitor testing anomalies throughout the testing window.

5.7.1 Changes in Student Performance

Changes in student scores between administration years are examined using a regression model to check for outliers. For these between-year comparisons, students' current-year scores are regressed on their test scores from the previous year and on the number of days between the two years' test-end dates (to control for the instruction time between the two test scores).

A large score gain or loss in student scores between administration years is detected by examining the residuals for outliers. The residuals are computed as the observed value minus the regression model's predicted value. The studentized residuals are computed to detect unusual residuals. An unusual increase or decrease in student scores between administration years is flagged when the absolute value of the studentized residual is greater than 3.

The residuals of students are also aggregated for a testing session, TA, and school. The system flags any unusual changes in an aggregate performance between administrations and/or years based on the average of the residuals in the aggregate unit (e.g., testing session, TA, school). For each aggregate unit, a t value is computed and flagged when $|t|$ is greater than 3,

$$t = \frac{\sum_{i=1}^n \hat{e}_i / n}{\sqrt{\frac{s^2}{n} + \frac{\sum_{i=1}^n \sigma^2 (1 - h_{ii})}{n^2}}}$$

where s is the standard deviation of residuals in an aggregate unit; n is the number of students in an aggregate unit (e.g., testing session, TA, school), σ^2 is the MSE from the regression, and $\hat{e}_i$ is the residual for the i th student.

The variance of average residuals in the denominator is estimated in two components, conditioning on the true residual e_i , $var(E(\hat{e}_i|e_i)) = s^2$ and $E(var(\hat{e}_i|e_i)) = \sigma^2(1 - h_{ii})$. Following the law of total variance (Billingsley, 1995, p. 456),

$$var(\hat{e}_i) = var(E(\hat{e}_i|e_i)) + E(var(\hat{e}_i|e_i)) = s^2 + \sigma^2(1 - h_{ii}), \text{ hence,}$$

$$var\left(\frac{\sum_{i=1}^n \hat{e}_i}{n}\right) = \frac{\sum_{i=1}^n (s^2 + \sigma^2(1 - h_{ii}))}{n^2} = \frac{s^2}{n} + \frac{\sum_{i=1}^n (\sigma^2(1 - h_{ii}))}{n^2}.$$

The comparisons for the spring 2022 administration were not performed because there was no testing in spring 2021 due to the COVID-19 pandemic.

5.7.2 Test-Taking Time

The summative assessments are not timed, and thus individual test-taking times may vary across students. However, unusual test-taking times such as excessively shorter or longer test-taking times may indicate irregularities in test administration. An example of unusual test-taking time is a test record for an individual who scores very well on the test even though the average time spent is far less than that of students statewide. If students already know the answers to the questions, the test-taking time may be much shorter than the test-taking time for those who have no prior knowledge of the item content. Conversely, if a TA helps students by coaching them to change their responses during the test, the testing time could be longer than expected.

The state average testing time and standard deviation are computed based on all students available when the analysis was performed. Students and aggregate units are flagged if the test-taking time is different from the state average by three standard deviations or more, although the flagging criteria can be adjusted by an authorized user. The test-taking time analysis was performed and evaluated for the spring 2022 administration.

5.7.3 Inconsistent Item Response Pattern (Person Fit)

In item response theory (IRT) models, person-fit measurement is used to identify test takers whose response patterns are improbable given an IRT model. If a test has psychometric integrity, little irregularity is expected in the item responses of individuals who respond to items fairly and honestly.

If a test taker has prior knowledge of some test items (or is provided answers during the exam), he or she will respond correctly to those items at a higher probability than indicated by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student. However, if a student has prior knowledge of the entire test content, this will not be detected based on the person-fit index, although the item response time index might flag such a student.

The person-fit index is based on all item responses in a test. An unlikely response to a single test question may not result in a flagged person-fit index. Of course, not all unlikely patterns indicate cheating, as in the case of a student who is able to guess a significant number of correct answers. Therefore, the evidence of person-fit index should be evaluated along with other irregular factors to determine possible testing irregularities. The number of flagged students is summarized for every testing session, TA, and school.

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985), and Sotaridona, Pornel, and Vallejo (2003), an aberrant response pattern is defined as a deviation from the expected item score model. Snijders (2001) showed that the distribution of l_z is asymptotically normal (i.e., with an increasing number of administered items). Even at shorter test lengths of 8 or 15 items, the “asymptotic error probabilities are quite reasonable for nominal Type I error probabilities of 0.10 and 0.05” (Snijders, 2001).

Sotaridona et al. (2003) report promising results of using l_z for systematic flagging of aberrant response patterns. Students with l_z values less than -3 are flagged. Aggregate units are flagged with t less than -3,

$$t = \frac{\text{Average } l_z \text{ values}}{\sqrt{s^2/n}},$$

where s is the standard deviation of l_z values in an aggregate unit and n is number of students in the aggregate unit. The person-fit analysis was performed and evaluated for the spring 2022 administration.

5.7.4 Item-Response Change

Students are allowed to revisit items as many times as they wish within a session and may also mark items to be revisited prior to completing the session. However, excessively high rates of response change, especially high rates of item score increases (i.e., response changes from wrong to right), may indicate irregularities in test administration. For example, test administrators (TAs) could review students’ responses and either coach them to modify their responses or keep the session active and change responses themselves.

To identify irregular patterns of response change, the item score for the final response to each item and the penultimate response if one exists are examined, and the number of instances in which the item score increases are counted.

The average and standard deviation of positive item score changes are computed based on all students available when the analysis was performed. Students and aggregate units are flagged if the number of positive item score changes is larger than the state average by three standard deviations or more, although the flagging criteria can be adjusted by an authorized user. The item-response analysis was performed and evaluated for the spring 2022 administration.

5.7.5 Observed Online Test-Taking Time

The Smarter Balanced assessments and the WCAS are not timed, and an individual student may need more or less time overall. The length of a test session is determined by SCs and TAs who are knowledgeable about the class periods in the school’s instructional schedule and the timing needs associated with the assessments. Students should be allowed extra time if they need it, but SCs or TAs must use their best professional judgment when allowing students extra time. Students should be actively engaged in responding productively to test questions.

During the online tests, item response time is captured as the item page time (the length of time that each item page is presented) in milliseconds. Discrete items appear on the screen one at a time. For items associated with a stimulus, the page time is the time spent on all items associated with the stimulus because all associated items appear on the screen together in ELA and mathematics. In the WCAS, page time is time spent on one page, regardless of item count on that page, as an item associated with a stimulus may appear on its own page under specific conditions (e.g., locked items) while the rest of the associated items for that same stimulus appear on another page together. For each student, the total time taken to finish the test was computed by summing up the page times.

Tables 5.19 and 5.20 present average testing time and testing time at percentiles for the overall test, the CAT component, and the PT component for the online Smarter Balanced assessments. Table 5.21 presents the same information for the WCAS online tests.

Table 5.19: Smarter Balanced ELA Test-Taking Time, Spring 2022 Administration

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test							
3	2:44	1:43	3:29	3:51	4:19	4:58	6:05
4	2:52	1:47	3:40	4:03	4:31	5:11	6:18
5	2:51	1:46	3:36	3:58	4:26	5:06	6:17
6	2:31	1:31	3:06	3:24	3:48	4:22	5:25
7	2:30	1:24	3:06	3:22	3:42	4:11	5:05
8	2:31	1:20	3:06	3:22	3:41	4:09	4:58
HS	2:31	1:19	3:08	3:22	3:39	4:04	4:50
CAT Component							
3	0:58	0:33	1:11	1:17	1:25	1:37	1:57
4	0:58	0:32	1:10	1:16	1:23	1:34	1:55
5	0:58	0:32	1:10	1:16	1:24	1:35	1:55
6	1:08	0:35	1:22	1:28	1:37	1:48	2:10
7	1:06	0:32	1:19	1:25	1:32	1:43	2:02
8	1:06	0:31	1:20	1:26	1:33	1:43	2:02
HS	1:11	0:33	1:26	1:32	1:40	1:50	2:09
PT Component							
3	1:46	1:24	2:21	2:39	3:02	3:35	4:31
4	1:55	1:28	2:33	2:52	3:16	3:49	4:45
5	1:53	1:26	2:30	2:48	3:11	3:44	4:42
6	1:23	1:08	1:47	2:01	2:19	2:46	3:36
7	1:25	1:04	1:49	2:02	2:17	2:40	3:23
8	1:25	0:59	1:49	2:01	2:16	2:37	3:15
HS	1:20	0:56	1:44	1:54	2:07	2:25	2:59

Table 5.20: Smarter Balanced Mathematics Test-Taking Time, Spring 2022 Administration

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test							
3	1:27	0:55	1:48	1:58	2:13	2:34	3:10
4	1:26	0:53	1:46	1:57	2:10	2:30	3:06
5	1:37	1:03	1:59	2:12	2:28	2:52	3:34
6	1:20	0:46	1:38	1:46	1:58	2:14	2:43
7	1:08	0:36	1:22	1:29	1:38	1:50	2:12
8	1:15	0:40	1:32	1:39	1:49	2:02	2:26
HS	1:20	0:44	1:40	1:48	1:58	2:12	2:38
CAT Component							
3	0:50	0:33	1:03	1:09	1:17	1:30	1:52
4	0:52	0:33	1:04	1:11	1:19	1:32	1:52
5	0:52	0:32	1:04	1:11	1:19	1:30	1:52
6	0:45	0:26	0:55	1:00	1:06	1:15	1:31
7	0:44	0:23	0:53	0:58	1:03	1:11	1:25
8	0:47	0:25	0:58	1:03	1:08	1:16	1:31
HS	0:48	0:26	1:00	1:05	1:11	1:20	1:34
PT Component							
3	0:36	0:29	0:46	0:52	1:00	1:11	1:31
4	0:34	0:28	0:43	0:48	0:55	1:05	1:24
5	0:45	0:39	0:56	1:04	1:14	1:28	1:55
6	0:35	0:28	0:44	0:49	0:55	1:05	1:23
7	0:24	0:19	0:30	0:33	0:38	0:44	0:56
8	0:28	0:21	0:35	0:39	0:44	0:51	1:04
HS	0:32	0:24	0:41	0:46	0:51	1:00	1:15

Table 5.21: WCAS Test-Taking Time, Spring 2022 Administration

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75th	80th	85th	90th	95th
5	1:30	0:51	1:50	1:59	2:12	2:31	3:05
8	1:04	0:30	1:17	1:22	1:29	1:39	1:56
11	1:03	0:29	1:17	1:22	1:28	1:36	1:51

5.7.6 Prevention and Recovery of Disruptions in Test Delivery System

CAI is continuously improving our ability to protect our systems from interruptions. CAI's TDS is designed to ensure that student responses are captured accurately and stored on more than one server in case of a failure. CAI architecture, described below, is designed to recover from a failure of any component with little interruption. Each system is redundant, and critical student response data is transferred to a different data center each night.

CAI has developed a unique monitoring system that is very sensitive to changes in server performance. Most monitoring systems provide warnings when something is going wrong. CAI's does, too, but it also provides warnings when any given server is performing differently from its performance over the few hours prior, or differently from the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing CAI to detect potential problems, investigate them, and mitigate them *before* a failure. On multiple occasions, this has enabled CAI to adjust and/or replace equipment before any problems occurred.

CAI has also implemented an escalation procedure that enables us to alert clients within minutes of any disruption. The emergency alert system notifies by text message CAI's executive and technical staff, who then immediately join a telephone conference call to understand the problem.

The section below describes CAI system architecture and how it recovers from device failures, Internet interruptions, and other problems.

5.7.7 High-Level System Architecture

CAI architecture provides the redundancy, robustness, and reliability required by a large-scale testing program. CAI's general approach, which has been adopted by Smarter Balanced as standard policy, is pragmatic and well supported by CAI's architecture.

Any system built around an expectation of flawless performance of computers or networks within schools and districts is bound to fail. The CAI system is designed to ensure that the testing results and experience are able to respond robustly to such inevitable failures. Thus, CAI's TDS is designed to protect data integrity and to prevent student data loss at every point in the process.

The key elements of the testing system, including the data integrity processes, at work at each point in the system are described below. Fault tolerance and automated recovery are built into every component of the system, as described.

Student Machine

Student responses are conveyed to CAI servers in real time as students respond. Long responses, such as essays, are saved automatically at configurable intervals (usually set to one minute), so that student work is not at risk during testing.

Responses are saved asynchronously, with a background process on the student machine waiting for confirmation of successfully stored data on the server. If confirmation is not received within the designated time (usually set to 30–90 seconds), the system will prevent the student from doing any more work until connectivity is restored. The student is offered the choice of asking the system to try again or pausing the test and returning at a later time. For example:

- If connectivity is lost and restored (“silently restored”) within the designated time period, the student may be unaware of the momentary interruption.

- If connectivity cannot be silently restored, the student is prevented from testing and given the option of logging out or retrying the save.
- If the system fails completely, upon logging back in the system the student returns to the item at which the failure occurred.

In short, data integrity is preserved by confirmed saves to CAI's servers and prevention of further testing if confirmation is not received.

Test Delivery Satellites

The test delivery satellites communicate with student machines to deliver items and receive responses. Each satellite is a collection of web and database servers. Each satellite is equipped with redundant array of independent disks (RAID) systems to mitigate the risk of disk failure. Each response is stored on multiple independent disks.

One server serves as a backup hub for every four satellites. This server continually monitors and stores all changed student response data from the satellites, creating an additional copy of the real-time data. In the unlikely event of failure, data are completely protected. Satellites are automatically monitored and, upon failure, they are removed from service. Real-time student data are immediately recoverable from the satellite, backup hub, or hub (described below), with backup copies remaining on the drive arrays of the disabled satellite.

If a satellite fails, students will exit the system. The automatic recovery system enables them to log in again within seconds or minutes of the failure, without data loss. This process is managed by the hub. Data will remain on the satellites until the satellite receives notice from the demographic and history servers that the data are safely stored on those disks.

Hub

Hub servers are redundant clusters of database servers with RAID drive systems. Hub servers continuously gather data from the test delivery satellites and their mini-hubs and store that data as described. This real-time backup copy remains on the hub until the hub receives notification from the demographic and history servers that the data have reached the designated storage location.

Demographic and History Servers

The demographic and history servers store student data for the duration of the testing window. They are clustered database servers, also with RAID subsystems, providing redundant capability to prevent data loss in the event of server or disk failure. At the normal conclusion of a test, these servers receive completed tests from the test delivery satellites. Upon successful completion of the storage of the information, these servers notify the hub and satellites that it is safe to delete student data.

QA System

The QA system gathers data used to detect cheating, monitors real-time item function, and evaluates test integrity. Every completed test runs through the QA system, and any anomalies (such as unscored or missing items, unexpected test lengths, or other unlikely issues) are flagged and a notification immediately goes out to CAI psychometricians and project team.

Database of Record

The Database of Record (DoR) is the final storage location for the student data. These clustered database servers with RAID systems hold the completed student data.

5.7.8 Automated Backup and Recovery

Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure the safety, security, and integrity of all data. This set of systems and processes is designed to provide complete data integrity and prevent loss of student data. Redundant systems at every point, real-time data integrity protection and checks, and well-considered real-time backup processes prevent loss of student data, even in the unlikely event of system failure.

5.7.9 Other Disruption Prevention and Recovery

The CAI system is designed to be extremely fault-tolerant. The system can withstand failure of any component with little to no interruption of service. One way that this robustness is achieved is through redundancy. Key redundant systems are as follows:

- The hosting provider has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely.
- The hosting provider has multiple redundancies in the flow of information to and from CAI's data centers by partnering with nine different network providers. Each fiber carrier must enter the data center at separate physical points, protecting the data center from a complete service failure caused by an unlikely network cable cut.
- On the network level, there are redundant firewalls and load balancers throughout the environment.
- There is redundant power and switching within all server cabinets.
- Data are protected by nightly backups. CAI completes a full weekly backup and incremental nightly backups. Should a catastrophic event occur, CAI is able to reconstruct real-time data using the data retained on the TDS satellites and hubs.
- The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or if they will need to rerun the backup.

CAI's TDS is hosted in an industry-leading facility, with redundant power, cooling, state-of-the-art security, and other features that protect the system from failure. The system itself is redundant at every component, and the unique design ensures that data is always stored in at least two locations in the event of failure. The engineering that led to this system protects the student responses from loss.

SUMMARY

The Smarter Balanced assessments and WCAS tests are administered online for most students in Washington, and on paper for a small population of students who lack internet access or have an IEP, 504 Plan, or other similar learning plan that require a paper for braille, large print, or standard print forms in ELA, mathematics, or WCAS. Spanish print forms are also available in mathematics and WCAS.

In both online and paper-pencil tests, the role, responsibility, and training required for key personnel involved with the administrations were well documented and communicated to schools and districts. All school personnel who serve as TAs, for example, are required to attend district-developed training sessions and sign security paperwork at the end of training. School-level personnel and decision-making teams, used the *Guidelines* and local-decision making processes to prepare and provide students with embedded and non-embedded features to access the tests.

Maintaining test security and test integrity is of high priority in all tests. There are built-in system-level security measures to ensure that personal information is secured and transferred data are not altered in any way. Staff of different roles are assigned different levels of access to the system. TAs are also trained in how to prepare the room for tests, including seating arrangements, and in the reporting of improprieties. The vendor also monitors testing response time and response patterns to detect irregularities.

6. ACHIEVEMENT-LEVEL SETTING

6.1 OVERVIEW

The process of achievement-level setting is designed to identify a “cut score,” or minimum test score, that is required to identify achievement level for students. Achievement-level setting generally requires a panel of subject matter experts and others with relevant perspectives (e.g., teachers, school administrators, parents). Several methodologies exist to collect panelists’ determinations and to translate their results appropriately into cut scores.

There was an achievement-level setting convened in 2018 for Washington Comprehensive Assessment of Science (WCAS) in grades 5, 8 and HS and remain the same. Cut scores and expected skill level in Smarter Balanced assessments remained the same as those set in 2015. This chapter presents the achievement-level setting process employed for Smarter Balanced assessments and WCAS conducted in earlier years for reference.

6.2 SMARTER BALANCED ASSESSMENTS

In 2014, Smarter Balanced facilitated participation from teachers, parents, higher education faculty, business leaders, and other community members from all of the Smarter Balanced states in a highly inclusive, consensus-based process that asked participants to closely examine assessment content and detailed Achievement Level Descriptors to determine threshold scores for each achievement level. Detailed information from Smarter Balanced on this processes can be found on the Smarter Balanced website (<https://validity.smarterbalanced.org/scoring/>). At their meeting on January 7 and 8, 2015, members of Washington’s State Board of Education approved the cut scores recommended by the Smarter Balanced Assessment Consortium that established the threshold scale scores for four achievement levels. Cut scores and expected skill level in Smarter Balanced assessments remain the same as those set in January 2015.

In addition to approving these threshold scale scores, Washington’s State Board of Education also established an initial “equal impact” approach to setting the minimum high school graduation scores on the Smarter Balanced English language arts (ELA) and mathematics tests (now known as the graduation pathway cut scores). The impact of cut scores on students in 2016 and later years was thus approximately equal to the impact on students of exit exams during the previous few years.

Starting in 2017–18, as a result of legislative action, OSPI administered the high school summative tests to grade 10 students. Smarter Balanced provided the cut scores for grade 10 ELA and mathematics tests, which were approved by Washington’s State Board of Education.

6.3 WCAS

The Bookmark procedure was used to set achievement standards for the WCAS in 2018. Introduced in 1999, the Bookmark procedure has been widely used across the United States for achievement-level setting (Mitzel, Lewis, Patz, & Green, 2001). The procedure requires panelists to work through an online test booklet in which the items have been ordered from easiest to hardest based on student performance data. Panelists are asked to place a bookmark in the ordered booklet to demarcate each performance standard. In the Washington achievement-level setting meetings, bookmarks were placed with the assumption that the borderline students will perform successfully at a given

achievement level with a probability of at least 0.50 for the grades 5, 8, and high school tests. The cut score for a particular performance standard is derived by averaging the corresponding bookmarks across panelists for that performance standard.

In addition to the Bookmark procedure, the contrasting groups method was used to provide additional information for the achievement-level setting process. For the contrasting groups study, participating teachers from around the state were asked to rate their students after receiving training concerning the meaning of the new Achievement-Level Descriptors (ALDs) for the respective grade. Based on their understanding of the ALDs and their students' classroom performance, the teachers were asked to rate their students into one of the three categories: Basic, Proficient, or No Basis when the teachers decided that they did not have enough information to rate the students. The contrasting group information was compiled before the achievement-level setting meeting. Two raw score distributions were produced: one distribution for the students who were rated Basic by their teachers, and one distribution for those who were rated Proficient by their teachers. The range in which the two distributions intersected was converted into the page ranges in the test-level Ordered Item Booklet, and this information was provided to the achievement-level setting committee to facilitate setting the final cut pages.

The Washington State Board of Education approved the following achievement-level setting process for grades 5, 8, and 11 science in August 2018.

- Achievement-level setting committee meeting:
 - Panelists took the test for the subject that they were meeting on.
 - Panelists were presented the ALDs.
 - Panelists were presented with the contrasting group study results.
 - Panelists provided the first round of rating.
 - Panelists were presented with the percentages regarding who would score at or above each achievement level given the cut scores.
 - Panelists provided the second round of rating.
 - Panelists were presented with the proportion of students taking the test who correctly responded to the item on each page of the online booklet.
 - Panelists provided the third round of rating.

A more detailed description of the achievement-level setting procedure and results on grade 5, 8, and 11 science tests is provided in the *Achievement Level Setting Technical Report, Washington Comprehensive Assessment of Science (WCAS), Grades 5, 8, and 11*, available by request from OSPI's website <https://www.k12.wa.us/student-success/testing/state-testing/scores-and-reports/technical-reports>.

6.4 CUT SCORES

The cut scores obtained as a result of the standard-setting process are on the ability or theta scale; the scores are then translated into scale scores, for which the ranges may vary. For all WCAP assessments, scale scores are mapped into four achievement levels (i.e., Level 1, Level 2, Level 3, Level 4) using three cut scores.

For the Smarter Balanced assessments, the scaled cut score varies by grade level because scores are vertically linked across grades. ALDs provide a description of content-area knowledge and skills that students at each achievement level are expected to possess. The ELA and mathematics ALDs are available on the Smarter Balanced website at <https://portal.smarterbalanced.org/library/en/mathematics-alds-and-college-content-readiness-policy.pdf> and <https://portal.smarterbalanced.org/library/en/elaliteracy-alds-and-college-content-readiness-policy.pdf>.

For the WCAS, the cut score for Level 2 is 650 for every grade; this means that a student must earn a score of 650 or higher to achieve a Level 2 classification. The cut score for the Proficient Level 3 is 700 for every grade; this means that a student must earn a score of 700 or higher to achieve a Level 3 classification. The cut score for Level 4 is derived using a linear function of theta and scale score for the Level 2 and Level 3 cut. The WCAS ALDs are available on OSPI's website at <https://www.k12.wa.us/student-success/testing/state-testing/scores-and-reports/achievement-level-descriptors>.

The theta cuts and the corresponding scale score cuts for Smarter Balanced and the WCAS are presented in Tables 6.1 and 6.2.

Table 6.1: WCAP Cut Scores—Smarter Balanced Assessments

Content Area	Level 2		Level 3		Level 4	
	Theta	Scale Score	Theta	Scale Score	Theta	Scale Score
Smarter Balanced ELA G–3	–1.646	2367	–0.888	2432	–0.212	2490
Smarter Balanced ELA G–4	–1.075	2416	–0.410	2473	0.289	2533
Smarter Balanced ELA G–5	–0.772	2442	–0.072	2502	0.860	2582
Smarter Balanced ELA G–6	–0.597	2457	0.266	2531	1.280	2618
Smarter Balanced ELA G–7	–0.340	2479	0.510	2552	1.641	2649
Smarter Balanced ELA G–8	–0.247	2487	0.685	2567	1.862	2668
Smarter Balanced ELA HS	–0.205	2491	0.807	2577	1.979	2678
Smarter Balanced Mathematics G–3	–1.689	2381	–0.995	2436	–0.175	2501
Smarter Balanced Mathematics G–4	–1.310	2411	–0.377	2485	0.430	2549
Smarter Balanced Mathematics G–5	–0.755	2455	0.165	2528	0.808	2579
Smarter Balanced Mathematics G–6	–0.528	2473	0.468	2552	1.199	2610
Smarter Balanced Mathematics G–7	–0.390	2484	0.657	2567	1.515	2635
Smarter Balanced Mathematics G–8	–0.137	2504	0.897	2586	1.741	2653
Smarter Balanced Mathematics HS	0.228	2533	1.245	2614	2.291	2697

Table 6.2: WCAP Cut Scores—WCAS

Content Area	Level 2		Level 3		Level 4	
	Theta	Scale Score	Theta	Scale Score	Theta	Scale Score
WCAS Grade 5	-1.24418	650	-0.48273	700	0.81311	785
WCAS Grade 8	-0.81903	650	-0.07857	700	0.88031	765
WCAS HS	-1.79726	650	-1.07733	700	0.22897	791

SUMMARY

Smarter Balanced assessments and the WCAS are criterion-based. Achievement level setting is designed to identify a “cut score,” or minimum test score, that is required to identify a student at a particular achievement level.

There was an achievement level setting in 2018 for grades 5, 8, and 11 WCAS and in 2014 for Smarter Balanced. Cut scores (in scale scores matrix) and expected skill level in all WCAP assessments remained the same as those adopted in previous achievement-level setting meetings. All achievement-level setting meetings mentioned followed widely accepted procedures to ensure that statistics and test data were error-free, and appropriate expectations were set for each achievement level.

7. SCORING

The Smarter Balanced Assessment Consortium (SBAC) provided the vertically scaled item parameters by linking across all grades using common items in adjacent grades. All scores are estimated based on these item parameters. The Smarter Balanced assessments consisted of computer-adaptive tests (CATs) and fixed-form, randomly-assigned performance tasks. Because of the CAT and depending on the items presented, two students having the same raw score are likely to receive different scale scores in a test. Further details on scoring for the Smarter Balanced tests can be found in the *Smarter Balanced Scoring Specifications for Summative and Interim Assessments* document at https://technicalreports.smarterbalanced.org/scoring_specs/book/scoringspecs.html.

The fixed-form Washington Comprehensive Assessment of Science (WCAS) is scored by the number-correct method. In this approach, a student's number-correct score (or raw score) is converted to a scale score. Two students with the same raw score will have the same scale score. The conditional standard error of measurement for every possible scale score in a form is calculated as well.

The following sections describe conversion tables, achievement levels, attemptedness rules, proficiency range for each content category, and handscoring.

7.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The Washington Comprehensive Assessment Program (WCAP) assessments are scored using maximum likelihood estimation (MLE). The likelihood function for generating the MLEs is based on a mixture of items types.

Indexing items by i , the likelihood function based on the j th person's score pattern for I items is

$$L_j(\theta_j | \mathbf{z}_j, \mathbf{a}, b_1, \dots, b_k) = \prod_{i=1}^I p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}),$$

where $b_i = (b_{i,1}, \dots, b_{i,m_i})$ for the i th item's step parameters, m_i is the maximum possible score of this item, a_i is the discrimination parameter for item i , z_{ij} is the observed item score for the person j , and k indexes step of the item i .

For Smarter Balanced assessments, the probability $p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})$ takes either the form of a two-parameter logistic (2PL) model for items with one point or the form based on the generalized partial credit model (GPCM) for items with two or more points.

In the case of items with one score point, $m_i = 1$,

$$p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = \begin{cases} \frac{\exp(Da_i(\theta_j - b_{i,1}))}{1 + \exp(Da_i(\theta_j - b_{i,1}))} = p_{ij}, \text{ if } z_{ij} = 1 \\ \frac{1}{1 + \exp(Da_i(\theta_j - b_{i,1}))} = 1 - p_{ij}, \text{ if } z_{ij} = 0 \end{cases};$$

in the case of items with two or more points,

$$p_{ij}(z_{ij}|\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = \begin{cases} \frac{\exp(\sum_{k=1}^{z_{ij}} D a_i(\theta_j - b_{i,k}))}{s_{ij}(\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})}, & \text{if } z_{ij} > 0 \\ \frac{1}{s_{ij}(\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})}, & \text{if } z_{ij} = 0 \end{cases},$$

where $s_{ij}(\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = 1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l D a_i(\theta_j - b_{i,k}))$, and $D = 1.7$.

For the WCAS, the probability $p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})$ takes either the form of a one-parameter logistic (1PL) model for items with one point or the form based on the generalized partial credit model (GPCM) for items with two or more points. The difference between 1PL and 2PL modes is that $a_i = 1$ for the 1PL model.

Conditional Standard Error of Measurement

With MLE, the standard error (SE) for student j is:

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

where $I(\theta_j)$ is the test information for student j , calculated as

$$I(\theta_j) = \sum_{i=1}^I D^2 a_i^2 \left(\frac{\sum_{l=1}^{m_i} l^2 \exp(\sum_{k=1}^l D a_i(\theta_j - b_{i,k}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l D a_i(\theta_j - b_{i,k}))} - \left(\frac{\sum_{l=1}^{m_i} l \exp(\sum_{k=1}^l D a_i(\theta_j - b_{i,k}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l D a_i(\theta_j - b_{i,k}))} \right)^2 \right),$$

where m_i is the maximum possible score point (starting from 0) for the i th item, D is the scale factor, 1.7, $a = 1$ for the WCAS. The SE is calculated based only on the answered item(s) for both complete and incomplete tests. Since the SE is based on specific theta, it is also called conditional standard error of measure (CSEM).

7.2 THETA TO SCALE SCORE TRANSFORMATION

The student's performance in each content-area test is summarized in an overall test score referred to as a *scale score*. The number of items a student answers correctly and the difficulty of the items presented are used to estimate theta scores. Theta scores are linearly transformed to scale scores using the formula $SS = a * \theta + b$. Scale scores from different sets of items within a test can be meaningfully compared. For Smarter Balanced assessments, the scaling constants a and b are provided by SBAC. Since Smarter Balanced assessments are vertically scaled, there is one slope and one intercept for each subject of English language arts/literacy (ELA) and mathematics. Because all ELA or mathematics tests are on the same scale, the ELA test scores or the mathematics test scores can also be compared across tested grades within each subject. For the WCAS, a and b for each test were decided after standard setting. Table 7.1 lists the scaling constants.

Table 7.1: Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA	3-8, HS	85.8	2508.2
Mathematics	3-8, HS	79.3	2514.9
WCAS Grade 5	5	65.66	731.70
WCAS Grade 8	8	67.53	705.31
WCAS HS	11	69.45	774.82

Standard errors of the MLEs are transformed to be placed onto the reporting scale. This transformation is:

$$SE_{ss} = a * SE_{\theta},$$

where SE_{ss} is the standard error of the ability estimate on the reporting scale, SE_{θ} is the standard error of the ability estimate on the θ scale, and a is the same slope of the scaling constant that transforms θ to the reporting scale.

The scale scores are mapped into four achievement levels using three achievement standards (cut scores). Tables 7.2 and 7.3 provide the three scale score cuts for each test.

Table 7.2: Scale Score Cuts—Smarter Balanced

Grade	ELA			Mathematics		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
3	2367	2432	2490	2381	2436	2501
4	2416	2473	2533	2411	2485	2549
5	2442	2502	2582	2455	2528	2579
6	2457	2531	2618	2473	2552	2610
7	2479	2552	2649	2484	2567	2635
8	2487	2567	2668	2504	2586	2653
10	2491	2577	2678	2533	2614	2697

Table 7.3: Scale Score Cuts—WCAS

Subject	Grade	Level 2	Level 3	Level 4
WCAS	5	650	700	785
WCAS	8	650	700	765
WCAS	11	650	700	791

7.3 CONVERSION TABLES FOR WCAS

One nature of PCM is the relationship of the one-to-one correspondence between raw scores and the theta scores for fixed-form tests. As such, for each fixed-form test, it is possible to generate the conversion from each raw score to a theta score. When applying the transformation rules, a theta score, scale score, and raw score can be mapped interchangeably in a one-to-one relationship. For

the WCAS, the conversion table for each test is presented in Appendix C, Conversion Tables for State-Specific Tests.

7.4 LOWEST/HIGHEST OBTAINABLE SCORES

Although the observed score is measured more precisely in a CAT than in a fixed-form test, especially for high- and low-performing students, if the item pool does not include easy or difficult items to measure low- and high-performing students, the standard error could be large at the low and high ends of the ability range. OSPI adhered to the Smarter Balanced decision to truncate extremely unreliable student ability estimates. Tables 7.4 and 7.5 present the lowest obtainable theta/scale score (LOT/LOSS) and the highest obtainable theta/scale score (HOT/HOSS) in both theta and scale score metrics. Estimated thetas lower than LOT or higher than HOT are truncated to the LOT and HOT values and assign LOSS and HOSS associated with the LOT and HOT. LOT and HOT were applied to all tests and all scores (total and reporting category scores). The standard error for LOT and HOT is computed using the LOT and HOT ability estimates given the administered items.

Table 7.4: Lowest and Highest Obtainable Scores—Smarter Balanced

Subject	Grade	Theta Metric		Scale Score Metric	
		LOT	HOT	LOSS	HOSS
ELA	3	-5.9110	3.5332	2001	2811
	4	-5.5500	4.1826	2032	2867
	5	-5.2670	4.7546	2056	2916
	6	-5.0000	5.0000	2079	2937
	7	-4.9660	5.3119	2082	2964
	8	-4.7925	5.6063	2097	2989
	HS	-4.7305	6.1096	2102	3032
Mathematics	3	-5.6030	3.1219	2071	2762
	4	-5.3601	4.0264	2090	2834
	5	-5.3012	4.7426	2095	2891
	6	-5.1942	5.0000	2103	2911
	7	-5.1311	5.6630	2108	2964
	8	-5.0681	6.0272	2113	2993
	HS	-5.0000	7.1896	2118	3085

Table 7.5: Lowest and Highest Obtainable Scores—WCAS

Test	Theta Metric		Scale Score Metric	
	LOT	HOT	LOSS	HOSS
WCAS Grade 5 Online	-5.43	5.00	375	1060
WCAS Grade 8 Online	-5.34	5.25	345	1060
WCAS Grade 11 Online	-5.54	5.98	390	1190

7.5 SCORING ALL CORRECT AND ALL INCORRECT CASES

In the item response theory (IRT) maximum likelihood (ML) ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. For all correct and all incorrect cases, the highest obtainable scores (HOT and HOSS) or the lowest obtainable scores (LOT and LOSS) were assigned in the 2014–15 Smarter Balanced administration. Since the 2015–16 administration for Smarter Balanced and the 2017–18 WCAS administration, all incorrect and all correct cases were scored by either adding 0.5 to or subtracting 0.5 from an item score with the smallest item discrimination parameter among the administered operational items for a student.

7.6 RULES FOR CALCULATING STRENGTHS AND WEAKNESSES FOR REPORTING CATEGORIES

7.6.1 Claim Scores for Smarter Balanced Assessments

For the spring 2022 assessment, Washington adopted the adjusted blueprint in both ELA and mathematics. Because the number of items per claim was too small, the reliability was too low to report scores, thus claim scores were not generated for the spring 2022 tests.

7.6.2 Reporting Area Proficiency Range for the WCAS

The WCAS include reporting area scores. Unless indicated otherwise, reporting area and subscale scores are synonymous in this report. Different from the test-level scoring, a student's performance at each reporting area is not indicated by the four achievement levels. Instead, for the WCAS, proficiency in each reporting area is measured by comparing the achievement to the proficiency range of that reporting area. The following steps were used to calculate the proficiency range and student achievement for reporting areas:

1. Construct the raw-to-theta-to scale-score conversion table using the item parameters of items belonging to a reporting area.
2. Identify the smallest theta score that is greater than or equal to Level 3 theta cut (Proficient, scale score 700), and the smallest theta score that is greater than or equal to the Advanced (Level 4) theta cut. The raw scores associated with these two theta scores are, respectively, the lower bound and the upper bound raw scores of the proficiency range.
3. Divide the lower bound raw score, and the upper bound raw score by the total raw score points of the reporting area. The two calculated percentages are the lower and the upper bound of the proficiency ("At Standard") range.
4. To assess student performance, divide the total raw score earned by the raw score of the reporting area. Round the attained percentage to the nearest whole number.
5. If the rounded percentage attained by the student falls within the proficiency range, the student is "At Standard" in that reporting area. If the rounded percentage attained by the student falls above the proficiency range, the student is "Above Standard" in that reporting area. Otherwise, the student is "Below Standard".

Table 7.6 contains the proficiency ranges at each reporting area for each WCAS test.

Table 7.6: Reporting Area Level Summary for WCAS, Form A

Subject	Reporting Area	Theta Range	Max Raw Score	Below Standard (%)	At Standard (%)	Above Standard (%)
WCAS G5	Practices and Crosscutting Concepts in Earth and Space Science	-4.08 ~ 3.94	12	<50	>=50 and <=67	>67
	Practices and Crosscutting Concepts in Life Science	-4.51 ~ 4.04	12	<50	>=50 and <=67	>67
	Practices and Crosscutting Concepts in Physical Science	-4.91 ~ 4.10	14	<50	>=50 and <=64	>64
WCAS G8	Practices and Crosscutting Concepts in Earth and Space Science	-4.10 ~ 3.72	12	<58	>=58 and <=67	>67
	Practices and Crosscutting Concepts in Life Science	-4.43 ~ 4.05	16	<56	>=56 and <=69	>69
	Practices and Crosscutting Concepts in Physical Science	-3.96 ~ 4.75	14	<50	>=50 and <=64	>64
WCAS G11	Practices and Crosscutting Concepts in Earth and Space Science	-4.91 ~ 3.76	12	<50	>=50 and <=67	>67
	Practices and Crosscutting Concepts in Life Science	-4.16 ~ 3.93	15	<33	>=33 and <=53	>53
	Practices and Crosscutting Concepts in Physical Science	-4.55 ~ 5.31	18	<28	>=28 and <=44	>44

7.7 ATTEMPTEDNESS RULE

Students must attempt the test for it to be scored. In Smarter Balanced assessments, all tests with at least one CAT item and one PT item answered are considered “attempted.” If a student logged onto both the CAT and the PT parts of the test, but no items are answered, the student is considered as having participated. These tests will be included in the data file, but no scores will be computed.

- Attemptedness rules for CAT:
 - o N (not attempted) = responded to zero items
 - o Y (attempted) = responded to at least one item
- Attemptedness rules for PT:
 - o N (not attempted) = responded to zero items
 - o Y (attempted) = responded to at least one item

In Smarter Balanced assessments, all tests are scored if the tests meet the following rules of attemptedness:

- CAT attemptedness = Y and PT attemptedness = Y

For the WCAS, a test is attempted when the student provides responses to at least two items, regardless of whether they are operational items, field-test (pilot) items, or non-scoring items. A valid item response is non-blank for machine-scored items, and a score or a condition code other than blank for hand-scored items. Condition codes are letter codes assigned to responses that cannot be scored, for example, random keystrokes or symbols, and non-legible responses.

Attempted tests are scored and the condition codes, including blanks, are set to zero. If the two responses are both non-scoring and/or field-test items, the achievement score would be zero.

7.8 TARGET SCORES FOR SMARTER BALANCED ASSESSMENTS

The target-level reports are impossible to produce for a single test, because the number of items included per target (i.e., group of related standards) is too small to produce a reliable score at the target level. Similarly, for fixed-form tests such as the WCAS, there are too few items at reporting levels beyond the Reporting Areas described above to reliably report student performance. When aggregated across multiple students' tests, however, the adaptive Smarter Balanced tests may see a class of 20 students respond to 10 or 15 different items measuring a given target.

Due to the sampling nature of the Smarter Balanced blueprint and adaptive algorithm (details available online at <http://www.smarterapp.org/documents/AdaptiveAlgorithm.pdf>), target scores should not be interpreted to represent the breadth of standards in a given target or the breadth of the skills described in those standards. It is possible that, in the scenario above, the 10 or 15 items that the 20 students saw measured only a single standard within the given target or, further, measured the same skill within that standard. Target data can be combined with other, local information about student performance with the standards to generate a more complete picture about student strengths and weaknesses with content articulated in the standards.

Target scores are computed for attempted tests based on the responded items. Target scores are computed in each claim (four claims) for ELA and in Claim 1 only for mathematics.

Target scores are computed in two ways: (1) target scores relative to a student's overall estimated ability (θ), and (2) target scores relative to the proficiency standard (Level 3 cut).

7.8.1 Target Scores Relative to Student's Overall Estimated Ability

The expression $p_{ij} = p(z_{ij} = 1)$ represents the probability that student j responds correctly to item i (z_{ij} represents the j th student's score on the i th item). For items with one score point, the 2PL IRT model is used to calculate the expected score on item i for student j with estimated ability $\hat{\theta}_j$ as:

$$E(z_{ij}) = \frac{\exp(Da_i(\hat{\theta}_j - b_i))}{1 + \exp(Da_i(\hat{\theta}_j - b_i))}.$$

For items with two or more score points, using the generalized partial credit model, the expected score for student j with estimated ability $\hat{\theta}_j$ on an item i with a maximum possible score of m_i is calculated as

$$E(z_{ij}) = \sum_{l=1}^{m_i} \frac{l \exp(\sum_{k=1}^l Da_i(\hat{\theta}_j - b_{i,k}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l Da_i(\hat{\theta}_j - b_{i,k}))}.$$

For each item i , the residual between observed and expected score for each student is defined as:

$$\delta_{ij} = z_{ij} - E(z_{ij}).$$

Residuals are summed for items within a target. The sum of residuals is divided by the total number of points possible for items within the target, T .

$$\delta_{jT} = \frac{\sum_{i \in T} \delta_{ji}}{\sum_{i \in T} m_i}.$$

For an aggregate unit, a target score is computed by averaging individual student target scores for the target across all students in the aggregate unit.

$$\bar{\delta}_{Tg} = \frac{1}{n_g} \sum_{j \in g} \delta_{jT}, \text{ and } se(\bar{\delta}_{Tg}) = \sqrt{\frac{1}{n_g(n_g-1)} \sum_{j \in g} (\delta_{jT} - \bar{\delta}_{Tg})^2},$$

where n_g is the number of students who responded to any of the items that belong to the target T for an aggregate unit g . If a student did not happen to see any items on a particular target due to the sampling nature of the blueprint and the adaptive algorithm, the student is NOT included in the n_g count for the aggregate.

A statistically significant difference from zero in these aggregates is used to report the group of students performance as better, worse, or similar to the test as a whole on this target. In some cases, insufficient information will be available, and that will be indicated, as well.

For target level strengths/weakness, the following are reported:

- If $\bar{\delta}_{Tg} \geq +1 * se(\bar{\delta}_{Tg})$, then performance is reported as better than on the rest of the test.
- If $\bar{\delta}_{Tg} \leq -1 * se(\bar{\delta}_{Tg})$, then performance is reported as worse than on the rest of the test.
- Otherwise, performance is reported as similar to performance on the test as a whole.
- If $se(\bar{\delta}_{Tg}) > 0.2$, data are insufficient.

