California Department of Education Assessment Development & Administration Division

California Assessment of Student Performance and Progress Smarter Balanced 2019–2020 Technical Report

Submitted June 2, 2021

By ETS

Contract No. CN150012

Table of Contents

List of Tables

List of Figures

Acronyms and Initialisms Used in the *CAASPP Smarter Balanced Technical Report*

Term	Definition
2PL	two-parameter logistic
AERA	American Educational Research Association
AI	artificial intelligence
AIR	American Institutes for Research
APA	American Psychological Association
ASL	American Sign Language
CAA	California Alternate Assessment
CAASPP	California Assessment of Student Performance and Progress
CAC	California Assessment Conference
CAI	Cambium Assessment, Inc.
CALPADS	California Longitudinal Pupil Achievement Data System
CalTAC	California Technical Assistance Center
CAT	computer adaptive test
CCR	*California Code of Regulations*
CCSS	Common Core State Standards
CDE	California Department of Education
CDS	county/district/school
CERS	California Educator Reporting System
CI	confidence interval
COVID-19	novel coronavirus disease 2019
CR	constructed response
CRESST	Center for Research on Evaluation, Standards, & Student Testing
CSEM	conditional standard error of measurement
CSU	California State University
DEI	Data Entry Interface
DIF	differential item functioning
EAP	Early Assessment Program
EC	*Education Code*
EL	English learner
ELA	English language arts/literacy
eSKM	Enterprise Score Key Management
GPCM	generalized partial credit model
HOSS	highest obtainable scale score
HOT	highest obtainable theta
ICC	intraclass correlation
IEP	individualized education program
IRT	item response theory
JAWS®	Job Access With Speech
ISAAP	Individual Student Assessment Accessibility Profile

Table of Acronyms and Initialisms *(continuation)*

Term	Definition
LEA	local educational agency
LOSS	lowest obtainable scale score
LOT	lowest obtainable theta
MI	Measurement Incorporated
MLE	maximum likelihood estimation
NAEP	National Assessment of Educational Progress
NCME	National Council on Measurement in Education
NEL	not tested English learner
NTE	not tested medical emergency
ORS	Online Reporting System
OTI	Office of Testing Integrity
PAR	Psychometric Analysis & Research
PGE	parent/guardian exemption
PIN	problem item notification
PISA	Program for International Student Assessment
PT	performance task
QWK	quadratic weighted kappa
SBE	State Board of Education
SCOE	Sacramento County Office of Education
SEM	standard error of measurement
SFTP	secure file transfer protocol
SGID	School and Grade Identification sheet
SS	scale score
SSID	Statewide Student Identifier
SSR	Student Score Report
STAIRS	Security and Test Administration Incident Reporting System
TCC	test characteristic curve
TDS	test delivery system
TIF	test information function
TIPS	Technology and Information Processing Services
TOMS	Test Operations Management System
USC	United States Code
WER	writing extended response

Chapter 1: Introduction

1.1. Background

In October 2013, Assembly Bill 484 established the California Assessment of Student Performance and Progress (CAASPP) as the new student assessment system that replaced the Standardized Testing and Reporting program. The primary purpose of the CAASPP System of assessments is to assist teachers, administrators, and students and their parents/ guardians by promoting high-quality teaching and learning through the use of a variety of item types and assessment approaches. These tests provide the foundation for the state's school accountability system.

The Smarter Balanced Summative Assessments for English language arts/literacy (ELA) and mathematics were administered during the 2019–2020 CAASPP administration as a result of California's participation in the Smarter Balanced Assessment Consortium. This technical report describes the results of that administration.

In 2019–2020, the CAASPP System comprised the following assessments:

- Smarter Balanced assessments and tools:
  - Summative Assessments—Online assessments for ELA and mathematics in grades three through eight and grade eleven
  - Interim Assessments—Optional resources developed for grades three through eight and grade eleven designed to inform and promote teaching and learning by providing information that can be used to monitor student progress toward mastery of the Common Core State Standards (CCSS) that may be administered to students at any grade level
  - Digital Library (now Tools for Teachers)—Professional development materials and instructional resources designed to help teachers use formative assessment processes for improved teaching and learning in all grades
- California Alternate Assessments (CAAs) for ELA and mathematics in grades three through eight and grade eleven
- Science assessments in grades five and eight and high school (grade ten, eleven, or twelve; these are the California Science Test and the CAA for Science)
- The California Spanish Assessment, optional for eligible students in grades three through eight and high school and designed to measure a student's Spanish competency in reading, writing mechanics, and listening, as well as to serve as a high school measure suitable to be used in part for the California Seal of Biliteracy

The CAASPP Smarter Balanced tests are presented as online assessments. Braille, large-print, and standard paper–pencil versions of the Smarter Balanced assessments are made available to individual students within a local educational agency (LEA) whose need to take a paper–pencil assessment is documented in a student's individualized education program (IEP) or Section 504 plan. Students who repeatedly experience difficulty accessing the online assessments because of technical issues that cannot be resolved within two weeks may be allowed to take a standard paper–pencil test, upon approval by the California Department of Education (CDE). The paper–pencil versions are fixed forms (i.e., a test where students are given a fixed set of questions irrespective of the student's responses or ability) that also include the components of the online assessment such as constructed-response (CR) items and performance tasks (PTs).

More background information about the CAASPP System can be found on the CAASPP Description – *CalEdFacts* web page at https://www.cde.ca.gov/ta/tg/ai/cefcaaspp.asp.

1.2. Test Purposes

The purposes of the Smarter Balanced assessment system are to provide teachers with information and the tools they need to improve teaching and learning and to prepare students for college and career readiness. The Smarter Balanced Summative Assessments, which are aligned with the California CCSS for ELA and mathematics, form one component of the Smarter Balanced assessment system. The summative assessments are comprehensive, end-of-year tests of grade-level learning that measure students' progress toward college and career readiness.

1.3. Test Content

Smarter Balanced Summative Assessments are composed of two required components: a computer adaptive test (CAT) and a PT. A student's final scale score is calculated by combining the student's responses to both components.

1.3.1. Computer Adaptive Test

The computer-adaptive portion of the test is designed to present items of difficulty to match the ability of each student, as indicated by the responses the student provided to previous test items. By adapting to the student's ability as the assessment is being taken, the CAT presents an individually tailored set of questions that is appropriate for each student. As a result, it provides more accurate scores for all students across the full range of the achievement continuum. Compared with a fixed-form assessment—that is, a test where all students are given the same questions, regardless of their responses or ability—a CAT requires fewer questions to obtain an equally precise estimate of a student's ability.

At the beginning of the test, the test delivery system (TDS) assumes that the student is of average ability and presents an item that is appropriate for an average student. During the test, if a student gives an incorrect answer, the TDS will follow up with an easier question or a group of questions; if the student answers correctly, the next question or next group of questions will be slightly more difficult. As the adaptive test continued, the next question or group of questions was based on the student's answers to all previous questions—the student's responses to the current item and previous items determined the pathway to a subsequent item.

Because the answers on items used to estimate the student's ability were machine-scored, the student's performance on the items already administered was known immediately, and the successive items were selected to adapt to the estimated ability of the student. The CAT selected questions based on a student's responses, scored the responses, and revised its estimate of the student's ability. This process continued until the test content outlined in the test's blueprint was covered.

The CAT required a large pool of test questions statistically calibrated on a common scale to cover the ability range. For the Smarter Balanced Online Summative Assessments, the test question statistics were obtained mainly from the spring 2013–2014 field test. Each year, new items are field-tested and added to the Smarter Balanced item pools.

1.3.2. Performance Tasks

The PT is a nonadaptive portion of a Smarter Balanced content-area assessment designed to provide students with an opportunity to demonstrate their ability to apply knowledge and higher-order thinking skills to explore and analyze a complex, real-world scenario.

Some PT responses are machine-scored, others are human scored. Scores are later combined with CAT results for the student’s final score.

1.4. Intended Population

At each grade level, the Smarter Balanced Summative Assessments for ELA and mathematics were administered to between 435,000 and 473,000 students during the 2018–2019 administration. However, on March 18, 2020, Governor Gavin Newsom signed an order suspending the CAASPP for all students in California (Office of Governor Gavin Newsom, 2020). Because of the suspension of testing, a vast majority of the intended population did not have the opportunity to take the Smarter Balanced assessments during the 2019–2020 administration.

All students enrolled in grades three through eight and grade eleven are required to take part in the Smarter Balanced Summative Assessments unless they are eligible to participate in the alternate assessments (*California Code of Regulations*, Title 5 [5 *CCR*] Education, Division 1, Chapter 2, Subchapter 3.75, Article 1*,* Section 851.5). English learner (EL) students who are in their first 12 months of attending school in the United States are exempt from taking the ELA portion of the assessment. EL students are defined as follows:

> “English learner students are those students for whom there is a report of a primary language other than English on the state-approved Home Language Survey **and** who, on the basis of the state approved oral language (grades kindergarten through grade twelve) assessment procedures and literacy (grades three through twelve only), have been determined to lack the clearly defined English language skills of listening comprehension, speaking, reading, and writing necessary to succeed in the school’s regular instructional programs.”[1]

EL students within their first 12 months of enrollment in a U.S. school who choose to participate in taking the ELA assessment are included in the calculation of the percent of students tested, but their scores are excluded from all aggregate calculations so long as the student record has an include indicator of “E” and the condition code “NEL” (not tested English learner). The condition code “NEL” is automatically assigned based on the following criteria when the student’s English language acquisition status is “EL,” the student has been enrolled in a U.S. school for less than twelve months, and one of the following criteria is true:

- Was enrolled and did not test
- *or* Student tested but did not meet attemptedness/completion status
- *or* Test is a force-complete

For students with significant cognitive disabilities, the decision to administer the Smarter Balanced Summative Assessments or the CAAs is made by their IEP team. Parents/

“English Learner (EL) Students (Formerly Known as Limited-English-Proficient or LEP),” from the CDE Glossary of Terms web page at https://bit.ly/33gqx38

Guardians may submit a written request to have their child exempted from taking any or all parts of the Smarter Balanced Summative Assessments or CAAs.

Students whose parents/guardians submit a written request are exempted from taking the tests (*Education Code [EC]* Section 60615). Additionally, students who were not tested due to a medical emergency are also be exempt.

1.5. Intended Use and Purpose of Test Scores

The results of tests within the CAASPP System are used for two primary purposes as described in *EC* sections 60602.5(a) and (a)(4). (Excerpted from the *EC* Section 60602 web page at https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=EDC&division=4.&title=2.&part=33.&chapter=5.&article=1 [outside source].)

> "60602.5(a) It is the intent of the Legislature in enacting this chapter to provide a system of assessments of pupils that has the primary purposes of assisting teachers, administrators, and pupils and their parents; improving teaching and learning; and promoting high-quality teaching and learning using a variety of assessment approaches and item types. The assessments, where applicable and valid, will produce scores that can be aggregated and disaggregated for the purpose of holding schools and local educational agencies accountable for the achievement of all their pupils in learning the California academic content standards."
>
> "60602.5(a)(4) Provide information to pupils, parents and guardians, teachers, schools, and local educational agencies on a timely basis so that the information can be used to further the development of the pupil and to improve the educational program."

In other words, results for tests within the CAASPP System are used for two primary purposes:

1. To communicate students' progress in achieving the state's academic standards to students, parents and guardians, and teachers
2. To inform decisions that teachers and administrators make about improving the educational program

Sections 60602.5(c) and (d) provide additional information regarding use and purpose of test scores for the system of assessments:

> "60602.5(c) It is the intent of the Legislature that parents, classroom teachers, other educators, pupil representatives, institutions of higher education, business community members, and the public be involved, in an active and ongoing basis, in the design and implementation of the statewide pupil assessment system and the development of assessment instruments."
>
> "60602.5(d) It is the intent of the Legislature, insofar as is practically feasible and following the completion of annual testing, that the content, test structure, and test items in the assessments that are part of the statewide pupil assessment system become open and transparent to teachers, parents, and pupils, to assist stakeholders in working together to demonstrate improvement in pupil academic achievement. A planned change in annual test content, format, or design should be made available to educators and the public well before the beginning of the school year in which the change will be implemented."

1.6. Testing Window and Impact of the Novel Coronavirus Disease 2019 Pandemic

The Smarter Balanced Summative Assessments for grades three through eight and grade eleven are administered within a testing window pursuant to 5 *CCR,* sections 855(a)(1), 855(a)(2), 855(b), and 855(c). For the 2019–2020 CAASPP administration, the state testing window started on January 14 and was intended to end on July 15, 2020. However, most of the schools in California halted in-person instruction after March 13, 2020. Then, on March 18, 2020, Governor Gavin Newsom signed an order suspending the CAASPP for all students in California (Office of Governor Gavin Newsom, 2020).

Table 1.1 lists the total number of students who took the Smarter Balanced assessments before the testing window was closed in 2019–2020. In the table, a force complete test record indicates that a student was missing one or more testing opportunities at the end of the testing window. Force completions happened after the LEA testing window closed and were preceded by test expirations of incomplete tests from the TDS. An expired test opportunity occurred when a student started, but did not complete, a test opportunity from the TDS before the set expiration period. This might have occurred during an LEA's administration window or at the end of the state window, depending on the expiration rules.