7.8.2 Target Scores Relative to Proficiency Standard (Level 3 Cut)

The expression $p_{ij} = p(z_{ij} = 1)$ represents the probability that student j responds correctly to item i (z_{ij} represents the j th student's score on the i th item). For items with one score point, the 2PL IRT model is used to calculate the expected score on item i for student j with $\theta_{Level\ 3\ cut}$ as:

$$E(z_{ij}) = \frac{\exp(Da_i(\theta_{Level\ 3\ cut} - b_i))}{1 + \exp(Da_i(\theta_{Level\ 3\ cut} - b_i))}.$$

For items with two or more score points, using the generalized partial credit model, the expected score for student j with *Level 3 cut* on an item i with a maximum possible score of m_i is calculated as

$$E(z_{ij}) = \sum_{l=1}^{m_i} \frac{l \exp(\sum_{k=1}^l Da_i(\theta_{Level\ 3\ cut} - b_{i,k}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l Da_i(\theta_{Level\ 3\ cut} - b_{i,k}))}.$$

For each item i , the residual between observed and expected score for each student is defined as

$$\delta_{ij} = z_{ij} - E(z_{ij}).$$

Residuals are summed for items within a target. The sum of residuals is divided by the total number of points possible for items within the target, T .

$$\delta_{jT} = \frac{\sum_{i \in T} \delta_{ji}}{\sum_{i \in T} m_i}.$$

For an aggregate unit, a target score is computed by averaging individual student target scores for the target across all students in the aggregate unit.

$$\bar{\delta}_{Tg} = \frac{1}{n_g} \sum_{j \in g} \delta_{jT}, \text{ and } se(\bar{\delta}_{Tg}) = \sqrt{\frac{1}{n_g(n_g-1)} \sum_{j \in g} (\delta_{jT} - \bar{\delta}_{Tg})^2},$$

where n_g is the number of students who responded to any of the items that belong to the target T for an aggregate unit g . If a student did not happen to see any items on a particular target due to the sampling nature of the blueprint and the adaptive algorithm, the student is NOT included in the n_g count for the aggregate.

A statistically significant difference from zero in these aggregates is used to report the group of students performance as better, worse or similar to the proficiency standard (i.e., the Level 3 cut score) on this target. In some cases, insufficient information will be available, and that will be indicated, as well.

For target level strengths/weakness, the following are reported:

- If $\bar{\delta}_{Tg} \geq +1 * se(\bar{\delta}_{Tg})$, then performance is reported as *above* the proficiency standard.
- If $\bar{\delta}_{Tg} \leq -1 * se(\bar{\delta}_{Tg})$, then performance is reported as *below* the proficiency standard.
- Otherwise, performance is reported as *at/near* the proficiency standard.
- If $se(\bar{\delta}_{Tg}) > 0.2$, data are insufficient.

7.9 HANDSCORING

For the WCAP assessments, CAI provided the automated electronic scoring, and Measurement Incorporated (MI) provided all handscoring. In ELA, short-answer (SA) items and full-write items are hand-scored. In mathematics and science, SA items are hand-scored. Additionally, some additional constructed response items other than SA are hand-scored.

Both automated electronic scoring and handscoring was used to score ELA, mathematics, and science items. Item-specific scoring rubrics are written during item development. The scoring rubrics are then reviewed by content experts, along with the item content, as a part of the item review meetings. A central aspect of the validity of test scores is the degree to which scoring rubrics are related to the appropriate Learning Standards. A key facet of reliability is whether scoring rules are applied faithfully during scoring sessions. The following procedures are used to score the WCAP items and apply to all content areas that include open-ended questions calling for constructed responses. These procedures are used for both field-test items and operational items.

7.9.1 Rangefinding

Rangefinding refers to the process of creating scoring rubrics and accompanying training sets of responses for constructed response items that cannot be machine-scored.

Rangefinding for WCAS

MI scoring staff assembled groups of responses that exemplified the different score points represented in rubrics. Once examples of all of the score points were identified, packets or anchor and practice sets were put together for each item. These sets were annotated and copied for use at rangefinding, which was conducted on multiple dates and in various locations depending on the subject. The pilot rangefinding committees consisted of Washington state educators, OSPI staff members, CAI test development staff, and MI scoring staff. Operational rangefinding is conducted the first time an item is used operationally with a group consisting of OSPI staff members and MI scoring staff, as described in the section below.

Each committee began with a review of the item and the rubric. Copies of the student response anchor sets were presented to the committees, one item at a time. The committees reviewed and scored several student samples together to ensure that everyone was interpreting the rubric consistently. Committee members then went on to score responses independently, and those scores were discussed until a consensus was reached. Responses for which a good agreement rate was attained were used in training the scorers. Discussions of the responses used rubric language, assuring OSPI and all involved that the score point examples clearly illustrated the specific requirements of each score level. MI staff made notes of how and why the committees arrived at score point decisions, and this information was used by the scoring directors in scorer training. Annotations for the score on each of the responses were recorded and approved by the committee.

OSPI, MI, and CAI staff discussed rubric edits that the committees suggested. Changes were then made by OSPI and approved by the committee. OSPI and the committee went through the prepared practice sets and scored them individually. These scores were discussed to reach consensus regarding the true score of each response. Any changes to the annotations were made in accordance with the rubric. If additional responses were required to adequately represent all score points, these were pulled by MI scoring staff and approved by OSPI. Any changes to rubrics were then made by OSPI and approved by MI staff and OSPI assessment content specialists. These final rubrics were used by MI staff to train scorers.

Training Materials Review for WCAS

All scoring training materials being carried over from a previous contract/previous administration were reviewed prior to use in the operational test administration. OSPI provided MI scoring staff with all training materials, including rubrics, anchor sets, practice sets, qualification sets, validity papers, non-scorable codes/definitions, and scoring director notes from previous rangefinding meetings (when available). MI and OSPI staff first reviewed these materials individually. For items that were being used operationally for the first time, for example, MI staff selected responses from the 2022 administration to construct the qualifying sets and validity sets. Then, a series of conference calls/web meetings were held during which OSPI walked MI scoring staff through the materials with the purpose of providing additional scoring information, solidifying training notes, and confirming the responses to appear in the training materials for the operational items.

7.9.2 Handscoring for Smarter Balanced Assessments

Constructed response short-answer (SA) items and essay (i.e., full write) items in ELA and SA items in mathematics for the summative assessments administered by Cambium Assessment Inc. (CAI) are routed to Measurement Incorporated (MI) for scoring. MI provides handscoring using human raters. For the 2021–22 summative operational item pool, there were a total of 436 SA items and 198 essay items in ELA and 345 items in mathematics. Table 7.7 shows the number of items by grade and subject.

Table 7.7. Number of Hand-Scored Items in 2021–22 Smarter Balanced Summative Item Pool, by Grade and Subject

Grade	ELA/L		Mathematics
	Short Answer	Essay	
3	13	25	46
4	17	29	52
5	15	30	74
6	69	22	52
7	70	30	35
8	76	33	41
HS	176	29	45
Total	436	198	345

All guidelines for handscoring responses were specified by Smarter Balanced. Outlined below is the handscoring process MI followed in spring 2022 in accordance with the Smarter Balanced guidelines. This process applied to the scoring of all student constructed responses for ELA SA and essay items and mathematics SA items.

Rater Selection

MI has developed a pool of over three thousand raters experienced in scoring the Smarter Balanced assessments. MI first recruited qualified raters who had experience scoring these assessments. Recent advancements in rater evaluation practices have allowed MI to estimate rater accuracy parameters for experienced Smarter Balanced raters; these data were used to recruit the most historically accurate raters. Once recruited, experienced raters were assigned to the content area and grade band(s) with which they were most experienced.

To supplement this pool, MI also recruited raters with experience successfully scoring other large-scale assessments. MI assigned those raters to the grade level, subject area, and item type for which they were most qualified based on their performance on similar projects. Returning raters were selected based on experience and performance, as well as attendance, punctuality, and cooperation with work procedures and MI policies. MI maintains evaluations and performance data for all staff who work on each scoring project in order to determine employment eligibility for future projects. Finally, MI targeted recruitment of new raters as needed, in an effort to continue to identify talent across the country that will best fulfill the handscoring requirements.

All raters possessed, at a minimum, a four-year college degree. MI collected proof of degree for all raters as a condition of employment. All raters resided in the United States, and properly completed Form I-9 to verify their identity and employment authorization. Raters' I-9 forms are retained on file as required by law and made available for inspection by authorized government officers as needed. MI is an equal-opportunity employer, and believes that a diverse work force is of the utmost

importance. When hiring, MI strives to ensure the work force is diverse across age, ethnicity, gender, and other demographic groups.

In selecting team leaders who will monitor the raters, MI scoring leadership reviewed records of all returning staff. They looked for people who were experienced team leaders with a record of good performance on previous projects, and they also considered raters who had been recommended for promotion to the team leader position.

MI requires all handscoring project staff (scoring directors, team leaders, raters, and clerical staff) to sign a confidentiality/nondisclosure agreement before receiving any training or viewing any secure project materials. The employment agreement indicates that no participant in training and/or scoring may reveal information about the test, the scoring criteria, or the scoring methods to any person.

Rater Training and Scoring

All raters hired to score the Smarter Balanced assessments were trained using the rubric(s), anchor sets, and training/qualifying sets provided by Smarter Balanced. These sets were created during the original field-test scoring in 2014 and approved by Smarter Balanced. The same anchor sets are used each year. Additionally, MI conducts an annual review of the rater agreement and scoring materials in order to inform the development of item-specific, supplemental training materials. Supplemental materials are developed each summer and implemented in the subsequent operational administration.

Once hired, raters were assigned to a scoring group that corresponds to the subject/grade that they were deemed best suited to score (based on work history, results of the placement assessments, and performance on past scoring projects). Raters were trained to score a specific item group of either SA (research, brief write, reading, and mathematics) or essay (i.e., full-write) items. Within each item group, raters were divided into teams supervised by team leaders and a scoring director. Each scoring director, team leader, and rater was assigned a unique number for easy identification of their scoring work throughout the scoring session. The number of items an individual rater scores was minimized to allow the rater to quickly develop experience scoring responses to a given set of items.

All raters, regardless of experience, were required to train on all anchor and training sets. Following training, all raters were required to pass the qualification sets in order to prove that they understood and could apply the criteria accurately. Until a rater had trained and qualified successfully, the rater was not permitted to score any student responses. MI carefully orchestrated training so that raters understood that all scoring decisions must be grounded in the training materials. In addition, raters learned how to navigate the anchor set, developed the knowledge and flexibility needed to evaluate or escalate a variety of responses, and retained the necessary consistency to score all responses accurately.

In order to begin working, all scoring personnel logged in to MI's secure Scoring Resource Center (SRC). SRC includes all online training modules, serves as the portal to MI's Virtual Scoring Center (VSC) interface, and maintains the data repository of all scoring reports used for rater monitoring. MI's training system (VSC Train) provides a remote, secure application for training both team leaders and raters. VSC Train provided each trainee with a training lesson for each item that allowed the trainee to complete the following steps:

- 1) Review the anchor set(s)
- 2) Score the practice set(s)

- 3) Review an annotated version of the practice set(s) after submitting scores
- 4) Score the qualification sets

Training design varied slightly depending on Smarter Balanced item type:

- ELA essay: Raters trained and qualified on a baseline training lesson for a grade and writing purpose (e.g., grade 3 narrative, grade 6 argumentative, etc.). After qualifying on the baseline, raters then completed qualifying sets for each item in that grade and purpose. Raters could only score those items for which they have passed the qualifying set.
- ELA brief write, reading, and research SA: Raters trained and qualified on a baseline lesson within a specific grade band and target. Qualification on the baseline lesson qualified the rater to score all items in that grade band and target.
- Mathematics SA: Raters trained and qualified on baseline lessons within a specific grade band. Qualification on a baseline lesson qualified the rater to score that item and all items associated with it; for items with no associated items, training was for the specific item.

Rater training time varied by grade and content area. Training for SA brief write, reading, research, and mathematics items could typically be accomplished in one day, while training for essay items took up to five days to complete. Raters generally worked 6.5 hours per day, excluding breaks. Evening shift raters worked 3.75 hours, excluding breaks.

In addition to item-specific information, a variety of substantive procedural and policy information was provided to each trainee during training. This included information about “alert” responses and non-scorable responses, as well as instructions for how to communicate with leadership during handscoring. This ensured that raters were fully prepared to hand-score responses and were also aware of all responsibilities and scoring requirements before they were allowed to begin scoring.

Each trainee’s practice and qualification results were reported to the team leaders and scoring director. Scoring leadership reviewed each trainee’s results, paying particular attention to frequently mis-scored responses.

Following training, all training materials remained available to raters throughout scoring via the VSC Score Resource Library. This library included the item and rubric, the annotated anchor and practice sets, and any supplemental materials that were required to ensure accurate completion of the scoring effort.

When scoring, raters had access only to those items for which they had successfully trained and qualified. The handscoring system sorts individual student responses into small sets of 5-10, grouped by item. When a rater is qualified to score multiple items, this approach eases cognitive load by presenting the rater with a scoring set in which all responses relate to the same item.

Raters were trained to recognize non-scorable responses, and these responses were systematically routed to scoring supervisors for final condition-code assignment per Smarter Balanced requirements. For some item types, such as essays, condition-code responses were scored by scoring experts trained to specialize in the scoring of these types of responses.

An “alerts” procedure was explained to raters during training sessions, where raters are trained to recognize “alerts” in their various forms, including those for suicide, criminal activity, alcohol or

drug use, extreme depression, violence, rape, sexual or physical abuse, self-harm, intent to harm others, and neglect.

Multiple strategies were employed to minimize rater bias during scoring. First, raters did not have access to any student identifiers. Unless the students signed their names, wrote about their hometowns, or in some way provided other identifying information as part of their response, the raters had no knowledge of student characteristics. Second, all raters were trained using Smarter Balanced–provided materials, which were approved as unbiased examples of responses at the various score points. Training involved constant comparisons with the rubric and anchor papers so that raters’ judgments were based solely on the scoring criteria. Finally, following training, a cycle of diagnosis and feedback was maintained to identify any issues. Specifically, raters were closely monitored during scoring, and any instances of raters making scoring decisions based on anything except the criteria were discussed with the raters. After this feedback had been provided, raters were further monitored, and if any continue to exhibit bias after receiving a reasonable amount of feedback, they were dismissed.

Finally, a series of automated score verifications were implemented to further ensure the accuracy of scores. For example, a blank check was conducted, which reset scores when a condition code of “blank” was assigned to a response that had one or more characters in the response string (e.g., a response comprised of spaces or tabs). In this case, only after three independent raters had assigned a condition code of “blank” to a response that appeared blank, but which included characters in the response string, was the score recorded. A similar check was run when a score or condition code other than “blank” was assigned to a response that included no characters in the response string. Automatic resetting of double-scored responses when two raters assign non-adjacent scores, mismatched condition codes, or a combination of a condition code and a numeric score provided an additional score verification. In addition to automatically resetting and rescored these responses, the raters’ information was captured in a report and reviewed by scoring directors, one of many tools used to determine retraining needs.

Rater Statistics and Monitoring

At a minimum, 10-15% (depending on state contractual requirements) of the hand-scored responses received blind double reads. Additionally, 5% of the responses scored comprised pre-approved validity responses. MI’s VSC system automatically and randomly routed the requisite number of responses to raters for second reads and validity in an inconspicuous manner. Raters had no means of discerning whether they were scoring a first read, a second read, or a validity response. This system also prohibited raters from being eligible to score second reads for responses they had already scored.

MI’s VSC scoring system randomly seeds validity responses among operational responses during scoring. A small set of validity responses is provided by Smarter Balanced for all vendors to use, and these are supplemented with responses selected and approved by MI scoring management. The “true” scores for these responses are entered into a validity database. Validity responses are indistinguishable from operational responses.

VSC reports provided real-time reports throughout the scoring effort. These reports were available for access by handscoring management. Inter-rater reliability reports provide the percentage of exact, adjacent, and non-adjacent agreement for scorable responses. Validity performance reports provide the percentage of exact, adjacent, and non-adjacent agreement for validity responses and were used

to monitor drift. Score point frequency distribution reports provide the percentage per score point and include the mean and standard deviation for each item.

Years of Smarter Balanced handscoring has allowed MI to amass a longitudinal dataset of rater performance data. MI's rater monitoring system uses validity responses calibrated to fit a unidimensional Item Response Theory (IRT) model for each content area/item type. Extensive metrics (inter-rater reliability, calibrated validity, and sub-pools for monitoring drift) calculated by the monitoring system were used to ensure accuracy and productivity throughout the handscoring of a project. The system generated automated measures of rater performance drawing on validity, IRR, and other performance data. Raters and scoring managers received daily, automated messages summarizing raters' performance, ensuring all handscoring staff were aware of current performance and any issues that required attention. Additional outputs were also provided in manager-level reports and used to identify raters who required retraining and/or removal due to issues with accuracy and/or production. These data allowed scoring management to direct scoring leaders in review of specific VSC reports in order to determine the specific areas of attention required for any raters.

The monitoring system afforded the objective, dynamic identification of the most accurate and productive raters, referred to as "advanced raters." Advanced rater status changed daily based on current rater performance to ensure that any rater drift did not negatively impact scoring accuracy. Advanced rater status was a precondition for conducting second readings.

Team leaders spot-checked (i.e., read behind) raters' scoring to ensure that the raters were on target, and conducted one-on-one retraining sessions to address any problems found. At the beginning of the project, team leaders read behind every rater every day; they became more selective about the frequency and number of read-behinds as raters became more proficient at scoring.

Rater Retraining and Dismissal

Retraining was an ongoing process once scoring is underway. Daily analysis of the rater status reports enabled management personnel to identify individual or group retraining needs. When it became apparent that a whole team or group was having difficulty with a particular type of response, large group training sessions were conducted.

When read-behinds or daily statistics identified a rater who could not maintain acceptable agreement rates, the rater was retrained and monitored by scoring leadership personnel. Raters are released from the project if retraining is unsuccessful. In these situations, all items scored by a rater during the timeframe in question were identified, reset, and released back into the scoring pool. The aberrant rater's scores were deleted, and the responses were redistributed to other qualified raters for rescoring.

7.9.3 Handscoring for WCAS

Rater selection, rater training and scoring, rater statistics and monitoring, rater retraining and dismissal sections described above were also applied to the handscoring items of the WCAS.

For handscoring items in the WCAS, student responses on a given test were scored independently and by multiple scorers. All responses for science were read once; 15% second reads were also conducted. The second reads were randomly chosen by the imaging system at the item/prompt level. The score from the first rater (R1) was the final item score. The scoring director assigned the pre-

defined condition code to responses that were identified as non-scorable condition codes (except blanks).

When science item handscoring was completed, MI scoring staff would compile reviews of the field-test items. These reviews would be submitted to OSPI assessment content specialists.

For WCAS grades 5, 8, and 11, raters are also given blind validity responses to score throughout the project at a rate of 10%. The validity selection process begins first by identifying an item as either anchor (previously operational) or non-anchor (previously field tested). If an item is identified as an anchor item, validity responses from the previous administration of the assessment are carried forward and placed in the validity “pool” for that given item. The “true” scores or scores the responses have received previously, are carried forward and are not changed. If the item is a non-anchor item, MI scoring staff select 75–100 responses from “live” responses (responses from the current administration) after rangefinding, and OSPI provides final approval to make up the validity pool for all newly operational items.

The science assessment staff from OSPI reviews the item validity agreement statistics on a regular basis and consults with MI scoring directors about retraining or clarification of the true score for the validity responses as needed.

7.9.4 Rater Agreements

Rater inter-rater reliability (IRR) was computed based only on scorable responses (numeric scores) scored by two independent raters. Non-scorable responses (e.g., off-topic, off-purpose, or foreign-language responses) that were scored by scoring leadership—and not by two independent raters—were excluded from IRR computations. For the hand-scored items, the human-human agreement was computed based on 2021–22 Washington summative assessments.

In ELA, essay (i.e., full write) item responses were scored in three dimensions: conventions (0–2 rubric), evidence/elaboration (1–4 rubric), and organization/purpose (1–4 rubric). All ELA SA items were scored using a 0–2 rubric. Mathematics SA and other handscored items were scored using 0–1, 0–2, or 0–3 rubrics. The handscored items on the WCAS were scored using 0–1 or 0–2 rubrics. Condition codes are scored as zero.

For the WCAS, as a fixed-form test, there were 3 handscored items on the grade 5 test, 1 handscored item on the grade 8 test, and 2 handscored items on the grade 11 test. In every grade level, the ELA PT includes one full write item. ELA SA items may appear on an ELA PTs in all grade levels and on an ELA CATs only in grades 6–8 and high school. Math SA and other handscored items may appear only on the Math PTs in all grade levels. In an ELA CAT, because items are selected adapting to a student’s ability while meeting the test blueprint, item usages vary across items. Tables 7.8–7.11 provide a summary of the human-human IRR based on items with a sample size greater than 50. The IRR is presented with mean of percent exact agreement, minimum and maximum percent exact agreements, combined percent exact and adjacent agreement, and the mean, minimum and maximum quadratic weighted kappa (QWK). The average number of responses, as well as minimum and maximum number of responses to a given item are presented as well.

The quadratic weighted Kappa coefficient is computed by:

$$\text{Quadratic Weighted Kappa} = 1 - (\sum_{ij} w_{ij} a_{ij} / \sum_{ij} w_{ij} c_{ij})$$

Where w is the weight defined as d/p^2 , d is the points discrepancy between the two raters, and p is the maximum point of the item; a is the observed frequency in the cell ij th, and c is the expected frequency in the cell ij th.

Table 7.8: Interrater Agreement—ELA Smarter Balanced for Full-Write Items

Grade	Dimension	Number of Items	Number of Responses			%Exact			%(Exact+ Adjacent)	QWK		
			Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	Conventions	25	314.9	222	382	60.3	54.1	66.7	97.5	0.54	0.42	0.65
	Evid/Elab	25	314.9	222	382	62.3	52.7	74.1	96.7	0.61	0.42	0.76
	Org/Purp	25	314.9	222	382	62.1	51.8	74.1	96.7	0.61	0.41	0.76
4	Conventions	29	314.8	228	367	55.5	47.0	64.6	95.3	0.52	0.38	0.67
	Evid/Elab	29	314.8	228	367	60.0	51.4	70.3	96.1	0.64	0.47	0.78
	Org/Purp	29	314.8	228	367	59.9	48.6	69.7	96.2	0.64	0.45	0.77
5	Conventions	29	338.4	238	379	60.3	50.4	68.6	97.5	0.49	0.27	0.63
	Evid/Elab	29	338.4	238	379	58.5	52.7	64.3	96.9	0.66	0.53	0.73
	Org/Purp	29	338.4	238	379	59.4	50.4	67.0	97.1	0.67	0.51	0.73
6	Conventions	22	428.5	322	471	59.9	52.0	67.1	97.3	0.54	0.42	0.61
	Evid/Elab	22	428.5	322	471	65.9	47.1	74.7	98.3	0.69	0.50	0.78
	Org/Purp	22	428.5	322	471	65.6	47.1	73.8	98.4	0.69	0.45	0.77
7	Conventions	30	334.4	272	363	63.9	56.0	72.8	97.9	0.51	0.31	0.67
	Evid/Elab	30	334.4	272	363	62.5	51.5	72.7	97.6	0.67	0.58	0.76
	Org/Purp	30	334.4	272	363	63.4	53.2	73.3	97.7	0.68	0.56	0.77
8	Conventions	33	314.2	247	347	67.1	48.8	78.1	98.3	0.51	0.33	0.62
	Evid/Elab	33	314.2	247	347	61.1	47.6	72.1	97.7	0.66	0.54	0.77
	Org/Purp	33	314.2	247	347	61.1	46.4	72.1	97.9	0.67	0.59	0.74
HS	Conventions	29	402.9	383	425	71.1	63.4	78.1	98.6	0.60	0.41	0.67
	Evid/Elab	29	402.9	383	425	61.2	48.2	71.4	98.5	0.71	0.54	0.79
	Org/Purp	29	402.9	383	425	61.3	47.9	70.7	98.6	0.71	0.53	0.79

Legend: Evid/Elab: Evidence/Elaboration, Org/Purp: Organization/Purpose

Table 7.9: Interrater Agreement—ELA Smarter Balanced for Short-Answer Items

Grade	Number of Items	Number of Responses			%Exact			%(Exact+ Adjacent)	QWK		
		Mean	Min	Max	Mean	Min	Max		Mea	Min	Max
3	13	426.2	413	442	68.9	62.4	76.8	100.0	0.70	0.60	0.77
4	17	375.1	369	381	68.6	58.3	77.0	100.0	0.71	0.58	0.79
5	15	385.9	377	396	67.1	56.1	81.7	100.0	0.72	0.63	0.86
6	37	502.5	51	2182	70.3	58.8	85.4	100.0	0.66	0.33	0.86
7	44	442.0	72	2159	68.8	55.3	83.2	100.0	0.66	0.44	0.80
8	48	424.4	92	1367	69.6	55.8	83.9	100.0	0.68	0.48	0.80
HS	91	283.6	51	620	68.7	49.0	86.8	100.0	0.70	0.44	0.90

Table 7.10: Interrater Agreement—Mathematics Smarter Balanced

Grade	Score Point Range	Number of Items	Number of Responses			%Exact			% (Exact+ Adjacent)	QWK		
			Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	0–1	8	575.1	443	677	92.7	91.2	95.2	100.0	0.85	0.80	0.89
4	0–1	10	524.4	493	604	87.8	80.9	95.4	100.0	0.69	0.56	0.87
5	0–1	9	472.4	453	504	91.7	81.9	98.1	100.0	0.70	0.37	0.96
6	0–1	12	480.9	317	703	97.0	93.7	100.0	100.0	0.69	0.25	1.00
7	0–1	10	603.9	401	739	95.1	86.6	98.9	100.0	0.77	0.35	0.95
8	0–1	15	720.2	690	762	91.9	82.1	98.3	100.0	0.77	0.57	0.96
HS	0–1	15	897.4	100	1023	92.9	87.2	99.6	100.0	0.75	0.63	0.99
3	0–2	32	595.0	132	731	90.2	78.3	99.3	100.0	0.92	0.84	0.97
4	0–2	38	496.8	130	602	88.6	77.8	99.8	100.0	0.88	0.40	1.00
5	0–2	57	475.4	161	555	88.5	75.8	98.8	100.0	0.87	0.51	0.97
6	0–2	40	673.9	643	742	88.0	73.9	97.9	100.0	0.85	0.72	0.98
7	0–2	24	625.8	566	710	91.6	83.1	97.1	100.0	0.87	0.60	0.97
8	0–2	26	702.1	671	760	90.1	82.2	99.2	100.0	0.87	0.72	0.99
HS	0–2	22	884.8	550	1011	90.6	74.7	99.3	100.0	0.87	0.52	0.99
3	0–3	6	425.5	277	633	91.4	88.5	95.0	100.0	0.96	0.94	0.98
4	0–3	4	517.8	485	597	85.1	82.5	87.3	100.0	0.93	0.91	0.94
5	0–3	8	444.4	298	546	88.1	84.6	97.3	100.0	0.90	0.78	0.96
7	0–3	1	625.0	625	625	87.5	87.5	87.5	100.0	0.90	0.90	0.90
HS	0–3	7	946.1	917	988	87.1	78.6	91.0	100.0	0.90	0.88	0.92

Table 7.11: Interrater Agreement—WCAS

Subject	Item Position	Points Possible	Number Read Twice	% Exact Agreement	% (Adjacent + Exact Agreement)	% Non-Adjacent Agreement	Kappa
WCAS G5 Form A	14	1	11,256	100	100	0	0.9953
	25	1	11,312	100	100	0	0.9951
	30	1	10,878	96	100	0	0.8450
WCAS G8 Form A	27	1	11,364	100	100	0	0.9875
WCAS G11 Form A	21	2	8,245	99	100	0	0.9896
	31	2	7,938	95	100	0	0.9688

7.10 TEST RESULTS

Two sets of spring 2022 test results are provided, one for accountability and the other for graduation. Test results over time are presented in Appendix H. Due to the disruptions caused by Covid-19, there is not historical data for spring 2020 or spring 2021 as no tests were administered at those times.

Tests for Accountability

Appendix D presents the numbers of students, means, and standard deviations of scale scores for each test. Appendix E presents, for these same tests, the percentages of students by achievement

level. As stated earlier, Level 3 or above is considered proficient. Because Smarter Balanced assessments and the state-specific WCAS were scaled differently, the average scale scores cannot be compared.

Tests for Graduation Pathways

Appendix F presents the number of students, means, and standard deviations of scale scores for Smarter Balanced math and ELA tests that state-level legislation allows students to use to meet their graduation pathway, one of several requirements for graduation in Washington. WCAS is not included in this appendix as there is no testing pathway for graduation related to the WCAS. Appendix G presents, for these same tests, the percentage distribution of students by achievement levels.

SUMMARY

Smarter Balanced assessments consist of CATs and fixed-form, randomly-assigned performance tasks. In the CAT, depending on the items presented, two students having the same raw score would likely receive different scale scores. The fixed-form WCAS is scored by the number-correct method. In this approach, two students with the same raw score do have the same scale score.

Both Smarter Balanced tests and the WCAS have clearly stated rules on the handling of extreme scores (all correct or all incorrect), scoring of incomplete tests, and the definition of whether a student has attempted the test (see more information in the scoring specifications).

Some items in both Smarter Balanced tests and the WCAS needed to be handscored. The vendor that conducts handscoring follows a set of approved rules and procedures that govern the recruiting, training, monitoring, read-behind, and, if needed, the re-training and dismissal of human raters. As a result, the rater agreement is at 95% or higher for the exact and adjacent agreement and at 0.49 or higher for the quadratic weighted Kappa in Smarter Balanced assessments. For the WCAS, the interrater reliability index was at least 0.84.

8. RELIABILITY

Reliability refers to the consistency in test scores. For fixed-length tests, reliability can also refer to the internal consistency of test items. In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming the error variance is the same for all scores. Reliability is evaluated in terms of the standard errors of measurement (SEM). Within the item response theory (IRT) framework, measurement error varies based on ability. The amount of precision in estimating achievement can be determined by the test information, which describes the amount of information provided by the test at each score point along the ability continuum. Test information is a value that is the inverse of the measurement error of the test; the larger the measurement error, the less test information is being provided. In computer-adaptive tests (CATs), items administered vary among students, so the amount of measurement error differs from one test to another, which yields the conditional standard error of measurement (CSEM).

In this chapter, the evidence of reliabilities—score reliability, internal consistency reliability, SEM, CSEM, classification accuracy, and consistency of achievement-level assignments—is computed for the Washington Comprehensive Assessment Program (WCAP) assessments.

8.1 SMARTER BALANCED ASSESSMENTS

The reliability evidence of the Smarter Balanced summative tests is provided with marginal reliability, SEM, and classification accuracy and consistency in each achievement level.

8.1.1 Marginal Reliability

Marginal reliability was computed for the scale scores and took into account the varying measurement errors across the ability range. Marginal reliability is a measure of the overall reliability of an assessment based on the average CSEM, estimated at different points on the ability scale, for all students.

The marginal reliability ($\bar{\rho}$) is defined as

$$\bar{\rho} = [\sigma^2 - \left(\frac{\sum_{i=1}^N CSEM_i^2}{N} \right)] / \sigma^2,$$

where N is the number of students, $CSEM_i$ is the CSEM of the scale score for student i , and σ^2 is the variance of the scale scores. The higher the reliability coefficient, the greater the precision of the test.

Another way to examine test reliability is with the SEM. In IRT, SEM is estimated as a function of test information provided by a given set of items that make up the test. In computer-adaptive testing (CAT), items administered vary among all students, so the SEM also can vary among students, which yields CSEM. The average CSEM can be computed as

$$\text{Average CSEM} = \sigma \sqrt{1 - \bar{\rho}} = \sqrt{\sum_{i=1}^N CSEM_i^2 / N}.$$

The smaller the value of average CSEM, the greater the accuracy of test scores.

Test Reliability

Table 8.1 presents the marginal reliability coefficients and the average CSEM for the total scale scores. The reliability indexes for the total scores are at 0.87 or above in ELA and 0.84 or above in mathematics, indicating that the Smarter Balanced assessments have high reliability.

Table 8.1: Marginal Reliability for Smarter Balanced ELA and Mathematics

Grade	Number of Items Specified in Test Blueprint		Marginal Reliability	N	Scale Score Mean	Scale Score SD	Average CSEM
	Min	Max					
ELA							
3	22	22	0.88	76,355	2425.63	101.21	35.42
4	22	22	0.87	75,944	2470.10	102.30	37.57
5	22	22	0.88	77,054	2507.11	105.50	37.07
6	24	24	0.88	76,258	2516.57	101.95	35.65
7	24	24	0.88	78,145	2553.66	109.51	37.78
8	24	24	0.88	79,659	2565.80	110.25	37.62
HS	24	24	0.88	88,682	2608.78	119.49	41.61
Mathematics							
3	22	23	0.91	76,703	2432.25	97.05	28.92
4	20	23	0.91	76,164	2472.84	97.55	29.41
5	21	23	0.90	77,298	2494.84	104.12	33.56
6	22	23	0.89	76,429	2505.38	116.11	39.28
7	21	23	0.88	78,114	2523.38	121.48	42.50
8	21	23	0.87	79,593	2534.25	128.93	46.08
HS	22	24	0.84	98,726	2547.06	131.15	52.51

Reliability by Student Group

Tables 8.2 and 8.3 show the marginal reliability coefficients for each of the student groups: including gender and ethnicity groups. As shown in the tables, the reliability coefficients are similar across student groups but somewhat lower for multilingual learner (ML) and Individuals with Disabilities Education Act (IDEA) student groups, a large percentage of whom received Level 1 with large SEMs.

Table 8.2: Marginal Reliability Coefficients for Overall and by Student Group: ELA

Student Group	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
	Rel	CSEM	Rel	CSEM	Rel	CSEM	Rel	CSEM	Rel	CSEM	Rel	CSEM	Rel	CSEM
All Students	0.88	35.42	0.87	37.57	0.88	37.07	0.88	35.65	0.88	37.78	0.88	37.62	0.88	41.61
Gender														
Female	0.88	35.19	0.86	37.30	0.87	36.96	0.88	35.45	0.88	37.29	0.88	37.23	0.87	41.13
Male	0.88	35.65	0.86	37.83	0.88	37.16	0.88	35.83	0.88	38.23	0.88	38.00	0.88	42.04
Ethnic Group														
African American	0.85	36.07	0.84	38.01	0.86	37.46	0.86	36.42	0.87	38.91	0.87	38.88	0.87	43.11
Amer. Indian or Alaskan	0.83	39.18	0.83	39.74	0.85	37.80	0.85	37.32	0.86	41.54	0.86	39.30	0.85	43.63
Asian	0.88	35.44	0.86	38.06	0.87	38.41	0.87	36.37	0.87	37.78	0.87	37.75	0.87	41.99
Hispanic	0.85	36.50	0.84	38.21	0.86	36.85	0.86	35.96	0.86	38.76	0.87	38.34	0.87	42.20
Pacific Islander	0.81	36.96	0.82	39.01	0.85	36.89	0.84	35.99	0.85	40.15	0.85	39.39	0.84	43.20
White	0.87	34.73	0.85	37.03	0.86	36.89	0.86	35.25	0.87	37.02	0.87	37.02	0.87	41.03
Multiple	0.88	35.02	0.86	37.43	0.87	37.03	0.88	35.57	0.88	37.44	0.88	37.38	0.87	41.34
ML														
Yes	0.80	38.08	0.77	39.99	0.79	38.15	0.76	38.55	0.76	43.41	0.77	42.80	0.77	46.50
No	0.87	34.88	0.86	37.16	0.86	36.91	0.87	35.28	0.87	37.11	0.87	37.05	0.86	40.99
IDEA														
Yes	0.84	38.38	0.83	40.97	0.85	39.02	0.82	38.82	0.82	43.29	0.82	42.17	0.81	46.02
No	0.87	34.95	0.86	37.04	0.86	36.77	0.87	35.18	0.87	36.94	0.87	36.98	0.87	40.98
Section 504														
Yes	0.87	34.81	0.84	36.82	0.86	36.40	0.86	35.21	0.86	37.05	0.87	36.93	0.86	40.73
No	0.88	35.43	0.87	37.59	0.88	37.09	0.88	35.67	0.88	37.80	0.88	37.65	0.88	41.67
Economically Disadvantaged														
Yes	0.85	35.69	0.84	37.51	0.85	36.48	0.85	35.48	0.86	38.06	0.86	37.77	0.86	42.03
No	0.86	34.40	0.84	36.92	0.84	37.29	0.85	35.39	0.86	36.81	0.86	36.87	0.85	40.92

Note. Rel: Marginal reliability

Table 8.3: Marginal Reliability Coefficients for Overall and by Student Group: Mathematics

Student Group	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
	Rel	CSEM	Rel	CSEM	Rel	CSEM	Rel	CSEM	Rel	CSEM	Rel	CSEM	Rel	CSEM
All Students	0.91	28.92	0.91	29.41	0.90	33.56	0.89	39.28	0.88	42.50	0.87	46.08	0.84	52.51
Gender														
Female	0.91	28.73	0.90	29.10	0.89	33.16	0.88	38.73	0.87	42.53	0.87	45.70	0.83	51.69
Male	0.91	29.10	0.91	29.71	0.90	33.92	0.89	39.78	0.88	42.49	0.88	46.45	0.85	53.27
Ethnic Group														
African American	0.88	30.77	0.88	32.15	0.85	37.51	0.83	45.60	0.82	47.90	0.79	52.89	0.73	59.60
Amer. Indian or Alaskan	0.87	33.08	0.86	35.18	0.83	39.66	0.81	49.38	0.79	50.00	0.78	54.43	0.68	63.14
Asian	0.92	29.19	0.92	29.49	0.91	31.63	0.91	36.24	0.91	38.40	0.91	41.74	0.90	44.24
Hispanic	0.88	31.13	0.87	31.38	0.85	36.51	0.83	43.96	0.82	47.23	0.80	51.03	0.74	58.79
Pacific Islander	0.85	32.73	0.85	34.46	0.83	39.91	0.79	49.66	0.77	52.41	0.76	54.38	0.68	62.11
White	0.91	27.36	0.90	27.86	0.89	31.55	0.88	35.96	0.88	39.50	0.87	43.03	0.85	48.96
Multiple	0.91	28.30	0.91	28.68	0.90	33.00	0.89	38.43	0.88	41.64	0.87	44.71	0.85	51.40
ML														
Yes	0.85	33.38	0.83	35.12	0.77	41.69	0.72	54.27	0.66	57.66	0.65	61.94	0.55	70.10
No	0.91	27.93	0.91	28.33	0.90	32.18	0.89	36.96	0.88	40.43	0.87	44.08	0.85	50.18

IDEA														
Yes	0.88	35.97	0.86	37.25	0.82	43.63	0.77	55.24	0.73	58.96	0.70	61.16	0.57	72.03
No	0.91	27.69	0.91	28.06	0.89	31.78	0.89	36.49	0.88	39.70	0.87	43.76	0.84	49.47
Section 504														
Yes	0.91	27.54	0.90	27.63	0.89	32.19	0.88	36.52	0.87	40.26	0.86	44.58	0.84	49.41
No	0.91	28.94	0.91	29.46	0.90	33.60	0.89	39.37	0.88	42.58	0.87	46.15	0.84	52.72
Economically Disadvantaged														
Yes	0.89	29.40	0.88	29.93	0.86	34.82	0.84	42.04	0.83	45.44	0.81	49.16	0.75	57.38
No	0.90	26.53	0.90	26.93	0.90	29.80	0.89	33.94	0.89	37.25	0.89	40.75	0.87	46.24

Reliability by Claim

For the spring 2022 assessments, claim scores were not generated, so reliability of claim scores was not computed.

8.1.2 Conditional Standard Error of Measurement

Figures 8.1 and 8.2 present plots of the scale score CSEM across the range of ability. The item selection algorithm selected items efficiently, matching to each student's ability while matching to the test blueprints.

Overall, the standard error curves suggest that students are measured with a high degree of precision given that the standard errors are consistently low. However, larger standard errors are observed at the lower ends of the score distribution relative to the higher ends. This occurs because the item pools currently have a shortage of easy items that more precisely measure student performance at the low end of the score distribution. Content experts should use this information to consider how to further target and populate item pools.