Table 1.1 Total Number of Students Taking the Tests in the 2019–2020 Administration

Subject	Grade	Registered	Total Started	Force Complete	Total Expired	Total Submitted
ELA	Grade 3	453,327	229	84	26	119
ELA	Grade 4	453,661	175	32	26	117
ELA	Grade 5	445,509	369	205	22	142
ELA	Grade 6	462,745	624	497	16	111
ELA	Grade 7	468,415	1,108	611	23	474
ELA	Grade 8	484,703	946	738	10	198
ELA	Grade 11	471,751	23,342	9,028	1,595	12,719
Mathematics	Grade 3	453,327	126	24	3	99
Mathematics	Grade 4	453,661	96	25	5	66
Mathematics	Grade 5	445,509	164	54	9	101
Mathematics	Grade 6	462,745	164	116	2	46
Mathematics	Grade 7	468,415	144	77	5	62
Mathematics	Grade 8	484,703	769	199	54	516
Mathematics	Grade 11	471,751	14,377	3,700	667	10,010

Because of a lack of data, group score reports based on student demographic information were not reported. The official aggregations of the Smarter Balanced test results at the school or LEA level were not available; refer to chapter 7 for detailed information.

In addition, the following psychometric analyses could not be conducted and reported:

- Item exposure analyses
- Analysis on time spent on items and the test
- Reliability analysis, including overall test score reliability and student group reliability

- Interrater reliability for CR items, including rating agreement and reliability indexes for both human scoring and artificial intelligence scoring
- Consistency and accuracy of the performance level classifications
- Correlation analysis
- Claim scoring analysis
- Student group analyses, including all demographic groups such as gender, ethnicity, socioeconomic status, English proficiency level, and participation in special education programs

Chapter 8 provides details on the procedures of those analyses.

1.7. Significant CAASPP Developments in 2019–2020

1.7.1. Updated Accessibility Resources

The following changes were made to the list of Smarter Balanced accessibility resources:

- Illustration glossaries for mathematics items were made available for selected construct-irrelevant terms. These glossaries are an embedded and non-embedded designated support.
- Somali and Hmong languages were made available as translation glossaries for mathematics items.
- Unified English Braille Technical was made available for the mathematics assessment.
- Burmese was added as an embedded translation glossary available as a designated support for the mathematics assessment.
- “Medical supports” replaced the name for the “medical device” as a non-embedded designated support for all assessments.

1.8. Groups and Organizations Involved with the CAASPP System

1.8.1. State Board of Education

The State Board of Education (SBE) is the state agency that establishes educational policy for kindergarten through grade twelve in the areas of standards, instructional materials, assessment, and accountability. The SBE adopts textbooks for kindergarten through grade eight, adopts regulations to implement legislation, and has the authority to grant waivers of the *EC*.

In addition to adopting the rules and regulations for itself, its appointees, and California’s public schools, the SBE is also the state educational agency responsible for overseeing California’s compliance with programs that meet the requirements of the federal Every Student Succeeds Act and the state’s Public School Accountability Act, which measure the academic performance and progress of schools on a variety of academic metrics (CDE, 2020c).

1.8.2. California Department of Education

The CDE oversees California's public school system, which is responsible for the education of more than 6,160,000 children and young adults in more than 10,500 schools.[2] California aims to provide a world-class education for all students, from early childhood to adulthood. The CDE serves the state by innovating and collaborating with educators, school staff, parents/guardians, and community partners which together, as a team, prepares students to live, work, and thrive in a highly connected world.

Within the CDE, it is the Instruction & Measurement branch that oversees programs promoting improved student achievement. Programs include oversight of statewide assessments and the collection and reporting of educational data (CDE, 2020b).

1.8.3. California Educators

A variety of California educators, including teachers and school administrators—who were selected based on their qualifications, experiences, demographics, and geographic locations—were invited to participate in various aspects of the assessment process prior to the current administration. This included defining the purpose and scope, test design, item development, and standard setting. In 2019–2020, California educators were involved in scoring of the Smarter Balanced Summative Assessment CR items.

1.8.4. Smarter Balanced Assessment Consortium

The Smarter Balanced Assessment Consortium is a public agency governed by a consortium of states, of which California is a member. The consortium created an online assessment system aligned to the CCSS that is comprised of year-end summative assessments and optional interim assessments (Smarter Balanced, n.d.). Smarter Balanced provided the collection of test items in the item bank as well as provided access to the Digital Library (now Tools for Teachers). The Digital Library is a tool offered by the Smarter Balanced Assessment Consortium. It provides an online collection of resources to help teachers improve classroom-based assessment practices.

1.8.5. Contractors

1.8.5.1. Primary Testing Contractor—ETS

The CDE and the SBE contract with ETS to administer and report the CAASPP Smarter Balanced assessments. As the primary testing contractor, ETS has overall responsibility for working with the CDE to implement and maintain an effective assessment system and coordinating ETS' work with its subcontractors. Activities conducted directly by ETS include, but are not limited to, the following:

- Providing management of the program activities
- Supporting and training counties, LEAs, and direct funded charter schools
- Providing tiered help desk support to LEAs
- Constructing, producing, and controlling the quality of test booklets and related test materials
- Hosting and maintaining a website with resources for LEA CAASPP coordinators

[2] Fingertip Facts on Education in California – *CalEdFacts* web page at https://www.cde.ca.gov/ds/ad/ceffingertipfacts.asp

- Developing, hosting, and providing support for the Test Operations Management System (TOMS)
- Processing student test assignments
- Processing orders and shipment of test materials
- Servicing all aspects of CR scoring for the CAASPP Smarter Balanced Summative Assessments
- Producing and distributing score reports
- Developing a score reporting website
- Completing all psychometric procedures

1.8.5.2. Subcontractor—Cambium Assessment, Inc.

ETS also monitors and manages the work of Cambium Assessment, Inc. (CAI; formerly American Institutes for Research [AIR]), subcontractor to ETS for the CAASPP System of online assessments. Activities conducted by CAI include

- providing the CAI proprietary TDS, including the Student Testing Interface, Test Administrator Interface, secure browser, and practice and training tests;
- hosting and providing support for its TDS and Online Reporting System (ORS)[3], a component of the overall CAASPP Assessment Delivery System;
- scoring machine-scorable items; and
- providing high-level technology help desk support to LEAs for technology issues directly related to the TDS.

1.8.5.3. Subcontractor—Measurement Incorporated

ETS monitors and manages the work of Measurement Incorporated (MI), a subcontractor to ETS for the CAASPP System. MI uses its AI scoring system to score some of the CR items for the CAASPP Smarter Balanced Online Summative Assessments.

1.9. Systems Overview and Functionality

1.9.1. Test Operations Management System

TOMS is the password-protected, web-based system used by LEAs to manage all aspects of CAASPP testing. TOMS serves various functions, including, but not limited to, the following:

- Managing test administration windows
- Assigning and managing CAASPP online user roles
- Managing student test assignments and accessibility resources
- Ordering test materials
- Viewing and downloading reports

[3] The ORS will be replaced with the California Educator Reporting System (CERS) starting in January 2021.

- Reporting security incidents
- Providing a platform for authorized user access to secure materials, such as CAA *Directions for Administration,* student data and results, CAASPP user information, and access to the CAASPP Security and Test Administration Incident Reporting System/Appeals process

TOMS receives student enrollment data and LEA and school hierarchy data from the California Longitudinal Pupil Achievement Data System (CALPADS) via a daily feed. CALPADS is "a longitudinal data system used to maintain individual-level data including student demographics, course data, discipline, assessments, staff assignments, and other data for state and federal reporting."[4] LEA staff involved in the administration of the CAASPP assessments—such as LEA CAASPP coordinators, CAASPP test site coordinators, test administrators, and test examiners—are assigned varying levels of access to TOMS. For example, only an LEA CAASPP coordinator is given permission to set up the LEA's test administration window; a test administrator cannot download student reports. A description of user roles is explained more extensively in the *2019–2020 CAASPP Online Test Administration Manual* (CDE, 2020a).

1.9.2. Test Delivery System

The TDS is the means by which the statewide online assessments are delivered to students. CAT items are selected in the TDS according to an adaptive algorithm (AIR, 2014). Components of the TDS include

- the Test Administrator Interface, the web browser–based application that allows test administrators to activate student tests and monitor student testing;
- the Student Testing Interface, on which students take the test using the secure browser; and
- the secure browser, the online application through which the Student Testing Interface may be accessed. The secure browser prevents students from accessing other applications during testing.

1.9.3. Practice and Training Tests

The practice and training tests are provided to LEAs to prepare students and LEA staff for administration of the summative assessment. These tests simulate the experience of the Smarter Balanced online assessments. Unlike the summative assessments, the practice and training tests do not assess standards, gauge student success on the operational test, or produce scores. Students may access them using a web browser, although accessing them through the secure browser permits them to take the tests using the text-to-speech embedded accommodation and to test assistive technology.

The purposes of the training tests are to allow students and administrators to quickly become familiar with the user interface and components of the TDS as well as with the process of starting and completing a testing session. The purpose of the practice tests is to allow students and administrators to experience a grade-level assessment, grade-specific items and difficulty levels, PTs, and the format and structure of an operational assessment.

[4] From the CDE California Longitudinal Pupil Achievement Data System (CALPADS) web page at https://www.cde.ca.gov/ds/sp/cl/

1.9.4. Online Reporting System and California Educator Reporting System

Previously, there were two California online reporting systems: the ORS and the California Educator Reporting System (CERS). As of January 2021, the CERS is the single resource where LEA staff access student results from the summative and interim CAASPP assessments as well as results from the English Language Proficiency Assessments for California.

The ORS is the system used by LEAs to view preliminary student results from the CAASPP assessments. The primary purposes of the ORS are for LEAs to access completion data to determine which students need to complete testing or start testing, and for LEAs to access preliminary score reports that can provide claim-related data for schools within the LEA. Results in the ORS are preliminary and may not be used for accountability purposes.

The CERS allows educators to view their students' assessment results using grouping and other new features. For example, educators can create customized groups from assigned student groups; for interim assessments, specific assessment items can be viewed with student responses; and a distractor analysis feature can be used to identify student strengths and needs.

1.9.5. Constructed-Response Scoring Systems for ETS and Measurement Incorporated

CR items from the TDS were routed to either ETS' or MI's CR scoring systems based on the division of work between ETS and MI. CR items were scored by certified raters. A small percentage of CR items were deemed appropriate to be scored by the AI system and were routed for both AI scoring and human scoring for the purpose of producing agreement samples. More information regarding scoring of CR items is available in *Chapter 7: Scoring and Reporting*.

Targeted efforts were made to hire California educators for human scoring opportunities. Hired raters were provided in-depth training and certified before starting the human scoring process. Human raters were organized under a scoring leader and provided Smarter Balanced scoring materials such as anchor sets, scoring rubrics, validity samples, qualifying sets, and condition codes for unscorable responses within the interface. The quality control processes for CR scoring are explained further in *Chapter 9: Quality Control Procedures*.

1.10. Overview of the Technical Report

This technical report addresses the characteristics of the CAASPP Smarter Balanced Summative Assessment administered in spring 2020. The technical report contains nine additional chapters as follows:

- Chapter 2 presents an overview of the processes involved in a testing cycle for a Smarter Balanced Summative Assessment. This includes test administration, generation of test scores, and dissemination of score reports. It also includes information about the assignment of designated supports and accommodations.
- Chapter 3 discusses the procedures followed during the development of Smarter Balanced items to help ensure valid interpretation of test scores.
- Chapter 4 discusses the content and psychometric criteria that guide the construction of the Smarter Balanced Summative Assessments.

- Chapter 5 details the processes involved in the administration of the 2019–2020 Smarter Balanced Summative Assessments. It also describes the procedures followed by ETS to maintain test security throughout the test administration process.
- Chapter 6 discusses the standard setting process outlined by Smarter Balanced.
- Chapter 7 summarizes the types of scores and score reports that are produced at the end of each administration of the Smarter Balanced Summative Assessments.
- Chapter 8 summarizes the results of the analyses performed on the data resulting from the 2019–2020 administration. These include item response theory parameters from Smarter Balanced item pools; omission, expiration, and completion analyses; overall testing summaries as the means and standard deviations of scale scores and theta values; and the percentages of students at each performance level for each test. Other psychometric analyses, such as item exposure analyses, reliability analysis, consistency and accuracy of the performance level classifications, correlation analysis, student group reliability, interrater reliability for the human-scoring items, the agreement between human scoring and AI scoring, claim scoring analysis, and student group analyses were not conducted and so are not reported because of the constraints of sample size.
- Chapter 9 highlights the quality control processes used at various stages of administration of the Smarter Balanced assessments.
- Chapter 10 discusses the various procedures used to gather information to improve the Smarter Balanced assessments as well as strategies to implement possible improvements.

References

American Institutes for Research. (2014). *Smarter Balanced adaptive item selection algorithm design report*. Washington, DC: American Institutes for Research. http://www.smarterapp.org/documents/AdaptiveAlgorithm-Preview-v3.pdf

California Code of Regulations, Title 5, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, Section 851.5. (n.d.) ~~https://govt.westlaw.com/calregs/Document/I7907D7786C424BEFAC22247312385DFC?viewType=FullText&originationContext=documenttoc&transitionType=CategoryPageItem&contextData=(sc.Default)~~

California Department of Education. (2020a). *CAASPP online test administration manual, 2019–2020 test administration.* Sacramento, CA: California Department of Education. https://bit.ly/3eXYBXc

California Department of Education. (2020b, August). *Organization.* https://www.cde.ca.gov/re/di/or/

California Department of Education. (2020c, October). *State Board of Education responsibilities.* https://www.cde.ca.gov/be/ms/po/sberesponsibilities.asp

Office of Governor Gavin Newsom. (2020). *Governor Newsom issues executive order to suspend standardized testing for students in response to COVID-19 outbreak* [Press release]. https://www.gov.ca.gov/2020/03/18/governor-newsom-issues-executive-order-to-suspend-standardized-testing-for-students-in-response-to-covid-19-outbreak/

Smarter Balanced Assessment Consortium. (n.d.). *Smarter assessments.* https://smarterbalanced.org/our-system/assessments/

Chapter 2: Overview of CAASPP Smarter Balanced Processes

This chapter provides an overview of the processes involved in a testing cycle for a Smarter Balanced Summative Assessment, emphasizing test administration, psychometric analyses, generation of test scores, and dissemination of score reports. The Smarter Balanced Assessment Consortium developed the summative assessments and ETS administered, scored, and reported the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced assessments.

2.1. Item Development

All items in the Smarter Balanced operational item bank for the first year of testing were developed and revised during the pilot and field test periods. Thereafter, Smarter Balanced items are developed dynamically. New items are developed and field-tested by being embedded in the operational tests. Each year, some new items are added into the Smarter Balanced operational item banks, and some poorly performing items are removed from the item banks. During item development, item and performance task (PT) specifications provide guidance on how to translate the Smarter Balanced content specifications into actual assessment items (Smarter Balanced, 2016, 2017, and 2018b). Guidelines for bias and sensitivity, accessibility and accommodations, and style help item developers and reviewers ensure consistency and fairness across the item development process. These specifications and guidelines from Smarter Balanced were reviewed by member states, local educational agencies (LEAs), higher education professionals, and other stakeholders (Smarter Balanced, 2016). For more information regarding the item response theory (IRT) methodology used by Smarter Balanced to form the basis for new item development, test equating, and computer adaptive testing, refer to chapter 9 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016).

2.1.1. Item Format

The Smarter Balanced assessments include the following online item formats:

- Selected response
- Constructed response
- Technology enhanced

Formats for these item types are described in more detail in subsection *7.1.3 Types of Item Responses*.

2.1.2. Item Specifications

The item specifications describe the characteristics of the items that should be written to measure each content standard. Items of the same type should consistently measure the content standards in the same way. The *Smarter Balanced Item and Task Specifications* were given to item developers to help ensure that the tests are measuring the intended constructs without influence from extraneous factors. These documents contain item specification tables and provide item writers with definitions of the constructs that are intended to support the claims of measurement and clear direction regarding the types of evidence needed for students to demonstrate their knowledge and skills (Smarter Balanced, 2016, 2017, and 2018b); note that because these specifications were reorganized following the initial development, their publication dates were updated.

2.2. Test Assembly

2.2.1. Test Length

2.2.1.1. Operational Testing

The CAASPP online summative assessments for English language arts/literacy (ELA) and mathematics are composed of two portions: the computerized-adaptive test (CAT) and the PT. The number of PT items that a student is administered depends on the particular PT a student is assigned randomly at the student level. The number of CAT items encountered in an individual testing session may vary from student to student. The length of the CAT portion is determined by the termination rule of the CAT engine, which includes the following conditions:

1. Administer at least a specified minimum number of items in each reporting category and overall
2. Achieve a target level of precision on the overall test score
3. Achieve a target level of precision on all reporting categories

The termination rule of CAASPP assessments is discussed in more detail in the *Smarter Balanced Adaptive Item Selection Algorithm Design Report* (American Institutes for Research [AIR], 2014).

2.2.1.2. Field Testing

Field test PTs have been embedded into the Smarter Balanced operational tests since the 2016–2017 administration. Students who were assigned an embedded field test PT were not assigned an operational PT. Instead, they were assigned a CAT version with additional items for the purpose of reporting claim results. For ELA, these students received three additional items. For mathematics, these students received two additional items. Refer to *Enhanced Computer Adaptive Testing (CAT) Blueprints for Students Participating in the 2019–2020 Smarter Balanced Embedded Field Test of Performance Tasks (PTs)* in *Appendix 2.A: Smarter Balanced Blueprints* for the number of CAT items with embedded field test PTs in the blueprints (Smarter Balanced, 2018a).

2.2.2. Test Blueprints

2.2.2.1. Operational Items

Blueprints represent a set of constraints and specifications to which each test form must conform. Each grade band—grades three through five, grades six through eight, and grade eleven—of the Smarter Balanced assessments includes a separate blueprint (appendix 2.A) with criteria including, but not limited to,

- whether the test is adaptive or fixed form,
- termination conditions for the segment,
- content constraints such as minimum or maximum number of items administered in each content category, and
- nonnested content constraints such as priority weights for a group of items.

2.2.2.2. Field Test Items

Because there were embedded field test PTs administered in 2019–2020, the blueprints for the field test are provided specifically, along with the blueprints for the operational tests provided in appendix 2.A. The PTs that are field-tested do not contribute to score reporting. Instead, the additional operational CAT items as shown in the field test blueprints are counted into score reporting.

2.2.3. Item Selection

In the CAT portion of each assessment, items are presented to a student according to the adaptive algorithm mapped onto the test blueprint (AIR, 2015). Use of the adaptive algorithm in 2015–2016 testing and simulation studies in the following years are discussed in the report *Smarter Balanced Summative Assessments Testing Procedures for Adaptive Item-Selection Algorithm* (AIR, 2015; Smarter Balanced, 2019).

For more information regarding test length, refer to *Chapter 5: Test Administration*; the test blueprints are provided in appendix 2.A.

2.3. Test Administration

The Smarter Balanced Summative Assessments are administered online using the secure browser and test delivery system (TDS), ensuring a secure, confidential, standardized, consistent, and appropriate administration for students.

2.3.1. Test Security and Confidentiality

All tests within the CAASPP System are secure. For the Smarter Balanced Summative Assessment administration, every person having access to test materials maintains the security and confidentiality of the tests. ETS' internal Code of Ethics requires that all test information, including tangible materials (such as test booklets, test questions, and test results), confidential files, processes, and activities are kept secure. To ensure security for all tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI). A detailed description of the OTI and its mission is presented in subsection *5.2.1 ETS' Office of Testing Integrity (OTI)* in *Chapter 5: Test Administration*.

In the pursuit of enforcing secure practices, ETS strives to safeguard the various processes involved in a test development and administration cycle. Those processes are listed next. The practices related to each of the following security processes are discussed in detail in chapter 5:

- Procedures to maintain standardization of test security
- Security of electronic files using a firewall
- Transfer of scores via secure data exchange
- Data management in the secure database
- Statistical analysis on secure servers
- Student confidentiality
- Student test results

2.3.2. Procedures to Maintain Standardization

ETS takes all necessary measures to ensure the standardization of administration of the Smarter Balanced Summative Assessments. The measures for standardization include, but are not limited to, the aspects described in the following subsections.

2.3.2.1. Test Administrators

The Smarter Balanced Summative Assessments are administered in conjunction with the other assessments that compose the CAASPP System. ETS employs processes to ensure the standardization of an administration cycle; these processes are discussed in more detail in section *5.4 Procedures to Maintain Standardization*.

Staff at LEAs involved in the CAASPP Smarter Balanced administration include LEA CAASPP coordinators, CAASPP test site coordinators, and test administrators. The responsibilities of each of the staff members are described in the *CAASPP Online Test Administration Manual* (California Department of Education [CDE], 2020c).

2.3.2.2. Test Directions

Several series of instructions regarding the CAASPP administration are compiled in detailed manuals and provided to the LEA staff. Such documents include, but are not limited to, the following:

- ***CAASPP Online Test Administration Manual*—**This is a manual that provides test administration procedures and guidelines for LEA CAASPP coordinators and CAASPP test site coordinators, as well as the script and *Directions for Administration* to be followed exactly by test administrators during a testing session (CDE, 2020c). (Refer to *5.4.4.2 CAASPP Online Test Administration Manual* in chapter 5 for more information.)
- ***CAASPP and ELPAC Test Operations Management System User Guide*—**This is a manual that provides instructions for TOMS allowing LEA staff, including LEA CAASPP coordinators and CAASPP test site coordinators, to perform a number of tasks including setting up test administrations, adding and managing users, assigning tests, and configuring online student test settings (CDE, 2020d). (Refer to *5.4.4.3 CAASPP and ELPAC Test Operations Management System User Guide* in chapter 5 for more information.)

2.4. Fairness and Accessibility

All students enrolled in grades three through eight and grade eleven are required to participate in the Smarter Balanced mathematics assessment, except for students with the most significant cognitive disabilities who meet the criteria for the California Alternate Assessment (CAA) for Mathematics based on alternate achievement standards (approximately 1 percent or less of the student population). The decision to assign a student to take an alternate assessment is made by the student’s individualized education program (IEP) team.

All students enrolled in grades three through eight and grade eleven are required to participate in the Smarter Balanced for ELA, except the following:

- Students with the most significant cognitive disabilities who meet the criteria for the CAA for ELA alternate assessment based on alternate achievement standards (approximately 1 percent or less of the student population) take the CAA for ELA. The decision to assign a student to take an alternate assessment is made by the student’s IEP team.
- English learner (EL) students who are within their first 12 months of enrollment in a U.S. school as determined after April 15 of the previous school year have a one-time exemption from the Smarter Balanced for ELA assessment. These students may instead participate in the English Language Proficiency Assessments for California.

The treatment of incomplete tests and test-taking situations is illustrated in table 7.8 in subsection *7.4.1.3 Scoring of Incomplete Cases*.

2.5. Universal Tools, Designated Supports, and Accommodations

All public school students participate in the CAASPP System of assessments, including students with disabilities and EL students. Additional resources are sometimes needed for these students. The CDE provides a full range of assessment resources for all students, including those who are EL students and students with disabilities. There are four different categories of student accessibility resources in the California assessment accessibility system, including universal tools, designated supports, accommodations, and unlisted resources that are permitted for use in CAASPP online assessments. These are listed in the CDE web document *Matrix One: California Assessment of Student Performance and Progress Accessibility Resources* (CDE, 2019). [5]

Universal tools are available to all students. These resources may be turned on and off when embedded as part of the technology platform for the online CAASPP assessments on the basis of student preference and selection.

Designated supports are available to all students when determined as needed by an educator or team of educators, with parent/guardian and student input as appropriate, or when specified in the student's IEP or Section 504 plan.

Accommodations must be permitted on CAASPP assessments for all eligible students when specified in the student's IEP or Section 504 plan.

Unlisted resources are non-embedded and made available if specified in the eligible student's IEP or Section 504 plan and only on approval by the CDE.

Assignment of designated supports and accommodations to individual students based on student need is made in TOMS by the LEA CAASPP coordinator or CAASPP test site coordinator, either through individual assignment through the student's profile in TOMS; by uploading of settings for multiple students that were either selected and entered into a macro-enabled template called the Individual Student Assessment Accessibility Profile (ISAAP) Tool that created an upload file; or entered into a template without macros. These designated supports and accommodations were delivered to the student through the TDS at the time of testing. Refer to section *1.9 Systems Overview and Functionality* in *Chapter 1: Introduction* for more details regarding this system.

Appendix 2.B presents counts and percentages of students assigned designated supports, accommodations, or unlisted resources for PTs and CAT, respectively, during the 2019–2020 CAASPP Smarter Balanced administration. The majority of students do not use any designated supports, accommodations, or unlisted resources.

2.5.1. Resources for Selection of Accessibility Resources

The full list of the universal tools, designated supports, and accommodations that are used in CAASPP online assessments, including Smarter Balanced assessments, is documented in Matrix One (CDE, 2019). Most embedded and non-embedded universal tools, designated supports, and accommodations listed in parts 1, 2, and 3 of Matrix One are available for the

[5] This technical report is based on the version of Matrix One that was available during the 2019–2020 CAASPP administration. Note that Matrix One has since been combined with the English Language Proficiency Assessments for California Matrix Four to form a single accessibility resources matrix, the California Assessment Accessibility Resources Matrix (CDE, 2020c).

CAST through the online testing interface or, in the case of non-embedded resources, from the school or LEA. Part 5 of Matrix One includes approved unlisted resources that are available. School-level personnel, IEP teams, and Section 504 teams used Matrix One when deciding how best to support the student's test-taking experience.

The Smarter Balanced Assessment Consortium's *Usability, Accessibility, and Accommodations Guidelines* ("*Guidelines*") (Smarter Balanced, 2020) aids in the selection of universal tools, designated supports, and accommodations deemed necessary for individual students.[6] The *Guidelines* apply to all students and promote an individualized approach to the implementation of assessment practices. The *Guidelines* are intended to provide Smarter Balanced policy regarding universal tools, designated supports, and accommodations. Another manual, the *Smarter Balanced Usability, Accessibility, and Accommodations Implementation Guide* (Smarter Balanced, 2014), provides suggestions for implementation of these resources.