Figure 8.1: CSEM for Smarter Balanced ELA

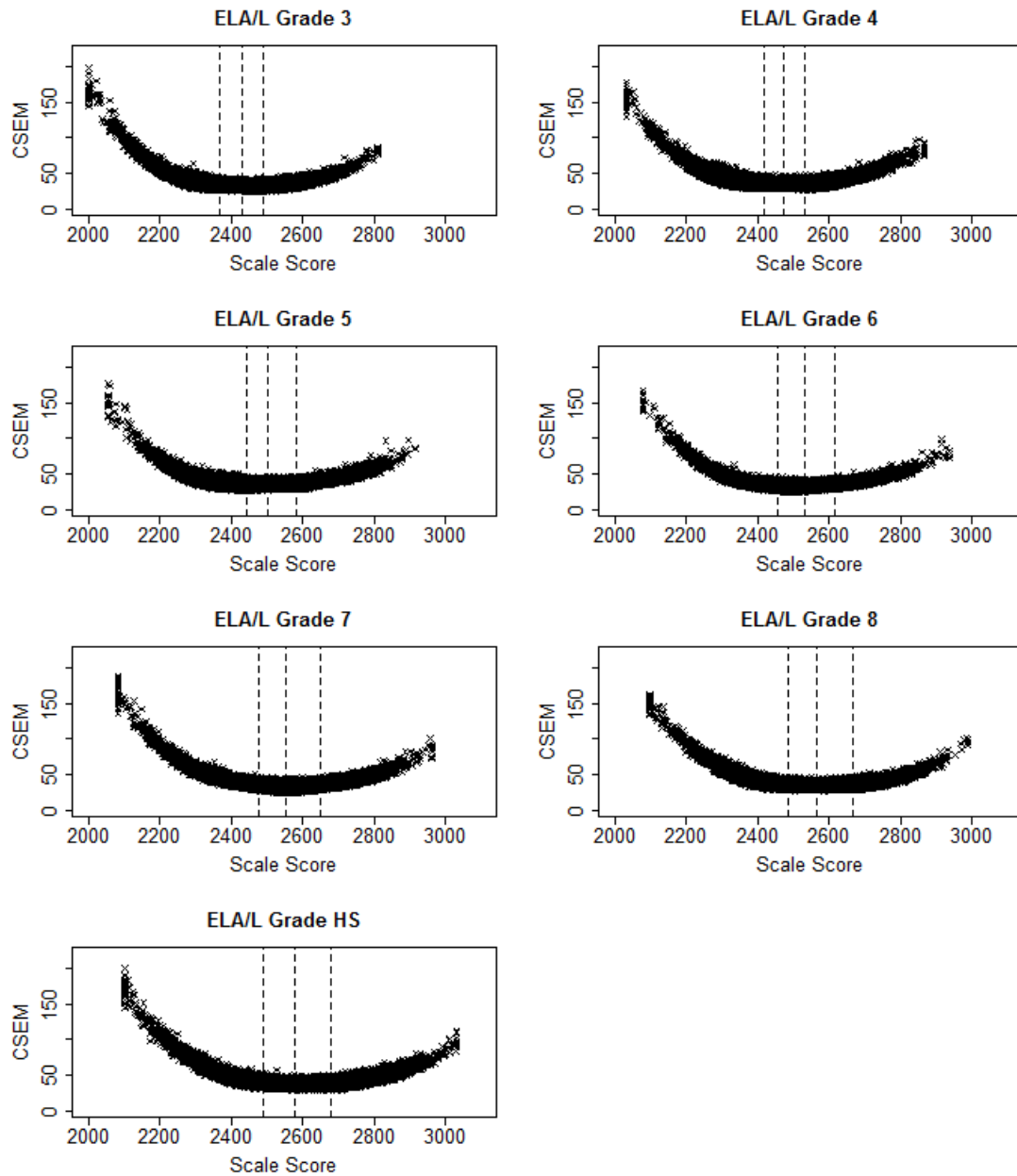
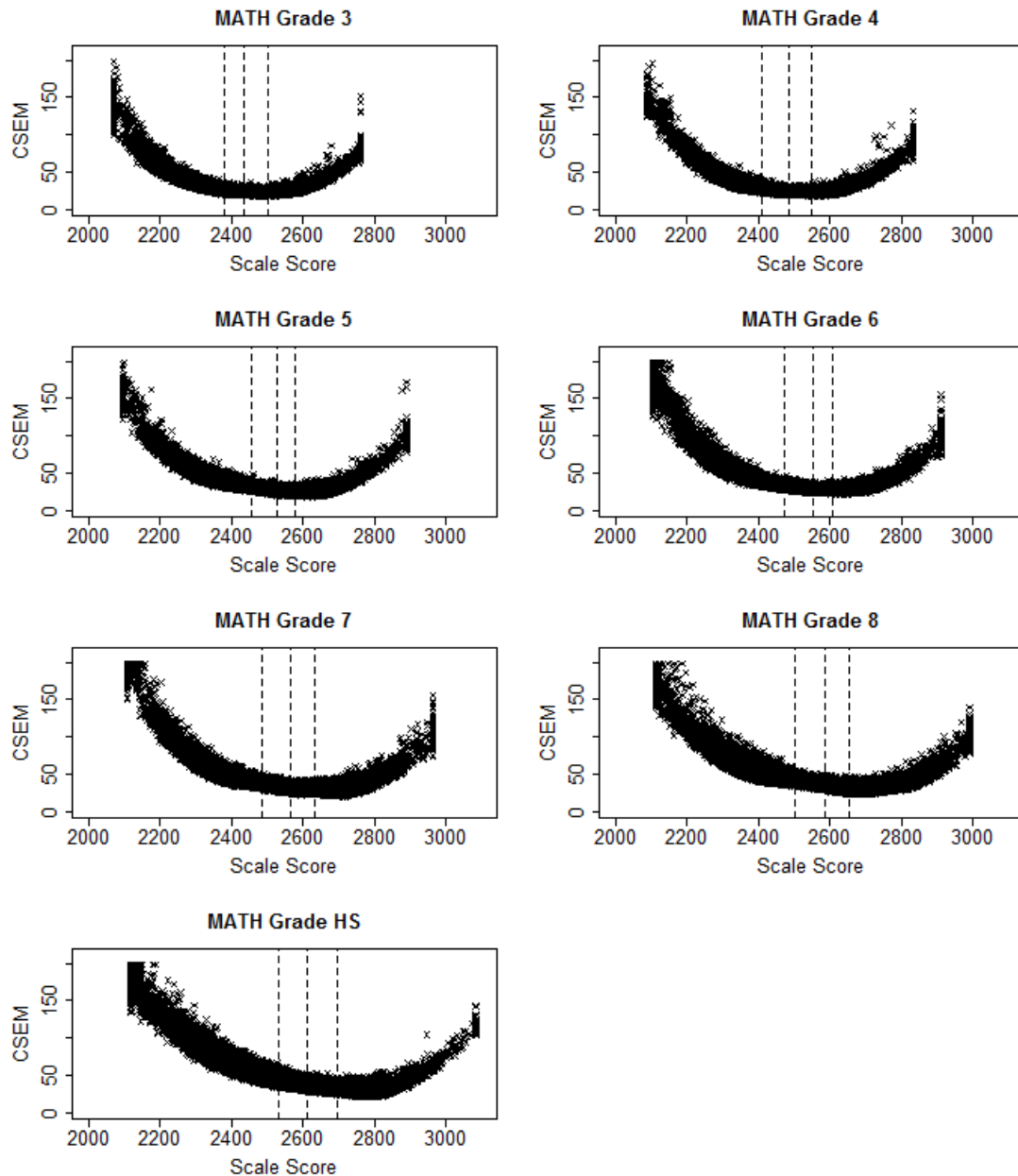


Figure 8.2: CSEM for Smarter Balanced Mathematics



8.1.3 Classification Accuracy and Consistency

When student performance is reported in terms of achievement levels, a reliability of achievement classification is computed in terms of the probabilities of accurate and consistent classification of students as specified in Standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The indexes consider the accuracy and consistency of classifications.

The classification index can be examined in terms of classification accuracy and classification consistency. Classification accuracy refers to the agreement between the classifications based on the

form actually taken and the classifications that would be made on the basis of the test takers' true scores if their true scores were knowable. Classification consistency refers to the agreement between the classifications based on the form actually taken (adaptively administered items) and the classifications that would be made on the basis of an alternate form (another set of adaptively administered items given the same ability)—that is, the percentages of students who are consistently classified in the same achievement levels on two equivalent test forms.

In reality, true ability is unknowable, and students do not take an alternate, equivalent form; the classification accuracy and the classification consistency are therefore estimated on the basis of students' item scores, the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with a measurement error.

For the i th student, the student's estimated ability is $\hat{\theta}_i$ with SEM of $se(\hat{\theta}_i)$, and the estimated ability is distributed as $\hat{\theta}_i \sim N(\theta_i, se^2(\hat{\theta}_i))$, assuming a normal distribution, where θ_i is the unknown true ability of the i th student. The probability of the true score at achievement level l based on the cut scores c_{l-1} and c_l is estimated as

$$\begin{aligned} p_{il} &= p(c_{l-1} \leq \theta_i < c_l) = p\left(\frac{c_{l-1} - \hat{\theta}_i}{se(\hat{\theta}_i)} \leq \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} < \frac{c_l - \hat{\theta}_i}{se(\hat{\theta}_i)}\right) \\ &= p\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)} < \frac{\hat{\theta}_i - \theta_i}{se(\hat{\theta}_i)} \leq \frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) = \Phi\left(\frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) - \Phi\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)}\right). \end{aligned}$$

Instead of assuming a normal distribution of $\hat{\theta}_i \sim N(\theta_i, se^2(\hat{\theta}_i))$, the above probability is estimated directly using the likelihood function.

The likelihood function of theta given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut point (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point. If a student with estimated theta is below the cut point, a probability of at or above the cut point is an estimate of the chance that this student is misclassified as below the cut, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

The probability of the i th student being classified at achievement level l ($l = 1, 2, \dots, L$) based on the cut scores cut_{l-1} and cut_l , given the student's item scores $\mathbf{z}_i = (z_{i1}, \dots, z_{ij})$ and item parameters $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_J)$, using the J administered items, can be estimated as

$$\begin{aligned} p_{il} &= P(cut_{l-1} \leq \theta_i < cut_l | \mathbf{z}, \mathbf{b}) = \frac{\int_{cut_{l-1}}^{cut_l} L(\theta | \mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{z}, \mathbf{b}) d\theta} \text{ for } l = 2, \dots, L-1, \\ p_{i1} &= P(-\infty < \theta_i < cut_1 | \mathbf{z}, \mathbf{b}) = \frac{\int_{-\infty}^{cut_1} L(\theta | \mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{z}, \mathbf{b}) d\theta} \\ p_{iL} &= P(cut_{L-1} \leq \theta_i < \infty | \mathbf{z}, \mathbf{b}) = \frac{\int_{cut_{L-1}}^{\infty} L(\theta | \mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{z}, \mathbf{b}) d\theta} \end{aligned}$$

where the likelihood function, based on general IRT models, is

$$L(\theta|\mathbf{z}_i, \mathbf{b}) = \prod_{j \in d} \left(z_{ij} c_j + \frac{(1-c_j) \exp(z_{ij} D a_j (\theta - b_j))}{1 + \exp(D a_j (\theta - b_j))} \right) \prod_{j \in p} \left(\frac{\exp(D a_j (z_{ij} \theta - \sum_{k=1}^{z_{ij}} b_{jk}))}{1 + \sum_{m=1}^{K_j} \exp(D a_j (\sum_{k=1}^m (\theta - b_{jk})))} \right),$$

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = (a_j, b_j, c_j)$ if the j th item is a dichotomous item, and $\mathbf{b}_j = (a_j, b_{j1}, \dots, b_{jK_j})$ if the j th item is a polytomous item; a_j is the item's discrimination parameter (for Rasch model, $a_j = 1$), c_j is the guessing parameter (for Rasch and 2PL models, $c_j = 0$), and D is 1.7 for non-Rasch models and 1 for Rasch model.

Classification Accuracy

Using p_{il} , a $L \times L$ table can be constructed as

$$\begin{pmatrix} n_{a11} & \cdots & n_{a1L} \\ \vdots & \vdots & \vdots \\ n_{aL1} & \cdots & n_{aLL} \end{pmatrix}$$

where $n_{alm} = \sum_{pl_i=l} p_{im}$. n_{alm} is the expected number of students at achievement level lm , pl_i is the i th student's achievement level, and p_{im} are the probabilities of the i th student being classified at achievement level m . In the above table, the row represents the observed level and the column represents the expected level.

The classification accuracy (CA) at level l ($l = 1, \dots, L$) is estimated by

$$CA_l = \frac{n_{all}}{\sum_{m=1}^L n_{alm}},$$

and the overall classification accuracy is estimated by

$$CA = \frac{\sum_{l=1}^L n_{all}}{N},$$

where N is the total number of students.

Classification Consistency

Using p_{il} , which is similar to accuracy, another $L \times L$ table can be constructed by assuming the test is administered twice independently to the same student group, hence:

$$\begin{pmatrix} n_{c11} & \cdots & n_{c1L} \\ \vdots & \vdots & \vdots \\ n_{cL1} & \cdots & n_{cLL} \end{pmatrix}$$

where $n_{clm} = \sum_{i=1}^N p_{il} p_{im}$. p_{il} and p_{im} are the probabilities of the i th student being classified at achievement level l and m , respectively, based on observed scores and hypothetical scores from an equivalent test form.

The classification consistency (CC) at level l ($l = 1, \dots, L$) is estimated by

$$CC_l = \frac{n_{cll}}{\sum_{m=1}^L n_{clm}},$$

and the overall classification consistency is

$$CC = \frac{\sum_{l=1}^L n_{cll}}{N}.$$

The analysis of the classification index is performed based on overall scale scores. Table 8.6 provides classification accuracy and consistency for Smarter Balanced assessments.

Table 8.6: Smarter Balanced Classification Accuracy and Consistency

Grade	Achievement Level	ELA		Mathematics	
		Accuracy	Consistency	Accuracy	Consistency
3	Overall	0.75	0.66	0.78	0.70
	L1	0.88	0.81	0.86	0.80
	L2	0.61	0.49	0.65	0.52
	L3	0.57	0.46	0.72	0.62
	L4	0.86	0.79	0.88	0.82
	Proficiency Cut Point	0.91	0.87	0.92	0.89
4	Overall	0.73	0.65	0.79	0.71
	L1	0.88	0.80	0.88	0.82
	L2	0.53	0.42	0.72	0.62
	L3	0.56	0.45	0.71	0.61
	L4	0.85	0.78	0.87	0.81
	Proficiency Cut Point	0.90	0.86	0.92	0.89
5	Overall	0.74	0.66	0.79	0.71
	L1	0.87	0.80	0.88	0.83
	L2	0.56	0.45	0.69	0.58
	L3	0.65	0.55	0.62	0.50
	L4	0.84	0.76	0.88	0.82
	Proficiency Cut Point	0.90	0.87	0.93	0.90
6	Overall	0.75	0.66	0.79	0.71
	L1	0.88	0.81	0.90	0.85
	L2	0.65	0.54	0.69	0.59
	L3	0.69	0.59	0.62	0.50
	L4	0.82	0.71	0.86	0.79
	Proficiency Cut Point	0.91	0.87	0.92	0.88
7	Overall	0.76	0.67	0.78	0.70
	L1	0.88	0.80	0.89	0.84
	L2	0.63	0.52	0.67	0.56
	L3	0.72	0.63	0.65	0.53
	L4	0.83	0.73	0.87	0.79
	Proficiency Cut Point	0.91	0.87	0.91	0.87
8	Overall	0.76	0.67	0.77	0.69
	L1	0.87	0.80	0.87	0.82
	L2	0.66	0.55	0.63	0.51
	L3	0.72	0.64	0.61	0.49
	L4	0.82	0.72	0.88	0.81
	Proficiency Cut Point	0.91	0.87	0.92	0.89
HS	Overall	0.76	0.68	0.78	0.70
	L1	0.86	0.78	0.89	0.84
	L2	0.65	0.53	0.61	0.50

Grade	Achievement Level	ELA		Mathematics	
		Accuracy	Consistency	Accuracy	Consistency
	L3	0.70	0.61	0.67	0.54
	L4	0.85	0.78	0.88	0.79
	Proficiency Cut Point	0.92	0.88	0.92	0.88

For spring 2022, the overall classification index ranged from 0.73 to 0.79 for the accuracy and from 0.65 to 0.71 for the consistency across all grades and subjects. For achievement levels, the classification index is higher in L1 and L4 than in L2 and L3. The higher classification index at L1 and L4 is due to the intervals used to compute the classification probability to classify students into L1 $[-\infty, \text{L2 cut}]$ or L4 $[\text{L4 cut}, \infty]$ being wider than the intervals used in L2 $[\text{L2 cut}, \text{L3 cut}]$ and L3 $[\text{L3 cut}, \text{L4 cut}]$. The misclassification probability tends to be higher for narrower intervals. The classification index at the proficiency cut point is high, ranging from 0.90 to 0.93 for the accuracy and from 0.86 to 0.90 for the consistency.

Accuracy of classifications is higher than the consistency of classifications in all achievement levels. The accuracy is higher than the consistency because the accuracy is based on one test with a measurement error and the true score while the consistency is based on two tests with measurement errors.

8.2 WASHINGTON COMPREHENSIVE ASSESSMENT OF SCIENCE (WCAS)

The reliability evidence of the WCAS tests is provided with reliability, SEM, CSEM and classification accuracy and consistency in each achievement level.

8.2.1 Internal Consistency

Internal consistency reliability indicators demonstrate how well items within a test are related. For the fixed-form WCAS, internal consistency can be estimated by Cronbach's coefficient alpha, which is a lower-bound estimate of test reliability. It is considered as a measure of scale reliability. Alpha coefficients range from 0 to 1. The closer an alpha is to 1, the more reliable the test is. An alpha of 0.8 or above is considered acceptable for tests with modest test lengths, such as the WCAP tests.

Cronbach's coefficient alpha was computed as:

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{i=1}^n \sigma_i^2}{\sigma_x^2} \right]$$

where n is the sample size, σ_i^2 is the raw score variance for item i . σ_x^2 is the variance of the total raw scores. Cronbach alpha for each test overall, by student groups and by reporting areas, is computed and provided in Tables 8.7–8.9. At the population level, the alpha coefficients were above 0.75 except for ML in grades 5, 8, and 11, which were 0.72, 0.75, and 0.69, respectively.

Table 8.7: Grade 5 WCAS Form A Test Reliability Estimates

		Sample Size	Maximum Possible Raw Score	Alpha Coefficient	Scale Score SEM
Total		76,273	35	0.88	27.80
Gender					
Male		39,070	35	0.89	27.80
Female		37,123	35	0.88	27.75
Ethnic Group					
American Indian/Alaskan		917	35	0.84	27.98
Asian		6,747	35	0.89	27.80
Black		3,665	35	0.84	27.77
Hispanic		19,465	35	0.84	27.72
Non-Hispanic White		37,420	35	0.87	27.76
Native Hawaiian and/or Pacific Islander		1,010	35	0.80	27.92
Multiple Races		6,865	35	0.88	27.81
Unknown/Missing		184	35	0.86	27.73
Program					
Multilingual Learner	Yes	9,738	35	0.72	28.53
	No	66,371	35	0.87	27.76
Special Education	Yes	9,803	35	0.86	27.61
	No	66,299	35	0.88	27.69
Migrant	Yes	1,444	35	0.77	28.10
	No	68,031	35	0.88	27.77
Economically Disadvantaged	Yes	29,061	35	0.84	27.67
	No	35,181	35	0.87	27.74

Table 8.8: Grade 8 WCAS Form A Test Reliability Estimates

		Sample Size	Maximum Possible Raw Score	Alpha Coefficient	Scale Score SEM
Total		78,466	40	0.91	25.83
Gender					
Male		40,259	40	0.91	25.84
Female		37,922	40	0.90	25.78
Ethnic Group					
American Indian/Alaskan		960	40	0.87	25.48
Asian		6,943	40	0.91	26.36
Black		3,396	40	0.87	25.46
Hispanic		20,443	40	0.88	25.48
Non-Hispanic White		38,669	40	0.90	25.92
Native Hawaiian and/or Pacific Islander		1,034	40	0.86	25.36
Multiple Races		6,871	40	0.91	25.81
Unknown/Missing		150	40	0.90	26.09
Program					
Multilingual Learner	Yes	7,339	40	0.75	26.04
	No	70,994	40	0.90	25.90
Special Education	Yes	8,823	40	0.87	25.29
	No	69,514	40	0.90	25.88
Migrant	Yes	1,697	40	0.83	25.56
	No	73,765	40	0.91	25.86
Economically Disadvantaged	Yes	32,403	40	0.88	25.53
	No	39,658	40	0.90	26.11

Table 8.9: Grade 11 WCAS Form A Test Reliability Estimates

		Sample Size	Maximum Possible Raw Score	Alpha Coefficient	Scale Score SEM
Total		55,946	45	0.87	26.67
Gender					
Male		29,320	45	0.89	26.57
Female		26,420	45	0.85	26.69
Ethnic Group					
American Indian/Alaskan		671	45	0.82	27.77
Asian		4,746	45	0.89	26.36
Black		2,222	45	0.82	28.01
Hispanic		14,258	45	0.83	27.67
Non-Hispanic White		29,012	45	0.87	26.26
Native Hawaiian and/or Pacific Islander		654	45	0.82	28.90
Multiple Races		4,268	45	0.87	26.38
Unknown/Missing		115	45	0.84	27.39
Program					
Multilingual Learner	Yes	4,968	45	0.69	30.48
	No	50,879	45	0.87	26.33
Special Education	Yes	5,498	45	0.83	27.18
	No	50,349	45	0.87	26.43
Migrant	Yes	1,316	45	0.79	29.18
	No	53,639	45	0.87	26.61
Economically Disadvantaged	Yes	22,476	45	0.83	27.45
	No	31,293	45	0.87	26.15

8.2.2 Standard Error of Measurement

The SEM estimates how precisely a test can measure students' true abilities, named true scores. True scores are unknown because no test can perfectly provide a reflection of student true abilities. SEM is directly related to the reliability of a test. The larger the SEM, the lower the reliability of a test and the less precise the test scores are. SEM on the scale score scale is computed as follows:

$$SEM = SD\sqrt{1 - R_{xx'}}$$

where SD represents the standard deviation of the scale score distribution, and $R_{xx'}$ is the estimated Cronbach's alpha coefficient.

SEM can be used to construct the score band that a test-taker's true score is expected to fall within. Assuming normal distribution, plus and minus two times the SEMs will produce a score band in which, approximately 95% of the time, the test-taker's true score will fall.

The scale score SEMs are presented in Table 8.10.

Table 8.10: Reporting Area Reliabilities by Test, WCAS, 2022 Administration

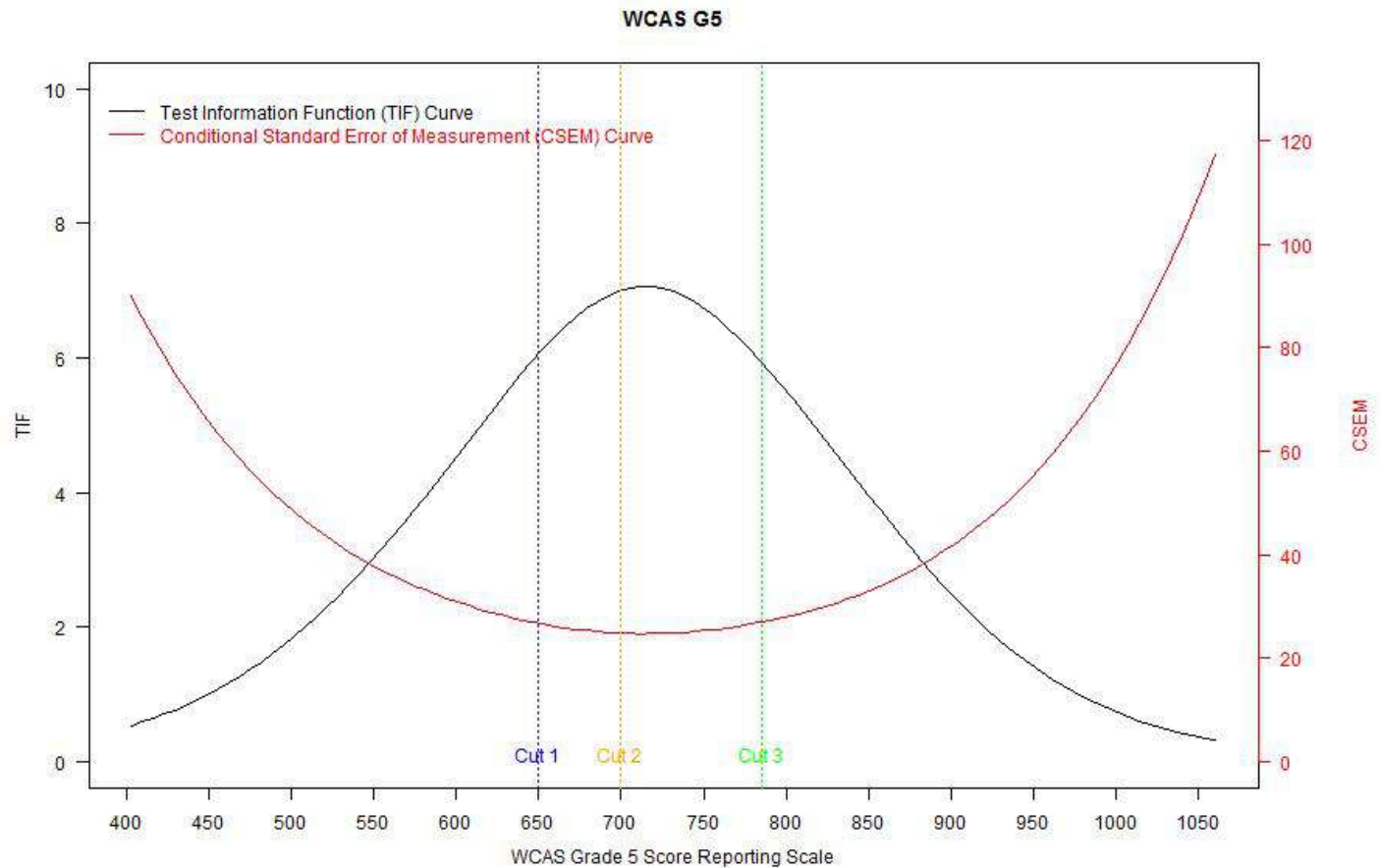
Subject	Reporting Area	N Count	N Item	Max	Alpha	SEM
WCAS G5 Form A	Practices and Crosscutting Concepts in Earth & Space Science	76,273	9	12	0.75	50.04
	Practices and Crosscutting Concepts in Life Science	76,273	12	12	0.79	50.38
	Practices and Crosscutting Concepts in Physical Science	76,273	11	14	0.67	47.79
WCAS G8 Form A	Practices and Crosscutting Concepts in Earth & Space Science	78,466	11	12	0.78	50.58
	Practices and Crosscutting Concepts in Life Science	78,466	12	16	0.78	43.79
	Practices and Crosscutting Concepts in Physical Science	78,466	11	14	0.79	46.95
WCAS HS Form A	Practices and Crosscutting Concepts in Earth & Space Science	55,946	10	12	0.75	52.88
	Practices and Crosscutting Concepts in Life Science	55,946	12	15	0.71	48.42

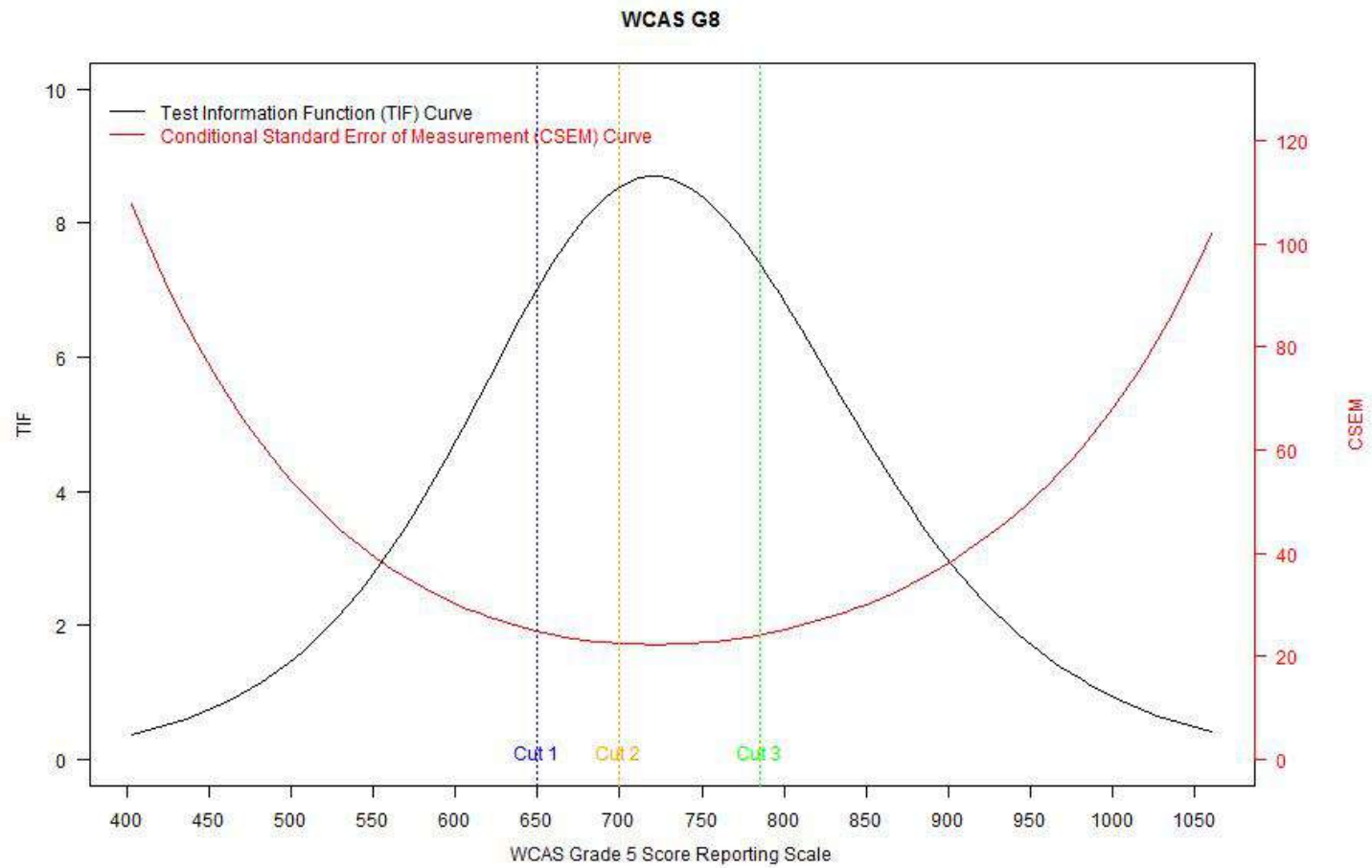
Subject	Reporting Area	N Count	N Item	Max	Alpha	SEM
	Practices and Crosscutting Concepts in Physical Science	55,946	14	18	0.66	46.18

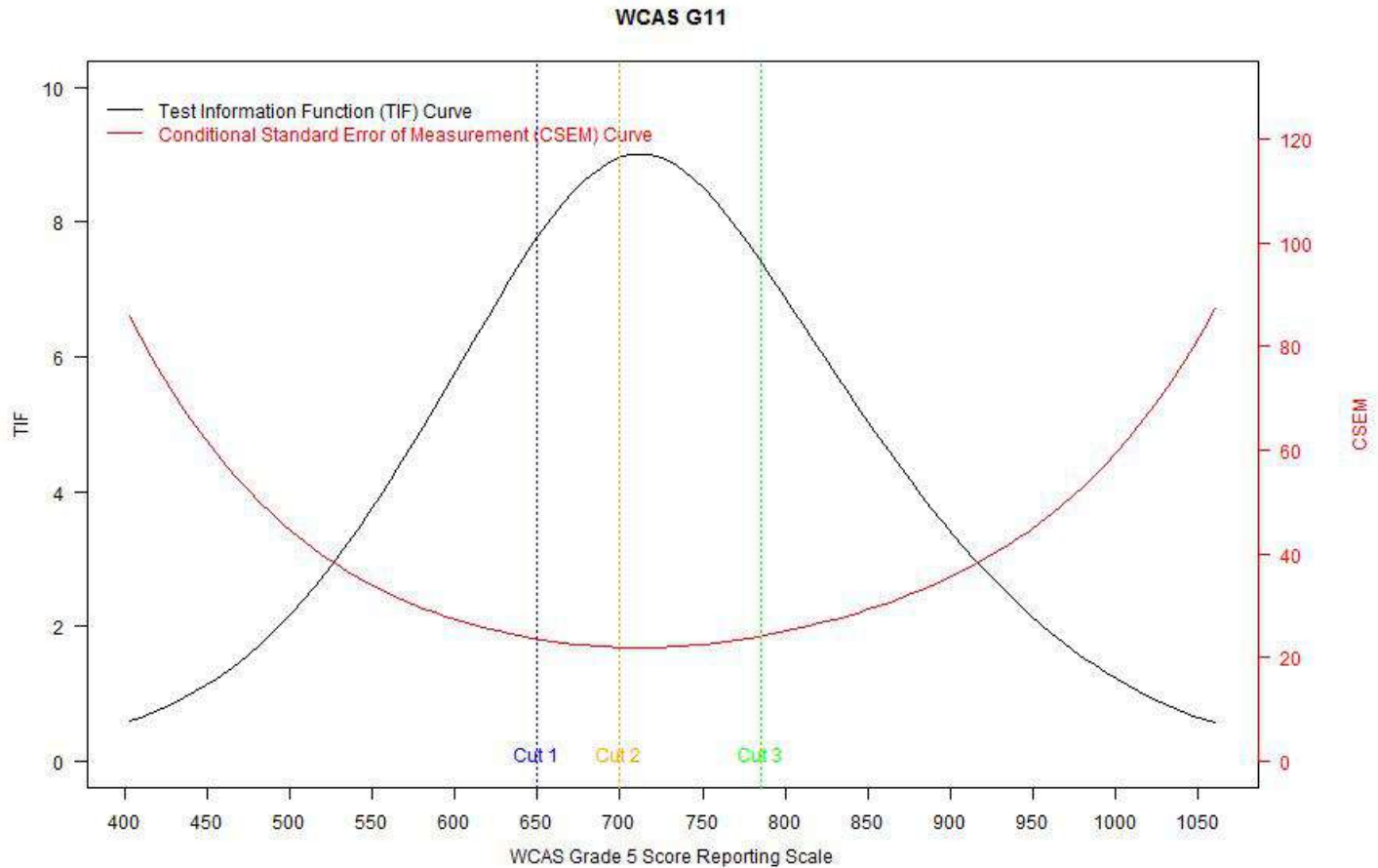
8.2.3 Conditional Standard Error of Measurement

As stated in Section 7.1, the CSEM is the inverse of the square root of the test information function (TIF) conditioned on each specific theta point on the logit scale. Along the logit scale, typically, CSEMs are smaller toward the center of a scale, where more items and more test information are available, and larger toward both ends of the scale, where fewer items and less test information are available. The TIF and CSEM plots for the WCAS are presented in Figure 8.3.

Figure 8.3: TIF and CSEM for the WCAS







8.2.4 Classification Accuracy and Consistency

The computation for classification accuracy and consistency for the WCAS is the same as the computation for Smarter Balanced assessments, except that 1PL is used to compute probability. The formulas for Smarter Balanced assessments are presented in Section 8.1.3.

The results for the WCAS are provided in Table 8.11. The results show that both consistency and accuracy at the cut point for proficiency or better are quite high, ranging between 0.86 and 0.93.

Table 8.11: Classification Consistency and Accuracy

Subject	N	Overall Accuracy	Overall Consistency	Proficiency Cut Point Accuracy	Proficiency Cut Point Consistency
WCAS G5 Form A	76,273	0.75	0.67	0.91	0.88
WCAS G8 Form A	78,466	0.78	0.71	0.93	0.90
WCAS G11 Form A	55,946	0.76	0.67	0.90	0.86

SUMMARY

Reliability in Smarter Balanced assessments is captured by the marginal reliability measure, derived from the CSEM. The marginal reliability measures for all students range from 0.87 to 0.88 in ELA and from 0.84 to 0.91 in mathematics. The CSEM curves of Smarter Balanced tests did not show any abnormalities, with the lower end of thetas showing larger standard errors. The overall classification accuracy of Smarter Balanced assessment ranges between 0.73 and 0.76 in ELA and 0.77 and 0.79 in mathematics. The overall classification consistency ranges between 0.65 and 0.68 in ELA and 0.69 and 0.71 in mathematics. Classification accuracy and classification consistency at the proficiency level cut were high, ranging between 0.90 and 0.93 for classification accuracy and ranging between 0.86 and 0.89 for classification consistency.

The scale reliability for the WCAS tests, which are fixed-form, is estimated using Cronbach's Alpha. For the 2022 administration, the Alpha coefficients were at or above 0.80 except for ML and Migrant students. Classification accuracy and classification consistency at the proficiency level cut were also high, ranging between 0.86 and 0.93

9. VALIDITY

Validity refers to the degree to which evidence and theory support the interpretation of test scores for the proposed uses of tests. (America Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). It is the central concern underlying test development, administration, scoring, reporting, and the uses and interpretations of test scores. The validity of an intended interpretation of test scores relies on all of the evidence accrued about the technical quality of a testing system, including test development and construction procedures, test score reliability, accurate scaling and equating, procedures for setting meaningful achievement standards, standardized test administration and scoring procedures, and attention to fairness for all test-takers. The appropriateness and usefulness of the General Summative Assessments depends on the assessments meeting the relevant standards of validity.

This chapter focuses on presenting additional validity evidence that has been gathered for Smarter Balanced assessments and the state-specific WCAS.

9.1 SMARTER BALANCED ASSESSMENTS

For Smarter Balanced assessments, validity evidence provided in this chapter includes:

- Test Content
- Relations to Other Variables
- Student Ability vs. Test Difficulties

Evidence on test content validity is provided with the blueprint match rates for the delivered tests. Evidence on relations to other variables is provided with relationships between course grades and performance on the Smarter Balanced tests. Lastly, the empirical distribution of the Washington student scale scores and the distribution of item difficulty parameters are provided.

Some of the evidence on standardized test administration, scoring procedures, and attention to fairness for all test-takers is provided in other chapters.

9.1.1 Evidence on Test Content

The Smarter Balanced summative assessment includes two components: a computer-adaptive test (CAT) and a performance task (PT). For the CAT, each student receives a different set of items, adapting to his or her ability. For the PT, each student is randomly administered a fixed-form test. All PTs adhere to the same blueprint design.

In the adaptive item-selection algorithm, item selection takes place in two discrete stages: blueprint satisfaction and match-to-ability. The Smarter Balanced blueprints specify a range of items to be administered in each claim, content domain/standards, and targets. Moreover, blueprints constrain the DOK and item and passage types. For DOK constraints, the Smarter Balanced blueprint specifies the minimum number of items, not the maximum. In blueprints, all content blueprint elements are configured to obtain a strictly enforced range of items administered. The algorithm also seeks to satisfy target-level constraints, but these ranges are not strictly enforced. In ELA, the blueprints also specify the number of passages in reading and listening claims (Claims 1 and 3, respectively).

Tables 9.1 and 9.2 present the percentages of tests aligned with the ELA test blueprint constraints for items in claims, targets and DOK, and passages in Claims 1 and 3. All ELA tests met all spring 2022 blueprint requirements.

Tables 9.3 and 9.4 provide the percentages of tests aligned with the test blueprint constraints for the mathematics CAT for claims, DOK, and target constraints. In mathematics, the tests met all blueprint requirements, except for in grade 7. In grade 7, the violations were in target sets of B and C and target sets of E and F in claim 1. Violations involved administering one to two items fewer or one item more than required.

Table 9.1: Percentage of ELA Delivered Tests Meeting Blueprint Requirements for Each Claim and Number of Passages Administered (Grades 3–5)

Claim	Content Category/Target	Required Items/Passages	%BP Match		
			Grade 3	Grade 4	Grade 5
1	Literary Text	4	100	100	100
	Target 2: Central Ideas	1–3	100	100	100
	Target 4: Reasoning and Evaluation				
	Targets 1, 3, 5, 6, & 7	1–3	100	100	100
	Long Literary Text Passage	1	100	100	100
	Short Literary Text Passage				
	Informational Text	4	100	100	100
	Target 9: Central Ideas	1–3	100	100	100
	Target 11: Reasoning and Evaluation				
	Targets 8, 10, 12, 13, & 14	1–3	100	100	100
	Long Informational Text Passage	1	100	100	100
	Short Informational Text Passage				
	DOK 2	≥ 4	100	100	100
	DOK 3 or 4	≥ 1	100	100	100
2	Writing	4	100	100	100
	Target 1, 3, or 6: Organization/Purpose	1	100	100	100
	Target 1, 3, or 6: Evidence/Elaboration	1	100	100	100
	Target 8: Language and Vocabulary Use	1	100	100	100
	Target 9: Edit/Clarify	1	100	100	100
	DOK 2 or higher	≥ 2	100	100	100
3	Listening	4	100	100	100
	Target 4: Listen/Interpret	4	100	100	100
	DOK 2 or higher	≥ 2	100	100	100
	Listening Passage	2	100	100	100
4	Research	4	100	100	100
	Target 2: Interpret and Integrate Information	1–2	100	100	100
	Target 3: Analyze Information/Sources	1–2	100	100	100
	Target 4: Use Evidence	1–2	100	100	100

Table 9.2: ELA Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Number of Passages Administered (Grades 6–8, HS)

Claim	Content Category/Target	Required Items/Passages in G6–8	Required Items/Passages in HS	%BP Match			
				Grade 6	Grade 7	Grade 8	HS
1	Literary Text	4	4	100	100	100	100
	Target 2: Central Ideas	1–3	1–3	100	100	100	100
	Target 4: Reasoning and Evidence			100	100	100	100
	Targets 1, 3, 5, 6, and 7	1–3	1–3	100	100	100	100
	Target 2 or 4 Short Text	0–1	0–1	100	100	100	100
	Long Literary Text Passage	1	1	100	100	100	100
	Informational Text	6	6	100	100	100	100
	Target 9: Central Ideas	2–4	2–4	100	100	100	100
	Target 11: Reasoning and Evidence			100	100	100	100
	Targets 8, 10, 12, 13, and 14	2–4	2–4	100	100	100	100
	Target 9 or 11 Short Text	0–1	0–1	100	100	100	100
	Long Informational Text Passage	1	1	100	100	100	100
	Short Informational Text Passage	1	1	100	100	100	100
	DOK 1	≤ 3	≤ 2	100	100	100	100
	DOK 3 or higher	≥ 1	≥ 2	100	100	100	100
2	Writing	4	4	100	100	100	100
	Target 1, 3, or 6: Organization/Purpose	1	1	100	100	100	100
	Target 1, 3, or 6: Evidence/Elaboration	1	1	100	100	100	100
	Target 8: Language and Vocabulary Use	1	1	100	100	100	100
	Target 9: Edit/Clarify	1	1	100	100	100	100
	DOK 2	≥ 2	≥ 2	100	100	100	100
3	Listening	4	4	100	100	100	100
	Target 4: Listen/Interpret	4	4	100	100	100	100
	DOK 2 or higher	≥ 2	≥ 2	100	100	100	100
	Listening Passage	2	2	100	100	100	100
4	Research	4	4	100	100	100	100
	Target 2: Analyze and Integrate Information	1–2	1–2	100	100	100	100
	Target 3: Evaluate Information/Sources	1–2	1–2	100	100	100	100
	Target 4: Use Evidence	1–2	1–2	100	100	100	100

Table 9.3: Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 3–5)

Claim	Content Domain	Grade 3		Grade 4		Grade 5	
		Required Items	%BP Match	Required Items	%BP Match	Required Items	%BP Match
1	Overall	10	100	10	100	10	100
	DOK 2 or higher	≥ 4	100	≥ 4	100	≥ 4	100
	<i>Priority Cluster</i>	7	100				
	Targets B, C, G, I	3	100				
	Targets D, F	3	100				
	Target A	1	100				
	<i>Supporting Cluster</i>	3	100				
	Targets E, J, K	2	100				
	Target H	1	100				
	<i>Priority Cluster</i>			7	100		
	Targets A, E, F			3	100		
	Target G			2	100		
	Target D			1	100		
	Target H			1	100		
	<i>Supporting Cluster</i>			3	100		
	Targets I, K			1	100		
	Targets B, C, J			1	100		
	Target L			1	100		
	<i>Priority Cluster</i>					7	100
	Targets E, I					3	100
	Target F					2	100
	Targets C, D					2	100
	<i>Supporting Cluster</i>					3	100
	Targets J, K					2	100
	Targets A, B, G, H					1	100
2&4	Overall	3	100	3	100	3	100
	DOK 3 or higher	≥ 1	100	≥ 1	100	≥ 1	100
	2. Target A	0–1	100	0–1	100	0–1	100
	2. Targets B, C, D	0–1	100	0–1	100	0–1	100
	4. Targets A, D	0–1	100	0–1	100	0–1	100
	4. Targets B, E	0–1	100	0–1	100	0–1	100
3	Overall	4	100	4	100	4	100
	DOK 3 or higher	≥ 1	100	≥ 1	100	≥ 1	100
	Targets A, D	1–2	100	1–2	100	1–2	100
	Targets B, E	1–2	100	1–2	100	1–2	100
	Targets C, F	1	100	1	100	1	100

Table 9.4: Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 6–8)

Claim	Content Domain	Grade 6		Grade 7		Grade 8	
		Required Items	%BP Match	Required Items	%BP Match	Required Items	%BP Match
1	Overall	9–10	100	10	100	10	100
	DOK 2 or higher	≥ 4	100	≥ 4	100	≥ 4	100
	<i>Priority Cluster</i>	6–7	100				
	Targets E, F	3	100				
	Target A	1–2	100				
	Targets B, G	0–2	100				
	Target D	1	100				
	<i>Supporting Cluster</i>	3	100				
	Targets C, H, I, J	3	100				
	<i>Priority Cluster</i>			7	99		
	Targets A, D			4	100		
	Targets B, C			3	99		
	<i>Supporting Cluster</i>			3	99		
	Targets E, F			2	99		
	Targets G, H, I			1	100		
	<i>Priority Cluster</i>					7	100
	Targets C, D					3	100
	Targets B, E, G					3	100
	Targets F, H					1	100
	<i>Supporting Cluster</i>					3	100
	Targets A, I, J					3	100
2&4	Overall	3	100	3	100	3	100
	DOK 3 or higher	≥ 1	100	≥ 1	100	≥ 1	100
	2. Target A	0–1	100	0–1	100	0–1	100
	2. Targets B, C, D	0–1	100	0–1	100	0–1	100
	4. Targets A, D	0–1	100	0–1	100	0–1	100
	4. Targets B, E	0–1	100	0–1	100	0–1	100
3-Calcul	Overall	3	100	4	100	4	100
	DOK 3 or higher	≥ 1	100	≥ 1	100	≥ 1	100
	Targets A, D	1–2	100	1–2	100	1–2	100
	Targets B, E	1–2	100	1–2	100	1–2	100
	Targets C, F, G	1	100	1	100	1	100
3-No Calcul	Overall	1	100				

Table 9.5: Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets (HS)

Claim	Content Domain	HS	
		Required Items	%BP Match
1	Overall	11	100
	DOK 2 or higher	≥ 4	100
	<i>Priority Cluster</i>	8	100
	Targets D, E	1–2	100
	Target F	0–1	100
	Targets G, H, I	2	100
	Target J	0–2	100
	Target K	0–2	100
	Targets L, M, N	2	100
	<i>Supporting Cluster</i>	3	100
	Target O	0–2	100
	Target P	0–2	100
	Targets A, B	0–1	100
	Target C	0–1	100
2&4	Overall	3	100
	DOK 3 or higher	≥ 1	100
	2. Target A	0–1	100
	2. Targets B, C, D	0–1	100
	4. Targets A, D	0–1	100
	4. Targets B, E	0–1	100
3-Calcul	Overall	3	100
	DOK 3 or higher	≥ 1	100
	Targets A, D	1–2	100
	Targets B, E	1–2	100
	Targets C, F, G	1	100
3-No Calcul	Overall	1	100

Table 9.6 summarizes the target coverage by claim, including the average and range of the number of unique targets administered in each CAT test. Although the target coverage varied somewhat across individual tests, all targets were covered at an aggregate level across all tests combined.