In addition to assigning accessibility resources individually and via file upload in TOMS, LEAs had the option of using the ISAAP Tool to assign resources to students. Smarter Balanced developed the ISAAP Tool to facilitate selection of the accessibility resources that match student access needs for the Smarter Balanced assessments*.* The CAASPP ISAAP Tool was used by LEAs in conjunction with the *Guidelines* as well as with state regulations and policies (such as Matrix One) related to assessment accessibility as a part of the ISAAP process *and the CAASPP and ELPAC Accessibility Guide for Online Testing* (CDE, 2020a)*.* LEA personnel, including IEP and Section 504 plan teams, used the CAASPP 2019–2020 ISAAP Tool to facilitate the selection of designated supports and accommodations for students.

2.5.2. Delivery of Accessibility Resources

Universal tools, designated supports, and accommodations can be delivered as either embedded or non-embedded resources. Embedded resources are digitally delivered features or settings available as part of the technology platform for the online CAASPP assessments. Examples of embedded resources include the braille language resource, color contrast, and closed captioning for ELA listening items.

Non-embedded resources are available, when provided by the LEA, for both online and paper–pencil CAASPP assessments. These resources are not part of the technology platform for the computer-administered CAASPP tests. Examples of non-embedded resources include magnification, noise buffers, and the use of a scribe.

Refer to subsection *5.6.1 Universal Tools, Designated Supports, and Accommodations for Students with Disabilities* for a detailed description of the accessibility resources available to students taking the Smarter Balanced assessments.

2.5.3. Unlisted Resources

An unlisted resource is an instructional resource that a student regularly uses in daily instruction, assessment, or both, that has not been previously identified as a universal tool, designated support, or accommodation. Matrix One included an inventory of unlisted resources that were already identified and were preapproved (CDE, 2019). During the

[6] This technical report is based on the version of the *Usability, Accessibility, and Accommodations Guidelines* that was available during the 2019–2020 CAASPP administration.

2019–2020 CAASPP administration, an LEA CAASPP coordinator or CAASPP test site coordinator had the option to submit a web form in TOMS to request such a resource for an eligible student. The resource was specified in the eligible student's IEP or Section 504 plan and only was assigned with the CDE's approval.

Unlisted resources are non-embedded resources that are made available if specified in the eligible student's IEP or Section 504 plan and only upon approval by the California Department of Education. Unlisted resources that change the construct of an assessment and are approved will be flagged as causing a change in construct. Test results for a student using an unlisted resource that was approved but changed the construct of what was being tested was considered invalid for accountability purposes. The student's score status would be changed to "Invalid" and the student's scale score would be reported but appear on the Student Score Report (SSR) with an asterisk and a footnote that the test was administered under conditions that resulted in a score that may not be an accurate representation of the student's achievement.

2.6. Scores

For information regarding score specifications and score reports, refer to *Chapter 7: Scoring and Reporting*.

2.6.1. Score Reporting

TOMS is a secure website hosted by ETS that permits LEA users to manage aspects of CAASPP test administration such as test assignment and the assignment of test settings. It also provides a secure means for LEA CAASPP coordinators to download Student Score Reports (SSRs) as PDF files.

Another means of viewing CAASPP scores was the Online Reporting System (ORS), a secure website that provides authorized users with interactive and cumulative online reports for ELA and mathematics at the student, school, and LEA levels. The ORS provides three types of score reports: an individual SSR, a school report, and an LEA report. Refer to subsection *7.6.1 Online Reporting* for details about TOMS and the ORS and subsection *7.6.3 Types of Score Reports* for the content of each type of score report.

Note that the California Educator Reporting System was not available during 2019–2020 testing and is not discussed.

2.6.2. Aggregation Procedures

To provide meaningful results to the stakeholders, CAASPP scores for a given grade are aggregated at the school, LEA or direct funded charter school, county, and state levels. State-level results are available on the Test Results for California's Assessments website at https://caaspp-elpac.cde.ca.gov/caaspp/. The aggregated scores are presented for all students or selected demographic student groups.

Aggregate scores are generated by combining student scores. They can be created by combining results at the state, LEA or direct funded charter school, or school level; combining for all students; or by combining results for students who represent selected demographic student groups.

Aggregation procedures used to present CAASPP Smarter Balanced results are described in section *7.5 Overview of Score Aggregation Procedures* of this report. Because of an extremely small sample of students who completed the tests, aggregated score results by demographic groups will not be reported this year.

2.7. Calibration and Scaling

IRT methods are ideally suited to the assessments and measurement goals of Smarter Balanced in both establishing a common scale and ongoing maintenance of the program. The purpose of calibration, equating, and scaling using IRT methods is to place item difficulty and student ability estimates at all grade levels in each content area onto a common theta scale. As a result, scores on different versions of the same test are statistically adjusted to compensate for any differences in difficulty between the test versions.

The Common Core State Standards were developed with the intent of supporting inferences concerning a student's change in achievement (i.e., progress) as demonstrated by performance on the corresponding assessments. *Vertical scaling* is an approach that places test scores across grades onto a common scale. A vertical scale is a single scale for scores on tests at different grade levels of the same content area. Reporting scores on a vertical scale allows student progress to be tracked for a particular content area across grade levels; it is expected that students' proficiency increases across different levels of the assessment. An advantage of vertical scaling is that progress expectations concerning the establishment of achievement levels across grades can be inspected and ordered by standard setting panelists.

All items used on the Smarter Balanced Online Summative Assessments were calibrated within grade and vertically scaled during the 2013–2014 Smarter Balanced field test phase (Smarter Balanced, 2016). These activities supported the creation of scale scores.

The basic steps in the process of scaling the scores in each content area—ELA or mathematics—are as follows:

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

The reported test scores for the 2019–2020 administration of the Smarter Balanced assessments were expressed on the baseline scale. The baseline scale was defined following the 2013–2014 Smarter Balanced field test administration first. Items developed in later years were linked to the baseline scale after being field tested.

2.7.1. Calibration

Unidimensional IRT models were used for calibration. Based on the psychometric research conducted during the pilot and field test phases by the Smarter Balanced Assessment Consortium, the two-parameter logistic (2PL) model (Birnbaum,1968) and the generalized partial credit model (GPCM) (Muraki, 1992) were chosen for calibration. Refer to equation 7.1 in subsection *7.4.1.1 Theta Scores* for the 2PL model and GPCM formulas.

Item parameter calibration software, model-to-data fit, and evaluation of vertical scale anchor items are described in more detail in chapter 6 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016). The summary statistics describing the distribution of item difficulty and discrimination parameter estimates at each grade level for the 2019–2020 administration item pool are available in appendix 8.A.

2.7.2. Horizontal Scaling

Item parameters derived for the Smarter Balanced assessments were linked during the Smarter Balanced field test administration by concurrently calibrating items within each grade for each content area. The calibration approach relied on a hybrid of the “common items” approach and the “randomly equivalent groups” linking approach. The common items approach requires that items and tasks partially overlap and be administered to different student samples. For the randomly equivalent groups approach, the test items presented to different student samples are considered as comparably “on scale” by virtue of the random equivalence of the groups. The horizontal linking design incorporated both types of approaches and was accomplished by assembling test versions with partially overlapping test content and randomly assigning the test versions to students.

2.7.3. Vertical Scaling

After the grade-specific horizontal scaling was conducted for a content area, a separate, cross-grade, vertical scaling was conducted by Smarter Balanced consortium using common items (vertical linking items). To implement the vertical scaling, representative sets of off-grade items were administered to some students in the next lower adjacent grade—for example, a set of grade five items was administered to some students in grade four.

Vertical linking item sets were intended to sample the construct that included both the CAT and PT components and associated item types as well as claims that conformed to the test blueprint. Linking items from the lower grade were administered to the upper-adjacent-grade–level students. Content experts designated a target grade for each item and a minimum and maximum grade designation. A set of PTs was given on-grade; the same set was administered off-grade for vertical linking.

The vertical scaling was undertaken separately for ELA and for mathematics, using grade six as the base grade. Grade seven was linked to grade six, and then grade eight was linked to grade seven, and so forth, until grade eleven was placed onto the vertical scale. Likewise, grade five was linked to grade six, grade four was linked to grade five, and so forth, until grade three was placed onto the vertical scale (refer to figure 2.1).

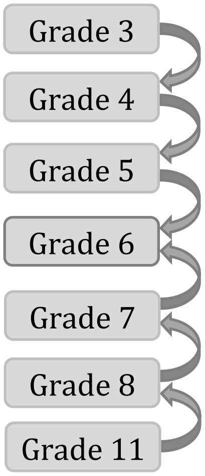

Figure 2.1 Vertical scaling

Once the Smarter Balanced horizontal and vertical scales were established, the remaining items (i.e., the entire calibration item pool including the noncommon items) were linked onto this final scale in each grade and content area.

2.7.4. Vertical Scale Evaluation

The results of vertical scaling were evaluated using a number of methods. Refer to the section *Vertical Scale Evaluation* in *Chapter 9 Field Test Design, Sampling, and Administration* in the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016). This source includes the following information:

- Correlation of difficulties of common items across grade levels
- Changes in test difficulty across grades
- Comparison of mean scale scores across grades
- Comparison of scale scores associated with achievement levels across grades
- Comparison of overlap and separation of scale score distributions across grades
- Comparison of variability in scale scores within and across grades

References

American Institutes for Research. (2014). *Smarter Balanced adaptive item-selection algorithm design report*. Washington, DC: Jon Cohen and Larry Albright. http://www.smarterapp.org/documents/AdaptiveAlgorithm-Preview-v3.pdf

American Institutes for Research. (2015). *Smarter Balanced summative assessments testing procedures for adaptive item selection algorithm, 2014–2015 test administrations, English language arts/literacy (ELA), grades three–eight and grade eleven, and mathematics, grades three–eight and grade eleven.* Washington, DC: American Institutes for Research. https://portal.smarterbalanced.org/library/en/testing-procedures-for-adaptive-item-selection-algorithm.pdf

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

California Department of Education. (2019). *Matrix one: California Assessment of Student Performance and Progress accessibility resources*. Sacramento, CA: California Department of Education. https://bit.ly/3h0EkD2

California Department of Education. (2020a). *CAASPP and ELPAC accessibility guide for online testing.* Sacramento, CA: California Department of Education. https://bit.ly/3usOzny

California Department of Education. (2020b). *CAASPP and ELPAC Test Operations Management System user guide*. Sacramento, CA: California Department of Education. https://bit.ly/2QKKJaV

California Department of Education. (2020c). *CAASPP online test administration manual, 2019–2020 test administration.* Sacramento, CA: California Department of Education. https://bit.ly/3eXYBXc

California Department of Education. (2020d). *California Assessment Accessibility Resources Matrix*. Sacramento, CA: California Department of Education. https://bit.ly/3vHmKbm

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations implementation guide.* Los Angeles, CA: Smarter Balanced Assessment Consortium and National Center on Educational Outcomes. https://bit.ly/3xN8U9b

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: 2013–14 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf

Smarter Balanced Assessment Consortium. (2017). *ELA computer adaptive test (CAT) and performance task (PT) item specifications*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~ (This link in not active)

Smarter Balanced Assessment Consortium. (2018a). *Enhanced CAT blueprints for students participating in the 2017-18 Smarter Balanced embedded field test of performance tasks*. https://portal.smarterbalanced.org/library/en/enhanced-cat-blueprints-pt-eft.pdf

Smarter Balanced Assessment Consortium. (2018b). *Mathematics CAT and PT item specifications*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~ (This link is not active)

Smarter Balanced Assessment Consortium. (2019). *Smarter Balanced Assessment Consortium: 2017–18 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. http://portal.smarterbalanced.org/library/en/2017-18-summative-assessment-technical-report.pdf

Smarter Balanced Assessment Consortium. (2020). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines.* Los Angeles: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

Appendix 2.A: Smarter Balanced Blueprints

English Language Arts/Literacy (ELA) Summative Assessment Blueprints

Blueprint Table for ELA/Literacy Grades Three Through Five as of the 2019–2020 Test Administration

Claim/Score Reporting Category[7]	Content Category[8]	CAT Stimuli	PT Stimuli[9]	CAT Items[10]	PT Items[11]	Total Items by Claim
1. Reading	Literary	2	0	7–8	0	14–16
1. Reading	Informational	2	0	7–8	0	14–16
2. Writing	Organization/Purpose	0	1	3	1[12]	9[13]
2. Writing	Evidence/Elaboration	0	1	3	1[6]	9[7]
2. Writing	Conventions	0	1	3	1[6]	9[7]
3. Speaking/ Listening	Listening	3–4	0	8–9	0	8–9
4. Research	Research	0	1	8	1	9

[7] Each student receives an overall ELA/literacy score and four claim scores or subscores reported at the individual level.

[8] For more information on content categories, see the test blueprints documents at https://contentexplorer.smarterbalanced.org/test-development#content.

[9] Each student receives one PT, which includes a set of stimuli on a given topic.

[10] The CAT component of the test includes only machine-scored items.

[11] Each PT includes one research item which may be a machine-scored item or a short-text item. Each PT also has one full write that is scored across three traits: Organization/Purpose, Evidence/Elaboration, and Conventions. Short-text and full-write items are designed for hand scoring and may be artificial intelligence (AI) scored with an application that yields comparable results by meeting or exceeding reliability and validity criteria for hand scoring.

[12] For the purpose of this table, Writing PTs are noted as three separate “items”; however, the Writing PT score is derived from a single student response scored on three traits.

[13] Total items for Claim 2 include six CAT items and three items from the PT, as described in footnote 6.

Blueprint Table for ELA/Literacy Grades Six Through Eight as of the 2019–2020 Test Administration

Claim/Score Reporting Category[14]	Content Category[15]	CAT Stimuli	PT Stimuli[16]	CAT Items[17]	PT Items[18]	Total Items by Claim
1. Reading	Literary	1–2	0	4–7	0	14–17
1. Reading	Informational	2–3	0	10–12	0	14–17
2. Writing	Organization/Purpose	0	1	3	1[19]	9[20]
2. Writing	Evidence/Elaboration	0	1	3	1[6]	9[7]
2. Writing	Conventions	0	1	3	1[6]	9[7]
3. Speaking/ Listening	Listening	3−4	0	8−9	0	8−9
4. Research	Research	0	1	8	1	9

[14] Each student receives an overall ELA/literacy score and four claim scores or subscores reported at the individual level.

[15] For more information on content categories, see the content specifications document at https://contentexplorer.smarterbalanced.org/test-development#contentSpecs.

[16] Each student receives one PT, which includes a set of stimuli on a given topic.

[17] The CAT component of the test includes machine-scored items and short-text items. Up to two short-text items in Reading and one short-text item in Writing are designed for hand scoring and may be AI scored with an application that yields comparable results by meeting or exceeding reliability and validity criteria for hand scoring.

[18] Each PT includes one research item which may be a machine-scored item or a short-text item. Each PT also has one full write that is scored across three traits: Organization/ Purpose, Evidence/Elaboration, and Conventions. Short-text and full-write items are designed for hand scoring and may be AI scored with an application that yields comparable results by meeting or exceeding reliability and validity criteria for hand scoring.

[19] For the purpose of this table, Writing PTs are noted as three separate "items"; however, the Writing PT score is derived from a single student response scored on three traits.

[20] Total items for Claim 2 include six CAT items and three items from the PT, as described in footnote 6.

Blueprint Table for ELA/Literacy Grade Eleven as of the 2019–2020 Test Administration

Claim/Score Reporting Category[21]	Content Category[22]	CAT Stimuli	PT Stimuli[23]	CAT Items[24]	PT Items[25]	Total Items by Claim
1. Reading	Literary	1	0	4	0	15−16
1. Reading	Informational	3	0	11−12	0	15−16
2. Writing	Organization/Purpose	0	1	3	1[26]	9[27]
2. Writing	Evidence/Elaboration	0	1	3	1[6]	9[7]
2. Writing	Conventions	0	1	3	1[6]	9[7]
3. Speaking/ Listening	Listening	3−4	0	8−9	0	8−9
4. Research	Research	0	1	8	1	9

[21] Each student receives an overall ELA/literacy score and four claim scores or subscores reported at the individual level.

[22] For more information on content categories, see the content specifications document at https://contentexplorer.smarterbalanced.org/test-development#contentSpecs.

[23] Each student receives one PT, which includes a set of stimuli on a given topic.

[24] The CAT component of the test includes machine-scored items and short-text items. One or two short-text items in Reading and one short-text item in Writing are designed for hand scoring and may be AI scored with an application that yields comparable results by meeting or exceeding reliability and validity criteria for hand scoring.

[25] Each PT includes one research item which may be a machine-scored item or a short-text item. Each PT also has one full write that is scored across three traits: Organization/ Purpose, Evidence/Elaboration, and Conventions. Short-text and full-write items are designed for hand scoring and may be AI scored with an application that yields comparable results by meeting or exceeding reliability and validity criteria for hand scoring.

[26] For the purpose of this table, Writing PTs are noted as three separate “items”; however, the Writing PT score is derived from a single student response scored on three distinct traits.

[27] Total Items for Claim 2 include six CAT items and three items from the PT as described in footnote 6.

Smarter Balanced Mathematics Summative Assessment Blueprints

Blueprint Table for Mathematics Grades Three Through Five as of the 2019–2020 Test Administration

Claim/Score Reporting Category	Content Category[1]	CAT Stimuli	PT Stimuli	CAT Items[2]	PT Items[3]	Total Items by Claim[4]
1. Concepts and Procedures	Priority Cluster	0	0	13–15	0	17–20
1. Concepts and Procedures	Supporting Cluster	0	0	4–5	0	17–20
2. Problem Solving 4. Modeling and Data Analysis[5]	Problem Solving	0	1	6	2–4	8–10
2. Problem Solving 4. Modeling and Data Analysis[5]	Modeling and Data Analysis	0	1	6	2–4	8–10
3. Communicating Reasoning	Communicating Reasoning	0	1	8	0–2	8–10

[1] For more information on content categories, see the content specifications document at https://contentexplorer.smarterbalanced.org/test-development#contentSpecs.

[2] All CAT items in grades three through five are designed to be machine scored.

[3] Each PT contains four to six total items. Up to four PT items may require hand scoring.

[4] While the range for the total items by claim for Problem Solving and Modeling and Data Analysis and Communicating Reasoning indicates 8–10 items in each reporting category, the total number of items across these two reporting categories for any individual test event is 18–20.

[5] Claim 2 (Problem Solving) and Claim 4 (Modeling and Data Analysis) have been combined because of content similarity and to provide flexibility for item development. There are still four claims, but only three claim scores will be reported with the overall mathematics score.

Blueprint Table for Mathematics Grades Six Through Eight as of the 2019–2020 Test Administration

Claim/Score Reporting Category	Content Category[6]	CAT Stimuli[7]	PT Stimuli	CAT Items	PT Items[8]	Total Items by Claim[9]
1. Concepts and Procedures	Priority Cluster	0	0	12–15	0	16–20
1. Concepts and Procedures	Supporting Cluster	0	0	4–5	0	16–20
2. Problem Solving 4. Modeling and Data Analysis[10]	Problem Solving	0	1	6	2–4	8–10
2. Problem Solving 4. Modeling and Data Analysis[5]	Modeling and Data Analysis	0	1	6	2–4	8–10
3. Communicating Reasoning	Communicating Reasoning	0	1	8	0–2	8–10

[6] For more information on content categories, see the content specifications document at https://contentexplorer.smarterbalanced.org/test-development#contentSpecs.

[7] All CAT items in grades six through eight are designed to be machine scored.

[8] Each PT contains four to six total items. Up to four PT items may require hand scoring.

[9] While the range for the total items by claim for Problem Solving and Modeling and Data Analysis and Communicating Reasoning indicates 8–10 items in each reporting category, the total number of items across these two reporting categories for any individual test event is 18–20.

[10] Claim 2 (Problem Solving) and Claim 4 (Modeling and Data Analysis) have been combined because of content similarity and to provide flexibility for item development. There are still four claims, but only three claim scores will be reported with the overall mathematics score.

Blueprint Table for Mathematics Grade Eleven as of the 2019–2020 Test Administration

Claim/Score Reporting Category	Content Category[38]	CAT Stimuli[39]	PT Stimuli	CAT Items	PT Items[40]	Total Items by Claim[41]
1. Concepts and Procedures	Priority Cluster	0	0	14–16	0	19–22
1. Concepts and Procedures	Supporting Cluster	0	0	5–6	0	19–22
2. Problem Solving 4. Modeling and Data Analysis[42]	Problem Solving	0	1	6	2–4	8–10
2. Problem Solving 4. Modeling and Data Analysis[5]	Modeling and Data Analysis	0	1	6	2–4	8–10
3. Communicating Reasoning	Communicating Reasoning	0	1	8	0–2	8–10

[38] For more information on content categories, see the content specifications document at https://contentexplorer.smarterbalanced.org/test-development#contentSpecs.

[39] All CAT items in grade eleven are designed to be machine scored.

[40] Each PT contains four to six total items. Up to six PT items may require hand scoring.

[41] While the range for the total items by claim for Problem Solving and Modeling and Data Analysis and Communicating Reasoning indicates 8–10 items in each reporting category, the total number of items across these two reporting categories for any individual test event is 18–20.

[42] Claim 2 (Problem Solving) and Claim 4 (Modeling and Data Analysis) have been combined because of content similarity and to provide flexibility for item development. There are still four claims, but only three claim scores will be reported with the overall mathematics score.

Enhanced Computer Adaptive Testing (CAT) Blueprints for Students Participating in the 2019–2020 Smarter Balanced Embedded Field Test of Performance Tasks (PTs)

ELA/Literacy Blueprints for the Spring 2020 CAT Assessments with Embedded Field Test PTs

Claim/Score Reporting Category	Gr 3–5 CAT	Gr 3–5 PT	Gr 6–8 CAT	Gr 6–8 PT	Gr 11 CAT	Gr 11 PT
Reading	14–16	0	14–17	0	15–16	0
Writing	9	0	9	0	9	0
Listening	8–9	0	8–9	0	8–9	0
Research	9	0	9	0	9	0

Mathematics Blueprints for the Spring 2020 CAT Assessments with Embedded Field Test PTs

Claim/Score Reporting Category	Gr 3–5 CAT	Gr 3–5 PT	Gr 6–8 CAT	Gr 6–8 PT	Gr 11 CAT	Gr 11 PT
Concepts and Procedures	17–20	0	16–20	0	19–22	0
Problem Solving & Modeling and Data Analysis	8	0	8	0	8	0
Communicating Reasoning	8	0	8	0	8	0

Appendix 2.B: Special Services Summaries

Table 2.B.1 Special Services Summary for ELA PT, Grades Three Through Six—All Tested

Accessibility Resource	Grade 3 Number	Grade 3 Pct. of Total	Grade 4 Number	Grade 4 Pct. of Total	Grade 5 Number	Grade 5 Pct. of Total	Grade 6 Number	Grade 6 Pct. of Total
Embedded Accommodation—American Sign Language	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Accommodation—Braille	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Accommodation—Text-to-Speech (for Reading Passages only)	19	13.87	14	9.79	18	11.04	17	13.82
Non-Embedded Accommodation—Alternate Response Options	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Print on Demand	0	0.00	1	0.70	0	0.00	0	0.00
Non-Embedded Accommodation—Scribe (for ELA Writing)	2	1.46	1	0.70	0	0.00	0	0.00
Non-Embedded Accommodation—Speech-to-Text	0	0.00	0	0.00	0	0.00	2	1.63
Non-Embedded Accommodation—Word Prediction	1	0.73	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Color Contrast	1	0.73	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Masking	2	1.46	0	0.00	1	0.61	2	1.63
Embedded Designated Support—Mouse Pointer	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Permissive Mode	0	0.00	1	0.70	0	0.00	1	0.81
Embedded Designated Support—Print Size	0	0.00	1	0.70	0	0.00	0	0.00
Embedded Designated Support—Streamlining	1	0.73	0	0.00	3	1.84	0	0.00
Embedded Designated Support—Text-to-Speech (for ELA except for Reading Passages)	52	37.96	20	13.99	45	27.61	16	13.01
Embedded Designated Support—Turn Off Any Universal Tools	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Amplification	0	0.00	0	0.00	0	0.00	1	0.81
Non-Embedded Designated Support—Bilingual Dictionary	1	0.73	0	0.00	0	0.00	1	0.81
Non-Embedded Designated Support—Color Contrast	1	0.73	0	0.00	0	0.00	0	0.00

Table 2.B.1 *(continuation)*

Accessibility Resource	Grade 3 Number	Grade 3 Pct. of Total	Grade 4 Number	Grade 4 Pct. of Total	Grade 5 Number	Grade 5 Pct. of Total	Grade 6 Number	Grade 6 Pct. of Total
Non-Embedded Designated Support—Color Overlay	1	0.73	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Magnification	1	0.73	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Medical Device	0	0.00	1	0.70	0	0.00	0	0.00
Non-Embedded Designated Support—Noise Buffers	10	7.30	7	4.90	15	9.20	1	0.81
Non-Embedded Designated Support—Read Aloud (for ELA except Reading Passages)	6	4.38	5	3.50	2	1.23	2	1.63
Non-Embedded Designated Support—Scribe (for Reading and Listening)	3	2.19	2	1.40	0	0.00	0	0.00
Non-Embedded Designated Support—Separate Setting	15	10.95	16	11.19	22	13.50	14	11.38
Non-Embedded Designated Support—Simplified Test Directions	12	8.76	10	6.99	14	8.59	19	15.45
Non-Embedded Designated Support—Translated Test Directions	0	0.00	0	0.00	0	0.00	1	0.81
Other—Unlisted Resources	0	0.00	0	0.00	0	0.00	0	0.00
Other—Designated support or accommodation is in IEP	28	20.44	20	13.99	28	17.18	26	21.14
Other—Designated support or accommodation is in Section 504 plan	1	0.73	2	1.40	2	1.23	1	0.81

Table 2.B.2 Special Services Summary for ELA PT, Grades Seven, Eight, and Eleven—All Tested