Table 9.6: Average and Range of the Number of Unique Targets Assessed within Each Claim, Across All Delivered Tests

Grade	Total Targets in BP				Average				Range (Minimum – Maximum)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
English Language Arts/Literacy												
3	14	5	1	3	7.5	4.0	1.0	3.0	6-8	4-4	1-1	3-3
4	14	5	1	3	7.6	4.0	1.0	3.0	6-8	4-4	1-1	3-3
5	14	5	1	3	7.4	4.0	1.0	3.0	5-8	4-4	1-1	3-3
6	14	5	1	3	9.1	4.0	1.0	3.0	6-10	4-4	1-1	3-3
7	14	5	1	3	9.2	4.0	1.0	3.0	7-10	4-4	1-1	3-3

Grade	Total Targets in BP				Average				Range (Minimum – Maximum)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
English Language Arts/Literacy												
8	14	5	1	3	9.0	4.0	1.0	3.0	7-10	4-4	1-1	3-3
HS	14	5	1	3	8.3	4.0	1.0	3.0	5-10	4-4	1-1	3-3
Mathematics												
3	11	4	6	6	9.0	1.0	3.6	2.0	9-9	1-1	3-4	2-2
4	12	4	6	6	9.0	1.0	3.6	2.0	8-9	1-1	3-4	2-2
5	11	4	6	6	8.0	1.0	3.4	2.0	8-8	1-1	3-4	2-2
6	10	4	7	6	8.6	1.0	3.0	2.0	7-9	1-1	2-4	1-2
7	9	4	7	6	6.3	1.0	3.4	2.0	5-7	1-1	3-4	1-2
8	10	4	7	6	9.0	1.0	3.4	2.0	6-10	1-1	2-4	1-2
HS	16	4	7	6	9.8	1.0	3.3	2.0	7-12	1-1	2-5	1-2

A CAT algorithm constructs a test form unique to each student, targeting the student's level of ability and meeting the test blueprints. Consequently, the test forms will not be statistically parallel (e.g., equal test difficulty); however, scores from the test should be comparable, and each test form should measure the same content, albeit with a different set of test items. The blueprint match and target coverage results demonstrate that all test forms conform to the same content target, thus providing evidence of content comparability. In other words, while each form is unique with respect to its items, all forms align with the same curricular expectations set forth in the test blueprints.

9.1.2 Evidence on Relations to Other Variables

Validity evidence based on relations to other variables can address a variety of questions. At its core, this type of validity addresses the relationship between test scores and variables of interest that are derived outside the testing system. One type of validity evidence based on relations to other variables is evidence for convergent and discriminant validity (Campbell & Fiske, 1959). Evidence for convergent validity is based on the degree to which test scores correlate with other measures of the same attribute—scores from two tests measuring the same attribute should be correlated. Conversely, evidence for discriminant validity is obtained when test scores are not correlated with measures of construct irrelevant attributes.

In a Washington study, the Office of Superintendent of Public Instruction (OSPI) examined the relationship between students who completed the 2015 Smarter Balanced assessments and their highest course grade in English language arts (ELA) courses and in mathematics courses. These studies showed a strong relationship between the Smarter Balanced achievement levels and course grades.

Methodology

High school juniors were administered the Smarter Balanced ELA and mathematics assessments in spring 2015. This was the first operational administration of the Smarter Balanced assessments.

OSPI maintains a database of student grade history. For the 2015 high school juniors with Smarter Balanced test scores, OSPI extracted the course grades for the following courses: ELA9, ELA10, Algebra I, Algebra II, and Geometry. Student assessment scores were matched to the course grades

using a State Student Identifier (SSID) number. Only students with a valid attempt on the Smarter Balanced assessments were selected for this study.

For each content area assessment, cross-tabulations of student achievement level to course grade were created. Within each assigned course grade (e.g., “B” in Algebra I), the percentage of students who attained each achievement level was computed.

Results

Table 9.7 shows the relationship between course grades and performance on the Smarter Balanced ELA assessment. For ELA10, 89% of the students who received an A in the course were classified in Level 3 or 4. As shown in Table 9.9, the percentage of students who were classified in Level 3 or 4 decreased as the course grade decreased. Only 34% of the students failing ELA10 (Grade F) received a Level 3 or 4 on the ELA assessment. For ELA9, a similar relationship was observed.

Table 9.8 shows the relationship between course grades and performance on the Smarter Balanced mathematics assessment. Again, the percentage of students who were classified in Level 3 or 4 decreased as the course grade decreased. This was true for all three courses. In the mathematics courses, the percentage of “A” students who attained Level 3 or 4 on the assessment were lower than in the ELA courses.

Conclusion

The results of the study lend support to the Smarter Balanced cut scores. Students who demonstrate higher achievement in their courses as demonstrated through teacher grading tend to attain higher achievement levels on the Smarter Balanced assessment.

Table 9.7: Percentage of Students in Each Smarter Balanced Achievement Level by Course Grade in 2015, ELA

Subject	Percentage in Each Category						
	Grade	Total	Level 1	Level 2	Level 3	Level 4	Level 3 or 4
ELA10	A	28,634	3%	7%	32%	57%	89%
ELA10	B	21,771	7%	17%	46%	29%	75%
ELA10	C	12,847	13%	28%	42%	15%	57%
ELA10	D	5,237	20%	33%	36%	11%	46%
ELA10	F	2,928	31%	35%	28%	6%	34%
ELA9	A	28,281	3%	8%	32%	57%	88%
ELA9	B	21,310	7%	18%	45%	28%	74%
ELA9	C	12,436	14%	28%	42%	15%	56%
ELA9	D	4,692	20%	32%	36%	11%	47%
ELA9	F	2,126	29%	34%	29%	7%	35%

Table 9.8: Percentage of Students in Each Smarter Balanced Achievement Level by Course Grade in 2015, Mathematics

Subject	Percentage in Each Category						
	Grade	Total	Level 1	Level 2	Level 3	Level 4	Level 3 or 4
Algebra	A	7,762	28%	29%	27%	16%	43%
Algebra	B	8,232	50%	32%	15%	3%	18%
Algebra	C	8,313	66%	24%	8%	2%	10%
Algebra	D	3,332	79%	17%	3%	0%	4%
Algebra	F	2,631	88%	10%	1%	1%	2%
Algebra 2	A	333	17%	21%	32%	29%	62%
Algebra 2	B	290	31%	38%	24%	7%	31%
Algebra 2	C	182	55%	34%	9%	2%	12%
Algebra 2	D	100	59%	30%	11%	0%	11%
Geometry	A	8,399	15%	23%	35%	27%	62%
Geometry	B	8,455	37%	36%	23%	5%	28%
Geometry	C	7,672	59%	30%	10%	1%	11%
Geometry	D	3,739	73%	22%	4%	0%	4%
Geometry	F	2,287	84%	13%	2%	0%	2%

9.1.3 Student Abilities vs. Test Difficulties

When student abilities are well matched to test difficulties, the standard error of measurement (SEM) can be reduced. Therefore, it is desired that the difficulty of a test matches the student's ability. To examine this aspect of the test, Figures 9.1 and 9.2 display the empirical distribution of the Washington student scale scores in the spring 2022 administration and the distribution of the administered summative item difficulty parameters. Overall, the student ability distribution is generally shifted to the left in all grades and subjects, a pattern more pronounced in the mathematics for upper grades, indicating that the pool includes more difficult items relative to the ability of students in the tested population. The pool includes difficult items to accurately measure high-performing students but needs additional easy items to better measure low-performing students. The Smarter Balanced Assessment Consortium identified this in their technical reports (<https://validity.smarterbalanced.org/reports-and-specifications/>), stating in section 4.13, "Although there is a wide distribution of item difficulty, pools tend to be difficult in relation to the population and to the cut score that is typically associated with proficiency." The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool and to augment the pool in proportion to the test blueprint constraints (e.g., content, Depth of Knowledge [DOK], item type, item difficulties) to better measure low-performing students.

Figure 9.1: Student Ability–Item Difficulty Distribution for ELA

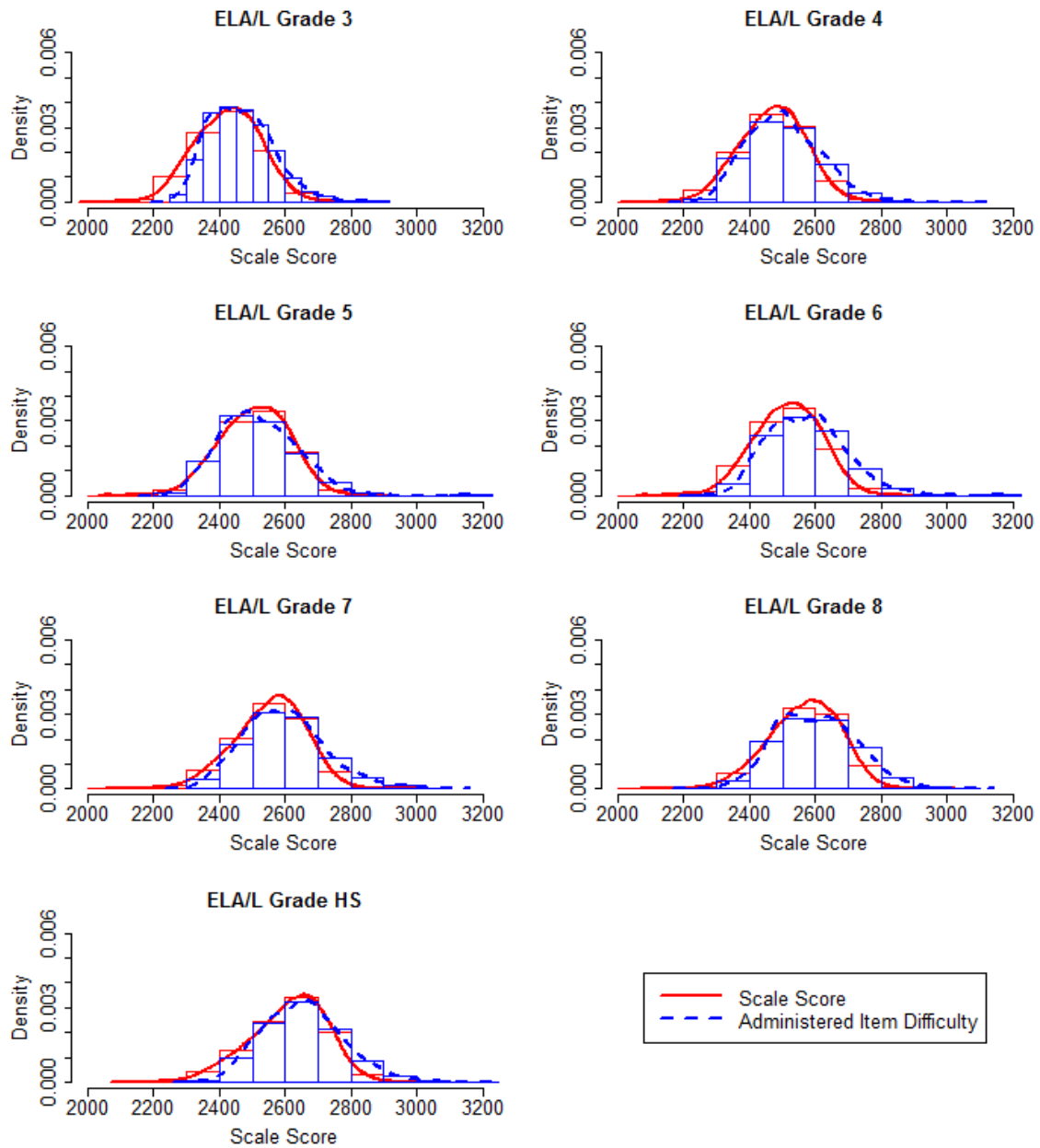
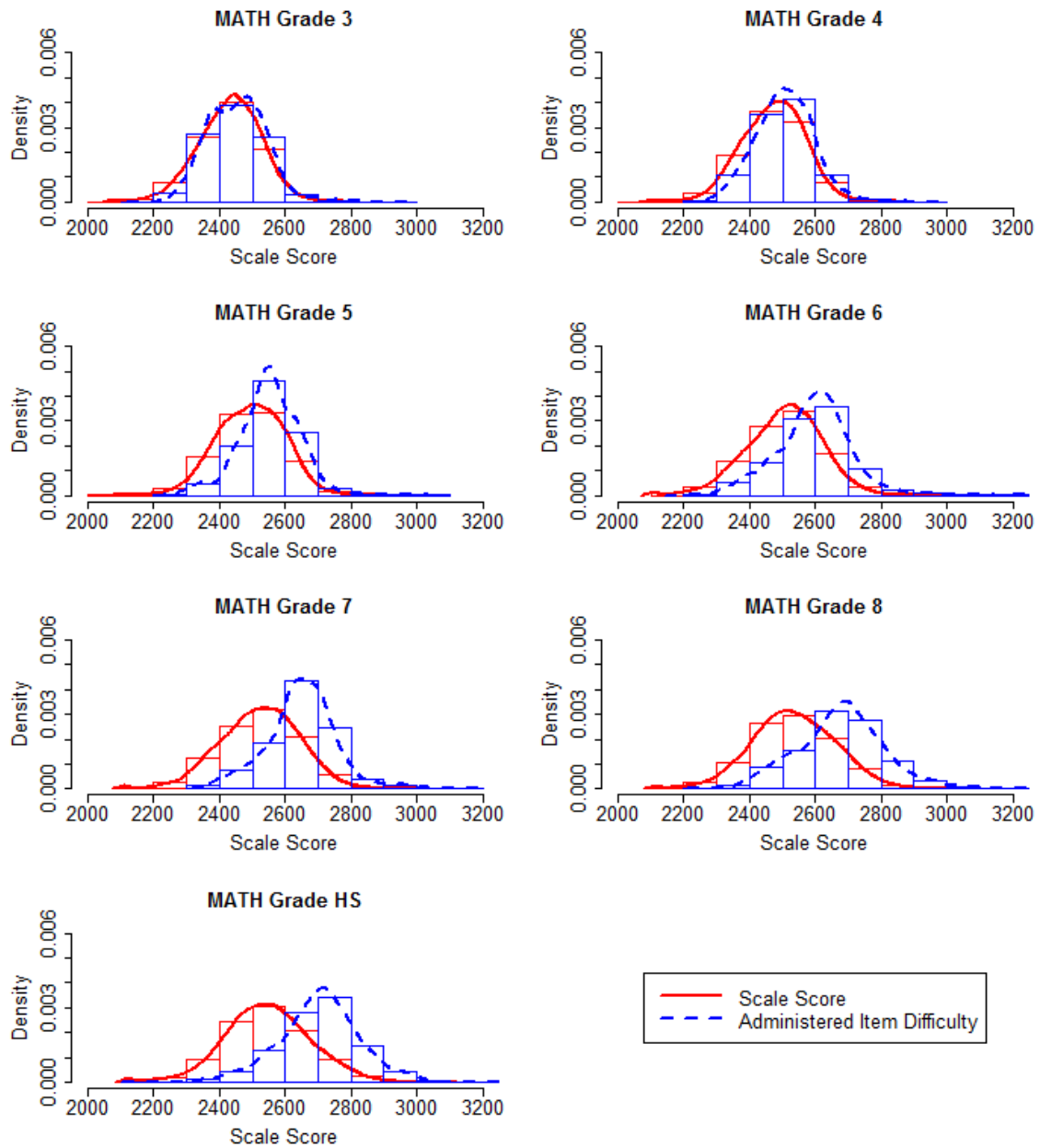


Figure 9.2: Student Ability–Item Difficulty Distribution for Mathematics



9.2 WASHINGTON COMPREHENSIVE ASSESSMENT OF SCIENCE (WCAS)

For the WCAS, validity evidence provided in this chapter includes:

- Internal Structure
- Relations to Other Variables

Some of the evidence on standardized test administration, scoring procedures, and attention to fairness for all test takers is provided in other chapters.

The analysis of internal structure of a test provides the information about the degree to which the relationships among items, the content standards, or test components conform to the construct on which the test score interpretations are based. For the WCAS, the test internal structures were examined using correlations among content standards and the principle component analysis methods. Evidence on relations to other variables is provided with relationships between course grades and performance on the WCAS.

9.2.1 Correlations Among Reporting Areas

To assess the strength of the interrelationships among the reporting areas, Pearson product-moment (PPM) correlation coefficients were computed:

$$r = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}}$$

where X_i is the score of reporting area X for examinee i ,

Y_i is the score of reporting area Y for examinee i ,

$\bar{X}$ is the mean sub-score of reporting area X ,

$\bar{Y}$ is the mean sub-score of reporting area Y , and

N is the total number of examinees.

For the WCAS, the correlations among reporting category scores, both observed (below diagonal) and corrected for attenuation (above diagonal) are presented in Table 9.9.

In these tables, reporting area level reliability is presented in the diagonal, the observed correlations are presented below the diagonal, and the disattenuated correlations are presented above the diagonal. Overall, the reliability coefficients, ranging from 0.66 to 0.79, show that the claims in the science assessments are moderate to high.

Table 9.9: Intercorrelations, WCAS 2022 Administration

Subject	Reporting Area	Earth and Space Science	Life Science	Physical Science
WCAS Grade 5 Form A	Practices and Crosscutting Concepts in Earth and Space Science	0.75	1.07*	1.02*

Subject	Reporting Area	Earth and Space Science	Life Science	Physical Science
	Practices and Crosscutting Concepts in Life Science	0.82	0.79	1.0*
	Practices and Crosscutting Concepts in Physical Science	0.72	0.73	0.67
WCAS Grade 8 Form A	Practices and Crosscutting Concepts in Earth and Space Science	0.78	1.0*	1.02*
	Practices and Crosscutting Concepts in Life Science	0.78	0.78	0.98
	Practices and Crosscutting Concepts in Physical Science	0.8	0.77	0.79
WCAS Grade 11 Form A	Practices and Crosscutting Concepts in Earth and Space Science	0.75	0.98	0.88
	Practices and Crosscutting Concepts in Life Science	0.72	0.71	0.96
	Practices and Crosscutting Concepts in Physical Science	0.62	0.66	0.66

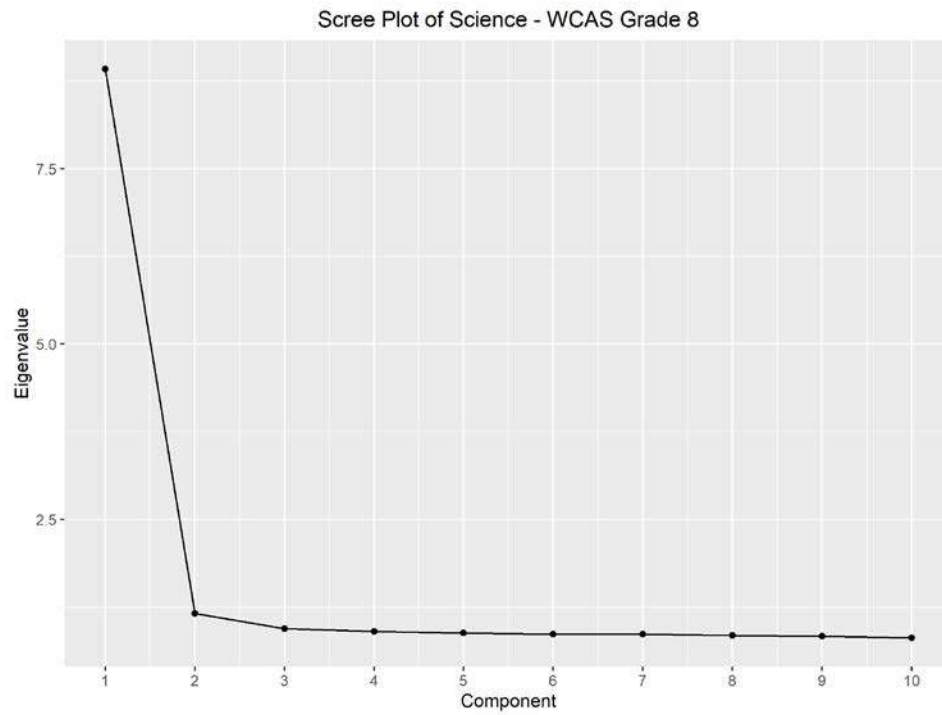
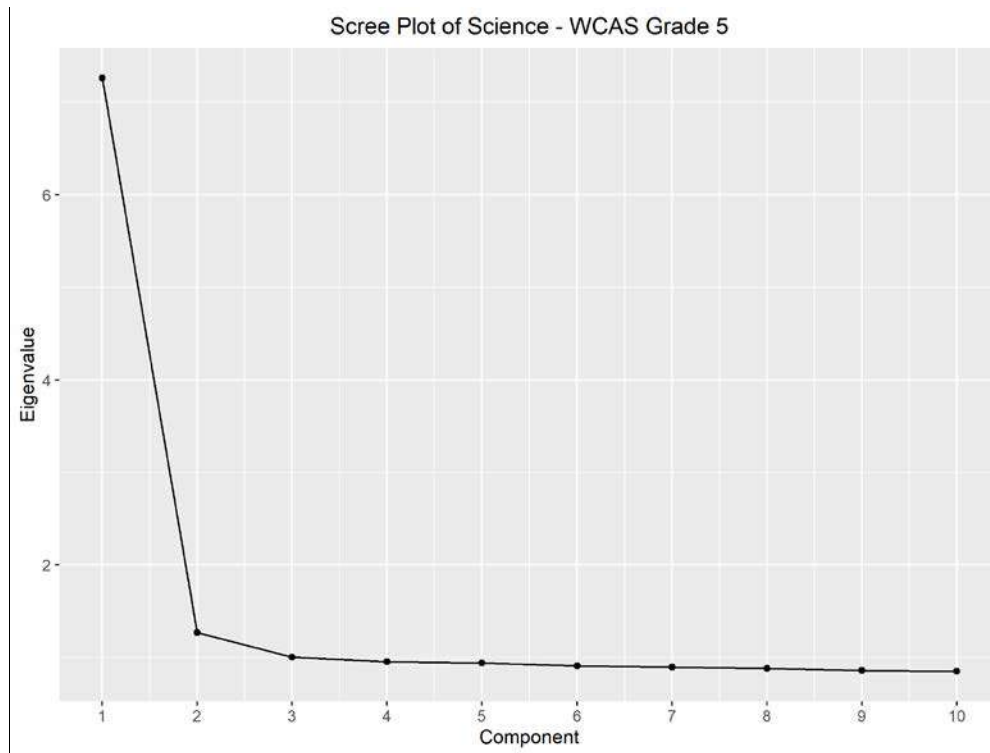
*Corrected correlations larger than 1.

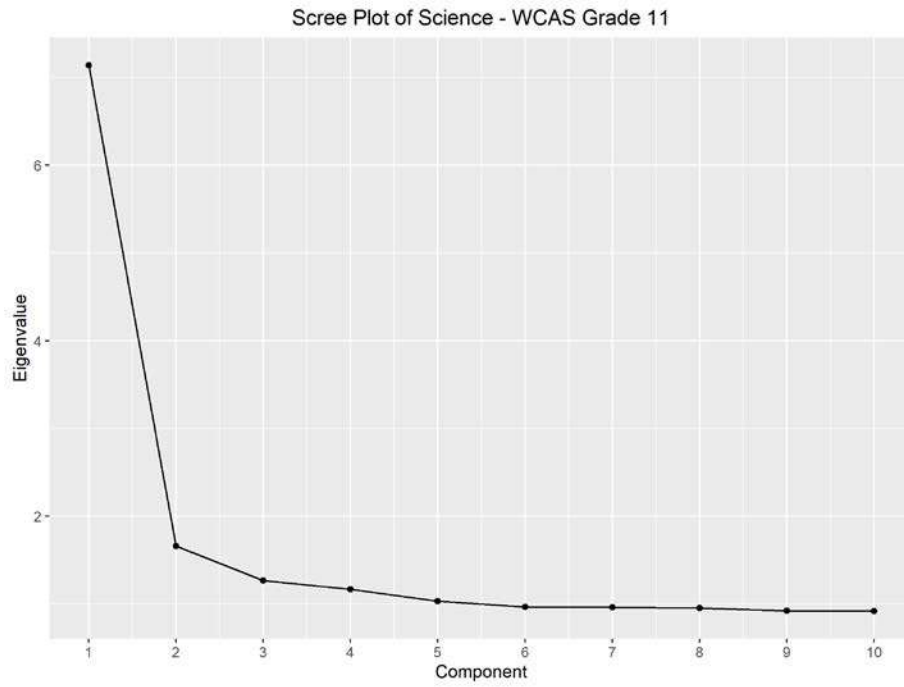
9.2.2 Dimensionality Analysis

One of the underlying assumptions of the PCM model is unidimensionality. The WCAS is designed to measure content standards within a specific content domain, which is referred to as one dimension. For example, the WCAS includes items designed to assess students' knowledge of four science content domains: Earth and Space Science (ESS), Life Science (LS), Physical Science (PS), and Engineering, Technology, and Applications of Science (ETS). These content domains represent different scientific knowledge and skills but are correlated to one test construct or dimension. For the WCAS, the dimensionality of each test was investigated. Principal components analysis (PCA) with an orthogonal rotation method (Jolliffe I.T., 2002; Cook, Kallen, & Amtmann, 2009) was used in the analysis.

The results are presented in scree plots, Figure 9.3. They show that the magnitude of the first eigenvalue is always much larger than the magnitude of the second factor in all tests, which indicates that the state-specific WCAS is unidimensional.

Figure 9.3: Scree Plots for WCAS





9.2.3 Evidence on Relations to Other Variables

Validity evidence based on relations to other variables can address a variety of questions. At its core, this type of validity addresses the relationship between test scores and variables of interest that are derived outside the testing system. One type of validity evidence based on relations to other variables is evidence for convergent and discriminant validity (Campbell & Fiske, 1959). Evidence for convergent validity is based on the degree to which test scores correlate with other measures of the same attribute—scores from two tests measuring the same attribute should be correlated. Conversely, evidence for discriminant validity is obtained when test scores are not correlated with measures of construct irrelevant attributes.

In a Washington study, the Office of Superintendent of Public Instruction (OSPI) compared the students grade 11 WCAS results, by achievement category, to teachers' course grades for the students.

Methodology

High school juniors (11th grade) were administered the WCAS in spring 2018. This was the first operational administration of the WCAS.

OSPI maintains a database of student grade history. OSPI used the student's last course grade in their 11th grade year for the study. For example, if a student had two semesters of grades reported for their 11th grade year, OSPI used the second semester grade. If a student had two "last" course grades, e.g., the student was taking two courses that were included in the list of Life science courses, OSPI used the higher course grade.

The course grades and assessment results were analyzed in five categories to represent the variety of science courses students took in grade 11:

- **Physical:** physical science, chemistry, physics, etc.
- **Life:** life science, biology, anatomy, genetics, etc.
- **EarthSpace:** earth science, astronomy, environmental science, etc.
- **Eng/Tech** (Engineering/Technology): technical science, engineering design, robotics, etc.
- **CTE** (Career and Technical Education): nursing, animal nutrition, agricultural biotechnology, etc.

Only students with a valid attempt on the grade 11 WCAS were selected for this study.

Results

Table 9.10 shows the relationship between course grades and performance on the WCAS. For all course categories, students who received an A in the course were classified in Level 3 or 4 at a rate of 52% to 76%, and only students who received a grade of B in Physical Science were classified in Level 3 or 4 within this rate range, at 55%. The percentage of students who were classified in Level 3 or 4 decreased in all course categories as the course grade decreased. Across all course categories, only 2% to 25% of the students with a course grade of F received a Level 3 or 4 on the WCAS.

**Table 9.10: Percentage of Students in Each WCAS Achievement Level
by Science Course Grade in 2018**

Subject	Course Grade	Total	Percentage in Each Category				
			Level 1	Level 2	Level 3	Level 4	Level 3 or 4
Physical	A	11897	9%	15%	40%	36%	76%
Physical	B	10497	20%	25%	39%	16%	55%
Physical	C	8429	31%	29%	32%	8%	40%
Physical	D	4855	38%	31%	26%	5%	31%
Physical	E	212	44%	33%	19%	3%	22%
Physical	F	2800	45%	29%	22%	3%	25%
Life	A	14399	11%	17%	41%	31%	72%
Life	B	12827	24%	28%	36%	12%	48%
Life	C	9416	36%	31%	27%	5%	32%
Life	D	5240	46%	30%	20%	3%	23%
Life	E	131	56%	25%	17%	1%	18%
Life	F	2439	55%	29%	14%	2%	16%
EarthSpace	A	3502	16%	19%	39%	25%	64%
EarthSpace	B	3073	30%	28%	32%	9%	41%
EarthSpace	C	2352	41%	31%	24%	4%	28%
EarthSpace	D	1331	47%	29%	20%	3%	23%
EarthSpace	E	28	50%	36%	7%	7%	14%
EarthSpace	F	816	55%	27%	15%	2%	17%
EngTech	A	4721	14%	18%	37%	31%	68%
EngTech	B	2608	28%	26%	33%	12%	45%
EngTech	C	1707	36%	28%	28%	7%	35%
EngTech	D	832	44%	29%	21%	5%	26%
EngTech	E	70	44%	36%	19%	1%	20%
EngTech	F	606	50%	28%	17%	4%	21%
CTE	A	8656	22%	26%	37%	15%	52%
CTE	B	5029	36%	30%	27%	6%	33%
CTE	C	3044	44%	31%	21%	4%	25%
CTE	D	1661	51%	28%	17%	2%	19%
CTE	E	37	54%	22%	19%	3%	22%
CTE	F	1011	84%	13%	2%	0%	2%

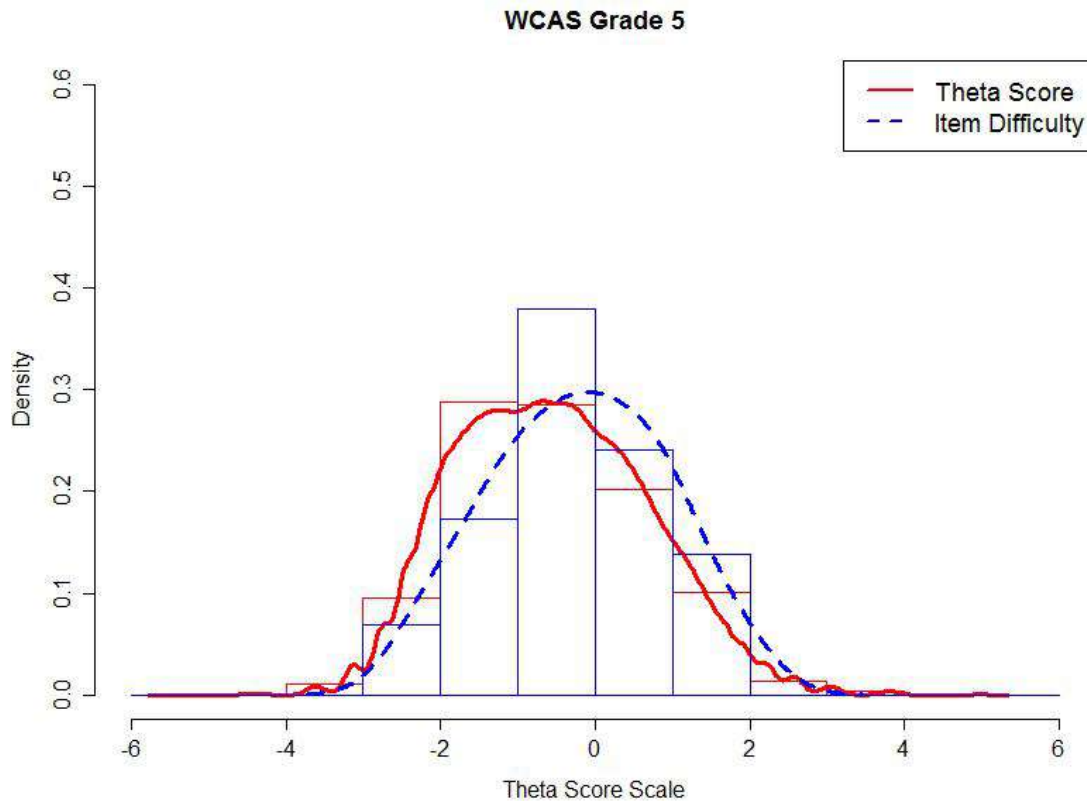
Conclusion

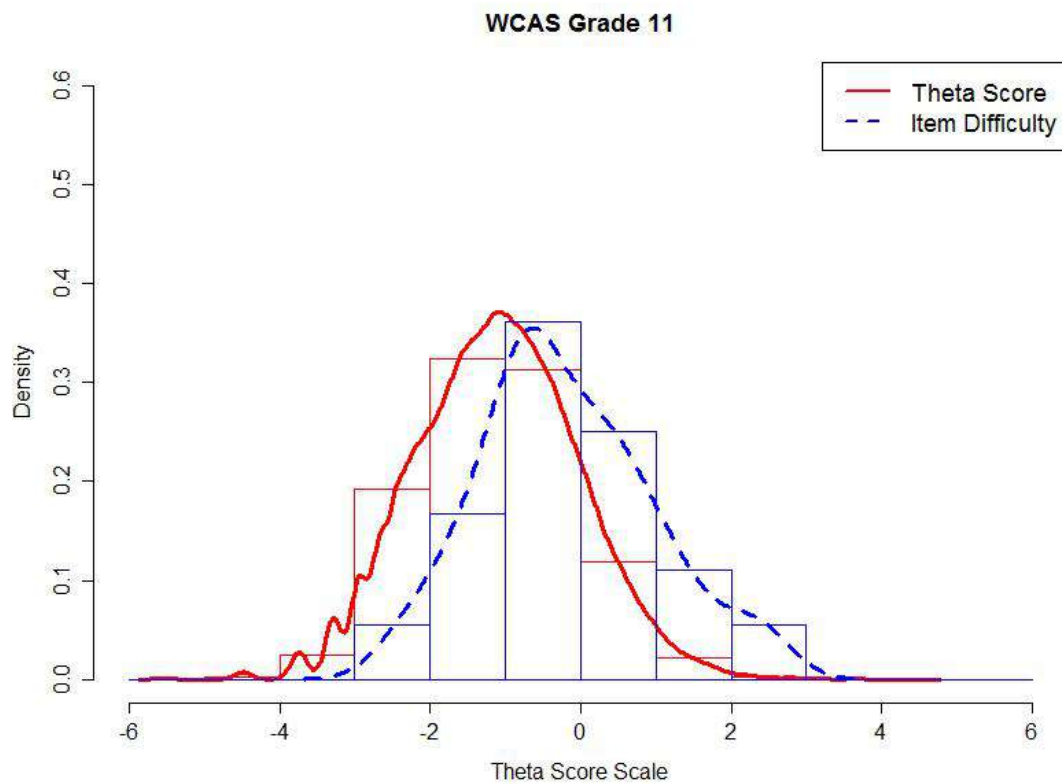
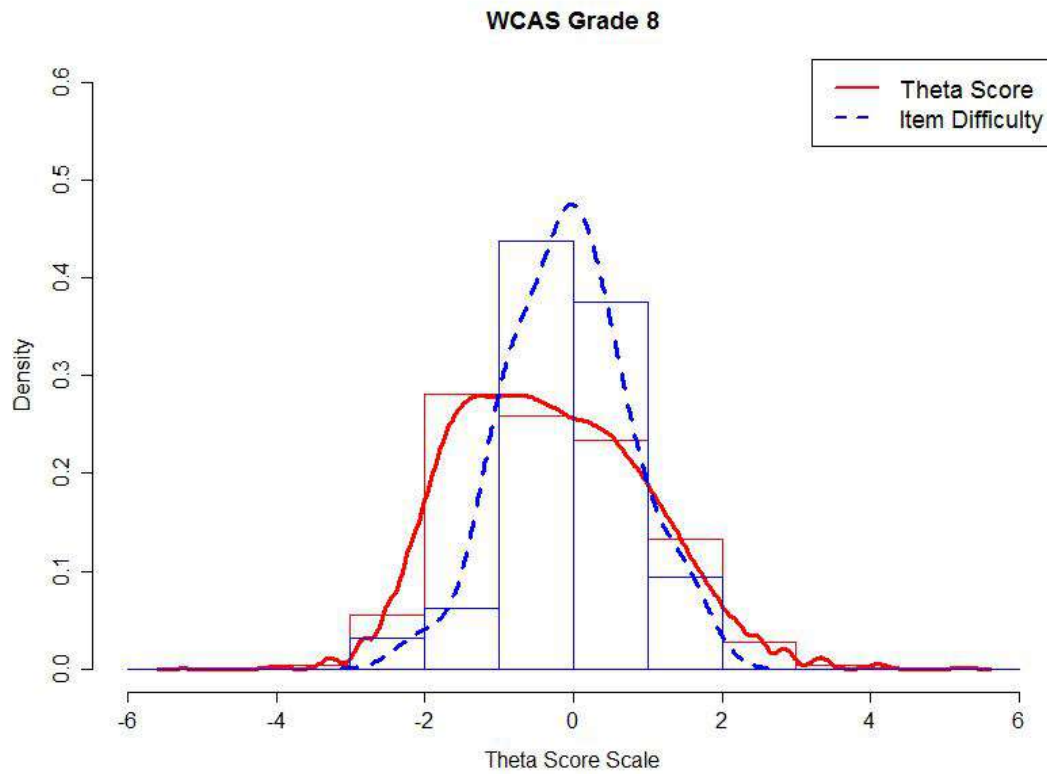
This study showed a strong positive relationship between the WCAS achievement levels and science course grades.

9.2.4 Student Abilities vs. Test Difficulties

When student abilities are well matched to test difficulties, the SEM can be reduced. It is desired that the difficulty of a test form matches student abilities. To examine this aspect of the test, Figure 9.4 shows the mapping of form difficulty with student abilities. The result shows that for science, the difficulty levels of the test forms are higher than the abilities of the students except for in high school, where it is about even.

Figure 9.4: WCAS Student Ability—Item Difficulty Distributions





SUMMARY

Validity is the central concern underlying test development, administration, scoring, reporting, and the uses and interpretations of test scores. The validity of an intended interpretation of test scores relies on all of the evidence accrued about the technical quality of a testing system, including test development and construction procedures, test score reliability, accurate scaling and equating, procedures for setting meaningful achievement standards, standardized test administration, scoring procedures, and attention to fairness for all test-takers. Such evidence has been presented in the earlier chapters of this report. This chapter presents additional information on the validity of the 2022 tests.

One measure of validity is the blueprint match rates for the delivered tests in adaptive tests. For Smarter Balanced ELA tests, all test met all blueprint requirement. In mathematics, the tests met all blueprint requirements except a few targets in grade 7.

The WCAS and Smarter Balanced tests both see some difference between student ability and test difficulty. The WCAS has reasonable correlations among the reporting areas, and the principal component analysis shows that the tests are unidimensional. Both the WCAS and Smarter Balanced test results correlate well to teacher-assigned grades.

10. REPORTING

Both Smarter Balanced Consortium assessments and the state-specific Washington Comprehensive Assessment of Science (WCAS) reported spring 2022 test results. The primary differences between the reporting of Smarter Balanced assessments and WCAS are in how claim and reporting area performances are measured and presented. The results were provided in two mediums: the Smarter Reporting System (SRS) and a Family Report sent to districts to distribute to families and students. This section includes information and examples for both Smarter Balanced and WCAS as both are presented in the same platform, SRS.