Accessibility Resource	Grade 7 Number	Grade 7 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 11 Number	Grade 11 Pct. of Total
Embedded Accommodation—American Sign Language	0	0.00	0	0.00	2	0.01
Embedded Accommodation—Braille	0	0.00	0	0.00	2	0.01
Embedded Accommodation—Text-to-Speech (for Reading Passages only)	49	10.04	37	18.23	169	1.21
Non-Embedded Accommodation—Alternate Response Options	0	0.00	0	0.00	1	0.01
Non-Embedded Accommodation—Print on Demand	0	0.00	0	0.00	1	0.01
Non-Embedded Accommodation—Scribe (for ELA Writing)	1	0.20	0	0.00	3	0.02
Non-Embedded Accommodation—Speech-to-Text	7	1.43	9	4.43	42	0.30
Non-Embedded Accommodation—Word Prediction	3	0.61	0	0.00	18	0.13
Embedded Designated Support—Color Contrast	8	1.64	12	5.91	5	0.04
Embedded Designated Support—Masking	19	3.89	26	12.81	30	0.21
Embedded Designated Support—Mouse Pointer	6	1.23	14	6.90	2	0.01
Embedded Designated Support—Permissive Mode	8	1.64	12	5.91	3	0.02
Embedded Designated Support—Print Size	8	1.64	13	6.40	19	0.14
Embedded Designated Support—Streamlining	7	1.43	14	6.90	26	0.19
Embedded Designated Support—Text-to-Speech (for ELA except for Reading Passages)	63	12.91	51	25.12	228	1.63
Embedded Designated Support—Turn Off Any Universal Tools	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Amplification	0	0.00	0	0.00	4	0.03
Non-Embedded Designated Support—Bilingual Dictionary	0	0.00	0	0.00	79	0.57
Non-Embedded Designated Support—Color Contrast	1	0.20	0	0.00	3	0.02
Non-Embedded Designated Support—Color Overlay	1	0.20	0	0.00	1	0.01
Non-Embedded Designated Support—Magnification	0	0.00	1	0.49	10	0.07
Non-Embedded Designated Support—Medical Device	0	0.00	0	0.00	0	0.00

Table 2.B.2 *(continuation)*

Accessibility Resource	Grade 7 Number	Grade 7 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 11 Number	Grade 11 Pct. of Total
Non-Embedded Designated Support—Noise Buffers	2	0.41	1	0.49	37	0.26
Non-Embedded Designated Support—Read Aloud (for ELA except Reading Passages)	7	1.43	4	1.97	49	0.35
Non-Embedded Designated Support—Scribe (for Reading and Listening)	2	0.41	0	0.00	4	0.03
Non-Embedded Designated Support—Separate Setting	31	6.35	5	2.46	459	3.28
Non-Embedded Designated Support—Simplified Test Directions	61	12.50	17	8.37	126	0.90
Non-Embedded Designated Support—Translated Test Directions	1	0.20	2	0.99	50	0.36
Other—Unlisted Resources	0	0.00	0	0.00	1	0.01
Other—Designated support or accommodation is in IEP	80	16.39	43	21.18	483	3.46
Other—Designated support or accommodation is in Section 504 plan	0	0.00	2	0.99	32	0.23

Table 2.B.3 Special Services Summary for Mathematics PT, Grades Three Through Six—All Tested

Accessibility Resource	Grade 3 Number	Grade 3 Pct. of Total	Grade 4 Number	Grade 4 Pct. of Total	Grade 5 Number	Grade 5 Pct. of Total	Grade 6 Number	Grade 6 Pct. of Total
Embedded Accommodation—American Sign Language	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Accommodation—Braille	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Abacus	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Alternate Response Options	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Calculator	0	0.00	0	0.00	0	0.00	9	19.15
Non-Embedded Accommodation—Multiplication Table	N/A	N/A	10	14.49	6	5.45	19	40.43
Non-Embedded Accommodation—Print on Demand	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Speech-to-Text	0	0.00	1	1.45	0	0.00	2	4.26
Non-Embedded Accommodation—Word Prediction	1	0.99	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—100s Number Table	0	0.00	3	4.35	1	0.91	5	10.64
Embedded Designated Support—Color Contrast	1	0.99	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Masking	2	1.98	0	0.00	1	0.91	1	2.13
Embedded Designated Support—Mouse Pointer	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Permissive Mode	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Print Size	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Streamlining	1	0.99	0	0.00	2	1.82	0	0.00
Embedded Designated Support—Text-to-Speech	50	49.50	19	27.54	36	32.73	12	25.53
Embedded Designated Support—Translated Test Directions (with Spanish Stacked Translation only)	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Translations (glossary)	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Turn Off Any Universal Tools	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Amplification	0	0.00	0	0.00	0	0.00	1	2.13

Table 2.B.3 *(continuation)*

Accessibility Resource	Grade 3 Number	Grade 3 Pct. of Total	Grade 4 Number	Grade 4 Pct. of Total	Grade 5 Number	Grade 5 Pct. of Total	Grade 6 Number	Grade 6 Pct. of Total
Non-Embedded Designated Support—Color Contrast	1	0.99	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Color Overlay	1	0.99	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Magnification	1	0.99	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Medical Device	0	0.00	1	1.45	0	0.00	0	0.00
Non-Embedded Designated Support—Noise Buffers	10	9.90	8	11.59	15	13.64	1	2.13
Non-Embedded Designated Support—Read Aloud	6	5.94	5	7.25	2	1.82	2	4.26
Non-Embedded Designated Support—Read Aloud (also for Spanish Stacked Translation)	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Scribe	3	2.97	3	4.35	0	0.00	0	0.00
Non-Embedded Designated Support—Separate Setting	14	13.86	14	20.29	20	18.18	10	21.28
Non-Embedded Designated Support—Simplified Test Directions	11	10.89	11	15.94	13	11.82	14	29.79
Non-Embedded Designated Support—Translated Test Directions	0	0.00	0	0.00	0	0.00	0	0.00
Other—Unlisted Resources	0	0.00	0	0.00	0	0.00	0	0.00
Other—Designated support or accommodation is in IEP	27	26.73	21	30.43	25	22.73	22	46.81
Other—Designated support or accommodation is in Section 504 plan	1	0.99	2	2.90	2	1.82	1	2.13

Table 2.B.4 Special Services Summary for Mathematics PT, Grades Seven, Eight, and Eleven—All Tested

Accessibility Resource	Grade 7 Number	Grade 7 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 11 Number	Grade 11 Pct. of Total
Embedded Accommodation—American Sign Language	0	0.00	0	0.00	3	0.03
Embedded Accommodation—Braille	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Abacus	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Alternate Response Options	0	0.00	0	0.00	1	0.01
Non-Embedded Accommodation—Calculator	14	21.54	56	10.00	160	1.52
Non-Embedded Accommodation—Multiplication Table	22	33.85	60	10.71	27	0.26
Non-Embedded Accommodation—Print on Demand	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Speech-to-Text	6	9.23	8	1.43	13	0.12
Non-Embedded Accommodation—Word Prediction	2	3.08	1	0.18	3	0.03
Non-Embedded Accommodation—100s Number Table	14	21.54	11	1.96	4	0.04
Embedded Designated Support—Color Contrast	7	10.77	13	2.32	76	0.72
Embedded Designated Support—Masking	18	27.69	27	4.82	333	3.16
Embedded Designated Support—Mouse Pointer	7	10.77	14	2.50	70	0.67
Embedded Designated Support—Permissive Mode	9	13.85	13	2.32	1	0.01
Embedded Designated Support—Print Size	8	12.31	14	2.50	53	0.50
Embedded Designated Support—Streamlining	8	12.31	13	2.32	15	0.14
Embedded Designated Support—Text-to-Speech	29	44.62	72	12.86	507	4.82
Embedded Designated Support—Translated Test Directions (with Spanish Stacked Translation only)	7	10.77	14	2.50	81	0.77
Embedded Designated Support—Translations (glossary)	0	0.00	0	0.00	6	0.06
Embedded Designated Support—Turn Off Any Universal Tools	0	0.00	0	0.00	0	0.00

Table 2.B.4 *(continuation)*

Accessibility Resource	Grade 7 Number	Grade 7 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 11 Number	Grade 11 Pct. of Total
Non-Embedded Designated Support—Amplification	0	0.00	0	0.00	3	0.03
Non-Embedded Designated Support—Color Contrast	0	0.00	0	0.00	1	0.01
Non-Embedded Designated Support—Color Overlay	0	0.00	0	0.00	1	0.01
Non-Embedded Designated Support—Magnification	0	0.00	1	0.18	7	0.07
Non-Embedded Designated Support—Medical Device	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Noise Buffers	0	0.00	2	0.36	27	0.26
Non-Embedded Designated Support—Read Aloud	3	4.62	4	0.71	20	0.19
Non-Embedded Designated Support—Read Aloud (also for Spanish Stacked Translation)	0	0.00	1	0.18	0	0.00
Non-Embedded Designated Support—Scribe	1	1.54	0	0.00	1	0.01
Non-Embedded Designated Support—Separate Setting	6	9.23	5	0.89	280	2.66
Non-Embedded Designated Support—Simplified Test Directions	16	24.62	46	8.21	397	3.77
Non-Embedded Designated Support—Translated Test Directions	0	0.00	7	1.25	40	0.38
Other—Unlisted Resources	0	0.00	0	0.00	0	0.00
Other—Designated support or accommodation is in IEP	24	36.92	66	11.79	369	3.51
Other—Designated support or accommodation is in Section 504 plan	0	0.00	3	0.54	19	0.18

Table 2.B.5 Special Services Summary for ELA, CAT Grades Three Through Six—All Tested

Accessibility Resource	Grade 3 Number	Grade 3 Pct. of Total	Grade 4 Number	Grade 4 Pct. of Total	Grade 5 Number	Grade 5 Pct. of Total	Grade 6 Number	Grade 6 Pct. of Total
Embedded Accommodation—American Sign Language	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Accommodation—Audio Transcript	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Accommodation—Braille	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Accommodation—Closed Captioning	0	0.00	2	1.40	0	0.00	0	0.00
Embedded Accommodation—Text-to-Speech (for Reading Passages only)	19	13.87	15	10.49	18	11.04	17	13.82
Non-Embedded Accommodation—Alternate Response Options	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Print on Demand	0	0.00	1	0.70	0	0.00	0	0.00
Non-Embedded Accommodation—Read Aloud (for ELA Reading Passages)	5	3.65	3	2.10	0	0.00	2	1.63
Non-Embedded Accommodation—Scribe (for ELA Writing)	2	1.46	2	1.40	0	0.00	0	0.00
Non-Embedded Accommodation—Speech-to-Text	0	0.00	0	0.00	0	0.00	2	1.63
Non-Embedded Accommodation—Word Prediction	1	0.73	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Color Contrast	1	0.73	1	0.70	0	0.00	0	0.00
Embedded Designated Support—Masking	2	1.46	1	0.70	1	0.61	2	1.63
Embedded Designated Support—Mouse Pointer	0	0.00	1	0.70	0	0.00	0	0.00
Embedded Designated Support—Permissive Mode	0	0.00	2	1.40	0	0.00	1	0.81
Embedded Designated Support—Print Size	0	0.00	2	1.40	0	0.00	0	0.00
Embedded Designated Support—Streamlining	1	0.73	1	0.70	3	1.84	0	0.00
Embedded Designated Support—Text-to-Speech (for ELA except for Reading Passages)	52	37.96	21	14.69	45	27.61	16	13.01
Embedded Designated Support—Turn Off Any Universal Tools	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Amplification	0	0.00	0	0.00	0	0.00	1	0.81

Table 2.B.5 *(continuation)*

Accessibility Resource	Grade 3 Number	Grade 3 Pct. of Total	Grade 4 Number	Grade 4 Pct. of Total	Grade 5 Number	Grade 5 Pct. of Total	Grade 6 Number	Grade 6 Pct. of Total
Non-Embedded Designated Support—Color Contrast	1	0.73	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Color Overlay	1	0.73	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Magnification	1	0.73	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Medical Device	0	0.00	1	0.70	0	0.00	0	0.00
Non-Embedded Designated Support—Noise Buffers	10	7.30	7	4.90	15	9.20	1	0.81
Non-Embedded Designated Support—Read Aloud (for ELA except Reading Passages)	6	4.38	5	3.50	2	1.23	2	1.63
Non-Embedded Designated Support—Scribe (for Reading and Listening)	3	2.19	3	2.10	0	0.00	0	0.00
Non-Embedded Designated Support—Separate Setting	15	10.95	16	11.19	21	12.88	14	11.38
Non-Embedded Designated Support—Simplified Test Directions	12	8.76	10	6.99	14	8.59	19	15.45
Non-Embedded Designated Support—Translated Test Directions	0	0.00	0	0.00	0	0.00	1	0.81
Other—Unlisted Resources	0	0.00	0	0.00	0	0.00	0	0.00
Other—Designated support or accommodation is in IEP	28	20.44	21	14.69	27	16.56	26	21.14
Other—Designated support or accommodation is in Section 504 plan	1	0.73	2	1.40	2	1.23	1	0.81

Table 2.B.6 Special Services Summary for ELA, CAT Grades Seven, Eight, and Eleven—All Tested

Accessibility Resource	Grade 7 Number	Grade 7 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 11 Number	Grade 11 Pct. of Total
Embedded Accommodation—American Sign Language	0	0.00	0	0.00	1	0.01
Embedded Accommodation—Audio Transcript	0	0.00	0	0.00	3	0.02
Embedded Accommodation—Braille	0	0.00	0	0.00	2	0.01
Embedded Accommodation—Closed Captioning	0	0.00	0	0.00	13	0.09
Embedded Accommodation—Text-to-Speech (for Reading Passages only)	49	10.04	38	18.72	172	1.23
Non-Embedded Accommodation—Alternate Response Options	0	0.00	0	0.00	1	0.01
Non-Embedded Accommodation—Print on Demand	0	0.00	0	0.00	1	0.01
Non-Embedded Accommodation—Read Aloud (for ELA Reading Passages)	11	2.25	3	1.48	36	0.26
Non-Embedded Accommodation—Scribe (for ELA Writing)	1	0.20	0	0.00	4	0.03
Non-Embedded Accommodation—Speech-to-Text	7	1.43	9	4.43	46	0.33
Non-Embedded Accommodation—Word Prediction	3	0.61	1	0.49	18	0.13
Embedded Designated Support—Color Contrast	8	1.64	13	6.40	5	0.04
Embedded Designated Support—Masking	19	3.89	28	13.79	30	0.21
Embedded Designated Support—Mouse Pointer	6	1.23	15	7.39	2	0.01
Embedded Designated Support—Permissive Mode	8	1.64	13	6.40	3	0.02
Embedded Designated Support—Print Size	8	1.64	14	6.90	19	0.14
Embedded Designated Support—Streamlining	7	1.43	15	7.39	26	0.19
Embedded Designated Support—Text-to-Speech (for ELA except for Reading Passages)	63	12.91	53	26.11	230	1.65
Embedded Designated Support—Turn Off Any Universal Tools	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Amplification	0	0.00	0	0.00	4	0.03
Non-Embedded Designated Support—Color Contrast	1	0.20	0	0.00	3	0.02
Non-Embedded Designated Support—Color Overlay	1	0.20	0	0.00	1	0.01

Table 2.B.6 *(continuation)*

Accessibility Resource	Grade 7 Number	Grade 7 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 11 Number	Grade 11 Pct. of Total
Non-Embedded Designated Support—Magnification	0	0.00	1	0.49	10	0.07
Non-Embedded Designated Support—Medical Device	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Noise Buffers	2	0.41	2	0.99	37	0.26
Non-Embedded Designated Support—Read Aloud (for ELA except Reading Passages)	7	1.43	5	2.46	50	0.36
Non-Embedded Designated Support—Scribe (for Reading and Listening)	2	0.41	0	0.00	4	0.03
Non-Embedded Designated Support—Separate Setting	31	6.35	6	2.96	462	3.31
Non-Embedded Designated Support—Simplified Test Directions	61	12.50	18	8.87	127	0.91
Non-Embedded Designated Support—Translated Test Directions	1	0.20	2	0.99	50	0.36
Other—Unlisted Resources	0	0.00	0	0.00	1	0.01
Other—Designated support or accommodation is in IEP	80	16.39	45	22.17	484	3.46
Other—Designated support or accommodation is in Section 504 plan	0	0.00	2	0.99	34	0.24

Table 2.B.7 Special Services Summary for Mathematics, CAT Grades Three Through Six—All Tested

Accessibility Resource	Grade 3 Number	Grade 3 Pct. of Total	Grade 4 Number	Grade 4 Pct. of Total	Grade 5 Number	Grade 5 Pct. of Total	Grade 6 Number	Grade 6 Pct. of Total
Embedded Accommodation—American Sign Language	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Accommodation—Braille	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Abacus	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Alternate Response Options	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Calculator	0	0.00	0	0.00	0	0.00	9	19.15
Non-Embedded Accommodation—Multiplication Table	N/A	N/A	10	14.49	6	5.45	19	40.43
Non-Embedded Accommodation—Print on Demand	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Speech-to-Text	0	0.00	1	1.45	0	0.00	2	4.26
Non-Embedded Accommodation—Word Prediction	1	0.99	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—100s Number Table	0	0.00	3	4.35	1	0.91	5	10.64
Embedded Designated Support—Color Contrast	1	0.99	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Masking	2	1.98	0	0.00	1	0.91	1	2.13
Embedded Designated Support—Mouse Pointer	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Permissive Mode	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Print Size	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Streamlining	1	0.99	0	0.00	2	1.82	0	0.00
Embedded Designated Support—Text-to-Speech	50	49.50	19	27.54	36	32.73	12	25.53
Embedded Designated Support—Translated Test Directions (with Spanish Stacked Translation only)	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Translations (glossary)	0	0.00	0	0.00	0	0.00	0	0.00
Embedded Designated Support—Turn Off Any Universal Tools	0	0.00	0	0.00	0	0.00	0	0.00

Table 2.B.7 *(continuation)*

Accessibility Resource	Grade 3 Number	Grade 3 Pct. of Total	Grade 4 Number	Grade 4 Pct. of Total	Grade 5 Number	Grade 5 Pct. of Total	Grade 6 Number	Grade 6 Pct. of Total
Non-Embedded Designated Support—Amplification	0	0.00	0	0.00	0	0.00	1	2.13
Non-Embedded Designated Support—Color Contrast	1	0.99	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Color Overlay	1	0.99	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Magnification	1	0.99	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Medical Device	0	0.00	1	1.45	0	0.00	0	0.00
Non-Embedded Designated Support—Noise Buffers	10	9.90	8	11.59	15	13.64	1	2.13
Non-Embedded Designated Support—Read Aloud	6	5.94	5	7.25	2	1.82	2	4.26
Non-Embedded Designated Support—Read Aloud (also for Spanish Stacked Translation)	0	0.00	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Scribe	3	2.97	3	4.35	0	0.00	0	0.00
Non-Embedded Designated Support—Separate Setting	14	13.86	14	20.29	20	18.18	10	21.28
Non-Embedded Designated Support—Simplified Test Directions	11	10.89	11	15.94	13	11.82	14	29.79
Non-Embedded Designated Support—Translated Test Directions	0	0.00	0	0.00	0	0.00	0	0.00
Other—Unlisted Resources	0	0.00	0	0.00	0	0.00	0	0.00
Other—Designated support or accommodation is in IEP	27	26.73	21	30.43	25	22.73	22	46.81
Other—Designated support or accommodation is in Section 504 plan	1	0.99	2	2.90	2	1.82	1	2.13

Table 2.B.8 Special Services Summary for Mathematics, CAT Grades Seven, Eight, and Eleven—All Tested

Accessibility Resource	Grade 7 Number	Grade 7 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 11 Number	Grade 11 Pct. of Total
Embedded Accommodation—American Sign Language	0	0.00	0	0.00	3	0.03
Embedded Accommodation—Braille	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Abacus	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Alternate Response Options	0	0.00	0	0.00	1	0.01
Non-Embedded Accommodation—Calculator	14	21.54	59	10.54	160	1.52
Non-Embedded Accommodation—Multiplication Table	22	33.85	64	11.43	27	0.26
Non-Embedded Accommodation—Print on Demand	0	0.00	0	0.00	0	0.00
Non-Embedded Accommodation—Speech-to-Text	6	9.23	8	1.43	13	0.12
Non-Embedded Accommodation—Word Prediction	2	3.08	1	0.18	3	0.03
Non-Embedded Accommodation—100s Number Table	14	21.54	11	1.96	4	0.04
Embedded Designated Support—Color Contrast	7	10.77	13	2.32	76	0.72
Embedded Designated Support—Masking	18	27.69	27	4.82	333	3.16
Embedded Designated Support—Mouse Pointer	7	10.77	14	2.50	70	0.67
Embedded Designated Support—Permissive Mode	9	13.85	13	2.32	1	0.01
Embedded Designated Support—Print Size	8	12.31	14	2.50	53	0.50
Embedded Designated Support—Streamlining	8	12.31	13	2.32	15	0.14
Embedded Designated Support—Text-to-Speech	29	44.62	77	13.75	505	4.80
Embedded Designated Support—Translated Test Directions (with Spanish Stacked Translation only)	7	10.77	16	2.86	79	0.75
Embedded Designated Support—Translations (glossary)	0	0.00	0	0.00	6	0.06
Embedded Designated Support—Turn Off Any Universal Tools	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Amplification	0	0.00	0	0.00	3	0.03
Non-Embedded Designated Support—Color Contrast	0	0.00	0	0.00	1	0.01

Table 2.B.8 *(continuation)*

Accessibility Resource	Grade 7 Number	Grade 7 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 11 Number	Grade 11 Pct. of Total
Non-Embedded Designated Support—Color Overlay	0	0.00	0	0.00	1	0.01
Non-Embedded Designated Support—Magnification	0	0.00	1	0.18	7	0.07
Non-Embedded Designated Support—Medical Device	0	0.00	0	0.00	0	0.00
Non-Embedded Designated Support—Noise Buffers	0	0.00	2	0.36	27	0.26
Non-Embedded Designated Support—Read Aloud	3	4.62	4	0.71	20	0.19
Non-Embedded Designated Support—Read Aloud (also for Spanish Stacked Translation)	0	0.00	1	0.18	0	0.00
Non-Embedded Designated Support—Scribe	1	1.54	0	0.00	1	0.01
Non-Embedded Designated Support—Separate Setting	6	9.23	5	0.89	280	2.66
Non-Embedded Designated Support—Simplified Test Directions	16	24.62	52	9.29	396	3.76
Non-Embedded Designated Support—Translated Test Directions	0	0.00	9	1.61	39	0.37
Other—Unlisted Resources	0	0.00	0	0.00	0	0.00
Other—Designated support or accommodation is in IEP	24	36.92	70	12.50	368	3.50
Other—Designated support or accommodation is in Section 504 plan	0	0.00	3	0.54	19	0.18

Chapter 3: Item Development

This chapter discusses the procedures followed during the development of Smarter Balanced items to help ensure valid interpretation of test scores.

3.1. Background

The Smarter Balanced Assessment Consortium, in coordination with its member states, developed innovative item types and authored items based on the Common Core State Standards. The Consortium used an iterative process involving higher education and kindergarten through grade twelve educators who were trained in item development, as well as state partners, professional item writers, and assessment vendors at various stages in the item development process.

3.2. Additional Information

More information regarding the item development process (including the qualifications of those involved), item development specifications, and content alignment studies undertaken by Smarter Balanced to produce item types and items for the assessment can be found in chapter 3 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016).

Reference

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: 2013–14 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf

Chapter 4: Test Assembly

The Smarter Balanced Summative Assessments were administered operationally as part of the California Assessment of Student Performance and Progress for the first time during the 2014–2015 school year. The summative assessments each consist of two parts: a computer adaptive test (CAT) and performance tasks. The Smarter Balanced Summative Assessments are constructed to measure students' performance relative to Common Core State Standards (CCSS). The assessments also are constructed to produce scores that meet professional standards for reliability and validity of test score interpretation. The content standards and desired psychometric attributes are used as the basis for assembling the test forms. This chapter discusses the content and psychometric criteria that guide the construction of the Smarter Balanced Summative Assessments.

4.1. Smarter Balanced Adaptive Item Selection Algorithm

This section describes the algorithm and the design for implementation of adaptive item selection for the Smarter Balanced test delivery system (TDS). The implementation builds extensively on the algorithm implemented in Cambium Assessment, Inc.'s (CAI's) TDS.

The general item selection approach is that the next item to be administered to a specific student is chosen on the basis of a function of three variables. The first variable is an index of the importance of the item for meeting the content requirements of the test. The other two variables are values of the item response theory item information functions in the region of the student's current ability estimate. One of these information functions is for the student's total score; the other is for the student's claim score.

More information about how each of these three measures is defined can be found in the *Smarter Balanced Adaptive Item-Selection Algorithm Design Report* (American Institutes for Research [AIR], 2015).

Values for these three measures are calculated to guide and support item selection. A value is computed for whether the item or group of items will be selected based on how well that item matches the target content, contributes to overall score information, and contributes to claim score information.

$$\text{Item Selection}=w_1\text{Content Match}+\ w_2\text{Overall Information}+w_3\text{Claim Information} \tag{4.1}$$

Refer to the Alternative Text for Equation 4.1 for a description of this equation.

This objective function is used to measure an item's contribution to each of these objectives. A higher value for "Content Match" means that an item is more important for meeting the content requirements. A higher value for "Overall Information" means that an item contributes more information to the estimation of the student's current overall ability.
A higher value for "Claim Information" means that an item contributes more information for estimating the student's current claim ability. Weights of these objectives can be adjusted to achieve the desired balance and optimize performance for a given item pool. This algorithm enables users to maximize information subject to the constraint that the blueprint is almost always met, with minimal exceptions.

4.1.1. Content Match

Each item or item group is characterized by its contribution to meeting the blueprint, given the items that have already been administered at any point. The contribution is based on the presence or absence of features specified in the blueprint.

The Smarter Balanced summative test blueprints describe the content of the English language arts/literacy (ELA) and mathematics summative assessments for all grades tested and the means by which that content is assessed. The summative online test blueprints reflect the depth and breadth of the performance expectations of the CCSS.

The test blueprints have information about the number of items and depth of knowledge for items associated with each claim and assessment target. Each test is described by a blueprint for both the overall test and each claim within the test.

Each blueprint has features referred to as *constraints*. Constraints define features such as the minimum and maximum number of items required in a specific content area. For example, a constraint might require a minimum of four and a maximum of six algebra items. The value of content match is highest for items with content that has not met its minimum constraint, decreases for items representing content for which the minimum number of items has been reached but the maximum has not, and becomes negative for items representing content that has met the maximum.