10.1 SMARTER REPORTING SYSTEM

The Smarter Reporting System (SRS) is a Smarter Balanced-hosted system that OSPI uses to present score information generated by Cambium and MI, contractors responsible for scoring, that pass results to SRS via secure transfer protocols. The spring 2022 results are the first summative results to appear in SRS; prior to spring 2022, summative results were reported in the Cambium-hosted Online Reporting System (ORS). Due to the disruptions caused by COVID-19, the decision was made to not import spring 2019 and earlier results from ORS into SRS. District staff have access to historical test results from 2019 and earlier through the Washington Assessment Management System.

The SRS generates a set of online score reports that describes student performance with the primary audience being educators. For Smarter Balanced English language arts (ELA) and mathematics assessments, student results are sent to SRS after all responses for a student's test are scored and results calculated. Results are usually available in SRS within 10 days after students complete the tests. The scores for the WCAS appear in SRS in mid-August after all post-equating work has concluded. In addition to each individual student's score report, SRS produces aggregate (i.e., student group) score reports. These student groups can be created flexibly by school and district staff to meet local needs.

Furthermore, Custom Aggregate reports are available to school- and district-level staff (e.g., principals, instructional coaches, district assessment coordinators) that contain the summary results for the selected student group, as well as aggregate unit(s) above the user's role. For example, a school-level user can generate a Custom Aggregate report that shows results for the school, the district, and the state. These Custom Aggregate reports can be created flexibly within the SRS user interface. Table 10.1 lists the permissions, including the types of online reports, users can generate and access in SRS depending on their role in TIDE.

SRS Sandbox

Smarter Balanced hosts a non-secure, open-access SRS Sandbox for Washington at <https://wasandbox.smarterreporting.org>. This Sandbox contains mock (i.e., not actual student) data and provides users the opportunity to explore the tools, features, and reports available within the live SRS. The Sandbox allows staff to explore SRS as different levels of users to see what different tools, features, and reports are available to those users. The Sandbox includes examples of Smarter Balanced math and ELA reports and WCAS reports. The Sandbox also includes a User Guide that describes all the navigation, tools, features, and reports available in the live SRS.

10.1.1 Types of Score Reports In SRS

The SRS is designed to provide educators with a variety of information regarding student performance on state tests. It endeavors to present test results in easy-to-read and -understand ways. Simple language is used so that users can quickly understand assessment results and make valid inferences about student achievement. In addition, SRS is designed to present student performance in a uniform format. For example, similar colors are used for groups of similar elements, such as achievement levels, throughout the design.

Once authorized users log in to SRS, they see the SRS home page. This home page presents different tools and access to reports based on the user's role in TIDE (see Table 10.1). SC and above-level users can use Administrator Tools such as Custom Aggregate Reports, search for results by student ID or school, and, if assigned, by Assigned Groups. TA level users can see results for Assigned Groups of students that an SC or above has assigned to that TA user as well as search for results by student ID.

Generally, SRS provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Detailed information about the online score reports and instructions on how to navigate the online score reporting system can be found in the *SRS User Guide*, located in the SRS Sandbox as well as within the live SRS.

Table 10.1: Permissions and Reports Available to TIDE Users by Role

Permission	Role in TIDE
Create custom student groups	TA or SC
Edit custom student groups	TA or SC
Delete custom student groups	TA or SC
View individual student results by assigned student group	TA or DC, DA, SC
View individual student results by district, school, and grade	STATE, DC, DA, SC
Search for students	TA or STATE, DC, DA, SC
View student test history	TA or STATE, DC, DA, SC
Export results as CSV	TA or STATE, DC, DA, SC
Print individual student reports	TA or STATE, DC, DA, SC
Print student group batch reports	TA or STATE, DC, DA, SC
Print school and grade batch reports	STATE, DC, DA, SC
Create/view/export custom aggregate reports	STATE, DC, DA, SC, IS
Review student results that have not been released by the state	STATE
Release student results to all users by the state	STATE
Edit instructional resource links	STATE, DC, DA
Create assigned student groups	DC, DA, SC
Edit assigned student groups	DC, DA, SC

Permission	Role in TIDE
Delete assigned student groups	DC, DA, SC

10.1.2 Group Reporting

The aggregate score reports at a selected aggregate level are provided for overall students and by student groups. Users can see student assessment results by any student group. Table 10.2 presents the types of student groups categories available in SRS.

Table 10.2: Types of Student Groups

Group	Categories
Gender	Male
	Female
	Non-binary
Race/Ethnicity	American Indian/Alaska Native
	Asian
	Black or African American
	Demographic Race Two or More Races
	Hispanic or Latino
	Native Hawaiian or Pacific Islander
	White
IEP	Yes
	No
English Learner	Yes
	No
504 Plan	Yes
	No
	Not Stated
Migrant Status	Yes
	No
	Not Stated
Assessment Grade	Grade 3
	Grade 4
	Grade 5
	Grade 6
	Grade 7
	Grade 8
	Grade 9
	Grade10
	Grade 11

10.1.3 Paper Report

The SRS provides the functionality for users to print out reports described above in paper form, including Individual Student Reports or ISRs.

Furthermore, the Office of Superintendent of Public Instruction (OSPI), through Cambium, provides districts with electronic Family Reports for each student to the districts for distribution to families. Details about Family Reports are provided in section 10.3.

10.2 SRS REPORT PAGES

Home Page

The first page users see when they log on to SRS is the Home Page. Depending on the user's role in TIDE, different features, tools, and reports are available on this page.

The home page provides access to Administrator Tools, search by student or school, and Assigned Groups for the user, if any. Exhibits 10.1 and 10.2 present sample home pages at the TA-user level and the district-user level.

Exhibit 10.1: Home Page—TA-user Level



Washington Sandbox
 Washington Reporting System Sandbox

[User Guide](#)
[Interpretive Guide](#)
[My Reports](#)
[Teacher](#)

Access Assessment Results

Search by Student

Assigned Groups 3

My Groups 0

Search by Group

Group Name	School	Subjects
Sample Elementary School Grade 3	Sample Elementary School	All
Sample Elementary School Grade 4	Sample Elementary School	All
Sample Elementary School Grade 5	Sample Elementary School	All

Welcome to the SRS Sandbox!

The [User Guide](#) describes features of the Smarter Reporting System and instructions for using each feature.

The [Interpretive Guide](#) is designed to help educators, parents, and other stakeholders interpret reports for the Smarter Reporting System.

Exhibit 10.2: Home Page—District Level

Smarter REPORTING Washington Sandbox
Washington Reporting System Sandbox

User Guide Interpretive Guide My Reports District Admin

Administrator Tools

Custom Aggregate Report
 Create a customized report of student performance.

District / School Exports
 Export data for analysis in another application.

Student Groups
 Create and manage student groups for teachers.

Instructional Resources
 Upload links to instructional resources in the system.

Access Assessment Results

Search by Student

Enter the Statewide Student Identifier (SSID) Search

Search by School Grade

Select Select Search

Assigned Groups 0

Contact your administrator for access to group-level test results.

Welcome to the SRS Sandbox!

The [User Guide](#) describes features of the Smarter Reporting System and instructions for using each feature.

[User Guide](#)

The [Interpretive Guide](#) is designed to help educators, parents, and other stakeholders interpret reports for the Smarter Reporting System.

[Interpretive Guide](#)

Note

A Sandbox is an environment that allows users to explore the features and functionality of the Smarter Reporting System using generated data and sample assessments.

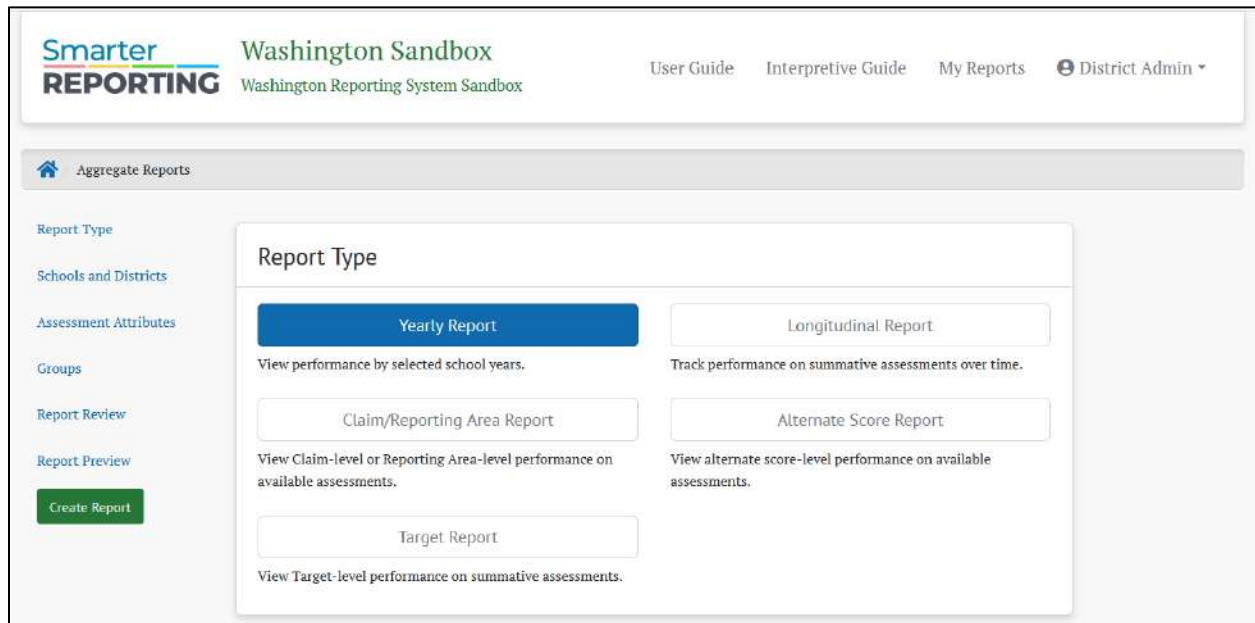
Smarter Balanced Summative assessment claim level scores were not reported in 2020-21 or 2021-22.

10.2.1 Custom Aggregate Reports

Users with SC or above roles in TIDE can access Custom Aggregate Reports. There are multiple report types available, as show in Exhibit 10.3 On each aggregate report, the report presents the

results for the user’s aggregate unit as well as the summary results for the state and aggregate unit above the selected aggregate, if any.

Exhibit 10.3: Custom Aggregate Reports



Yearly Reports

Yearly Reports summarize IAB, ICA, or summative assessment performance for student populations from one or more grade levels for one or more years. Yearly Reports also allow users to select different student groups to show in the report detail.

Exhibit 10.4 presents an example of a Yearly Report for ELA at the district level when a user includes gender in the report detail. This example and all example reports shown in this Technical Report are taken from the Sandbox so contain only mock data and no actual student data.

Exhibit 10.4: Subject Detail Page for ELA by Gender—District Level

Smarter REPORTING Washington Sandbox
Washington Reporting System Sandbox

User Guide Interpretive Guide My Reports District Admin

Aggregate Reports Custom Aggregate Report

Custom Aggregate Report

Create New Query Row count: 1 Show

Summative ELA Export

Column Order: Organization Assessment Grade Academic Year Group

2 empty rows Display value as: Percent Number Achievement levels: All Grouped

Organization	Assessment Grade	Academic Year	Group	Students Tested	Achievement Comparison	Average Scale Score ± Error Band	Level 1 Level	Level 2 Level	Level 3 Level	Level 4 Level
State WASHINGTON	5	2021-22	Overall	104		2528 ± 24	36%	6%	11%	45%
			Gender: Male	63		2524 ± 30	36%	6%	12%	44%
			Gender: Female	41		2534 ± 39	36%	7%	9%	46%
			Gender: Non-binary	0		-	-	-	-	-
District Sample District	5	2021-22	Overall	104		2528 ± 24	36%	6%	11%	45%
			Gender: Male	63		2524 ± 30	36%	6%	12%	44%
			Gender: Female	41		2534 ± 39	36%	7%	9%	46%
			Gender: Non-binary	0		-	-	-	-	-

Claim/Reporting Area Report

The Claim/Reporting Area Report provides the aggregate summaries on student performance in each claim (for Smarter Balanced math and ELA tests) and reporting area (for WCAS) of the summative test for a particular grade and subject.

In Spring 2022, due to the use of the adjusted blueprint, no Claim results were calculated for math or ELA Tests.

Exhibit 10.5 shows a sample district-level Reporting Area report for grade 8 WCAS.

Exhibit 10.5: Grade 8 WCAS Reporting Area Report

Smarter REPORTING Washington Sandbox Washington Reporting System Sandbox

User Guide Interpretive Guide My Reports District Admin

Aggregate Reports Custom Aggregate Report

Custom Aggregate Report

Create New Query Row count: 1 Show

Summative Science Export

Column Order Organization Assessment Grade Academic Year Claims/Reporting Area Group

0 empty rows Display value as Percent Number

Organization	Assessment Grade	Academic Year	Claims	Group	Students Tested	Achievement Comparison	Below Standard	At Standard	Above Standard
WASHINGTON	8	2021-22	Practices & Crosscutting Concepts in Physical Science	Overall	258		39%	29%	30%
			Practices & Crosscutting Concepts in Life Science	Overall	258		39%	30%	29%
			Practices & Crosscutting Concepts in Earth & Space Science	Overall	258		40%	28%	32%
Sample District	8	2021-22	Practices & Crosscutting Concepts in Physical Science	Overall	258		39%	29%	30%
			Practices & Crosscutting Concepts in Life Science	Overall	258		39%	30%	29%
			Practices & Crosscutting Concepts in Earth & Space Science	Overall	258		40%	28%	32%

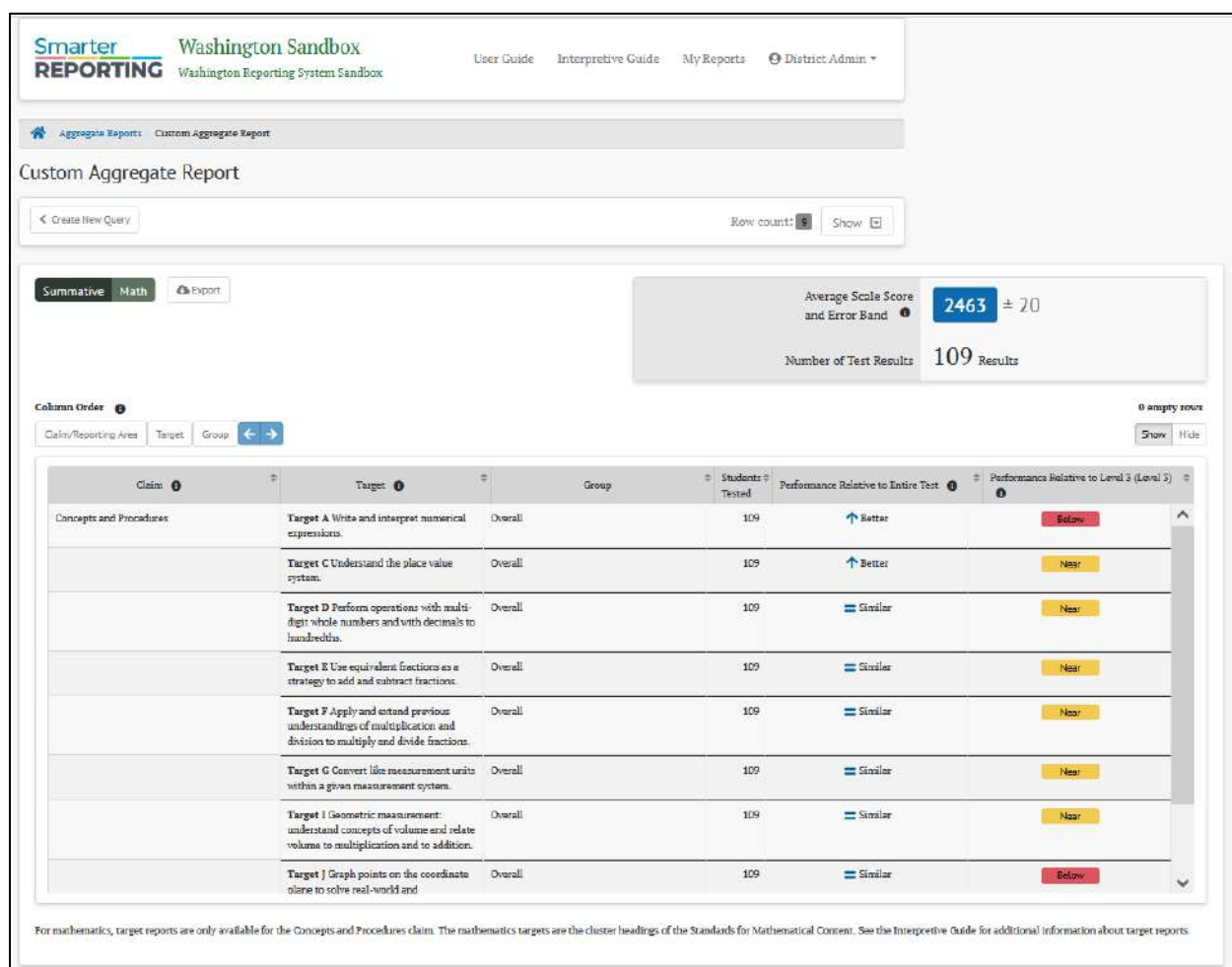
Target Report

The target report provides a yearly report of target performance for the Smarter Balanced math and ELA summative assessment by a student population (e.g., school or district) in a single year. Target reports are available for all ELA claims and the mathematics Concepts and Procedures claim (Claim 1) only. In Spring 2022, Target reports were generated for student groups. Target reports are not calculated for the WCAS.

The target report provides indicators for each target that are computed in two ways: performance relative to entire test and performance relative to Level 3. The reports also provide an average scale score and error band for the students in the group.

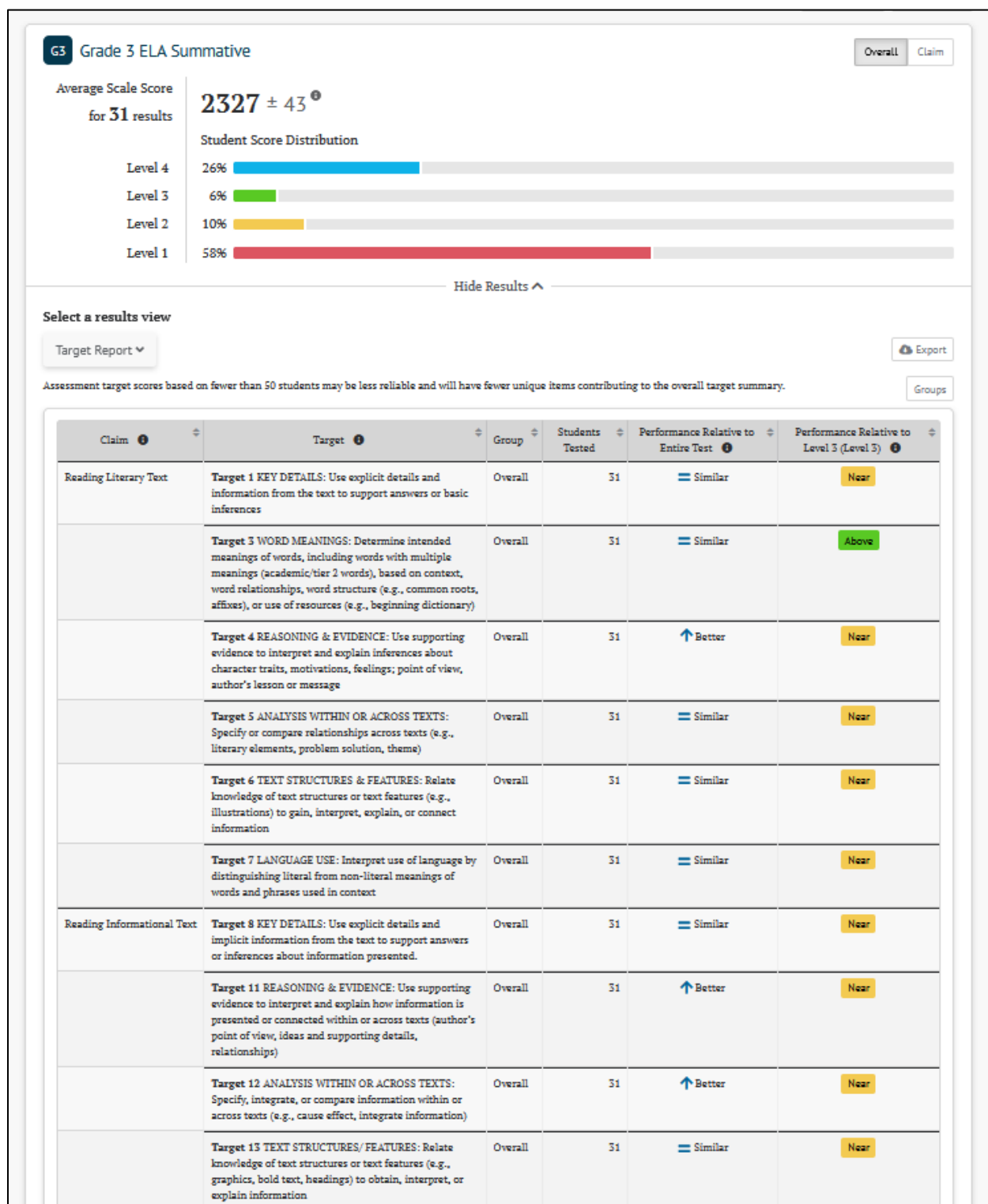
Exhibit 10.6 shows a sample Target Report for grade 5 math.

Exhibit 10.6: Target Report for Math Grade 5—Custom Aggregate Level



TA-level users can also generate Target reports for their assigned student groups. Exhibit 10.7 shows part of a grade 3 ELA Target report available to TA users.

Exhibit 10.7: Target Report for Grade 3 ELA—Teacher Level



Longitudinal Report

The longitudinal report tracks summative assessment performance for a single student population as they progress through different grades. In addition to presenting tabular data, it includes a line graph showing how the performance of the population changed from grade to grade.

Spring 2022 results are the first to appear in SRS, so there is no longitudinal data to present this year. Examples of longitudinal data will be included in subsequent years' technical reports.

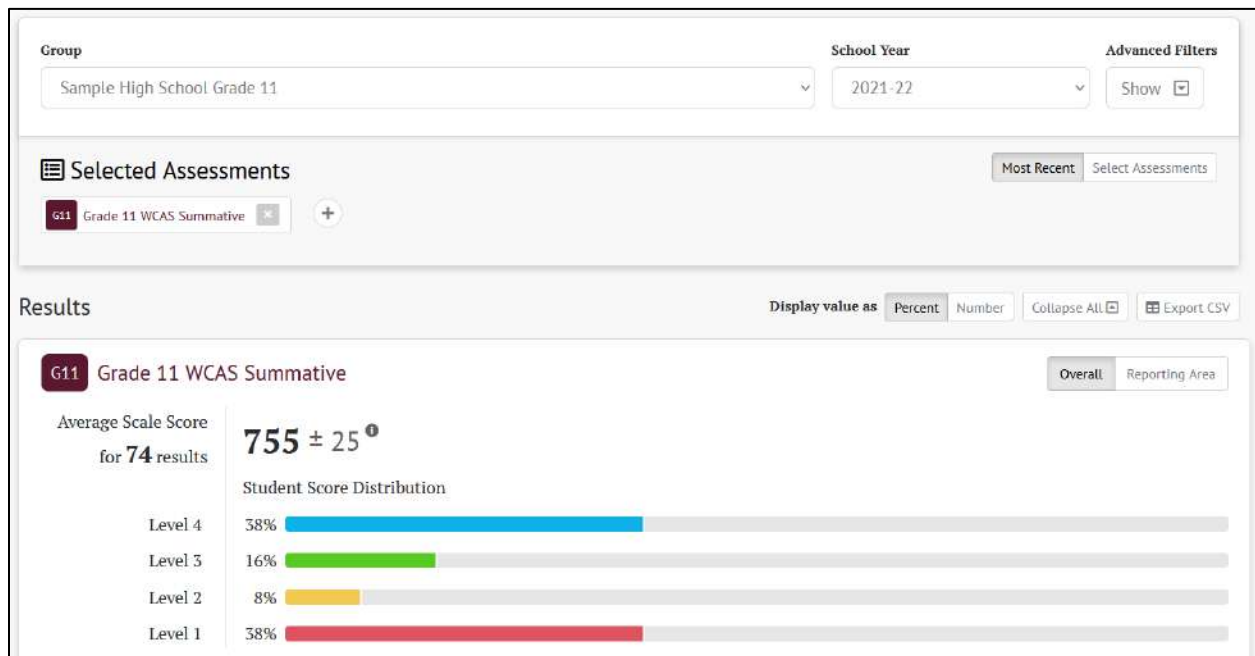
10.2.2 Assigned Student Groups Reports

For TA-level users, student test results are grouped by Assigned Groups, and a TA must be assigned a group by an SC or higher user prior to the TA being able to see student results in SRS.

From the Home page, users select one group from the Assigned Groups to view student results for that student group. Users can also create customized groups from the students in their Assigned Groups using the “My Groups” tool on the Home page.

Exhibit 10.8 shows a sample student group for Grade 11 WCAS results. The report includes the number of student tests, the average scale score and standard error of measure, and a distribution of results by level.

Exhibit 10.8: Student Group Results for Grade 11 WCAS



Further down, the report shows results for each student in the group, as shown in Exhibit 10.9. The results show the student's name, achievement level, and scale score.

Exhibit 10.9: Student Results for Grade 11 WCAS

Select a results view

Results By Student ▾

Student	Date	Session	Enrolled Grade	School	Status	Achievement Level	Scale Score / Error Band
Adams, Daphne	May 6, 2022	DAV-21ff	G11	Sample High School		Level 1	607
Allums, Glen	May 6, 2022	JON-2318	G11	Sample High School		Level 4	1090
Alonzo, Virginia	May 6, 2022	DAV-21ff	G11	Sample High School		Level 2	659
Anson, Cynthia	May 6, 2022	BRO-adb2	G11	Sample High School		Level 1	459
Archie, Shon	May 6, 2022	JON-c352	G11	Sample High School		Level 4	1043
Arnold, William	May 6, 2022	DAV-21ff	G11	Sample High School		Level 4	1073
Arteaga, Jonas	May 6, 2022	DAV-21ff	G11	Sample High School		Level 1	462
Bailey, Pamela	May 6, 2022	DAV-6c80	G11	Sample High School		Level 5	745
Bailey, Dorothy	May 6, 2022	DAV-6c80	G11	Sample High School		Level 4	1079
Baker, Terrance	May 6, 2022	DAV-6c80	G11	Sample High School		Level 1	614
Barnhart, Coy	May 6, 2022	JOH-57ac	G11	Sample High School		Level 5	756
Bartlett, Hillary	May 6, 2022	JON-fb83	G11	Sample High School		Level 1	417
Bast, William	May 6, 2022	WIL-cf2c	G11	Sample High School		Level 1	581

From this list of students, users can generate either a test history report for the student showing all tests the student has taken (See Exhibit 10.10) or an Individual Student Report, or ISR (see Exhibit 10.11).

Exhibit 10.10: Student Test History Report

Adams, Daphne

7000017966

Export CSV

Printable Reports

School Year

2021-22

Subject

All

Assessment Type

All

Advanced Filters

Show

Student Test History Report

Math

Collapse

G11

Grade 11 Math - Interim Comprehensive Assessment (ICA)

Jan 19, 2022

Level 4

1 result

G11

High School Math Performance Task (IAB)

Dec 20, 2021

Above Standard

1 result

G11

High School Math - Algebra and Functions I and II (IAB)

Nov 18, 2021

Above Standard

1 result

ELA

Collapse

G11

Grade 11 ELA - Interim Assessment (ICA)

Jan 19, 2022

Level 4

1 result

G11

High School ELA - Read Informational Texts (IAB)

Dec 27, 2021

Above Standard

1 result

G11

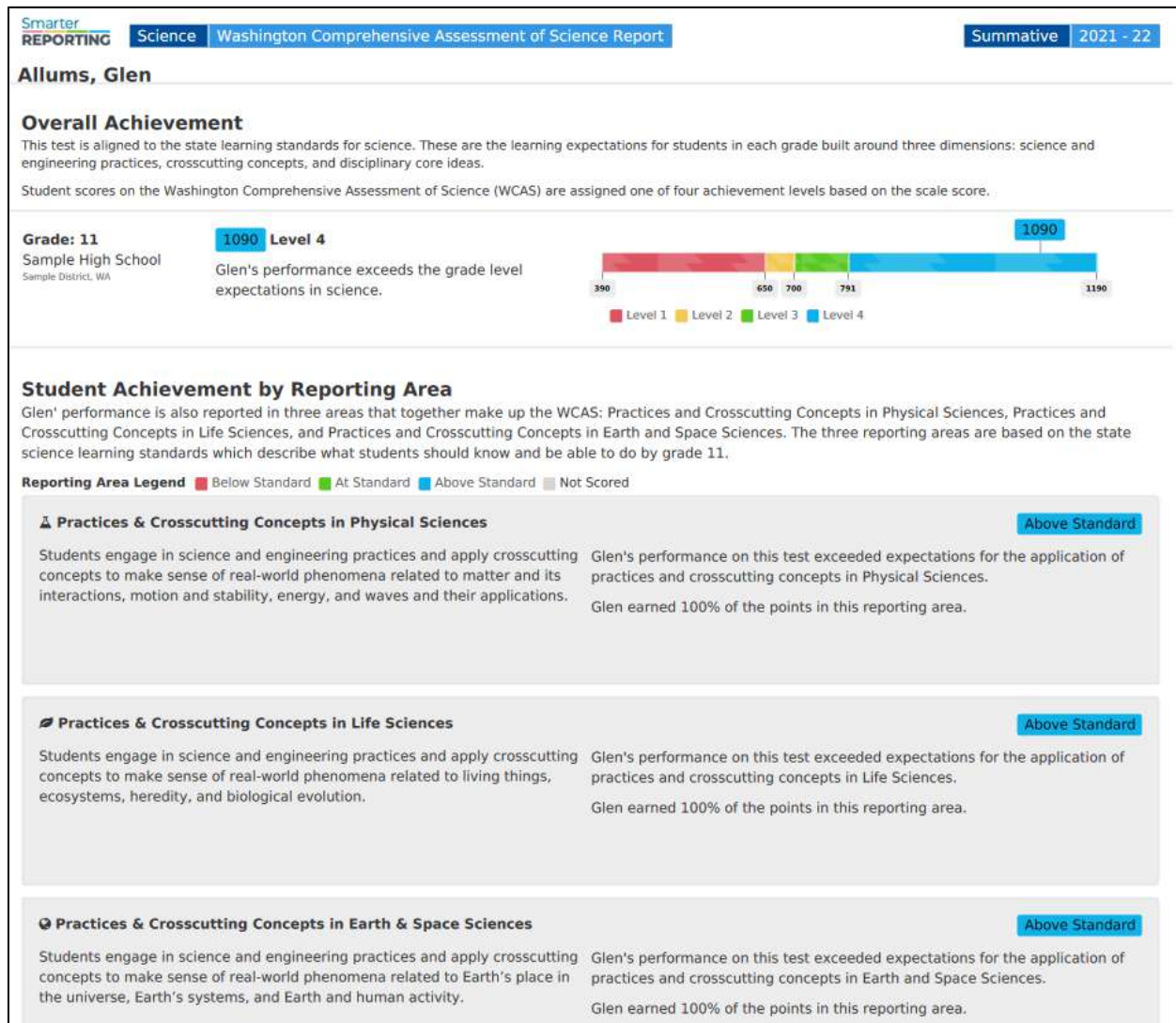
High School ELA Performance Task (IAB)

Dec 3, 2021

Above Standard

1 result

Exhibit 10.11: Student ISR for Grade 11 WCAS



ISRs

The Individual Student Report, or ISR, shows individual student performance on a selected test. The ISR shows (1) scale score (2) standard error of measurement (SEM) for math and ELA tests, (3) achievement level for the overall test, (4) achievement category in each Claim or Reporting Area (In spring 2022, no Claim results were calculated for math or ELA tests), and (5) writing performance descriptors in each dimension (ELA only).

10.3 ELECTRONIC FAMILY REPORT

The testing window closes in June, and Family Reports are generated in October for each participating student and provided to the student's school district. The Family Reports are generated by Cambium based on the approved specifications and the quality assurance procedures outlined in Section 11.5.2. CAI delivers PDF files of the family reports batched at the school level to OSPI electronically through a secure file transfer protocol site. OSPI is responsible for posting

these batched reports to districts in WAMS and splitting the PDF file to create student-level files that districts use to populate their parent/family portals.

In previous years, Cambium also printed and shipped 2 paper copies of students' Family Reports to districts. OSPI is working to transition the delivery of these reports to districts as only the electronic files described above that districts staff can load into local, secure family/parent portals or print and deliver paper copies to families if desired. The fall 2021 Family Reports were the first to be delivered to districts as electronic files, and that practice continued with the spring 2022 Family Reports. OSPI provided technical support for districts who wanted to load the electronic Family Reports into local, secure family/parent portals. OSPI also supported districts that wanted paper copies by printing a single gray-scale copy of students' Family Reports and shipping them to districts.

Examples of Family Reports are shown in Exhibit 10.12 for ELA and Exhibit 10.13 for the WCAS, which is 2 pages long. OSPI posts sample Family Reports for each administration online at <https://www.k12.wa.us/student-success/testing/state-testing/scores-and-reports/sample-score-reports>.

Exhibit 10.12: Smarter Balanced ELA Sample Electronic Family Score Report

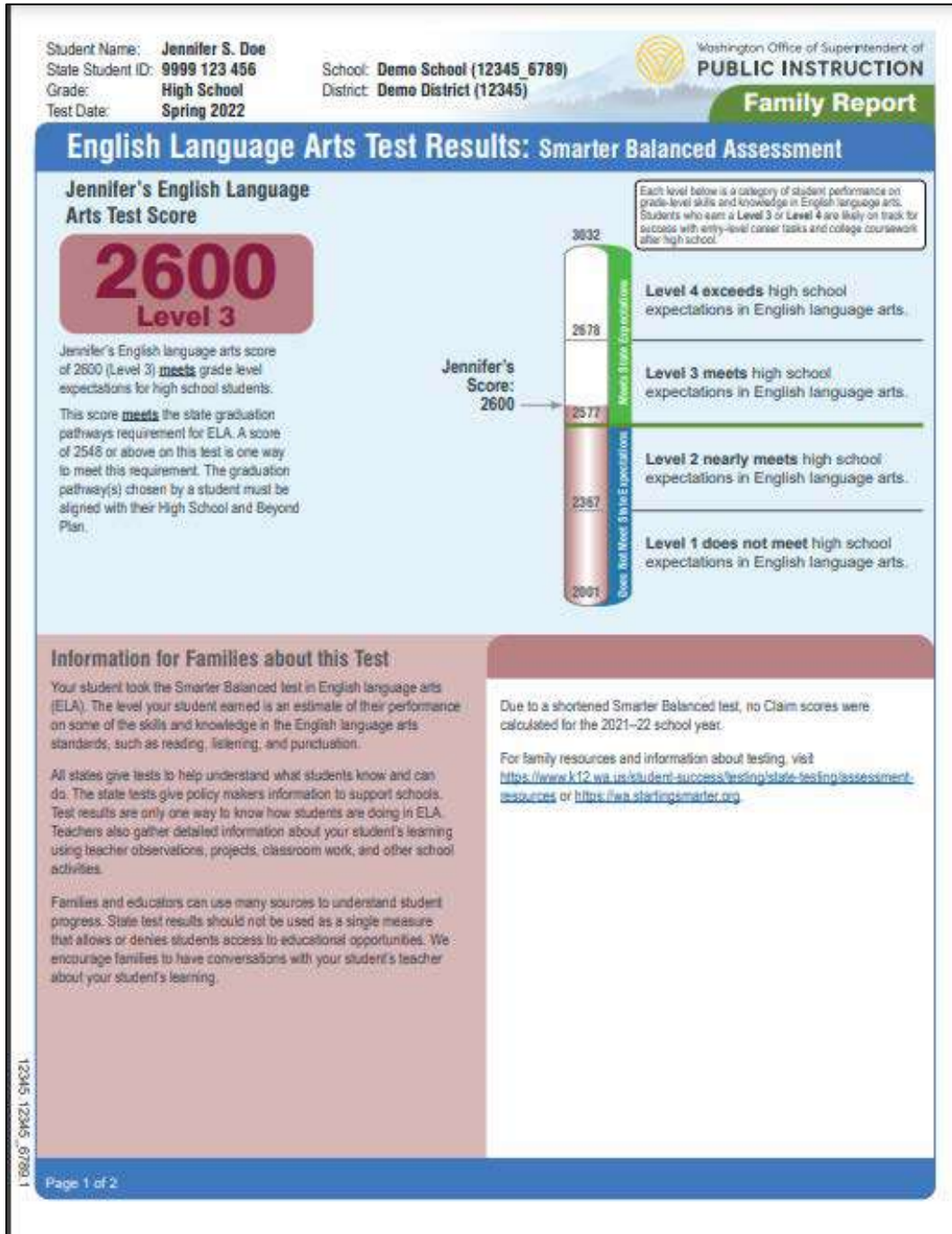
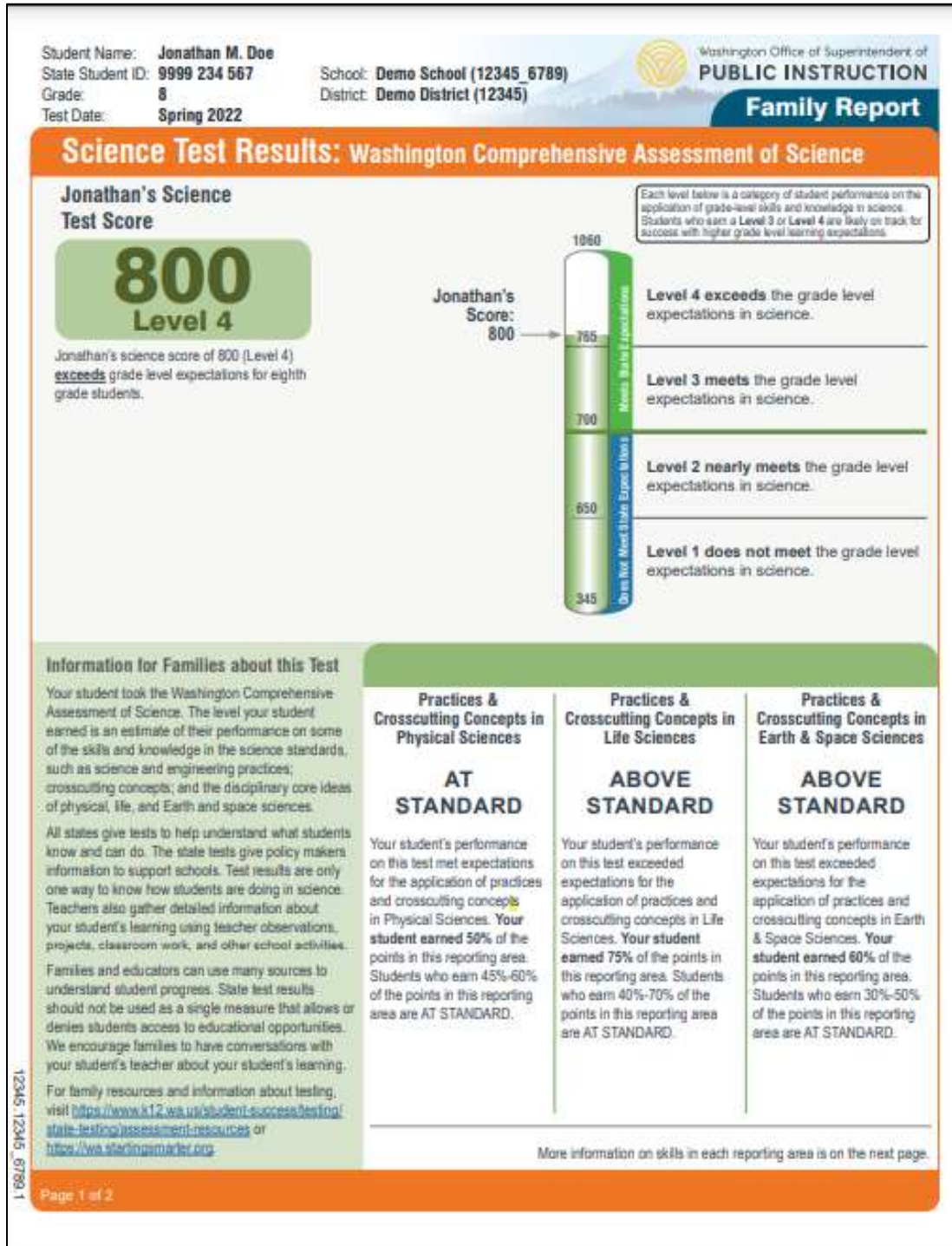


Exhibit 10.13: WCAS Sample Electronic Family Report



Student Name: Jonathan M. Doe State Student ID: 9999 234 567 Grade: 8 Test Date: Spring 2022	School: Demo School (12345_6789) District: Demo District (12345)	 Washington Office of Superintendent of PUBLIC INSTRUCTION Family Report
---	---	---

Science Test Results: Washington Comprehensive Assessment of Science

Reporting areas are broad statements of skills and knowledge students should know and be able to apply in science. Your student's performance in each reporting area contributes to the science test score.

The percentage of points needed to be AT STANDARD in each reporting area is determined from the percentage of points earned in each reporting area by students with a Level 3 test score. A student's performance is not required to be AT STANDARD or ABOVE STANDARD in all reporting areas to earn a Level 3 test score.

Skills that a student whose performance is AT STANDARD likely knows and is able to do in each science reporting area are below.

Practices & Crosscutting Concepts in Physical Sciences	Practices & Crosscutting Concepts in Life Sciences	Practices & Crosscutting Concepts in Earth & Space Sciences
A student whose performance is AT STANDARD: - Models how atoms are conserved during changes - Asks questions and investigates motion caused by contact and non-contact forces - Uses data to model energy in systems - Describes kinetic and thermal energy transfers - Models how waves travel in patterns, transfer energy, and interact - Designs devices to optimize collisions, forces, and energy transfers	A student whose performance is AT STANDARD: - Uses evidence to argue that organisms are systems of cells - Uses patterns to model the flow of energy and matter in an ecosystem and how organisms use energy and matter to survive - Uses models to understand how the structure and function of genes causes variations - Uses patterns in fossil data to compare organisms and infer evolutionary relationships - Evaluates solutions that stabilize ecosystems	A student whose performance is AT STANDARD: - Uses evidence to model Earth and other objects as part of a universe with movements controlled by gravity - Uses rock strata evidence to explain Earth's history - Models the cycling of matter and energy and explains changes in Earth's surface features and weather - Uses evidence to describe how human activities are affected by Earth's resources - Designs solutions to problems caused by using resources

12345_12345_6789_1

Page 2 of 2

10.4 INTERPRETATION OF REPORTED SCORES

A student's performance on a test is reported as a scale score and an achievement level for the overall test, and at an achievement category for each claim or reporting area. For spring 2022, the ELA and mathematics results are only available on the overall test; no claim-level results were calculated or reported. Students' scores and achievement levels are summarized at the aggregate levels. The next section provides a description of how to interpret these scores.

10.4.1 Scale Score

A scale score is used to describe how well a student performed on a test and is an estimate of a student's knowledge and skills of the standards as measured by the test. The scale score is transformed from a theta score, which is estimated based on mathematical models. Low scale scores can be interpreted as an indication that the student possess fewer of the knowledge and skills measured by the test. Conversely, high scale scores can be interpreted as an indication that the student has more knowledge and skills measured by the test.

Scale scores for both Smarter Balanced and the WCAS represent a continuum of student performance. And although there are cuts made along this continuum of scale scores that are categorized as different Achievement Levels, there is little inferable difference in performance between a student who is one point above a given cut than a student who is one point below that same cut even though those two students are reported in different achievement levels. For example, on the Spring 2022 grade 3 ELA test, one student who earned a scale score of 2433 would be categorized in Level 3 and a student who earned a scale score of 2431 would be categorized in Level 2. This is due to a cut between Level 3 and Level 2 had to be placed somewhere, but there is little inferable difference between a scale score of 2433 and a scale score of 2431.

This is due, in part, also to the SEM.

10.4.2 Standard Error of Measurement

A scale score (observed score on any test) is an estimate of the true score. If a student takes a similar test several times, the resulting scale score would vary across those times, sometimes being a little higher, sometimes a little lower, or sometimes the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test were administered several times.

The $\pm$ next to the student's scale score provides information about the certainty, or confidence, of the score's interpretation. The boundaries of the score band are one SEM above and below the student's observed scale score, representing a range of score values that is likely to contain the true score. For example, 2680 ± 10 indicates that, if a student were tested again, it is likely that the student would receive a score between 2670 and 2690. SEM can be different for students with the same scale score, depending on how closely the administered items match the student's ability.

When interpreting scale scores, it is recommended to consider the range of scale scores incorporating the SEM of the scale score. For example, if one student has a scale score of 2380 and an SEM of 20 and another student has a scale score of 2400 with an SEM of 20, there is overlap

between the range of possible scale scores for both students that can inform inferences made on those students' performances.

SEM is reported within SRS and provided to district staff in WAMS. SEM are not included on Family Reports to communicate that the scale score estimate reported for the student is the score used for Achievement Level determinations as well as state and federal reporting purposes and not the range of possible scale scores that the SEM might represent.

10.4.3 Achievement Level

Achievement levels are proficiency categories on a test that students fall into based on their scale scores. For both Smarter Balanced tests and the WCAS, scale scores are mapped into four achievement levels (i.e., Level 1, Level 2, Level 3, Level 4) using three achievement standards (cut scores). ALDs are a description of the content area knowledge and skills that students at each achievement level are expected to possess. Thus, achievement levels can be interpreted based on ALDs. Generally, students performing at Levels 3 and 4 are considered on track for success with higher grade-level learning expectations and, for the Smarter Balanced high school test, on track for success with entry-level career tasks and college coursework after high school. ALDs are available on the OSPI webpage at <https://www.k12.wa.us/student-success/testing/state-testing/scores-and-reports/achievement-level-descriptors>.

ALDs are reported in two different categories: Threshold and Range. Threshold ALDs describe skills that student likely have if they are just barely into a given level, i.e., their scale score is the minimum necessary for the achievement level. Range ALDs describe skills that students likely have at multiple scale scores within the achievement level. In this way, these two categories of ALDs are meant to articulate how one might interpret student performance for a student who is at the low end of a given achievement level differently from a student who is in the middle or high end of that same achievement level.

Therefore, in addition to the achievement level, the student's scale score with respect to where it falls within the range of scale scores for the achievement level, should be considered when interpreting student test results.

10.4.4 Achievement Category for Claims/Reporting Areas

For the spring 2022 Smarter Balanced assessments, claim scores were not generated or reported due to the adjusted test blueprint. An individual student is administered too few items in each claim to produce reliable scores.

Student performance on each reporting area for the WCAS is reported in three achievement categories: (1) Below Standard, (2) At Standard, and (3) Above Standard. A result of "Below Standard" means the student is likely able to demonstrate more skills from the Level 2 ALDs than the Level 3 ALDs as described in the WCAS grade-level ALDs, posted online at <https://www.k12.wa.us/student-success/testing/state-testing/scores-and-reports/achievement-level-descriptors>. "At Standard" means the student is likely able to demonstrate more skills from the Level 2 and 3 ALDs than the Level 4 ALDs, and "Above Standard" means the student is likely able to demonstrate many of the skills from the Level 4 ALDs.

As stated on the Family Reports for the WCAS, it is not necessary for a student to earn a “At Standard” in each Reporting Area to earn a Level 3 performance on the test overall; students can earn “Below Standard” in one or more Reporting Areas and still earn enough points on the test overall to earn a Level 3.

10.4.5 Achievement Category for Targets

For Smarter Balanced assessments, Target-level reports are produced for aggregate units (e.g., classroom, school, district) only. An individual student is administered too few items in a target to produce a reliable score for a target. Target results are not calculated for the WCAS.

The SRS reports two types of performance for each target: performance relative to entire test and performance relative to Level 3.

For target performance relative to the entire test, students’ observed performance on items within each target is compared to the students’ performance on the entire test. At the aggregate level, when observed performance within a target is greater than observed overall test performance, the target is reported as “Better” than overall test performance. Conversely, when observed performance within a target is below the observed overall test performance, the target is reported as “Worse.” Otherwise, the target is reported as “Similar.”

For target performance relative to the Level 3 cut, student performance on items within each target is compared to the Level 3 cut, the expected performance of students at the grade level. When observed performance within a target is greater than the proficiency cut, the target is reported as “Above.” Conversely, when observed performance within a target is below the proficiency cut, the target is reported as “Below.” Otherwise, the target is reported as “Near” or “At/Near.”

When interpreting Target reports, both categories of performance should be considered together. For example, consider a target where performance was “Worse” than performance on the test as a whole and “Above” relative to Level 3. In this case, student performance for that Target can be interpreted that student performance on that target was below the student’s performance on the test overall, but was above what was expected of students in the grade level with respect to being proficient with the skills and knowledge in the Target. Conversely, a result of “Better” and “Below” indicates that the student group performed better on the given Target than the test overall, but that performance was still not at the expected level of students for the grade level for those skills.

And while one might think to just look at the performance relative to Level 3 to determine if students have demonstrated performance that is expected of the grade level, this can only precisely be done for the “Above” and “Below” categories. Because the “Near” or “At/Near” category overlaps the Level 3 cut *in both directions, above and below*. Meaning that a target result of “Near” or “At/Near” could mean the observed performance was a little bit below or a little bit above what is expected of students at the grade level. There is no way to know, definitively, from Target reports if student performance was above the Level 3 cut (i.e., the expected performance of students in the grade). In combination with the “Worse,” “Similar,” and “Better” results for performance relative to the entire test, though, Targets with “Near” or “At/Near” can be loosely ranked as more of an area for growth (if categorized as “Worse”) or more of an area of strength (if categorized as “Better”).

Targets can provide some evidence to help address students' strengths and weaknesses as measured by the test. As with all test result information, Target results should be used in conjunction with other information about student learning when making instructional or program decisions, and it should be considered that student performance on each target is based on relatively few items, especially for a small group of students. One approach to interpreting target results might be to evaluate results over time, including a before-and-after evaluation of a specific instruction intervention designed to address the skills and knowledge in a given target. Another might be to consider target results in conjunction with classroom-observed student performance on specific instructional units or Smarter Balanced interims assessments, many of which are specifically aligned to a single, or at most three, Targets.

10.4.6 Aggregated Score

Students' scale scores and achievement levels can be, and for state and federal reporting are, aggregated to represent how a group of students perform on a test. When students' scale scores are aggregated, the aggregated scale scores can be interpreted as an estimate of knowledge and skills that a group of students possess. Given that student scale scores are estimates, the aggregated scale scores are also estimates and are subject to measures of uncertainty. In addition to the aggregated scale scores, the percentage of students in each achievement level are reported at the aggregate level to represent how well a group of students perform overall and by claim or reporting area.

10.5 APPROPRIATE USES FOR SCORES AND REPORTS

All states give tests to help understand what students know and can do. The state tests give policy makers information to support schools. While assessment results provide valuable information to understand students' performance, these scores and reports should be used with caution. It is important to note that scale scores are estimates of true scores and hence do not represent the precise measure for student performance. A student's scale score is associated with measurement error, and thus users need to consider measurement error when using student scores to make inferences about student achievement.

Moreover, assessment results should not be used as the only source of information, given that assessment results measured by a test provide limited information. For example, state test results should not be used as a single measure that allows or denies students access to educational opportunities. Test results should be used in conjunction with other sources of student achievement information, such as classroom assessment and teacher evaluation, when making decisions on student learning. Finally, when student performance is compared across groups, users need to take into account the group size. The smaller the group size, the larger the measurement error related to these aggregate data, thus more caution is required in interpretation.

Overall, assessment results are one source of information about what students know and are able to do in certain subject areas and can give information on whether students are on track to demonstrate knowledge and skills necessary for subsequent grade-level work and/or college and career readiness. Additionally, assessment results can be used to suggest groups of students' relative strengths and weaknesses in certain content areas.

Test information can provide a starting point or help narrow the focus for local educators' conversations and exploration of student performance with the standards. For example, achievement categories for claims/reporting areas can be used to suggest an individual student's relative strengths and weaknesses within a content area that teachers and schools can further explore and compare with other sources of information (classroom assessments, observations, projects, etc.).

Assessment results on groups of students' achievement on the test can be used to help teachers or schools make decisions on how to support students' learning. Aggregate score reports for teacher and school level can provide additional information about the strengths and weaknesses of students and can be utilized to improve teaching and student learning. For example, a group of students may have performed very well overall but possibly did not perform as well in several targets compared to their overall performance. In this case, teachers or schools can further explore strengths and weaknesses using local assessments and conversations as suggested by the assessment results. Further, by narrowing down the student performance result by student group, teachers and schools can focus their exploration of students' needs and improve teaching and student learning, particularly for students from disadvantaged student groups. Teachers can then provide additional instructions for these students to enhance their attainment of the intended student learning outcomes.

The Smarter Balanced mathematics and ELA and the WCAS tests are criterion-referenced tests, and assessment results are best used to compare student performance against the intended student learning outcomes. However, assessment results can be used to compare students' performance among different groups. Teachers can evaluate how their students perform compared with other students in schools and districts for overall scores and by claim or reporting area. Furthermore, longitudinal data, including year-over-year scale scores, can be used to describe individual students or groups of students performance over time. The scale score in the Smarter Balanced assessment is a vertical scale, which means scales are vertically linked across grades, and scores across grades are on the same scale. Therefore, scale scores are comparable across grades so that scale scores from one grade can be compared with the next. The WCAS assessment is not vertically linked, so scale scores are not comparable across grades.

SUMMARY

Smarter Balanced assessments and WCAS scores are reported online via the Smarter Balanced Reporting System (SRS) and through an electronic family report produced by Cambium. The Smarter Balanced Reporting system presents the scores after handscoring is completed; the electronic family reports are provided to districts and schools to provide to families.

In addition to student-level information, SRS provides aggregate reports at the student group, school, district, and state levels. At aggregate levels, Smarter Balanced math and ELA tests offer the option to view achievement category strengths and weaknesses for targets.

Smarter Balanced scale scores are vertically linked across grades so that they can be compared, unlike state-specific WCAS scores from different grades. The various reports available may be used to inform stakeholders regarding student performance and instructional strategies.

11. QUALITY CONTROL

Thorough quality control has been integrated into every aspect of the Washington Comprehensive Assessment Program (WCAP) assessments. From adaptive pool and test form constructions, to test booklet development and printing, to post-test score processing and analyses, the Office of Superintendent of Public Instruction (OSPI), CAI and the subcontractor Measurement Incorporated (MI) have built in multiple layers of reviews and verifications to ensure that outputs are of the highest quality. Aspects of this quality control have been discussed throughout this report. This chapter highlights some of these procedures.

For Smarter Balanced assessments that were administered as CATs, additional quality controls were conducted—such as pre-test simulations to ensure that the items selected met the selection criteria in terms of both item statistics and blueprint requirements. For more details of the quality control applicable for Smarter Balanced assessments, see the Smarter Balanced technical report (<https://validity.smarterbalanced.org/reports-and-specifications/>).

11.1 QUALITY CONTROL IN TEST CONFIGURATION

For online testing, the test configuration file contains the complete information required for test administration and scoring, such as the test blueprint specification, slopes, and intercepts for theta-to-scale score transformation, cut scores, and the item information (i.e., answer keys, item attributes, item parameters, passage information). For Smarter Balanced assessments, the configuration file contains all specifications for the CAT item selection algorithm and the scoring algorithm. For the Washington Comprehensive Assessment of Science (WCAS), the configuration file contains all of the items for each form and the scoring specification. The accuracy of the configuration file is checked and confirmed numerous times independently by multiple staff members prior to the testing window.

To verify the accuracy of the scoring engine, CAI uses simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the population. For Smarter Balanced assessments, the population includes all Smarter Balanced states. For the WCAS, the population is the Washington students who took the WCAS. The ability of each simulated student is used to generate a sequence of item response scores consistent with the underlying ability distribution. These simulations provide a check of form distributions (if administering multiple test forms) and test scores in fixed-form tests. For Smarter Balanced assessments, these simulations also provide a rigorous test of the adaptive algorithm.

Simulations are generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a very wide range of student response patterns. For Smarter Balanced assessments, the results of simulated test administrations are used to configure and evaluate the adequacy of the item selection algorithm used to administer the Smarter Balanced summative assessments. The purpose is to configure the adaptive algorithm to optimize item selection to meet blueprint specifications while targeting test information to student ability as well as checking the score accuracy.

For Smarter Balanced assessments, after the computer-adaptive test simulations, another set of simulations for the combined tests (adaptive test component plus a fixed-form performance task

component) are performed to check scores. Psychometricians compute scores using item responses from the simulation and compare their results to the scores from simulation. Their results have to match those from the scoring engine before the scoring engine is put to operational use.

11.1.1 Platform Review

CAI's Test Delivery System (TDS) supports a variety of item layouts. Each item goes through an extensive platform review on different operating systems, such as, Windows, Linux, and iOS to ensure that the item looks consistent in all systems. Some of the layouts have the stimulus and item response options/response area displayed side by side. In each of these layouts, both stimulus and response options have independent scroll bars.

Platform review is a process in which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on various platforms that are significantly different from one another.

Platform review is conducted by a team. The team leader projects the item as it was web approved in ITS, and team members, each behind a different platform, look at the same item to see that it renders as expected.

11.1.2 User Acceptance Testing and Final Review

Prior to deployment, the testing system and content are deployed to a staging server, where they are subject to user acceptance testing (UAT). UAT of the TDS serves both a software evaluation and content approval role. The UAT period provides OSPI with an opportunity to interact with the exact test with which the students will experience.

For both Smarter Balanced and WCAS tests, both internal and external UAT was conducted before the testing window opened. Detailed protocols were developed for TDS review process, and reviewers were given detailed instructions to note or report issues related to system functionality, items displaying, or scoring.

During the internal-Cambium UAT, CAI created pseudo tests that cover the entire range of possibilities of item responses and the complete set of scoring rules. The pseudo tests were then manually entered into TDS. When issues were found, CAI took immediate actions to solve them. When TDS was updated, the related pseudo cases could be re-entered to the system. The process was repeated until all issues were resolved.

Cambium provides a UAT environment for external UAT so that OSPI staff were able to conduct a hands-on review of the system prior to the testing window opening. UAT documents are created for each WCAP assessment to identify new and existing features along with test cases to ensure the system meets the client-configured specifications, complex business rules, and is functioning as expected. OSPI staff provide valuable feedback to CAI to make adjustments during the UAT review. OSPI approved TDS before the system was used for student testing.

CAI provides a small sample of test results from internal UAT efforts as a first-check of SRS. After review and approval of the sample, CAI sends the UAT test cases from OSPI's external UAT

to allow for the verification of the tests from the start in the administration through reporting. OSPI staff conduct UAT on the Smarter Reporting System, SRS. UAT documents are created by Smarter Balanced staff for each WCAP assessment with test cases to ensure the system meets the client-configured specifications and is presenting results from CAI's systems as expected. After UAT efforts are approved, CAI completes integration testing in production with SRS.

11.2 QUALITY ASSURANCE IN DOCUMENT PROCESSING

Scanning Accuracy

For paper tests, when test documents are scanned, a quality control sample of documents consisting of 10 test cases per document type (normally between 500 and 600 documents) was created so that all possible responses were verified, including various typical errors that required editing via MI's Data Inspection, Correction, and Entry (DICE) application program. This structured method of testing provided exact test parameters and a methodical way of determining that the output received from the scanner(s) was correct. MI staff carefully compared the documents and the data file created from them to further ensure that results from the scanner and editing process (validation and data correction) were accurate, and then transferred them to the CAI database.

11.3 QUALITY ASSURANCE IN DATA PREPARATION

CAI's TDS has a real-time quality-monitoring component built in. After a test is administered to a student, the TDS passes the resulting data to CAI's quality assurance (QA) system. QA conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item, keys for multiple-choice items, score points in each item, and total number of field-test items and operational items; and that the test record contains no data from items that have been invalidated.

The data is passed directly from the Quality Monitoring System (QMS) to the Database of Record (DoR), which serves as the repository for all test information, and from which all test information for reporting is pulled. For Washington, CAI provides the reporting data to SRS following the Smarter Reporting Test Results Transmission (TRT) format using an application programming interface (API). The data extract generator (DEG) is the tool that is used to pull data from the DoR for delivery to OSPI. CAI staff ensure that data in the extract files match the DoR prior to delivery to OSPI.

11.3.1 Quality Assurance in Handscoring

Double Scoring Rates, Agreement Rates, Validity Sets, and Ongoing Read-Behinds

MI's scoring process is designed to employ a high level of quality control. All scoring activities are conducted anonymously; at no time do scorers have access to student demographic information.

MI's Virtual Scoring Center (VSC) provides the infrastructure for extensive quality control procedures. Through the VSC platform, project leadership can perform spot checks (read-behinds) of each scorer to evaluate scoring performance; provide feedback and respond to questions; deliver

retraining and/or recalibration items on demand and at regularly scheduled intervals; and prevent scorers from scoring live responses in the event that they require additional monitoring.

Once scoring is underway, quality results are achieved by consistent monitoring of each scorer. The scoring director and team leaders read behind each scorer's performance every day to ensure that they are on target, and they conduct one-on-one retraining sessions when necessary. MI's QA procedures allow scoring staff to identify struggling scorers very quickly and to begin retraining immediately.

If through read-behinds (or data monitoring) it becomes apparent that a scorer is experiencing difficulties, they are given interactive feedback and mentoring on the responses that have been scored incorrectly. Retraining is an ongoing process throughout the scoring effort to ensure more accurate scoring. Daily analyses of the scorer status reports alert management personnel to individual or group retraining needs.

In addition to using validity responses as a qualification threshold, other validity responses are presented throughout scoring as ongoing checks for quality. VSC is capable of dynamically embedding calibration responses in scoring sets as individual items or in sets of whichever number of items is preferred by the State.

With the VSC program, the way in which student responses are presented prevents scorers from having any knowledge about which responses are being single- or double-read, or which responses are validity set responses.

11.3.2 Handscoring QA Monitoring Reports

MI generates detailed scorer status reports for each scoring project using a comprehensive system for collecting and analyzing score data. The scores are validated and processed according to the specifications set out by Smarter Balanced for Smarter Balanced assessments and OSPI for the WCAS. This allows MI to manage scorer quality and to take any corrective actions immediately. Updated real-time reports that show both daily and cumulative (project-to-date) data are available. These reports are available to states 24 hours a day via a secure website. Project leadership reviews these reports regularly. This mechanism allows project leadership to spot-check scores at any time and offer feedback to ensure that each scorer is on target.

11.3.3 Monitoring by OSPI

OSPI also directly observes MI activities virtually. MI provides virtual access to the training activities through the online training interface. OSPI monitors the scoring process through the Scoring Resource Center (SRC) with access to view and run specific reports during the scoring process.

11.3.4 Identifying, Evaluating, and Informing the State on Alert Responses

In addition to the processes enabled by CAI, MI also has a formal process for identifying when student responses reflect a possibly dangerous situation for the test taker. MI also flags potential security breaches that are identified and flagged during scoring. This process is used to notify state

clients of possible instances of teacher or proctor interference or student collusion. The alert procedure is habitually explained during scorer training sessions. Within the VSC system, if a scorer identifies a response that may require an alert, they flag or note that response as a possible alert and the system transfers the image to the scoring manager. Scoring management then decides if the response should be forwarded to the client for any necessary action or follow-up.

11.4 QUALITY ASSURANCE IN SCORING

To monitor the performance of the online delivery system during the test administration window, CAI statisticians examine the delivery demands, including the number of tests to be delivered, the length of the window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and CAI contracts for service in excess of this amount.

Once deployed, CAI's servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts CAI engineers at the first signs that trouble may be ahead. Applications log not only errors and exceptions, but also latency (timing) information for critical database calls. This information enables CAI to know instantly whether the system is performing as designed, or if it is starting to slow down or encounter a problem. In addition, latency data—such as data about how long it takes to load, view, or respond to an item—are captured for each assessed student. All of this information is logged as well, enabling CAI to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

A series of QA reports can also be generated at any time during the online assessment window, such as blueprint match rate, item exposure rate, and item statistics, for early detection of any unexpected issues. Any deviations from the expected outcome are flagged, investigated, and resolved. In addition to these statistics, a cheating analysis report is produced to flag any unlikely patterns of behavior in a testing session as discussed in Section 5.7.

Item statistics analysis reports allow psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational testing window. The item statistics analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. This report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation. The report is configurable and can be produced so that only items with statistics falling outside a specified range are flagged for reporting or to generate reports based on all items in the pool.

For the CAT component, other reports such as blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to the simulation results. The QA reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the testing window to ensure that test administrations conform to blueprint and that items are performing as anticipated.

Table 11.1 presents an overview of the QA reports.

Table 11.1: Overview of QA Reports

QA Reports	Purpose	Rationale
Item Statistics	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology-enhanced items)
Blueprint Match Rates	To monitor unexpectedly low blueprint match rates	Early detection of unexpected blueprint match issue
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (highly unused items/passages)	Early detection of any oversight in the blueprint specification
Response Change Analysis and Test Anomalies	To monitor testing irregularities	Early detection of testing irregularities

11.5 QUALITY ASSURANCE IN REPORTING

In the spring 2022 test administration, two types of score reports were produced: 1) Data reports available to district and school staff via the Smarter Balanced Reporting System (SRS), and 2) Family reports.

11.5.1 Student Data Files Quality Assurance

Scores for online assessments are assigned by automated systems in real time. For machine-scored portions of assessments, the machine rubrics are created and reviewed along with the items, and then validated and finalized during rubric validation following field testing. The review process “locks down” the item and rubric when the item is approved for web display (Web Approval). During operational testing, actual item responses are compared to expected item responses (given the item response theory [IRT] parameters), which can detect mis-keyed items, item score distribution, or other scoring problems. Potential issues are automatically flagged in reports available to CAI’s psychometricians.

The handscoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Handscored items are married up with the machine-scored items by CAI’s Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are checked by CAI’s QA system. The integrated scores are sent to CAI’s test scoring system, a real-time system that applies client-specific scoring rules and assigns scores from the calibrated items, including calculating achievement-level indicators, subscale scores, and other features, which then pass automatically to the reporting system and DOR. The scoring system is tested extensively prior to deployment, including hand checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

During the school year Smarter Balanced releases content update logs to notify service providers that an item requires an update and redeployment or a deactivation that may or may not require students who already saw the item to be rescored. If an item requires a rescore, CAI immediately deactivates the item to prevent additional students from receiving this item in their test. CAI and

OSPI then review the rescore options available based on the reason for the content update and the impact data. OSPI follows an approach to hold the student harmless, which prevents a student's score from going down as a result of a rescore. CAI processes the rescore and provides updated scores for reporting.

Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the “official” record is stored. Only after scores have passed the QA checks and are uploaded to the DOR are they passed via the nightly student data files to OSPI.

11.5.2 Data Reports in SRS Quality Assurance

When test results are first sent to SRS, they are embargoed meaning that only Smarter Balanced and OSPI staff can view results. Only after OSPI staff have confirmed that SRS is presenting individual and aggregate results correctly is the embargo lifted, allowing district and school users to see results.

11.5.3 Family Report Quality Assurance

Statistical Programming

The family reports contain custom programming and require rigorous QA processes to ensure their accuracy. All custom programming is guided by detailed and precise specifications in our reporting specifications document. Upon approval of the specifications, analytic rules are programmed, and each program is extensively tested on test decks and real data from other programs. The final programs are reviewed by two senior statisticians and one senior programmer to ensure that they implement the agreed-upon procedures. Custom programming is implemented independently by two statistical programming teams working from the specifications. Only when the output from both teams matches exactly are the scripts released for production. Quality control, however, does not stop there.

Much of the statistical processing is repeated, and CAI has implemented a structured software development process to ensure that the repeated tasks are implemented correctly and identically each time. Small programs, called macros, can be used to take specified data as input and produce data sets containing derived variables as output. Approximately 30 such macros reside in CAI's library for the grades 3–8 and high school program score reports. Each macro is extensively tested and stored in a central development server. Once a macro is tested and stored, changes to the macro must be approved by the director of score reporting and the director of psychometrics, as well as by the project directors for affected projects.

Each change is followed by a complete retesting with the entire collection of scenarios on which the macro was originally tested. The main statistical program is mostly made up of calls to various macros, including macros that read in and verify the data and conversion tables and the macros that do the many complex calculations. This program is developed and tested using artificial data generated to test both typical and extreme cases. Additionally, the program goes through a rigorous code review by a senior statistician.

Display Programming

The paper report development process uses graphical programming, which takes place in a Xerox-developed programming language called VIPP that allows virtually infinite control of the visual appearance of the reports. After designers at CAI create backgrounds, CAI's VIPP programmers write code that indicates where to place all variable information (data, graphics, and text) on the reports. The VIPP code is tested using both artificial and real data. CAI's data-generation utilities can read the output layout specifications and generate artificial data for direct input into the VIPP programs. This allows the testing of these programs to begin before the statistical programming is complete. In later stages, artificial data are generated according to the input layout and are run through the psychometric process and the score reporting statistical programs, and the output is formatted as VIPP input. This enables CAI to test the entire system. Programmed output goes through multiple stages of review and revision by graphics editors and the score reporting team to ensure that design elements are accurately reproduced and data are correctly displayed.

Once final data and VIPP programs are received, the CAI score reporting team reviews proofs that contain actual data based on CAI's standard QA documentation. In addition, CAI compares the data independently calculated by CAI psychometricians with the data on the reports. A large sample of reports is reviewed by several CAI staff members to make sure that all data are correctly placed on reports. All reports containing actual data are stored in a locked storage area. CAI provides student data files and individual student reports with sample districts for OSPI staff review. CAI will work closely with OSPI to resolve questions and correct any problems. The reports will not be delivered unless OSPI approves the sample reports and student data file. Once approved, CAI delivers electronic PDFs to OSPI as per the approved paper reporting specifications.

SUMMARY

Quality control is integrated into every aspect of the WCAP assessments and was fully employed for the spring 2022 tests. Prior to the opening of testing windows, simulations using test specifications as the actual tests were run to verify accuracy of the scoring engine, distribution of the test items, and alignment with the test blueprints. Test items were also reviewed by staff using various operating systems to detect any format inconsistency. After the testing windows closed, the handscoring vendor followed set procedures when selecting tests for second reads, monitoring scoring of individual scorers, and addressing issues detected. Before releasing test scores to students, all theta and scaled scores generated by the system were independently verified by staff. Presentation of results in SRS was reviewed and confirmed prior to release to local school and district users. Sample family reports were drawn and reviewed for all information contained—including student name, school and district, displayed graphics, scores achieved, and achievement levels with associated descriptors.

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.
- Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York: Wiley.
- Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *British Journal of Mathematical and Statistical Psychology*, 37, 1–21.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 37–46.
- Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *QUAL LIFE RES: Quality of Life Research*, 18(4), 447–460. doi:10.1007/s11136-009-9464-4.
- Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Fort Worth, TX: Harcourt Brace Jovanovich.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.
- Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67–86.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed Response and Differential Item Functioning: A Pragmatic Approach*. ETS Research Report Series, 1991(2), I-49. doi:10.1002/j.2333-8504.1991.tb01414.x.
- Haderlein, S. K., Saavedra, A. R., Polikoff, M. S., Silver, D., Rapaport, A., & Garland, M. (2021). Disparities in educational access in the time of COVID: Evidence from a nationally representative panel of American families. *AERA Open*, 7, 23328584211041350.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 13(4), 253–264.
- Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*, 23, 35–56.
- Jolliffe, I. T. (2002). *Principal Component Analysis, Second Edition*. Berlin: Springer.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.
- Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement*, 16(4), 247–260.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.
- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, 66(3), 331–342.
- Somes, Grant W. (1986). The Generalized Mantel–Haenszel Statistic. *The American Statistician*, 40(2), 106–108. doi: 10.1080/00031305.1986.10475369.
- Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing. *The Phillipine Statistician*, 52(1–4), 81–92.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 13(4), 265–276.

APPENDIX A

CLASSICAL ITEM ANALYSIS RESULTS AND DIF RESULTS FOR STATE-SPECIFIC TESTS

Table A-1: Grade 5 WCAS, Form A, Operational Classical Item Statistics, Spring 2022 Administration

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21005	75,796	0.45	0.61	0.39	0.61		-A	-A	-A	-A	-A
21275	75,796	0.56	0.44	0.56	0.44		+A	-A	-A	-A	-A
21276	75,796	0.38	0.34	0.66	0.34		-A	-A	-A	-A	+A
21272	75,796	0.46	0.50	0.50	0.50		-A	-A	-A	-A	-A
20770	75,796	0.50	0.46	0.45	0.18	0.38	-A	+A	+A	+A	+A
20641	75,796	0.55	0.61	0.39	0.61		+A	-A	-A	+A	+A
20642	75,796	0.52	0.28	0.72	0.28		+A	-A	-A	-A	-A
20643	75,796	0.52	0.22	0.78	0.22		+A	+B	+A	+A	-A
20644	75,796	0.45	0.43	0.57	0.43		-A	+A	+A	-A	-A
20597	75,796	0.52	0.23	0.77	0.23		-A	-A	-A	-A	+A
20598	75,796	0.38	0.48	0.52	0.48		-A	-A	-A	-A	-A
20599	75,796	0.27	0.27	0.73	0.27		-A	+A	+A	+A	+A
20601	75,796	0.34	0.21	0.79	0.21		-A	-A	-A	-A	+A
20603	75,796	0.45	0.16	0.84	0.16		+A	+A	+A	+A	+A
20857	75,796	0.62	0.37	0.63	0.37		+A	-A	-A	+A	+A
20858	75,796	0.51	0.30	0.70	0.30		+A	+A	-A	-A	-A
21033	75,796	0.37	0.30	0.70	0.30		+A	-A	+A	+A	+A
21036	75,796	0.64	0.57	0.17	0.51	0.31	+A	-A	-A	-A	-A
21037	75,796	0.37	0.65	0.10	0.50	0.40	+A	-A	+A	+A	+A
21041	75,796	0.54	0.62	0.38	0.62		-A	+A	+A	+A	+A
21044	75,796	0.45	0.75	0.25	0.75		-A	+A	+A	+A	+A
21031	75,796	0.66	0.65	0.21	0.28	0.51	+A	+A	-A	-A	-A
21032	75,796	0.53	0.27	0.55	0.37	0.09	-A	+A	+A	-A	-A
21034	75,796	0.49	0.40	0.60	0.40		-A	-A	-A	-A	-A
21035	75,796	0.56	0.41	0.38	0.42	0.20	+A	+A	+A	+A	-A
20843	75,796	0.55	0.72	0.28	0.72		+A	+A	+A	+A	+A

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
20844	75,796	0.54	0.73	0.27	0.73		+A	+A	+A	+A	-A
20855	75,796	0.09	0.13	0.87	0.13		-A	-A	-A	-A	+A
20863	75,796	0.64	0.56	0.44	0.56		-A	+A	+A	+A	-A

Table A-2: Grade 5 WCAS, Field-Test Classical Item Statistics, Spring 2022 Administration

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21421	10,812	0.50	0.61	0.39	0.61	-	-A	-A	-A	-A	-A
21422	10,812	0.23	0.17	0.83	0.17	-	+A	-A	-A	-A	-A
21423	10,812	0.45	0.37	0.36	0.53	0.11	-A	-A	-A	-A	-A
21424	10,812	0.38	0.21	0.79	0.21	-	-A	+A	-A	-A	-A
21425	10,812	0.40	0.14	0.86	0.14	-	-A	-A	-A	-A	-A
21426	10,850	0.59	0.47	0.53	0.47	-	-A	-A	-A	-A	+A
21427	10,850	0.27	0.33	0.39	0.57	0.04	-A	+A	-A	+A	-A
21428	2,436	0.21	0.19	0.81	0.19	-	+A	+A	-A	+A	-
21429	10,850	0.15	0.04	0.96	0.04	-	-A	+A	-A	+A	-A
21430	10,850	0.43	0.44	0.27	0.58	0.15	+A	-A	-A	+A	-A
21462	5,797	0.48	0.83	0.17	0.83	-	+A	-A	-A	-A	-
21467	5,761	0.49	0.54	0.46	0.54	-	+A	-A	-A	-A	-
21469	5,779	0.42	0.46	0.54	0.46	-	+A	-A	-A	-A	-
21470	5,769	0.48	0.56	0.44	0.56	-	+A	+A	-A	-A	-
21474	5,862	0.09	0.17	0.83	0.17	-	-A	-A	-A	-A	-
21489	10,818	0.50	0.31	0.69	0.31	-	+A	-A	-A	-A	+A
21491	10,818	0.55	0.38	0.43	0.37	0.20	+A	-A	-A	-A	-A

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21493	10,818	0.57	0.56	0.30	0.29	0.42	+A	+A	+A	+A	+A
21496	10,818	0.59	0.23	0.77	0.23	-	-A	-A	-A	-A	-A
21619	5,788	0.47	0.48	0.32	0.41	0.27	+A	-A	+A	-A	-
21627	5,913	0.53	0.69	0.15	0.33	0.52	-A	-A	-A	-A	-
21628	6,094	0.33	0.56	0.19	0.52	0.30	-A	-A	-A	-A	-
21630	5,748	0.51	0.53	0.47	0.53	-	+A	-A	-A	-A	-
21631	5,710	0.47	0.42	0.31	0.54	0.15	+B	+A	-A	-A	-
21632	10,827	0.07	0.08	0.92	0.08	-	-A	+A	+A	+A	+A
21633	2,059	0.46	0.22	0.65	0.26	0.08	+A	+B	+A	+A	-
21634	10,827	0.24	0.39	0.61	0.39	-	-A	-A	-A	-A	-A
21635	10,827	0.50	0.53	0.47	0.53	-	+A	-A	-A	-A	-C
21636	10,827	0.23	0.43	0.57	0.43	-	+A	+A	-A	-A	+A
21637	5,878	0.45	0.67	0.33	0.67	-	-A	-A	+A	-A	-
21639	5,852	0.49	0.48	0.52	0.48	-	+A	-A	-A	-A	-
21642	2,621	0.42	0.14	0.77	0.19	0.04	+A	+A	-A	-A	-
21645	5,845	0.60	0.54	0.46	0.54	-	-A	+A	-A	-A	-
21646	10,785	0.48	0.19	0.81	0.19	-	+A	-A	-A	-A	+A
21647	10,785	0.48	0.36	0.64	0.36	-	-A	-A	-A	-A	-A
21648	4,540	0.46	0.35	0.65	0.35	-	+A	+A	-A	-A	-
21649	10,785	0.48	0.47	0.24	0.56	0.19	+A	+A	-A	-A	-A
21650	10,785	0.39	0.53	0.47	0.53	-	-A	-A	+A	-A	+A
21651	10,871	0.25	0.28	0.53	0.38	0.09	+A	+A	+A	+A	+A
21653	10,833	0.56	0.58	0.42	0.58	-	+A	-A	-A	-A	-A
21654	10,833	0.32	0.22	0.78	0.22	-	+A	+A	-A	-A	-A
21655	10,833	0.42	0.60	0.40	0.60	-	+A	-A	-A	-A	+A
21656	10,833	0.22	0.29	0.71	0.29	-	+A	-A	-A	-A	+A
21657	10,833	0.36	0.41	0.59	0.41	-	+A	-A	-A	-A	+A
21659	10,871	0.40	0.52	0.48	0.52	-	+A	-A	-A	-A	-A

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21661	10,871	0.48	0.48	0.52	0.48	-	-A	+A	-A	-A	+A
21662	10,871	0.24	0.63	0.37	0.63	-	+A	+A	-A	-A	-A

**DIF Statistics are not calculated for demographic sample sizes <100*

Table A-3: Grade 8 WCAS, Form A, Operational Classical Item Statistics, Spring 2022 Administration

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21283	78,077	0.42	0.54	0.26	0.40	0.34	+A	-A	-A	-A	-A
20777	78,077	0.52	0.73	0.27	0.73		+A	+A	-A	+A	+A
20778	78,077	0.46	0.56	0.44	0.56		-A	-A	+A	-A	+A
20779	78,077	0.36	0.32	0.48	0.39	0.12	+A	+A	+A	+A	+A
21056	78,077	0.40	0.66	0.34	0.66		+A	-A	-A	-A	-A
21061	78,077	0.58	0.46	0.54	0.46		-A	-A	-A	-A	-A
21265	78,077	0.61	0.60	0.23	0.35	0.42	+A	+A	+A	+A	+A
21280	78,077	0.58	0.42	0.58	0.42		+A	-A	-A	-A	-A
21051	78,077	0.58	0.46	0.54	0.46		+A	+A	-A	-A	-A
20798	78,077	0.57	0.37	0.63	0.37		-A	-B	-A	-A	-A
20799	78,077	0.60	0.43	0.57	0.43		-A	-A	-A	+A	+A
20800	78,077	0.55	0.61	0.39	0.61		-A	-A	-B	-A	-A
20801	78,077	0.50	0.30	0.70	0.30		+B	+A	-A	-A	-A
20802	78,077	0.49	0.41	0.59	0.41		+A	+A	+A	+A	-A
20803	78,077	0.38	0.50	0.50	0.50		-A	+A	+A	+A	+A
20804	78,077	0.57	0.60	0.40	0.60		-A	-A	-A	-A	-A
20805	78,077	0.51	0.44	0.56	0.44		-A	+A	+A	+A	+A
20806	78,077	0.60	0.56	0.44	0.56		-A	-A	-A	-A	-A
20807	78,077	0.47	0.65	0.35	0.65		-A	-A	+A	-A	+A
21077	78,077	0.52	0.42	0.40	0.37	0.23	-A	-A	+A	+A	-A
21080	78,077	0.60	0.47	0.34	0.39	0.27	-A	-A	-A	+A	-A
21081	78,077	0.49	0.40	0.60	0.40		-A	+A	-A	+A	-A
21083	78,077	0.40	0.43	0.57	0.43		+A	-A	-A	-A	-A
21084	78,077	0.58	0.34	0.66	0.34		+A	+A	-A	+A	+A
20412	78,077	0.49	0.24	0.57	0.39	0.05	+A	+A	+A	-A	+A
20413	78,077	0.50	0.35	0.43	0.43	0.13	+A	+A	+A	+A	+A
20414	78,077	0.42	0.23	0.77	0.23		+A	+A	+A	-A	+A

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
20440	78,077	0.63	0.62	0.38	0.62		+A	+A	+A	+A	-A
21090	78,077	0.48	0.31	0.69	0.31		-A	+A	+A	+A	+A
21092	78,077	0.69	0.53	0.30	0.34	0.36	-A	+A	+A	+A	-A
21093	78,077	0.49	0.43	0.57	0.43		+A	+A	+A	+A	+A
21094	78,077	0.42	0.28	0.72	0.28		-A	+A	-A	-A	-A

Table A-4: Grade 8 WCAS, Field-Test Classical Item Statistics, Spring 2022 Administration

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21383	11,134	0.50	0.31	0.51	0.34	0.14	-A	+A	-A	+A	+A
21386	11,207	0.07	0.14	0.86	0.14	-	-A	+A	-A	-A	+A
21387	11,207	0.45	0.55	0.45	0.55	-	-A	-A	-A	-A	-A
21388	11,207	0.65	0.58	0.42	0.58	-	+A	-A	-A	-A	+A
21389	11,207	0.41	0.28	0.52	0.40	0.08	+A	+A	-A	+A	+A
21390	11,207	0.68	0.43	0.41	0.31	0.27	+A	+A	-A	-A	-A
21391	11,134	0.44	0.28	0.72	0.28	-	-A	-A	-A	-A	-A
21392	11,134	0.27	0.58	0.42	0.58	-	-A	-A	+A	-A	+A
21393	11,134	0.29	0.41	0.59	0.41	-	-A	-A	+A	-A	+A
21394	11,181	0.49	0.33	0.67	0.33	-	+A	+A	+A	+A	-A
21395	11,181	0.47	0.29	0.54	0.33	0.13	+A	-A	-A	-A	-A
21396	11,181	0.55	0.49	0.51	0.49	-	+A	+A	-A	+A	-A
21401	11,181	0.45	0.34	0.66	0.34	-	+A	+A	+A	+A	-A
21451	5,262	0.50	0.61	0.18	0.42	0.40	-A	+A	-A	-A	-
21453	5,366	0.17	0.10	0.90	0.10	-	+A	+A	-A	-A	-

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21454	5,194	0.60	0.41	0.41	0.35	0.24	+A	-A	+A	-A	-
21455	5,144	0.46	0.76	0.07	0.33	0.59	+A	+A	+A	-A	-
21456	5,199	0.34	0.36	0.64	0.36	-	-A	+A	+A	+A	-
21458	5,350	0.26	0.30	0.70	0.30	-	-A	+A	-A	-A	-
21475	5,164	0.38	0.52	0.24	0.47	0.29	-A	+A	-A	-A	-
21483	11,178	0.50	0.54	0.46	0.54	-	+A	-A	-A	-A	-C
21484	11,178	0.22	0.11	0.89	0.11	-	+A	+A	-A	-A	-A
21486	11,178	0.58	0.37	0.63	0.37	-	+A	+A	-A	+A	-A
21487	11,178	0.38	0.46	0.54	0.46	-	+A	-A	-A	-A	-A
21499	11,149	0.43	0.44	0.56	0.44	-	-A	+A	+A	-A	-A
21501	11,149	0.45	0.50	0.50	0.50	-	+A	+A	+A	+A	-A
21503	11,149	0.52	0.26	0.74	0.26	-	-A	+A	+A	+A	-A
21505	11,149	0.55	0.40	0.60	0.40	-	+A	+A	-A	+A	-A
21531	11,114	0.45	0.38	0.62	0.38	-	+A	-A	-A	-A	-A
21532	11,114	0.32	0.38	0.41	0.40	0.18	+A	-A	-A	-A	-A
21533	11,114	0.31	0.20	0.80	0.20	-	-A	-A	-A	-A	-A
21534	11,114	0.33	0.48	0.52	0.48	-	+A	+A	-A	-A	-A
21535	2,669	0.51	0.49	0.51	0.49	-	+A	+A	-	-A	-
21663	5,255	0.39	0.30	0.70	0.30	-	+A	-B	-A	-A	-
21673	5,199	0.50	0.74	0.26	0.74	-	+A	-A	-A	-A	-
21676	5,116	0.35	0.27	0.73	0.27	-	+A	+A	+A	+A	-
21677	5,226	0.53	0.25	0.75	0.25	-	+A	-A	+A	-A	-
21679	5,025	0.44	0.65	0.35	0.65	-	+A	+A	-A	+A	-
21691	5,250	0.26	0.34	0.66	0.34	-	+A	-A	-A	-A	-
21692	5,219	0.39	0.42	0.58	0.42	-	-A	-A	-B	-A	-
21714	11,114	0.54	0.65	0.19	0.31	0.50	+A	+A	-A	-A	-A
21715	11,114	0.45	0.66	0.34	0.66	-	+A	+A	+A	+A	-A
21716	3,290	0.47	0.46	0.54	0.46	-	+A	+A	+A	-A	-

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21717	11,114	0.30	0.26	0.50	0.47	0.03	+A	+A	+A	-A	-A
21718	11,114	0.25	0.25	0.75	0.25	-	-A	-A	-A	-A	+A
21724	5,108	0.46	0.18	0.82	0.18	-	-A	+A	-A	-A	-

**DIF Statistics are not calculated for demographic sample sizes <100*

Table A-5: Grade 11 WCAS, Form A, Operational Classical Item Statistics, Spring 2022 Administration

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
20814	55,727	0.27	0.28	0.72	0.28		-A	-A	-A	-A	-A
21298	55,727	0.42	0.13	0.87	0.13		-A	+A	-A	-A	-A
21310	55,727	0.28	0.15	0.85	0.15		-A	+B	+A	-A	+A
20809	55,727	0.13	0.09	0.84	0.14	0.02	-A	+A	+A	-A	+A
21098	55,727	0.28	0.19	0.81	0.19		-A	-A	+A	-A	-A
21108	55,727	0.50	0.20	0.80	0.20		-A	+A	+A	-A	-A
20706	55,727	0.57	0.71	0.14	0.30	0.56	-B	+A	-A	+A	-A
20707	55,727	0.43	0.52	0.48	0.52		+A	-A	-A	-A	-A
20821	55,727	0.53	0.41	0.59	0.41		-A	-A	-A	-A	-A
20822	55,727	0.57	0.66	0.34	0.66		+A	+A	+A	+A	-A
20823	55,727	0.62	0.57	0.43	0.57		-A	-A	-A	-A	-A
20824	55,727	0.35	0.19	0.67	0.29	0.04	-A	-A	+A	-A	-A
20825	55,727	0.24	0.14	0.86	0.14		-A	-A	-A	+A	-A
21135	55,727	0.56	0.59	0.41	0.59		-A	-A	+A	+A	+A
21136	55,727	0.45	0.49	0.51	0.49		-A	-A	-A	-A	+A
21145	55,727	0.62	0.57	0.43	0.57		+A	+A	+A	+A	+A
21127	55,727	0.36	0.22	0.78	0.22		+A	-A	+A	-A	-A
21129	55,727	0.52	0.24	0.69	0.14	0.17	-A	-A	+A	+A	+A
21131	55,727	0.51	0.45	0.55	0.45		+A	+A	+A	+A	+A
21132	55,727	0.62	0.40	0.49	0.23	0.28	+B	+A	+A	+A	-A
21134	55,727	0.39	0.39	0.61	0.39		-A	+A	-A	-A	+A
21168	55,727	0.39	0.20	0.80	0.20		-A	-A	+A	-A	+A
21169	55,727	0.28	0.26	0.74	0.26		-A	+A	+A	-A	-A
21170	55,727	0.43	0.39	0.61	0.39		-A	-B	-A	-A	-A
21173	55,727	0.64	0.56	0.22	0.43	0.35	+A	-A	-A	-A	-A
20544	55,727	0.54	0.36	0.38	0.52	0.10	+A	+A	+A	-A	-A
20545	55,727	0.55	0.34	0.66	0.34		+A	+A	+A	+A	-A

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P-value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
20551	55,727	0.26	0.42	0.58	0.42	0.01	+A	+A	+A	+A	+A
20554	55,727	0.29	0.25	0.75	0.25		+A	+A	+A	+A	+A
20555	55,727	0.30	0.06	0.89	0.10		+A	+A	+A	+A	+A
20556	55,727	0.31	0.04	0.96	0.04		-A	+A	+A	+A	-A
20704	55,727	0.29	0.31	0.69	0.31		+A	+A	+A	+A	+A
20705	55,727	0.29	0.22	0.78	0.22	0.24	+A	+A	+A	-A	-A
20771	55,727	0.34	0.74	0.26	0.74		-B	-A	-A	+A	+A
21126	55,727	0.27	0.43	0.57	0.43		+A	+A	+A	+A	-A
21130	55,727	0.57	0.38	0.48	0.28		+A	+A	+A	-A	-A

Table A-6: Grade 11 WCAS, Field-Test Classical Item Statistics, Spring 2022 Administration

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P-value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21322	4,321	0.48	0.37	0.63	0.37	-	+A	+A	-A	+A	-
21323	4,321	0.32	0.05	0.95	0.05	-	-A	+A	-A	-A	-
21324	4,321	0.39	0.59	0.20	0.42	0.39	+A	+A	-A	-A	-
21334	4,289	0.34	0.38	0.62	0.38	-	+A	-A	-A	-A	-
21337	4,289	0.44	0.18	0.82	0.18	-	-A	-A	-A	-A	-
21338	4,289	0.33	0.45	0.32	0.45	0.23	+A	+A	+A	-A	-
21342	4,294	0.34	0.43	0.57	0.43	-	-A	+A	+A	-A	-
21343	4,294	0.53	0.32	0.68	0.32	-	-A	-A	-A	-A	-
21344	4,294	0.06	0.42	0.58	0.42	-	+A	-A	+A	+A	-
21345	4,294	0.39	0.15	0.85	0.15	-	+A	-A	-A	-A	-
21354	4,282	0.43	0.43	0.33	0.48	0.19	+A	+A	-A	+A	-

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21355	4,282	0.25	0.48	0.52	0.48	-	-A	-A	-A	-A	-
21356	2,778	0.49	0.46	0.54	0.46	-	+A	+A	-A	-A	-
21357	4,282	0.54	0.67	0.33	0.67	-	+A	+A	-A	-A	-
21358	4,282	0.14	0.22	0.59	0.38	0.03	-A	-A	-A	-A	-
21536	1,982	0.51	0.43	0.33	0.48	0.19	-A	+A	-	-A	-
21537	1,972	0.47	0.45	0.31	0.47	0.22	-A	-A	-	-A	-
21538	2,036	-0.02	0.03	0.97	0.03	-	-A	-A	-	-A	-
21542	1,980	0.48	0.38	0.62	0.38	-	-A	-A	-	-A	-
21543	2,002	0.59	0.42	0.39	0.38	0.23	+A	+A	-	-A	-
21544	1,935	0.40	0.10	0.90	0.10	-	-A	+A	-	-A	-
21546	2,102	0.25	0.46	0.54	0.46	-	-A	-A	-	-A	-
21550	2,037	0.56	0.64	0.15	0.42	0.43	-A	-A	-	-A	-
21552	1,871	0.55	0.60	0.18	0.44	0.38	+A	-A	-	-A	-
21554	1,960	0.25	0.23	0.77	0.23	-	-B	-B	-	-A	-
21555	2,004	0.18	0.06	0.94	0.06	-	-A	-A	-	-A	-
21557	1,967	0.44	0.72	0.28	0.72	-	-A	-A	-	-A	-
21558	2,019	0.46	0.58	0.21	0.41	0.38	-A	-A	-	+A	-
21562	1,925	0.52	0.76	0.24	0.76	-	-A	-A	-	-A	-
21564	4,295	0.46	0.31	0.69	0.31	-	-A	-A	-A	-A	-
21565	4,295	0.32	0.49	0.51	0.49	-	-A	+A	+A	+A	-
21566	4,295	0.12	0.29	0.71	0.29	-	+A	-A	-A	-A	-
21567	4,295	0.46	0.34	0.66	0.34	-	-A	-A	+A	-A	-
21568	4,295	0.47	0.29	0.71	0.29	-	+A	+A	-A	-A	-
21569	4,295	0.52	0.39	0.61	0.39	-	+A	-A	+A	+A	-
21571	4,301	0.33	0.44	0.56	0.44	-	+A	-A	+A	+A	-
21572	4,301	0.48	0.63	0.37	0.63	-	-A	-B	-B	-A	-
21574	4,301	0.50	0.30	0.70	0.30	-	-A	-A	-B	-A	-
21577	4,301	0.39	0.59	0.41	0.59	-	-A	+A	+A	-A	-

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21578	4,301	0.42	0.36	0.64	0.36	-	+A	+A	-A	-A	-
21590	4,292	0.46	0.21	0.79	0.21	-	+A	-A	-A	-A	-
21592	2,503	0.56	0.46	0.54	0.46	-	+B	+A	-	-A	-
21594	4,292	0.49	0.45	0.55	0.45	-	+A	+A	-A	-A	-
21602	4,292	0.49	0.60	0.40	0.60	-	+A	-A	+A	-A	-
21603	4,273	0.43	0.39	0.61	0.39	-	+A	+A	-A	+A	-
21604	4,292	0.55	0.49	0.51	0.49	-	+A	-A	-A	-A	-
21605	4,273	0.56	0.32	0.68	0.32	-	+A	+A	+A	+A	-
21606	3,925	0.52	0.48	0.52	0.48	-	+B	+A	-A	-A	-
21607	4,273	0.23	0.19	0.81	0.19	-	-A	+A	+A	-A	-
21608	4,292	0.47	0.31	0.69	0.31	-	+A	+A	+A	-A	-
21609	4,273	0.50	0.27	0.63	0.20	0.17	-A	+A	-A	+A	-
21610	4,281	0.41	0.78	0.22	0.78	-	+A	+A	-A	-A	-
21611	4,281	0.54	0.47	0.53	0.47	-	+A	-A	-A	-A	-
21612	4,281	0.38	0.52	0.24	0.48	0.28	-B	+A	-A	-A	-
21613	4,281	0.46	0.21	0.79	0.21	-	-A	+A	+A	+A	-
21701	1,925	0.10	0.24	0.76	0.24	-	-A	-A	-	-A	-
21703	1,976	0.27	0.38	0.62	0.38	-	-A	+A	-	+A	-
21726	1,968	0.16	0.27	0.73	0.27	-	-B	-A	-	-A	-
21728	2,013	0.27	0.29	0.71	0.29	-	-A	-A	-	+A	-
21729	1,965	0.30	0.27	0.73	0.27	-	+A	+A	-	-A	-
21731	1,968	0.42	0.63	0.11	0.52	0.37	-A	-A	-	+A	-
21732	1,917	0.46	0.59	0.15	0.52	0.33	+A	+A	-	+A	-
21733	1,987	0.40	0.45	0.29	0.53	0.18	-A	-A	-	-A	-
21734	2,017	0.44	0.34	0.66	0.34	-	-B	+A	-	-A	-
21756	2,022	0.39	0.24	0.76	0.24	-	-A	+A	-	+A	-
21758	2,083	0.49	0.47	0.30	0.45	0.24	-A	-A	-	+A	-
21759	2,017	0.23	0.10	0.90	0.10	-	-A	+A	-	-A	-

ITS ID	Total N	Classical					DIF				
		Adjusted Point Polyserial/Biserial	Average Score (P- value)	Prop Score Point 0	Prop Score Point 1	Prop Score Point 2	Female/ Male	Asian/ White	African American /White	Hispanic/ White	Native American /White
21760	2,008	0.44	0.51	0.49	0.51	-	-A	-A	-	-A	-
21774	4,269	0.34	0.33	0.48	0.39	0.13	-B	-A	-A	-A	-
21775	4,269	0.41	0.21	0.79	0.21	-	-A	-A	-A	-A	-
21776	4,269	0.52	0.46	0.54	0.46	-	-A	-A	-A	-A	-
21777	4,269	0.47	0.23	0.62	0.30	0.08	-A	-A	+A	-A	-
21778	4,269	0.45	0.41	0.41	0.36	0.23	-A	+A	-A	+A	-
21783	2,069	0.54	0.44	0.34	0.44	0.23	-A	+A	-	-A	-
21787	4,283	0.39	0.46	0.24	0.59	0.17	-A	-A	-A	-A	-
21788	4,283	0.44	0.31	0.69	0.31	-	+A	+A	+A	-A	-
21789	4,283	0.46	0.56	0.44	0.56	-	+A	-B	-A	-A	-
21790	4,283	0.54	0.33	0.67	0.33	-	+A	-A	-A	-A	-
21791	4,283	0.43	0.27	0.73	0.27	-	-A	-A	-A	-A	-

**DIF Statistics are not calculated for demographic sample sizes <100*

APPENDIX B

IRT RESULTS FOR STATE-SPECIFIC TESTS

Table B-1: Grade 5 WCAS, Form A Operational Items, Spring 2022 Administration

ITS ID	Total N	Rasch Step Value: b_1	Rasch Step Value: b_2	Infit	Outfit
21005	75,796	-0.99		1.00	1.02
21275	75,796	-0.26		0.90	0.87
21276	75,796	0.30		1.12	1.24
21272	75,796	-0.58		1.03	1.03
20770	75,796	0.20	-0.80	1.43	1.66
20641	75,796	-1.14		0.88	0.85
20642	75,796	0.57		0.90	0.85
20643	75,796	1.04		0.89	0.73
20644	75,796	-0.08		1.04	1.13
20597	75,796	0.89		0.84	0.74
20598	75,796	-0.45		1.14	1.22
20599	75,796	0.77		1.25	1.60
20601	75,796	1.30		1.19	1.32
20603	75,796	1.66		0.97	0.82
20857	75,796	-0.15		0.79	0.72
20858	75,796	0.59		0.97	0.90
21033	75,796	0.51		1.13	1.20
21036	75,796	-2.29	0.31	0.87	0.88
21037	75,796	-2.94	-0.14	1.31	1.32
21041	75,796	-1.21		0.90	0.84
21044	75,796	-1.96		0.96	0.84
21031	75,796	-1.59	-1.03	0.84	0.82
21032	75,796	-0.20	1.92	1.02	1.12
21034	75,796	-0.02		0.99	0.97
21035	75,796	-0.98	0.87	1.07	1.07
20843	75,796	-2.18		1.00	0.82
20844	75,796	-2.18		0.97	0.79
20855	75,796	1.92		1.33	2.69
20863	75,796	-0.88		0.76	0.69

Table B-2: Grade 5 WCAS, Field-Test Items, Spring 2022 Administration

ITS ID	Total N	Rasch Step Value: b_1	Rasch Step Value: b_2	Infit	Outfit
21421	10,812	-1.17		0.88	0.82
21422	10,812	1.50		1.14	1.41
21423	10,812	-1.21	1.83	1.06	1.05
21424	10,812	1.16		0.96	1.18
21425	10,812	1.76		0.90	0.90
21426	10,850	-0.44		0.79	0.72
21427	10,850	-1.14	3.03	1.22	1.25
21428	2,436	1.13		1.18	1.25
21429	10,850	3.38		1.02	1.24
21430	10,850	-1.76	1.44	1.07	1.07
21462	5,797	-2.63		0.81	0.55
21467	5,761	-0.78		0.92	0.88
21469	5,779	-0.38		1.00	1.02
21470	5,769	-0.88		0.92	0.94
21474	5,862	1.42		1.31	1.70
21489	10,818	0.43		0.88	0.90
21491	10,818	-0.72	0.77	0.97	0.99
21493	10,818	-1.21	-0.64	0.95	1.06
21496	10,818	0.96		0.75	0.59
21619	5,788	-1.29	0.34	1.11	1.12
21627	5,913	-2.21	-0.98	0.92	0.93
21628	6,094	-2.17	0.35	1.25	1.26
21630	5,748	-0.73		0.89	0.85
21631	5,710	-1.47	1.38	1.05	1.04
21632	10,827	2.42		1.14	2.54
21633	2,059	0.40	1.59	1.08	1.01
21634	10,827	0.04		1.20	1.35
21635	10,827	-0.70		0.90	0.86
21636	10,827	-0.20		1.23	1.32
21637	5,878	-1.47		0.93	0.84
21639	5,852	-0.47		0.92	0.90
21642	2,621	1.01	2.11	0.99	1.07
21645	5,845	-0.79		0.78	0.72
21646	10,785	1.27		0.87	0.68
21647	10,785	0.17		0.91	0.91
21648	4,540	0.21		0.96	0.89
21649	10,785	-1.87	1.10	1.02	1.02
21650	10,785	-0.75		1.05	1.05
21651	10,871	-0.30	1.77	1.37	1.51
21653	10,833	-0.97		0.82	0.77
21654	10,833	1.07		1.06	1.25

ITS ID	Total N	Rasch Step Value: b_1	Rasch Step Value: b_2	Infit	Outfit
21655	10,833	-1.10		0.99	1.00
21656	10,833	0.61		1.21	1.52
21657	10,833	-0.09		1.08	1.11
21659	10,871	-0.69		1.01	1.02
21661	10,871	-0.51		0.93	0.89
21662	10,871	-1.25		1.18	1.24

Table B-3: Grade 8 WCAS, Form A Operational Items, Spring 2022 Administration

ITS ID	Total N	Rasch Step Value: b_1	Rasch Step Value: b_2	Infit	Outfit
21283	78,077	-1.37	0.23	1.41	1.47
20777	78,077	-2.02		1.03	1.06
20778	78,077	-0.43		1.03	1.05
20779	78,077	-0.05	1.75	1.43	1.78
21056	78,077	-1.18		1.08	1.11
21061	78,077	-0.07		0.88	0.83
21265	78,077	-1.43	-0.24	0.99	0.97
21280	78,077	0.11		0.88	0.81
21051	78,077	-0.14		0.88	0.83
20798	78,077	0.41		0.89	0.84
20799	78,077	-0.21		0.84	0.79
20800	78,077	-0.89		0.88	0.84
20801	78,077	1.52		1.36	1.39
20802	78,077	0.31		1.02	1.05
20803	78,077	-0.10		1.18	1.26
20804	78,077	-0.89		0.85	0.81
20805	78,077	0.01		0.98	0.95
20806	78,077	-0.68		0.84	0.77
20807	78,077	-1.13		0.99	0.97
21077	78,077	-0.60	0.82	1.23	1.26
21080	78,077	-0.91	0.63	1.06	1.06
21081	78,077	0.19		1.00	1.00
21083	78,077	0.03		1.14	1.19
21084	78,077	0.54		0.86	0.79
20412	78,077	0.08	3.02	1.06	1.18
20413	78,077	-0.55	1.74	1.17	1.20
20414	78,077	1.24		1.05	1.05
20440	78,077	-0.97		0.77	0.70
21090	78,077	0.71		0.99	1.00
21092	78,077	-1.15	-0.25	0.86	0.84
21093	78,077	-0.18		0.99	1.00
21094	78,077	0.67		1.01	1.05

Table B-4: Grade 8 WCAS, Field-Test Items, Spring 2022 Administration

ITS ID	Total N	Rasch Step Value: b_1	Rasch Step Value: b_2	Infit	Outfit
21453	5,366	2.40		1.15	1.63
21454	5,194	-0.46	0.79	0.94	0.94
21456	5,199	0.48		1.13	1.16
21458	5,350	0.81		1.21	1.43
21475	5,164	-1.55	0.61	1.26	1.27
21663	5,255	0.76		1.05	1.05
21455	5,144	-2.82	-0.99	1.01	1.02
21673	5,199	-1.68		0.83	0.74
21676	5,116	0.96		1.09	1.18
21679	5,025	-1.14		0.97	1.00
21691	5,250	0.55		1.21	1.39
21451	5,262	-1.90	-0.04	1.04	1.04
21677	5,226	1.03		0.86	0.75
21692	5,219	0.11		1.08	1.08
21724	5,108	1.57		0.90	0.79
21483	11,178	-0.57		0.91	0.90
21484	11,178	2.34		1.10	1.44
21487	11,178	-0.11		1.09	1.12
21383	11,134	-0.04	1.47	1.08	1.06
21391	11,134	0.89		0.97	0.99
21392	11,134	-0.73		1.19	1.25
21393	11,134	0.18		1.19	1.28
21499	11,149	0.02		1.02	1.06
21501	11,149	-0.30		0.98	0.98
21503	11,149	1.06		0.88	0.77
21505	11,149	0.19		0.86	0.81
21531	11,114	0.26		0.99	0.99
21532	11,114	-0.59	1.20	1.40	1.46
21533	11,114	1.40		1.10	1.23
21534	11,114	-0.25		1.13	1.18
21535	2,669	-0.26		0.89	0.88
21714	11,114	-1.56	-0.66	0.97	1.00
21715	11,114	-1.19		0.94	0.86
21716	3,290	-0.12		0.95	0.93
21717	11,114	-0.30	3.68	1.23	1.29
21718	11,114	1.13		1.17	1.48
21386	11,207	2.02		1.26	2.66
21387	11,207	-0.57		0.97	0.99
21388	11,207	-0.73		0.71	0.63
21389	11,207	-0.14	2.35	1.14	1.22
21390	11,207	-0.40	0.46	0.79	0.76

ITS ID	Total N	Rasch Step Value: b_1	Rasch Step Value: b_2	Infit	Outfit
21394	11,181	0.58		0.92	0.92
21395	11,181	0.08	1.59	1.10	1.26
21396	11,181	-0.29		0.86	0.84
21401	11,181	0.55		0.95	1.04
21486	11,178	0.33		0.82	0.79

Table B-5: Grade 11 WCAS, Form A Operational Items, Spring 2022 Administration

ITS ID	Total N	Rasch Step Value: b_1	Rasch Step Value: b_2	Infit	Outfit
20544	55,727	-1.96	1.00	0.97	0.96
20545	55,727	-0.42		0.82	0.74
20551	55,727	-0.82		1.19	1.25
20554	55,727	0.21		1.11	1.18
20555	55,727	1.36	3.30	0.98	0.96
20556	55,727	2.39		0.76	0.49
20704	55,727	-0.13		1.12	1.31
20705	55,727	0.43		1.10	1.18
20706	55,727	-2.59	-1.66	0.92	0.92
20707	55,727	-1.21		1.00	1.01
20771	55,727	-2.35		1.04	1.28
20809	55,727	0.84	2.63	1.20	2.20
20814	55,727	0.03		1.15	1.26
20821	55,727	-0.74		0.87	0.85
20822	55,727	-1.75		0.79	0.72
20823	55,727	-1.32		0.78	0.73
20824	55,727	-0.27	2.15	1.16	1.29
20825	55,727	1.28		1.23	1.39
21098	55,727	0.55		1.03	1.12
21108	55,727	0.51		0.83	0.75
21126	55,727	-0.74		1.17	1.31
21127	55,727	0.42		1.03	0.99
21129	55,727	0.52	-0.47	1.05	1.12
21130	55,727	-0.74	-0.37	0.96	1.00
21131	55,727	-0.87		0.91	0.89
21132	55,727	-0.56	-0.77	0.94	0.91
21134	55,727	-0.53		1.04	1.09
21135	55,727	-1.55		0.85	0.81
21136	55,727	-1.07		0.98	0.98
21145	55,727	-1.43		0.78	0.71
21168	55,727	0.50		0.92	0.96
21169	55,727	-0.05		1.06	1.15
21170	55,727	-0.72		0.98	0.99
21173	55,727	-2.03	0.61	1.26	1.23
21298	55,727	1.19		0.89	0.74
21310	55,727	1.04		1.03	1.31

Table B-6: Grade 11 WCAS, Field-Test Items, Spring 2022 Administration

ITS ID	Total N	Rasch Step Value: b_1	Rasch Step Value: b_2	Infit	Outfit
21322	4,321	-0.41		0.88	0.87
21323	4,321	2.26		0.88	0.71
21324	4,321	-2.39	-0.78	1.11	1.18
21334	4,289	-0.60		1.04	1.06
21337	4,289	0.63		0.89	0.73
21338	4,289	-1.80	0.01	1.21	1.23
21342	4,294	-0.72		1.04	1.04
21343	4,294	-0.17		0.81	0.74
21344	4,294	-0.70		1.33	1.46
21345	4,294	1.01		0.91	0.80
21354	4,282	-1.75	0.36	1.04	1.05
21355	4,282	-1.03		1.13	1.17
21356	2,778	-0.96		0.87	0.83
21357	4,282	-1.98		0.80	0.71
21358	4,282	-0.62	2.24	1.30	1.59
21536	1,982	-1.76	0.35	0.95	0.96
21537	1,972	-1.82	0.14	1.01	1.01
21538	2,036	3.03		1.08	3.16
21542	1,980	-0.52		0.90	0.86
21543	2,002	-1.40	-0.11	0.86	0.84
21544	1,935	1.49		0.86	0.76
21546	2,102	-0.95		1.13	1.20
21550	2,037	-2.75	-0.96	0.88	0.85
21552	1,871	-2.54	-0.66	0.90	0.89
21554	1,960	0.40		1.05	1.40
21555	2,004	2.06		1.01	1.39
21557	1,967	-2.23		0.88	0.82
21558	2,019	-2.26	-0.74	1.03	1.04
21562	1,925	-2.54		0.78	0.66
21564	4,295	-0.13		0.90	0.85
21565	4,295	-1.05		1.09	1.10
21566	4,295	0.02		1.27	1.53
21567	4,295	-0.27		0.88	0.91
21568	4,295	0.01		0.88	0.85
21569	4,295	-0.55		0.84	0.82
21571	4,301	-0.78		1.06	1.09
21572	4,301	-1.75		0.87	0.82
21574	4,301	-0.05		0.84	0.82
21577	4,301	-1.57		0.98	0.94

ITS ID	Total N	Rasch Step Value: b_1	Rasch Step Value: b_2	Infit	Outfit
21578	4,301	-0.40		0.94	0.94
21590	4,292	0.53		0.88	0.74
21592	2,503	-0.92		0.80	0.74
21594	4,292	-0.84		0.89	0.84
21602	4,292	-1.61		0.87	0.85
21603	4,273	-0.55		0.92	0.93
21604	4,292	-1.07		0.81	0.76
21605	4,273	-0.20		0.78	0.68
21606	3,925	-0.97		0.84	0.81
21607	4,273	0.65		1.06	1.32
21608	4,292	-0.13		0.89	0.84
21609	4,273	0.03	-0.25	0.94	0.93
21610	4,281	-2.65		0.88	0.91
21611	4,281	-0.97		0.83	0.79
21612	4,281	-2.26	-0.23	1.13	1.12
21613	4,281	0.52		0.86	0.74
21701	1,925	0.32		1.21	1.65
21703	1,976	-0.45		1.11	1.20
21726	1,968	0.10		1.19	1.40
21728	2,013	0.00		1.06	1.23
21729	1,965	0.07		1.06	1.10
21731	1,968	-3.23	-0.54	1.03	1.06
21732	1,917	-2.91	-0.42	0.98	0.99
21733	1,987	-2.03	0.42	1.05	1.06
21734	2,017	-0.31		0.93	0.91
21756	2,022	0.31		0.95	0.92
21758	2,083	-1.87	-0.05	1.00	1.00
21759	2,017	1.49		1.00	1.25
21760	2,008	-1.12		0.95	0.94
21774	4,269	-1.01	0.65	1.18	1.25
21775	4,269	0.49		0.90	0.89
21776	4,269	-0.92		0.83	0.79
21777	4,269	-0.32	1.12	0.93	0.97
21778	4,269	-1.20	-0.11	1.07	1.06
21783	2,069	-1.66	0.06	0.91	0.91
21787	4,283	-2.35	0.66	1.07	1.07
21788	4,283	-0.10		0.90	0.91
21789	4,283	-1.37		0.91	0.90
21790	4,283	-0.22		0.81	0.74
21791	4,283	0.15		0.92	0.87

APPENDIX C

CONVERSION TABLES FOR STATE- SPECIFIC TESTS

Table C-1: Grade 5 WCAS Online Raw Score (RS) to Scale Score (SS) Relationship with Conditional Standard Error of Measurement (CSEM), Form A, Spring 2022 Administration

Raw Score	Scale Score	CSEM	Proficiency Level
0	375	121	1
1	444	68	1
2	495	50	1
3	527	42	1
4	551	38	1
5	571	34	1
6	588	32	1
7	603	31	1
8	616	29	1
9	629	28	1
10	650	27	2
11	651	27	2
12	662	26	2
13	672	26	2
14	682	25	2
15	700	25	3
16	701	25	3
17	710	25	3
18	720	25	3
19	729	25	3
20	738	25	3
21	748	25	3
22	758	26	3
23	768	26	3
24	785	27	4
25	789	27	4
26	801	28	4
27	813	29	4
28	827	30	4
29	841	32	4
30	858	34	4
31	877	37	4
32	901	42	4
33	932	50	4
34	982	68	4
35	1060	121	4

Table C-2: Grade 8 WCAS Online Raw Score (RS) to Scale Score (SS) Relationship with Conditional Standard Error of Measurement (CSEM), Form A, Spring 2022 Administration

Raw Score	Scale Score	CSEM	Proficiency Level
0	345	124	1
1	435	69	1
2	485	50	1
3	515	41	1
4	537	37	1
5	555	33	1
6	570	31	1
7	584	29	1
8	596	28	1
9	607	27	1
10	617	26	1
11	626	25	1
12	636	25	1
13	650	24	2
14	653	24	2
15	661	23	2
16	669	23	2
17	677	23	2
18	685	23	2
19	693	23	2
20	700	23	3
21	708	23	3
22	716	23	3
23	724	23	3
24	732	24	3
25	740	24	3
26	749	24	3
27	765	25	4
28	767	25	4
29	777	26	4
30	787	27	4
31	798	28	4
32	810	29	4
33	823	30	4
34	837	32	4
35	854	35	4
36	873	38	4
37	897	43	4
38	930	52	4
39	983	71	4
40	1060	125	4

Table C-3: Grade 11 WCAS Online Raw Score (RS) to Scale Score (SS) Relationship with Conditional Standard Error of Measurement (CSEM), Form A, Spring 2022 Administration

Raw Score	Scale Score	CSEM	Proficiency Level
0	390	128	1
1	464	71	1
2	515	51	1
3	547	43	1
4	570	38	1
5	589	35	1
6	605	32	1
7	620	31	1
8	632	29	1
9	650	28	2
10	655	27	2
11	666	26	2
12	675	26	2
13	685	25	2
14	700	25	3
15	702	24	3
16	710	24	3
17	719	24	3
18	727	23	3
19	734	23	3
20	742	23	3
21	750	23	3
22	758	23	3
23	765	23	3
24	773	23	3
25	781	23	3
26	791	24	4
27	797	24	4
28	805	24	4
29	814	25	4
30	823	25	4
31	832	26	4
32	842	26	4
33	852	27	4
34	862	28	4
35	874	29	4
36	886	30	4
37	899	31	4
38	913	32	4
39	929	34	4
40	948	37	4
41	969	40	4
42	995	45	4
43	1029	53	4
44	1084	73	4
45	11190	129	4

APPENDIX D

SCALE SCORE SUMMARY FOR ACCOUNTABILITY

Table D-1: Grade 3 ELA Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	75,874	2,425.66	101.20
Gender			
Female	37,183	2,434.08	100.45
Male	38,627	2,417.53	101.23
Not Exclusively Male or Female	64	2,435.52	113.02
Ethnic Group			
American Indian/Alaskan Native	954	2,364.89	96.21
Asian	6,885	2,478.44	102.70
African American/Black	3,480	2,392.33	93.45
Latino/Hispanic	18,760	2,382.30	93.59
White	37,558	2,442.18	95.96
Pacific Islander	1,050	2,365.08	85.77
Multi-Racial	7,179	2,435.19	99.13
Race Unknown/Missing	8	2,437.25	88.55
Program			
Limited English	11,962	2351.67	85.33
Non-Limited English	63,912	2439.70	97.81
Non-Special Education	64,413	2436.53	97.84
Special Education	11,461	2364.42	97.92
Low Income	37,200	2384.45	91.61
Non-Low Income	38,674	2465.34	93.87
Migrant	1,450	2346.66	84.60

Table D-2: Grade 4 ELA Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	75,435	2,470.14	102.28
Gender			
Female	36,936	2,478.51	101.08
Male	38,421	2,462.03	102.79
Not Exclusively Male or Female	78	2,491.71	98.06
Ethnic Group			
American Indian/Alaskan Native	907	2,404.17	97.76
Asian	6,830	2,526.71	102.69
African American/Black	3,364	2,435.84	96.47
Latino/Hispanic	19,096	2,426.93	94.82
White	37,055	2,486.54	96.54
Pacific Islander	1,043	2,408.81	91.98
Multi-Racial	7,134	2,480.53	99.85
Race Unknown/Missing	6	2,411.67	76.46
Program			
Limited English	10,255	2,383.94	83.64
Non-Limited English	65,180	2,483.98	98.15
Non-Special Education	64,535	2,482.29	97.26
Special Education	10,900	2,398.03	101.65
Low Income	36,958	2,429.62	93.98
Non-Low Income	38,477	2,509.07	94.52
Migrant	1,564	2,398.74	88.75

Table D-3: Grade 5 ELA Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	76,571	2,507.11	105.51
Gender			
Female	37,304	2,517.89	103.48
Male	39,180	2,496.78	106.38
Not Exclusively Male or Female	87	2,543.92	104.42
Ethnic Group			
American Indian/Alaskan Native	948	2,436.29	98.22
Asian	6,726	2,567.19	105.44
African American/Black	3,686	2,466.57	101.29
Latino/Hispanic	19,543	2,464.58	98.25
White	37,694	2,524.28	99.38
Pacific Islander	999	2,447.68	95.58
Multi-Racial	6,972	2,515.60	103.62
Race Unknown/Missing	3	2,476.33	68.13
Program			
Limited English	9,190	2,407.09	83.17
Non-Limited English	67,381	2,521.07	100.65
Non-Special Education	65,440	2,521.18	99.35
Special Education	11,131	2,424.17	102.57
Low Income	37,625	2,465.20	97.38
Non-Low Income	38,946	2,547.67	96.86
Migrant	1,640	2,430.83	90.88

Table D-4: Grade 6 ELA Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	75,796	2,516.59	101.92
Gender			
Female	36,860	2,529.14	100.26
Male	38,742	2,504.50	102.05
Not Exclusively Male or Female	194	2,546.42	94.69
Ethnic Group			
American Indian/Alaskan Native	889	2,449.54	96.06
Asian	6,548	2,571.70	100.44
African American/Black	3,265	2,474.45	97.16
Latino/Hispanic	19,584	2,474.52	95.26
White	37,560	2,534.23	95.79
Pacific Islander	1,018	2,457.06	90.78
Multi-Racial	6,928	2,525.37	101.37
Race Unknown/Missing	4	2,442.00	116.81
Program			
Limited English	7,738	2,411.57	78.59
Non-Limited English	68,058	2,528.87	97.16
Non-Special Education	65,771	2,530.14	96.22
Special Education	10,025	2,427.41	93.09
Low Income	37,303	2,477.09	94.33
Non-Low Income	38,493	2,554.88	94.10
Migrant	1,759	2,450.28	90.67

Table D-5: Grade 7 ELA Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	77,658	2,553.64	109.49
Gender			
Female	37,497	2,567.67	106.16
Male	39,918	2,540.33	110.94
Not Exclusively Male or Female	243	2,575.13	102.89
Ethnic Group			
American Indian/Alaskan Native	961	2,482.98	110.34
Asian	6,640	2,615.43	103.31
African American/Black	3,485	2,510.39	105.95
Latino/Hispanic	20,325	2,509.83	104.63
White	38,293	2,571.92	102.49
Pacific Islander	990	2,485.98	103.72
Multi-Racial	6,960	2,563.61	107.23
Race Unknown/Missing	4	2,491.50	149.85
Program			
Limited English	7,169	2,433.83	89.91
Non-Limited English	70,489	2,566.21	103.61
Non-Special Education	67,708	2,568.64	101.87
Special Education	9,950	2,451.22	104.69
Low Income	37,855	2,512.18	104.38
Non-Low Income	39,803	2,593.08	99.19
Migrant	1,788	2,476.59	102.37

Table D-6: Grade 8 ELA Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	79,138	2,565.73	110.25
Gender			
Female	38,247	2,580.36	106.93
Male	40,587	2,551.69	111.54
Not Exclusively Male or Female	304	2,598.57	100.73
Ethnic Group			
American Indian/Alaskan Native	986	2,496.80	106.02
Asian	6,916	2,626.77	104.00
African American/Black	3,442	2,521.76	107.22
Latino/Hispanic	20,716	2,522.88	104.90
White	39,056	2,583.54	103.95
Pacific Islander	1,040	2,500.94	102.43
Multi-Racial	6,976	2,574.25	109.02
Race Unknown/Missing	6	2,541.83	92.88
Program			
Limited English	6,837	2,443.27	89.93
Non-Limited English	72,301	2,577.71	104.61
Non-Special Education	69,459	2,580.62	103.15
Special Education	9,679	2,458.49	99.82
Low Income	37,906	2,523.40	104.67
Non-Low Income	41,232	2,604.65	100.47
Migrant	1,864	2,491.40	100.26

Table D-7: High School ELA Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	74,003	2,622.95	115.61
Gender			
Female	35,674	2,637.37	110.14
Male	37,974	2,609.25	118.91
Not Exclusively Male or Female	355	2,640.22	114.92
Ethnic Group			
American Indian/Alaskan Native	869	2,553.68	111.47
Asian	6,449	2,677.18	106.20
African American/Black	3,155	2,574.37	116.71
Latino/Hispanic	18,410	2,574.89	114.91
White	38,167	2,642.51	107.16
Pacific Islander	844	2,546.95	110.26
Multi-Racial	6,109	2,635.12	112.22
Race Unknown/Missing	N/A	2,301.00	N/A
Program			
Limited English	5,842	2,477.60	99.59
Non-Limited English	68,161	2,636.04	107.71
Non-Special Education	66,164	2,637.52	107.35
Special Education	7,839	2,499.45	109.15
Low Income	32,473	2,577.03	114.75
Non-Low Income	41,530	2,658.95	102.81
Migrant	1,713	2,536.13	112.37

Table D-8: Grade 3 Mathematics Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	75,840	2,432.26	97.04
Gender			
Female	37,163	2,428.18	94.57
Male	38,613	2,436.18	99.21
Not Exclusively Male or Female	64	2,442.95	98.02
Ethnic Group			
American Indian/Alaskan Native	952	2,374.48	91.19
Asian	6,904	2,490.65	101.38
African American/Black	3,462	2,392.96	89.53
Latino/Hispanic	18,765	2,390.59	90.52
White	37,528	2,448.15	89.79
Pacific Islander	1,043	2,367.70	83.74
Multi-Racial	7,179	2,438.39	94.91
Race Unknown/Missing	7	2,426.14	115.57
Program			
Limited English	12,029	2,368.15	87.34
Non-Limited English	63,811	2,444.89	93.81
Non-Special Education	64,405	2,443.10	91.70
Special Education	11,435	2,370.79	103.42
Low Income	37,139	2,392.50	89.43
Non-Low Income	38,701	2,470.57	88.29
Migrant	1,457	2,365.80	81.79

Table D-9: Grade 4 Mathematics Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	75,366	2,472.90	97.53
Gender			
Female	36,905	2,467.68	93.87
Male	38,382	2,477.89	100.70
Not Exclusively Male or Female	79	2,487.87	85.19
Ethnic Group			
American Indian/Alaskan Native	904	2,410.95	93.88
Asian	6,844	2,537.73	101.55
African American/Black	3,357	2,429.68	91.90
Latino/Hispanic	19,096	2,430.69	88.41
White	37,010	2,488.66	90.34
Pacific Islander	1,041	2,408.33	88.37
Multi-Racial	7,108	2,480.20	94.73
Race Unknown/Missing	6	2,414.43	97.59
Program			
Limited English	10,295	2,397.77	84.58
Non-Limited English	65,071	2,485.37	93.87
Non-Special Education	64,501	2,484.54	91.89
Special Education	10,865	2,403.37	101.29
Low Income	36,899	2,432.86	88.67
Non-Low Income	38,467	2,511.43	89.85
Migrant	1,566	2,411.66	79.32

Table D-10: Grade 5 Mathematics Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	76,516	2,494.88	104.12
Gender			
Female	37,270	2,491.14	99.86
Male	39,159	2,498.42	107.90
Not Exclusively Male or Female	87	2,502.57	100.62
Ethnic Group			
American Indian/Alaskan Native	946	2,428.82	97.64
Asian	6,732	2,566.57	107.16
African American/Black	3,673	2,445.77	96.17
Latino/Hispanic	19,560	2,451.89	94.11
White	37,654	2,511.61	96.91
Pacific Islander	1,000	2,432.74	96.07
Multi-Racial	6,948	2,500.05	103.46
Race Unknown/Missing	3	2,458.33	91.25
Program			
Limited English	9,227	2,408.23	86.43
Non-Limited English	67,289	2,507.40	100.43
Non-Special Education	65,415	2,508.33	97.56
Special Education	11,101	2,415.01	106.03
Low Income	37,577	2,452.72	94.55
Non-Low Income	38,939	2,535.70	96.40
Migrant	1,642	2,427.47	87.79

Table D-11: Grade 6 Mathematics Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	75,741	2,505.41	116.04
Gender			
Female	36,835	2,501.82	112.34
Male	38,716	2,508.79	119.39
Not Exclusively Male or Female	190	2,514.51	109.64
Ethnic Group			
American Indian/Alaskan Native	894	2,434.19	112.40
Asian	6,562	2,582.69	119.48
African American/Black	3,261	2,447.25	109.56
Latino/Hispanic	19,574	2,455.65	107.44
White	37,507	2,525.90	105.84
Pacific Islander	1,023	2,424.23	109.38
Multi-Racial	6,916	2,511.31	114.58
Race Unknown/Missing	4	2,403.00	117.06
Program			
Limited English	7,772	2,396.05	101.92
Non-Limited English	67,969	2,518.63	110.53
Non-Special Education	65,751	2,520.62	107.86
Special Education	9,990	2,404.67	118.02
Low Income	37,258	2,458.90	107.98
Non-Low Income	38,483	2,550.59	105.25
Migrant	1,753	2,440.63	102.68

Table D-12: Grade 7 Mathematics Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	77,395	2,523.34	121.46
Gender			
Female	37,344	2,518.67	118.91
Male	39,811	2,527.67	123.75
Not Exclusively Male or Female	240	2,533.57	100.72
Ethnic Group			
American Indian/Alaskan Native	962	2,453.06	109.63
Asian	6,652	2,610.77	127.03
African American/Black	3,480	2,469.71	112.68
Latino/Hispanic	20,238	2,471.66	110.17
White	38,128	2,543.39	112.18
Pacific Islander	988	2,441.49	109.82
Multi-Racial	6,944	2,529.49	119.28
Race Unknown/Missing	3	2,484.00	145.12
Program			
Limited English	7,166	2,406.41	100.16
Non-Limited English	70,229	2,536.04	116.70
Non-Special Education	67,512	2,538.97	114.13
Special Education	9,883	2,415.93	115.77
Low Income	37,704	2,475.48	110.61
Non-Low Income	39,691	2,568.95	113.53
Migrant	1,786	2,449.99	104.71

Table D-13: Grade 8 Mathematics Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	78,872	2,534.21	128.91
Gender			
Female	38,100	2,532.54	124.84
Male	40,473	2,535.69	132.66
Not Exclusively Male or Female	299	2,546.64	120.63
Ethnic Group			
American Indian/Alaskan Native	979	2,457.20	116.24
Asian	6,918	2,625.79	139.65
African American/Black	3,447	2,469.99	116.57
Latino/Hispanic	20,666	2,481.83	114.02
White	38,877	2,554.80	120.36
Pacific Islander	1,031	2,448.96	112.25
Multi-Racial	6,948	2,539.81	126.50
Race Unknown/Missing	6	2,477.67	106.23
Program			
Limited English	6,846	2,413.68	105.28
Non-Limited English	72,026	2,546.39	124.74
Non-Special Education	69,239	2,550.35	122.48
Special Education	9,633	2,417.48	113.12
Low Income	37,781	2,482.83	114.66
Non-Low Income	41,091	2,581.56	123.05
Migrant	1,847	2,462.93	105.22

Table D-14: High School Mathematics Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	72,736	2,561.25	135.45
Gender			
Female	35,014	2,558.48	129.50
Male	37,380	2,563.80	140.83
Not Exclusively Male or Female	342	2,565.83	126.52
Ethnic Group			
American Indian/Alaskan Native	851	2,480.06	114.11
Asian	6,331	2,655.28	140.15
African American/Black	3,053	2,496.57	118.55
Latino/Hispanic	18,189	2,502.32	118.96
White	37,495	2,582.16	128.21
Pacific Islander	834	2,476.42	111.47
Multi-Racial	5,983	2,568.25	136.60
Race Unknown/Missing	N/A	2,392.00	74.30
Program			
Limited English	5,793	2,433.24	106.76
Non-Limited English	66,943	2,573.12	131.65
Non-Special Education	65,099	2,576.54	129.28
Special Education	7,637	2,430.14	114.80
Low Income	31,867	2,505.29	120.46
Non-Low Income	40,869	2,605.01	130.33
Migrant	1,710	2,470.76	107.95

Table D-15: Grade 5 WCAS Scale Score Means & Standard Deviations (SD) for Total and Student Groups

	Number Tested	Mean	SD
Total	76,132	695.59	80.86
Gender			
Female	37,081	693.80	78.90
Male	38,966	697.24	82.64
Not Exclusively Male or Female	85	717.00	81.75
Ethnic Group			
American Indian/Alaskan Native	933	648.28	70.97
Asian	6,704	736.61	83.96
African American/Black	3,663	657.42	69.61
Latino/Hispanic	19,460	660.11	68.91
White	37,456	711.91	78.33
Pacific Islander	993	642.04	63.40
Multi-Racial	6,920	702.01	81.03
Race Unknown/Missing	3	667.00	139.48
Program			
Limited English	9,182	623.32	53.97
Non-Limited English	66,950	705.99	78.74
Non-Special Education	65,106	704.30	78.55
Special Education	11,026	643.80	74.81
Low Income	37,378	662.98	70.10
Non-Low Income	38,754	727.14	78.01
Migrant	1,638	635.55	58.77

Table D-16: Grade 8 WCAS Scale Score Means & Standard Deviations (SD) for Total and Student Groups

	Number Tested	Mean	SD
Total	78,941	683.54	86.04
Gender			
Female	38,161	679.86	82.89
Male	40,489	686.84	88.78
Not Exclusively Male or Female	291	706.44	81.49
Ethnic Group			
American Indian/Alaskan Native	973	636.99	70.32
Asian	6,914	727.16	88.38
African American/Black	3,446	641.49	72.44
Latino/Hispanic	20,682	646.21	73.64
White	38,902	701.08	83.88
Pacific Islander	1,039	625.78	67.15
Multi-Racial	6,979	689.83	85.42
Race Unknown/Missing	6	637.67	87.34
Program			
Limited English	6,852	599.55	53.68
Non-Limited English	72,089	692.05	84.10
Non-Special Education	69,323	692.40	84.09
Special Education	9,618	619.30	71.50
Low Income	37,815	649.94	75.25
Non-Low Income	41,126	714.50	83.71
Migrant	1,841	626.32	63.56

Table D-17: Grade 11 WCAS Scale Score Means & Standard Deviations (SD) for Total and Student Groups

	Number Tested	Mean	SD
Total	57,068	696.15	77.54
Gender			
Female	26,958	693.81	72.57
Male	29,890	698.24	81.73
Not Exclusively Male or Female	220	699.60	74.60
Ethnic Group			
American Indian/Alaskan Native	695	658.61	68.76
Asian	4,820	731.53	81.22
African American/Black	2,374	655.29	72.64
Latino/Hispanic	14,631	664.93	69.85
White	29,447	710.53	73.90
Pacific Islander	702	639.68	75.90
Multi-Racial	4,397	702.81	76.06
Race Unknown/Missing	2	742.00	32.53
Program			
Limited English	4,851	617.43	60.21
Non-Limited English	52,217	703.81	74.72
Non-Special Education	51,053	703.21	75.41
Special Education	6,015	635.98	68.86
Low Income	25,139	667.80	71.35
Non-Low Income	31,929	718.51	74.87
Migrant	1,379	646.23	64.47

APPENDIX E

Percentage of Students by Achievement Level for Accountability

Table E-1: Grade 3 ELA Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	76,383	27.49	21.25	22.42	28.41	0.35	0.31
Gender							
Female	37,406	30.09	22.02	22.27	25.26	0.29	0.30
Male	38,912	24.97	20.51	22.57	31.44	0.41	0.32
Not Exclusively Male or Female	65	38.46	15.38	18.46	26.15	1.54	N/A
Ethnic Group							
American Indian/Alaskan Native	961	9.68	14.26	24.25	51.09	0.62	0.10
Asian	6,947	48.29	21.10	15.81	14.55	0.07	0.82
Black/African American	3,514	16.22	17.87	26.21	39.13	0.40	0.57
Hispanic/Latino	18,914	12.85	17.11	25.17	44.42	0.35	0.46
White	37,752	32.46	23.90	21.75	21.45	0.37	0.14
Pacific Islander	1,070	7.38	13.27	27.94	51.03	0.37	1.50
Two or More Races	7,217	30.65	22.09	22.17	24.58	0.48	0.04
Unknown/Missing	8	25.00	25.00	25.00	25.00	N/A	N/A
Program							
Limited English	12,203	5.69	10.70	24.35	58.70	0.21	1.76
Non-Limited English	64,180	31.64	23.25	22.05	22.65	0.38	0.04
Non-Special Education	64,817	30.42	22.70	22.53	23.98	0.27	0.35
Special Education	11,566	11.08	13.12	21.76	53.20	0.80	0.10
Low Income	37,517	12.91	17.95	25.89	42.70	0.45	0.40
Non-Low Income	38,866	41.57	24.42	19.06	14.61	0.27	0.23
Migrant	1,472	4.35	12.43	22.49	60.19	0.34	1.15

Table E-2: Grade 4 ELA Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	75,956	27.86	22.41	20.09	29.24	0.33	0.35
Gender							
Female	37,181	30.42	22.87	20.01	26.34	0.28	0.38
Male	38,696	25.39	21.97	20.15	32.04	0.38	0.33
Not Exclusively Male or Female	79	34.18	22.78	21.52	21.52	N/A	1.27
Ethnic Group							
American Indian/Alaskan Native	912	9.21	15.57	20.94	53.73	0.55	N/A
Asian	6,884	50.39	21.88	13.64	13.87	0.07	0.71
Black/African American	3,393	16.18	19.27	22.66	41.44	0.29	0.56
Hispanic/Latino	19,288	13.14	18.74	22.71	45.00	0.32	0.67
White	37,238	32.80	24.82	19.64	22.34	0.37	0.12
Pacific Islander	1,066	8.44	16.04	21.20	53.94	0.28	1.88
Two or More Races	7,169	31.05	23.60	20.04	24.84	0.43	0.06
Unknown/Missing	6	N/A	33.33	16.67	50.00	N/A	N/A
Program							
Limited English	10,526	3.92	9.76	20.50	65.30	0.25	2.33
Non-Limited English	65,430	31.72	24.45	20.02	23.44	0.35	0.04
Non-Special Education	64,961	30.86	24.11	20.48	24.21	0.27	0.39
Special Education	10,995	10.20	12.38	17.74	58.92	0.72	0.15
Low Income	37,253	13.59	18.89	23.55	43.48	0.39	0.40
Non-Low Income	38,703	41.60	25.80	16.75	15.53	0.28	0.31
Migrant	1,590	5.60	13.96	24.03	56.10	0.25	1.38

Table E-3: Grade 5 ELA Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	77,171	25.42	27.64	19.45	26.98	0.41	0.36
Gender							
Female	37,574	28.34	28.66	19.25	23.27	0.39	0.33
Male	39,508	22.62	26.65	19.67	30.54	0.44	0.39
Not Exclusively Male or Female	89	39.33	34.83	8.99	15.73	N/A	2.25
Ethnic Group							
American Indian/Alaskan Native	954	8.28	16.14	19.81	55.14	0.52	0.10
Asian	6,787	49.17	25.64	12.44	12.48	0.18	0.72
Black/African American	3,723	12.84	24.71	21.76	40.05	0.46	0.54
Hispanic/Latino	19,764	12.21	22.96	23.16	41.11	0.40	0.72
White	37,900	29.82	30.76	18.69	20.27	0.41	0.13
Pacific Islander	1,020	7.84	22.25	21.67	47.55	0.49	1.57
Two or More Races	7,020	27.46	29.79	18.32	23.76	0.64	0.04
Unknown/Missing	3	N/A	33.33	33.33	33.33	N/A	N/A
Program							
Limited English	9,475	2.11	9.76	20.41	66.93	0.38	2.63
Non-Limited English	67,696	28.68	30.14	19.32	21.39	0.42	0.05
Non-Special Education	65,930	28.40	29.94	19.76	21.48	0.34	0.41
Special Education	11,241	7.94	14.14	17.67	59.29	0.86	0.12
Low Income	37,986	12.03	23.75	23.10	40.52	0.48	0.47
Non-Low Income	39,185	38.40	31.41	15.91	13.86	0.35	0.26
Migrant	1,673	4.30	17.81	21.34	55.83	0.48	1.49

Table E-4: Grade 6 ELA Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	76,715	16.53	28.89	25.84	27.83	0.83	0.37
Gender							
Female	37,291	19.35	30.84	25.52	23.41	0.81	0.35
Male	39,225	13.83	26.98	26.16	32.10	0.84	0.39
Not Exclusively Male or Female	199	21.11	39.70	23.12	14.57	1.51	1.01
Ethnic Group							
American Indian/Alaskan Native	910	3.96	16.48	23.74	53.52	2.31	N/A
Asian	6,626	34.09	34.74	17.51	13.27	0.21	0.97
Black/African American	3,327	7.39	21.37	26.96	42.89	1.26	0.60
Hispanic/Latino	19,887	6.86	20.89	28.71	42.54	0.92	0.60
White	37,898	19.60	33.19	25.75	20.65	0.75	0.15
Pacific Islander	1,055	4.27	16.59	25.31	51.94	1.61	1.90
Two or More Races	7,008	18.61	29.84	25.91	24.54	1.06	0.09
Unknown/Missing	4	N/A	25.00	N/A	75.00	N/A	N/A
Program							
Limited English	8,052	0.97	5.50	20.13	72.26	0.78	3.12
Non-Limited English	68,663	18.36	31.63	26.51	22.62	0.83	0.05
Non-Special Education	66,503	18.59	31.74	26.63	22.27	0.69	0.42
Special Education	10,212	3.11	10.30	20.72	64.04	1.74	0.09
Low Income	37,874	6.76	21.95	28.86	41.23	1.09	0.42
Non-Low Income	38,841	26.06	35.65	22.90	14.76	0.57	0.32
Migrant	1,795	3.40	15.54	28.41	51.81	0.72	1.28

Table E-5: Grade 7 ELA Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	78,834	19.26	33.86	22.12	23.55	1.11	0.38
Gender							
Female	38,076	22.39	35.82	21.25	19.29	1.15	0.37
Male	40,511	16.31	31.99	22.93	27.61	1.07	0.40
Not Exclusively Male or Female	247	22.27	38.46	22.27	15.38	1.62	N/A
Ethnic Group							
American Indian/Alaskan Native	993	5.64	22.76	21.65	46.73	3.22	N/A
Asian	6,700	41.16	34.93	13.58	10.03	0.27	0.63
Black/African American	3,580	8.94	27.26	25.31	36.28	1.96	0.70
Hispanic/Latino	20,724	8.40	27.02	26.88	36.33	1.24	0.69
White	38,744	22.57	38.33	20.74	17.30	0.97	0.19
Pacific Islander	1,024	4.79	22.27	27.05	43.95	1.86	1.46
Two or More Races	7,065	21.51	34.96	21.59	20.48	1.42	0.07
Unknown/Missing	4	N/A	50.00	N/A	50.00	N/A	N/A
Program							
Limited English	7,515	0.92	7.64	21.48	68.37	1.17	3.43
Non-Limited English	71,319	21.20	36.62	22.19	18.83	1.10	0.06
Non-Special Education	68,636	21.60	37.04	22.21	18.12	0.93	0.42
Special Education	10,198	3.55	12.44	21.48	60.13	2.27	0.16
Low Income	38,615	8.72	27.90	26.40	35.32	1.52	0.45
Non-Low Income	40,219	29.39	39.58	18.00	12.26	0.71	0.33
Migrant	1,823	3.57	20.13	26.44	48.82	0.99	0.93

Table E-6: Grade 8 ELA Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	80,409	18.07	33.44	24.17	23.04	1.20	0.38
Gender							
Female	38,874	21.31	35.14	23.54	18.69	1.23	0.38
Male	41,221	14.96	31.78	24.78	27.23	1.17	0.37
Not Exclusively Male or Female	314	23.89	40.13	21.97	12.42	1.59	1.59
Ethnic Group							
American Indian/Alaskan Native	1,012	4.94	20.95	25.79	45.75	2.37	0.20
Asian	7,006	37.47	37.27	15.24	9.55	0.39	0.90
Black/African American	3,532	8.21	26.50	28.20	35.08	1.98	0.57
Hispanic/Latino	21,148	7.75	26.48	29.53	34.73	1.36	0.69
White	39,522	21.42	37.42	22.68	17.37	1.05	0.13
Pacific Islander	1,089	5.60	20.11	28.74	42.70	2.30	2.20
Two or More Races	7,094	19.65	35.49	22.31	20.89	1.62	0.04
Unknown/Missing	6	16.67	16.67	33.33	33.33	N/A	N/A
Program							
Limited English	7,216	0.65	6.54	23.66	67.18	1.41	3.84
Non-Limited English	73,193	19.78	36.09	24.22	18.69	1.18	0.04
Non-Special Education	70,483	20.30	36.57	24.26	17.76	1.02	0.43
Special Education	9,926	2.22	11.18	23.54	60.57	2.45	0.04
Low Income	38,760	7.86	26.66	28.93	34.66	1.78	0.43
Non-Low Income	41,649	27.56	39.74	19.73	12.24	0.66	0.34
Migrant	1,902	3.26	19.61	29.86	46.06	1.00	1.00

Table E-7: High School ELA Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	76,601	33.57	33.43	17.05	12.95	2.85	0.54
Gender							
Female	36,871	37.43	34.28	15.51	9.86	2.76	0.48
Male	39,350	29.88	32.65	18.51	15.88	2.91	0.59
Not Exclusively Male or Female	380	41.32	30.00	13.95	9.74	5.00	1.58
Ethnic Group							
American Indian/Alaskan Native	933	10.72	31.19	25.19	26.05	6.65	0.21
Asian	6,607	53.72	29.14	9.72	5.66	1.66	0.73
Black/African American	3,389	18.21	31.25	22.75	21.51	6.05	0.86
Hispanic/Latino	19,360	18.00	33.01	23.11	21.85	3.76	1.15
White	39,074	39.67	34.79	14.59	8.72	2.11	0.20
Pacific Islander	942	10.93	27.28	25.69	28.03	7.64	2.76
Two or More Races	6,295	37.47	33.15	15.77	10.68	2.84	0.11
Unknown/Missing	1	N/A	N/A	N/A	100.00	N/A	100.00
Program							
Limited English	6,512	1.78	12.67	28.64	51.15	4.87	5.42
Non-Limited English	70,089	36.52	35.35	15.97	9.40	2.66	0.09
Non-Special Education	68,263	37.06	35.32	15.88	9.10	2.49	0.59
Special Education	8,338	5.01	17.93	26.60	44.50	5.80	0.18
Low Income	34,189	18.65	32.53	23.10	21.21	4.25	0.77
Non-Low Income	42,412	45.60	34.15	12.16	6.30	1.71	0.36
Migrant	1,793	9.20	28.22	27.72	31.57	3.07	1.39

Table E-8: Grade 3 Mathematics Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	76,693	24.05	26.27	20.95	28.31	0.38	0.73
Gender							
Female	37,562	22.08	26.24	21.69	29.65	0.32	0.74
Male	39,065	25.94	26.32	20.24	27.03	0.44	0.72
Not Exclusively Male or Female	66	30.30	18.18	24.24	25.76	1.52	1.52
Ethnic Group							
American Indian/Alaskan Native	961	8.12	16.44	24.97	49.53	0.83	0.10
Asian	7,041	48.37	24.70	13.92	12.80	0.20	1.75
Black/African American	3,526	11.09	21.36	23.68	43.08	0.71	1.11
Hispanic/Latino	19,062	10.42	20.95	23.66	44.55	0.37	1.19
White	37,801	28.20	30.12	20.42	20.89	0.35	0.37
Pacific Islander	1,081	6.01	12.58	25.62	54.67	1.02	2.50
Two or More Races	7,213	25.80	27.49	20.89	25.40	0.39	0.08
Unknown/Missing	8	25.00	25.00	N/A	37.50	12.50	N/A
Program							
Limited English	12,622	6.13	14.59	22.81	56.07	0.40	4.30
Non-Limited English	64,071	27.59	28.58	20.59	22.84	0.38	0.03
Non-Special Education	65,168	26.43	27.98	21.45	23.78	0.33	0.85
Special Education	11,525	10.60	16.61	18.14	53.95	0.69	0.10
Low Income	37,686	10.52	21.67	24.37	42.89	0.51	0.94
Non-Low Income	39,007	37.13	30.72	17.66	14.23	0.25	0.53
Migrant	1,487	4.51	15.20	22.73	57.36	0.20	1.82

Table E-9: Grade 4 Mathematics Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	76,141	21.99	24.75	26.91	25.98	0.33	0.69
Gender							
Female	37,271	19.30	24.94	28.22	27.23	0.28	0.70
Male	38,788	24.57	24.57	25.62	24.82	0.38	0.66
Not Exclusively Male or Female	82	24.39	28.05	36.59	10.98	N/A	3.66
Ethnic Group							
American Indian/Alaskan Native	908	6.06	16.08	27.42	50.00	0.44	N/A
Asian	6,946	49.01	23.11	16.86	10.80	0.23	1.24
Black/African American	3,407	9.36	18.73	29.50	42.00	0.32	1.14
Hispanic/Latino	19,362	8.60	19.03	30.80	41.25	0.30	1.07
White	37,287	25.63	28.71	26.55	18.76	0.33	0.41
Pacific Islander	1,072	4.38	16.42	27.61	51.21	0.37	2.52
Two or More Races	7,152	23.77	26.45	26.61	22.61	0.50	0.11
Unknown/Missing	7	N/A	28.57	42.86	28.57	N/A	14.29
Program							
Limited English	10,840	4.31	10.06	27.48	57.74	0.42	4.61
Non-Limited English	65,301	24.93	27.19	26.81	20.71	0.32	0.03
Non-Special Education	65,183	24.24	26.78	27.60	21.09	0.27	0.78
Special Education	10,958	8.61	12.72	22.80	55.11	0.72	0.13
Low Income	37,371	9.22	19.48	30.97	39.89	0.40	0.86
Non-Low Income	38,770	34.31	29.84	23.00	12.58	0.26	0.52
Migrant	1,595	4.45	13.73	30.41	51.10	0.31	1.50

Table E-10: Grade 5 Mathematics Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	77,331	21.80	17.20	25.69	34.90	0.38	0.67
Gender							
Female	37,650	19.64	16.95	27.04	35.96	0.37	0.64
Male	39,591	23.85	17.42	24.39	33.92	0.39	0.70
Not Exclusively Male or Female	90	21.11	26.67	27.78	24.44	N/A	3.33
Ethnic Group							
American Indian/Alaskan Native	950	6.74	9.58	21.58	61.68	0.32	0.11
Asian	6,841	49.26	17.34	17.89	15.36	0.15	1.45
Black/African American	3,719	8.50	11.94	25.84	53.21	0.46	0.78
Hispanic/Latino	19,863	9.15	12.15	26.73	51.53	0.40	1.13
White	37,931	25.33	20.55	26.60	27.13	0.37	0.36
Pacific Islander	1,028	6.03	9.44	24.22	59.92	0.29	2.43
Two or More Races	6,996	23.14	18.24	26.09	31.92	0.60	0.09
Unknown/Missing	3	N/A	33.33	33.33	33.33	N/A	N/A
Program							
Limited English	9,767	2.87	4.85	19.43	72.44	0.40	5.13
Non-Limited English	67,564	24.54	18.99	26.59	29.48	0.38	0.03
Non-Special Education	66,129	24.20	18.85	27.12	29.50	0.31	0.77
Special Education	11,202	7.61	7.51	17.20	66.80	0.83	0.07
Low Income	38,091	9.13	12.59	26.98	50.77	0.49	0.86
Non-Low Income	39,240	34.10	21.68	24.43	19.50	0.27	0.49
Migrant	1,678	4.41	7.57	25.21	62.28	0.54	1.61

Table E-11: Grade 6 Mathematics Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	76,787	17.58	17.84	27.72	36.08	0.75	0.62
Gender							
Female	37,307	16.02	17.51	28.44	37.32	0.67	0.59
Male	39,283	19.08	18.14	27.00	34.95	0.81	0.64
Not Exclusively Male or Female	197	14.72	22.34	35.53	24.87	2.54	1.02
Ethnic Group							
American Indian/Alaskan Native	907	4.96	8.38	24.81	60.42	1.43	N/A
Asian	6,672	41.73	21.54	19.81	16.52	0.37	1.27
Black/African American	3,321	5.90	11.08	24.96	57.12	0.90	0.90
Hispanic/Latino	19,966	6.63	11.54	27.09	53.79	0.91	1.06
White	37,870	20.54	21.53	29.88	27.35	0.66	0.30
Pacific Islander	1,060	3.77	8.30	21.04	66.04	0.85	2.64
Two or More Races	6,987	19.06	18.23	28.05	33.71	0.93	0.09
Unknown/Missing	4	N/A	N/A	25.00	75.00	N/A	N/A
Program							
Limited English	8,295	2.01	4.05	14.53	78.47	0.90	5.40
Non-Limited English	68,492	19.46	19.51	29.32	30.94	0.73	0.04
Non-Special Education	66,628	19.57	19.60	29.45	30.73	0.62	0.70
Special Education	10,159	4.55	6.30	16.37	71.13	1.59	0.07
Low Income	37,917	6.92	12.27	27.62	52.18	0.96	0.78
Non-Low Income	38,870	27.97	23.28	27.82	20.37	0.54	0.46
Migrant	1,800	3.94	9.72	24.56	60.89	0.83	1.78

Table E-12: Grade 7 Mathematics Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	78,728	17.64	19.34	26.15	35.75	1.08	0.62
Gender							
Female	37,972	16.05	18.70	26.88	37.27	1.06	0.60
Male	40,510	19.14	19.92	25.45	34.36	1.09	0.63
Not Exclusively Male or Female	246	15.85	23.58	29.67	28.46	2.03	0.41
Ethnic Group							
American Indian/Alaskan Native	989	4.55	10.82	22.14	59.76	2.73	N/A
Asian	6,751	44.14	22.10	18.12	15.20	0.43	1.04
Black/African American	3,561	6.60	12.52	26.57	52.77	1.52	0.76
Hispanic/Latino	20,748	6.63	12.65	26.21	53.19	1.28	1.18
White	38,616	20.40	23.54	27.54	27.50	0.97	0.30
Pacific Islander	1,025	3.22	10.34	20.39	64.59	1.46	2.15
Two or More Races	7,034	19.05	19.33	27.28	33.11	1.19	0.09
Unknown/Missing	4	25.00	N/A	N/A	50.00	25.00	N/A
Program							
Limited English	7,741	1.82	3.46	14.13	79.06	1.51	5.92
Non-Limited English	70,987	19.37	21.07	27.47	31.03	1.03	0.04
Non-Special Education	68,618	19.66	21.28	27.85	30.26	0.93	0.69
Special Education	10,110	3.95	6.22	14.66	72.98	2.11	0.14
Low Income	38,563	6.89	13.53	27.09	51.01	1.44	0.79
Non-Low Income	40,165	27.97	24.93	25.26	21.10	0.73	0.45
Migrant	1,829	3.28	9.79	24.38	61.67	0.82	1.53

Table E-13: Grade 8 Mathematics Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	80,220	17.97	15.80	24.35	40.74	1.11	0.57
Gender							
Female	38,749	16.83	15.82	25.03	41.16	1.12	0.55
Male	41,161	19.03	15.77	23.69	40.40	1.09	0.58
Not Exclusively Male or Female	310	18.39	17.10	27.42	34.19	2.90	0.65
Ethnic Group							
American Indian/Alaskan Native	1,001	4.00	9.09	20.68	64.14	2.00	0.20
Asian	7,039	44.45	17.94	17.79	19.29	0.48	1.24
Black/African American	3,536	5.54	9.50	21.78	61.45	1.70	0.82
Hispanic/Latino	21,149	6.84	10.44	23.46	57.97	1.26	1.03
White	39,360	20.92	19.29	26.25	32.53	0.99	0.23
Pacific Islander	1,085	3.87	6.27	19.63	67.56	2.67	2.30
Two or More Races	7,044	18.78	15.84	25.55	38.47	1.32	0.04
Unknown/Missing	6	N/A	16.67	33.33	50.00	N/A	N/A
Program							
Limited English	7,405	1.89	3.04	10.99	82.39	1.65	5.90
Non-Limited English	72,815	19.60	17.10	25.71	36.51	1.06	0.02
Non-Special Education	70,389	20.07	17.44	26.04	35.43	0.99	0.64
Special Education	9,831	2.87	4.05	12.29	78.79	1.97	0.04
Low Income	38,653	6.94	10.56	23.88	57.04	1.55	0.70
Non-Low Income	41,567	28.22	20.68	24.79	25.59	0.70	0.44
Migrant	1,896	3.80	7.44	22.26	65.08	1.37	1.21

Table E-14: High School Mathematics Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	75,237	15.90	17.91	22.29	41.13	2.73	0.59
Gender							
Female	36,155	14.36	18.22	23.24	41.54	2.61	0.55
Male	38,708	17.35	17.63	21.39	40.77	2.81	0.62
Not Exclusively Male or Female	374	14.44	17.11	23.53	37.97	6.68	1.87
Ethnic Group							
American Indian/Alaskan Native	895	2.68	7.49	18.99	65.92	4.92	N/A
Asian	6,517	39.53	22.69	16.77	19.00	1.98	0.87
Black/African American	3,281	4.42	10.36	20.05	59.25	5.91	1.04
Hispanic/Latino	19,073	5.16	10.99	21.13	59.31	3.37	1.26
White	38,403	18.58	21.58	24.27	33.37	2.16	0.21
Pacific Islander	915	2.19	6.78	20.44	64.26	6.23	2.62
Two or More Races	6,150	17.53	18.59	21.33	39.90	2.62	0.10
Unknown/Missing	3	N/A	N/A	N/A	100.00	N/A	100.00
Program							
Limited English	6,526	1.26	2.76	9.12	82.01	4.80	6.44
Non-Limited English	68,711	17.29	19.35	23.54	37.25	2.54	0.04
Non-Special Education	67,158	17.61	19.67	23.82	36.45	2.42	0.65
Special Education	8,079	1.71	3.29	9.57	80.00	5.37	0.10
Low Income	33,401	5.50	11.56	21.13	57.91	3.84	0.75
Non-Low Income	41,836	24.20	22.98	23.22	27.73	1.85	0.46
Migrant	1,775	1.80	6.87	16.56	72.68	2.03	1.63

Table E-15: Grade 5 WCAS Percentage Meeting Standards for Total and Student Groups

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	76,618	17.76	34.06	22.33	25.82	0.03	0.60
Gender							
Female	37,314	16.54	34.51	23.18	25.74	0.04	0.58
Male	39,217	18.91	33.62	21.53	25.91	0.03	0.61
Not Exclusively Male or Female	87	25.29	37.93	20.69	16.09	N/A	2.30
Ethnic Group							
American Indian/Alaskan Native	934	5.57	20.13	25.70	48.50	0.11	N/A
Asian	6,799	34.78	36.39	15.44	13.35	0.03	1.37
Black/African American	3,688	5.78	26.08	26.27	41.81	0.05	0.62
Hispanic/Latino	19,665	6.36	25.49	27.86	40.24	0.05	1.00
White	37,585	22.18	39.24	20.41	18.14	0.02	0.32
Pacific Islander	1,017	3.34	19.47	24.78	52.41	N/A	2.36
Two or More Races	6,927	19.55	36.26	20.89	23.26	0.04	0.06
Unknown/Missing	3	33.33	N/A	N/A	66.67	N/A	N/A
Program							
Limited English	9,648	1.13	9.67	24.83	64.29	0.06	4.77
Non-Limited English	66,970	20.16	37.57	21.97	20.27	0.03	N/A
Non-Special Education	65,586	19.63	36.75	22.61	20.97	0.04	0.70
Special Education	11,032	6.64	18.03	20.67	54.63	0.03	0.03
Low Income	37,686	6.92	27.29	27.01	38.73	0.05	0.77
Non-Low Income	38,932	28.25	40.61	17.80	13.32	0.02	0.44
Migrant	1,662	1.62	16.61	27.38	54.27	0.12	1.32

Table E-16: Grade 8 WCAS Percentage Meeting Standards for Total and Student Groups

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	79,473	21.92	20.50	22.93	34.55	0.10	0.56
Gender							
Female	38,398	19.65	20.53	24.34	35.38	0.09	0.53
Male	40,779	24.01	20.42	21.61	33.84	0.12	0.59
Not Exclusively Male or Female	296	27.70	28.72	20.27	23.31	N/A	1.69
Ethnic Group							
American Indian/Alaskan Native	975	5.74	13.74	24.72	55.69	0.10	0.10
Asian	7,009	39.72	23.31	19.09	17.81	0.07	1.28
Black/African American	3,489	7.62	14.19	24.02	53.88	0.29	0.95
Hispanic/Latino	20,917	8.78	14.81	24.11	52.15	0.14	0.98
White	39,025	27.62	23.82	22.89	25.58	0.09	0.23
Pacific Islander	1,068	5.62	8.52	21.72	64.04	0.09	2.62
Two or More Races	6,984	23.48	22.14	22.79	31.54	0.04	0.03
Unknown/Missing	6	16.67	N/A	33.33	50.00	N/A	N/A
Program							
Limited English	7,322	1.35	3.47	13.41	81.48	0.29	6.13
Non-Limited English	72,151	24.01	22.23	23.89	29.78	0.09	N/A
Non-Special Education	69,837	24.20	22.20	23.78	29.72	0.09	0.64
Special Education	9,636	5.37	8.22	16.75	69.49	0.18	0.01
Low Income	38,132	9.92	15.61	24.22	50.08	0.16	0.67
Non-Low Income	41,341	32.99	25.01	21.73	20.22	0.05	0.47
Migrant	1,866	3.59	10.13	22.94	63.08	0.27	1.07

Table E-17: Grade 11 WCAS Percentage Meeting Standards for Total and Student Groups

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	57,366	11.78	42.15	22.97	22.98	0.12	0.40
Gender							
Female	27,080	9.50	43.92	24.26	22.24	0.08	0.37
Male	30,063	13.82	40.54	21.82	23.66	0.16	0.41
Not Exclusively Male or Female	223	13.00	44.84	21.08	20.63	0.45	0.90
Ethnic Group							
American Indian/Alaskan Native	695	2.73	28.63	29.64	38.99	N/A	N/A
Asian	4,858	24.58	46.15	16.53	12.72	0.02	0.76
Black/African American	2,399	3.00	29.18	26.39	41.23	0.21	0.83
Hispanic/Latino	14,772	3.70	32.58	28.40	35.18	0.14	0.81
White	29,508	14.69	47.74	21.03	16.45	0.09	0.12
Pacific Islander	720	1.94	22.92	26.25	48.61	0.28	2.22
Two or More Races	4,412	13.06	44.72	21.40	20.49	0.34	N/A
Unknown/Missing	2	N/A	100.00	N/A	N/A	N/A	N/A
Program							
Limited English	5,092	0.35	9.29	23.86	66.22	0.27	4.46
Non-Limited English	52,274	12.89	45.36	22.88	18.77	0.11	N/A
Non-Special Education	51,331	12.82	45.30	22.68	19.10	0.10	0.44
Special Education	6,035	2.90	15.41	25.43	55.94	0.31	0.02
Low Income	25,294	4.38	33.85	27.50	34.12	0.14	0.47
Non-Low Income	32,072	17.60	48.70	19.39	14.19	0.11	0.33
Migrant	1,395	1.15	24.16	28.75	45.88	0.07	1.08

APPENDIX F

SCALE SCORE SUMMARY FOR GRADUATION

Table F-1: ELA for Graduation Percentage Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	14,172	2,533.56	111.52
Gender			
Female	6,254	2,547.67	107.28
Male	7,742	2,521.88	113.37
Not Exclusively Male or Female	176	2,546.30	118.76
Ethnic Group			
American Indian/Alaskan Native	357	2,498.59	104.42
Asian	782	2,543.66	115.30
Black/African American	870	2,501.14	105.26
Hispanic/Latino	4,984	2,512.56	105.29
White	5,802	2,556.52	112.70
Pacific Islander	356	2,502.07	99.12
Two or More Races	1,002	2,549.50	110.99
Unknown/Missing	19	2,509.90	131.32
Program			
Limited English	2,922	2,463.76	91.02
Non-Limited English	11,250	2,551.69	109.19
Non-Special Education	11,038	2,551.74	108.55
Special Education	3,134	2,469.52	97.35
Low Income	8,880	2,516.28	107.44
Non-Low Income	5,292	2,562.55	112.24
Migrant	525	2,489.11	100.26

Table F-2: Mathematics for Graduation Percentage Scale Score Means & Standard Deviations (SD) for Total and Student Groups, Smarter Balanced

	Number Tested	Mean	SD
Total	25,380	2,505.62	107.61
Gender			
Female	12,381	2,509.13	100.62
Male	12,784	2,502.43	113.87
Not Exclusively Male or Female	215	2,493.15	107.44
Ethnic Group			
American Indian/Alaskan Native	509	2,458.05	102.71
Asian	1,383	2,536.66	103.00
Black/African American	1,318	2,471.48	102.58
Hispanic/Latino	7,958	2,481.83	103.48
White	11,875	2,524.89	106.57
Pacific Islander	472	2,463.56	103.67
Two or More Races	1,829	2,509.63	106.05
Unknown/Missing	36	2,483.72	119.53
Program			
Limited English	3,331	2,437.69	100.87
Non-Limited English	22,049	2,515.88	104.83
Non-Special Education	21,360	2,520.46	101.57
Special Education	4,020	2,426.78	104.41
Low Income	13,887	2,484.01	105.28
Non-Low Income	11,493	2,531.73	104.55
Migrant	740	2,464.52	102.59

APPENDIX G

PERCENTAGE OF STUDENTS BY ACHIEVEMENT LEVEL FOR GRADUATION

Table G-1: ELA for Graduation Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	15,772	8.86	23.52	26.36	31.12	10.14	N/A
Gender							
Female	6,954	10.02	26.29	26.96	26.66	10.07	N/A
Male	8,627	7.81	21.33	25.85	34.76	10.26	N/A
Not Exclusively Male or Female	191	14.14	21.99	27.23	28.80	7.85	N/A
Ethnic Group							
American Indian/Alaskan Native	419	3.58	16.47	24.58	40.57	14.80	N/A
Asian	865	10.75	26.82	23.70	29.13	9.60	N/A
Black/African American	994	3.32	18.61	24.55	41.05	12.47	N/A
Hispanic/Latino	5,565	5.12	19.59	28.32	36.53	10.44	N/A
White	6,378	12.95	27.81	25.40	24.80	9.03	N/A
Pacific Islander	400	3.75	16.00	29.25	40.00	11.00	N/A
Two or More Races	1,125	11.38	26.04	25.51	26.13	10.93	N/A
Unknown/Missing	26	11.54	11.54	19.23	34.62	26.92	N/A
Program							
Limited English	3,221	0.71	8.57	26.64	54.80	9.28	N/A
Non-Limited English	12,551	10.96	27.36	26.28	25.04	10.37	N/A
Non-Special Education	12,270	10.77	27.52	26.69	24.98	10.04	N/A
Special Education	3,502	2.17	9.51	25.19	52.63	10.51	N/A
Low Income	9,932	5.89	20.52	27.31	35.69	10.59	N/A
Non-Low Income	5,840	13.92	28.63	24.74	23.34	9.38	N/A
Migrant	577	1.39	18.20	25.65	45.75	9.01	N/A

Table G-2: Mathematics for Graduation Percentage Meeting Standards for Total and Student Groups, Smarter Balanced

Group	Number of Students	Tested		Not Tested			
		Meets Standard		Does Not Meet Standard		Percentage No Score	Percentage Exempt
		Percentage Level 4	Percentage Level 3	Percentage Level 2	Percentage Level 1		
Total	27,203	2.25	11.48	25.76	53.80	6.70	N/A
Gender							
Female	13,233	1.71	11.14	27.70	53.01	6.44	N/A
Male	13,738	2.79	11.84	23.92	54.51	6.94	N/A
Not Exclusively Male or Female	232	1.72	9.91	24.14	56.90	7.33	N/A
Ethnic Group							
American Indian/Alaskan Native	563	0.18	4.97	16.52	68.74	9.59	N/A
Asian	1,482	3.98	17.34	28.88	43.12	6.68	N/A
Black/African American	1,423	0.91	5.76	19.75	66.20	7.38	N/A
Hispanic/Latino	8,572	0.96	7.14	21.96	62.79	7.16	N/A
White	12,633	3.26	14.86	29.21	46.67	6.00	N/A
Pacific Islander	518	0.58	5.60	16.02	68.92	8.88	N/A
Two or More Races	1,971	2.13	11.97	27.45	51.24	7.20	N/A
Unknown/Missing	41	2.44	7.32	24.39	53.66	12.20	N/A
Program							
Limited English	3,602	0.44	2.94	10.11	78.98	7.52	N/A
Non-Limited English	23,601	2.53	12.79	28.15	49.96	6.58	N/A
Non-Special Education	22,821	2.54	13.22	28.76	49.07	6.40	N/A
Special Education	4,382	0.78	2.42	10.13	78.41	8.26	N/A
Low Income	14,929	1.16	7.78	22.16	61.91	6.98	N/A
Non-Low Income	12,274	3.58	15.99	30.14	43.93	6.36	N/A
Migrant	782	0.51	5.24	18.54	70.33	5.37	N/A

APPENDIX H

HISTORICAL DATA

Table H-1: Percentage Proficient, Smarter Balanced, 2015–22

Subject	Year	Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	Gr 11/HS
ELA	2015	53%	55%	58%	54%	58%	58%	65%*
	2016	55%	57%	61%	57%	59%	61%	65%
	2017	53%	56%	59%	56%	61%	59%	64%
	2018	56%	58%	60%	57%	61%	60%	65%**
	2019	56%	58%	61%	58%	62%	59%	66%
	2020***	-	-	-	-	-	-	-
	2021***	-	-	-	-	-	-	-
	2022****	48%	49%	52%	45%	53%	51%	63%
Mathematics	2015	57%	54%	49%	46%	49%	47%	29%*
	2016	59%	56%	49%	48%	50%	49%	41%
	2017	58%	54%	49%	48%	51%	49%	43%
	2018	58%	54%	49%	49%	50%	49%	38%**
	2019	58%	54%	49%	47%	50%	47%	33%
	2020***	-	-	-	-	-	-	-
	2021***	-	-	-	-	-	-	-
	2022****	50%	46%	38%	34%	36%	34%	28%

* In 2015, the high school census year for state and federal accountability was grade 11. The WCAP allowed students in grade 10 to test in ELA and mathematics toward state graduation requirements. Should those grade 10 students earn a Level 3 or Level 4 in a subject, they would not be expected to return and test in grade 11. The grade 11 testing population is comprised entirely of the students who did not earn a Level 3 or Level 4 as grade 10 students in the previous school year. Therefore, pass rates are substantially lower than would be observed were the entire cohort to test during a single, census administration in grade 11.

** Starting in 2018, the census year for state and federal accountability changed from grade 11 to grade 10.

*** Due to disruptions caused by Covid-19, there was no spring testing in 2020 or 2021.

**** The test blueprint used in spring 2022 was the Smarter Balanced adjusted blueprint which is different than blueprints used from 2015–19

Table H-2: Percentage Proficient, WCAS Grades 5, 8, and 11, 2018–22

	Year	Percentage
Grade 5	2018	56%
	2019	54%
	2020*	-
	2021*	-
	2022	52%
Grade 8	2018	55%
	2019	53%
	2020*	-
	2021*	-
	2022	43%
Grade 11	2018	46%
	2019	50%
	2020*	-
	2021*	-
	2022	55%

* Due to disruptions caused by Covid-19, there was no spring testing in 2020 or 2021.