Refer to the blueprints for the Smarter Balanced ELA and mathematics assessments provided in appendix 2.A for additional details.

4.1.2. Information

Every item has an overall information value within the CAT algorithm and an information value for each claim. Details on how information is calculated is provided in equations 7.7 through 7.11 in *7.4.3 Theta Scores Standard Error*.

Items with higher discrimination parameters offer more information and therefore are generally given preference in item selection. Because the overexposure of highly discriminating items is a test security risk, the item selection algorithm includes additional rules to control the exposure of the items that provide the highest measurement information (AIR, 2014).

4.2. Simulation Study

For the CAT, prior to opening the 2019–2020 operational testing window, CAI conducts simulations to evaluate and ensure the appropriate implementation and quality of the adaptive item selection algorithm and the scoring algorithm. The simulation tool allows manipulation of key blueprint and configuration settings to match the blueprint of the test and minimize measurement error. In this simulation study, the adaptive tests are administered in one segment (section) in ELA for all grades tested, and mathematics grades three through five and in two segments in mathematics grades six through eight and grade eleven, including calculator and no-calculator segments. Each segment is simulated separately.

The *Smarter Balanced Summative Assessments Testing Procedures for Adaptive Item-Selection Algorithm,* (AIR, 2015) presents the results of an examination of the robustness of the item-selection algorithm of the Smarter Balanced CAT administrations in ELA and mathematics for grades three through eight and grade eleven. The information provided by the simulations includes

- evaluation of the simulation step,
- the percentage of tests aligned with the test blueprints (blueprint match rates),
- the number of targets covered in the simulated forms,
- accuracy of ability estimates indicated by bias and precision of ability estimates indicated by standard error,
- item exposure rates,
- selection of off-grade items and corresponding psychometric properties, and
- exposure rates of embedded field test items.

The results of CAI's simulation study show the following:

- Across content areas and grade levels, 98 percent or more of the simulated tests covered the test blueprint.
- Scale scores were estimated precisely across the entire scale with the exception of scores near the highest obtainable scale score and the lowest obtainable scale score.
- The vast majority of items were exposed to students less than 20 percent of the time.
- The embedded field test item exposure rates were below one percent.

Table 4.1 contains characteristics of items students received particular to the content area tests.

Table 4.1 Item Distribution Characteristics from the CAI Simulation

Characteristic	ELA	Mathematics
Received off-grade items	11–55% of students in grades 3–8 only	16–54% of students in grades 4–8 and grade 11
Scored above standard, received above-grade items	4–18% of the students for grades 3–8 only	N/A
Scored as not meeting the standard, received below-grade items	38–50% of students in grades 4, 6, and 7 only	19–54% of students in grades 4–8 and grade 11

CAI concluded that content domain scores were comparable across the grades within the content area with respect to a certain content domain and that scores at various ranges of the score distribution were measured with good precision. The results also demonstrated that global item exposure was controlled to the extent that no items were used too often, off-grade items were administered according to criteria in the test specifications to students who were performing very well or very poorly on the test, and the field test items were distributed equally across multiple blocks within a test as intended for that grade and content area.

References

American Institutes for Research. (2014). *Smarter Balanced adaptive item selection algorithm design report*. Washington, DC: American Institutes for Research. http://www.smarterapp.org/documents/AdaptiveAlgorithm-Preview-v3.pdf

American Institutes for Research. (2015). *Smarter Balanced Summative Assessments testing procedures for adaptive item-selection algorithm.* Washington, DC: American Institutes for Research. https://portal.smarterbalanced.org/library/en/testing-procedures-for-adaptive-item-selection-algorithm.pdf

Accessibility Information

Alternative Text for Equation 4.1

Item Selection equal w sub 1 multiplied by Content Match plus w sub 2 multiplied by Overall Information plus w sub 3 multiplied by Claim Information.

Chapter 5: Test Administration

This chapter details the processes involved in the administration of the 2019–2020 Smarter Balanced Summative Assessments. It also describes the procedures followed by ETS to maintain test security throughout the test administration process. In particular, the testing window early closure, due to the novel coronavirus disease 2019 (COVID-19) pandemic, and the impact are discussed in this chapter.

5.1. Test Administration

The testing window for the 2019–2020 administration of the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced assessments was planned for January 14 through July 15, 2020. Specific test administration schedules within that window were determined locally pursuant to the *California Code of Regulations*, Title 5 (5 *CCR),* sections 855(a)(1), 855(a)(2), 855(b), and 855(c). However, because of the COVID-19 pandemic, most of the schools in California halted in-person instruction after March 13, 2020. Then, on March 18, 2020, Governor Gavin Newsom signed an order suspending the CAASPP for all students in California (Office of Governor Gavin Newsom, 2020). As a result, less than 1 percent of students started the test and less than 0.5 percent of students completed and submitted their tests on average; the exception was grade eleven, with about 2 percent completion.

ETS conducted on-site test administration workshops in various locations throughout California in January and February and produced webcasts and videos on helpful topics. In addition, ETS provided a number of test administration resources to schools and local educational agencies (LEAs). These resources included detailed information on topics such as technology readiness, test administration, test security, accommodations, using the test delivery system (TDS), and general testing rules. These resources are discussed in more detail in section *5.4 Procedures to Maintain Standardization*.

5.1.1. Test Delivery Sections

The test delivery sections correspond to the computer adaptive tests (CATs) and performance task (PT) portions of the assessments. CAT items are delivered dynamically based on the students' performance on the previous items; students typically are presented with many different items, and items presented to any two students may appear in different locations within the test. For a given PT, students are presented with the same items in the same order of presentation and associated test length. During the 2019–2020 administration, PTs were randomly assigned at the student level.

Criteria for the minimum number of items for each claim that are required in the operational blueprints and the embedded field test blueprints are provided in appendix 2.A.

5.1.1.1. Computer Adaptive Testing Administration

CAT assessments are assembled dynamically to obtain a unique test for each student from a defined item pool so that each student is given a unique, content-conforming test form. Item statistics based on item response theory are used to determine the administration and adaptation of test items based on student responses and ability; this information is incorporated into the delivery algorithm. The item selection algorithm is described in more detail in *4.1 Smarter Balanced Adaptive Item Selection Algorithm*, along with item exposure rates.

Item exposure control (e.g., Sympson & Hetter, 1985) can be used to ensure that uniform rates of item administration are achieved because it is not desirable to have some items presented to many students while other items are presented to relatively few students.

5.1.1.2. Performance Task Administration

Smarter Balanced Assessment Consortium item and task specifications assume online delivery of the items and tasks. Most tasks are long enough to warrant several administration sessions. Such sessions could be same-day, back-to-back sessions with short breaks between sessions. All tasks are administered in controlled classroom settings. Estimated time requirements for completing PTs and administration time are provided in the *CAASPP Online Test Administration Manual* (California Department of Education [CDE], 2020c).

Student directions for all tasks begin with an overview of the entire task that briefly describes the necessary steps. The overview gives students advanced knowledge of the scorable products or performances to be created (Khattri, Reeve, & Kane, 1998). Allowable teacher-student interactions for a task are standardized (i.e., carefully scripted or described in task directions for purposes of comparability, fairness, and security). Teachers are directed not to assist students in the production of their scorable products or presentations. Note that, during the 2019–2020 administration of Smarter Balanced online assessments, some students were assigned an embedded field test PT rather than the operational PT. Because the scores on the embedded field test PTs do not contribute to the reported scores, these students were assigned a CAT with additional items. Refer to *Appendix 2.A: Smarter Balanced Blueprints* for the number of CAT items in the blueprints for assessments with embedded field test PTs.

5.2. Test Security and Confidentiality

For the Smarter Balanced Online Summative Assessment administration, every person who works with the assessments, communicates test results, or receives testing information is responsible for maintaining the security and confidentiality of the tests, including CDE staff, ETS staff, ETS subcontractors, LEA assessment coordinators, school assessment coordinators, students, parents/guardians, teachers, and cooperative educational service agency staff. ETS’ Code of Ethics requires that all test information, including tangible materials (such as test items), confidential files (such as those containing personally identifiable student information), processes related to test administration (such as the configurations of secure servers), and activities are kept secure. ETS has systems in place that maintain tight security for test items and test results, as well as for student data. To ensure security for all the tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI), which is described in the next subsection.

All tests within the CAASPP System, as well as the confidentiality of student information, should be protected to ensure the validity, reliability, and fairness of the results. As stated in *Standard 7.9* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), “The documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session” (p. 128).

This section of the *CAASPP Smarter Balanced Technical Report* describes the measures intended to prevent potential test security incidents prior to testing and the actions that were taken to handle actual security incidents during or after testing using the Security and Test Administration Incident Reporting System (STAIRS) process.

5.2.1. ETS' Office of Testing Integrity

The OTI is a division of ETS that provides quality assurance services for all testing programs managed by ETS; this division resides in the ETS legal department. The Office of Professional Standards Compliance at ETS publishes and maintains *ETS Standards for Quality and Fairness* (ETS, 2014)*,* which supports the OTI's goals and activities. The *ETS Standards for Quality and Fairness* provides guidelines to help ETS staff design, develop, and deliver technically sound, fair, and beneficial products and services and help the public and auditors evaluate those products and services.

The OTI's mission is to

- minimize any testing security violations that can impact the fairness of testing,
- minimize and investigate any security breach that threatens the validity of the interpretation of test scores, and
- report on security activities.

The OTI helps prevent misconduct on the part of students and administrators, detects potential misconduct through empirically established indicators, and resolves situations involving misconduct in a fair and balanced way that reflects the laws and professional standards governing the integrity of testing. In its pursuit of enforcing secure practices, the OTI strives to safeguard the various processes involved in a test development and administration cycle. For the Smarter Balanced assessments, those processes included the following:

- Security of electronic files using a firewall
- Printing and publishing
- Test administration
- Test delivery
- Processing and scoring
- Data management
- Statistical analysis
- Student confidentiality

5.2.2. Procedures to Maintain Standardization of Test Security

Test security requires accounting for all secure materials—including online summative test items, paper–pencil tests, and student data—before, during, and after each test administration. The LEA CAASPP coordinator is responsible for keeping all electronic and paper–pencil test materials secure, keeping student information confidential, and making sure the CAASPP test site coordinators and test administrators are properly trained regarding security policies and procedures.

The CAASPP test site coordinator is responsible for mitigating test security incidents at the test site and for reporting incidents to the LEA CAASPP coordinator. If the test site administered paper–pencil tests, the CAASPP test site coordinator is also responsible for the return of any secure materials to the LEA CAASPP coordinator, who, in turn, is responsible for returning any materials to the Scoring and Processing Center.

The test administrator is responsible for reporting testing incidents to the CAASPP test site coordinator and securely destroying printed and digital media for items and passages generated by the print-on-demand feature of the TDS (CDE, 2020d and 2020e).

The following measures ensured the security of CAASPP System assessments administered in 2019–2020:

- LEA CAASPP coordinators and test site coordinators must have signed and submitted a “CAASPP Test Security Agreement for LEA CAASPP coordinators and CAASPP test site coordinators” form in the Test Operations Management System (TOMS) before ETS granted the coordinators access to TOMS (5 *CCR*, Section 859[a]).
- Anyone having access to the testing materials must have electronically signed and submitted a “Test Security Affidavit for Test Examiners, Test Administrators, Proctors, Translators, Scribes, and Any Other Person Having Access to CAASPP Tests” form to the CAASPP test site coordinator before receiving access to any testing materials (5 *CCR*, Section 859[c]).

In addition, it was the responsibility of every participant in the CAASPP System to report immediately any violation or suspected violation of test security or confidentiality. The test site coordinator reported to the LEA CAASPP coordinator. The LEA CAASPP coordinator reported to the CDE within 24 hours of the incident (5 *CCR*, Section 859[e]).

5.2.3. Security of Electronic Files Using a Firewall

A firewall is software that prevents unauthorized entry to files, email, and other organization-specific information. All ETS data exchanges and internal email remain within the ETS firewall at all ETS locations, ranging from Princeton, New Jersey, to San Antonio, Texas, to Concord and Sacramento, California.

All electronic applications that are included in TOMS remain protected by the ETS firewall software at all times. Because of the sensitive nature of the student information processed by TOMS, the firewall plays a significant role in maintaining assurance of confidentiality among the users of this information.

Refer to section *1.9 Systems Overview and Functionality* in *Chapter 1: Introduction* for more information on TOMS.

5.2.4. Transfer of Scores via Secure Data Exchange

Because of the confidential nature of test results, ETS currently uses secure file transfer protocol (SFTP) and encryption for all data file transfers; test data is never sent via email. SFTP is a method for reliable and exclusive routing of files. Files reside on a password-protected server that only authorized users can access. ETS shares an SFTP server with the CDE. On that site, ETS posts Microsoft Word and Excel files, Adobe Acrobat PDFs, or other document files for the CDE to review; the CDE returns reviewed materials in the same manner. Files are deleted upon retrieval.

The SFTP server is used as a conduit for the transfer of files; secure test data is stored only temporarily on the shared SFTP server. Industry-standard secure protocols are used to transfer test content and student data from the ETS internal data center to any external systems.

ETS enters information about the files posted to the SFTP server in a web form on a SharePoint website. A CDE staff member checks this log throughout the day to check the status of deliverables and downloads and deletes the file from the SFTP server when its status shows it has been posted.

5.2.5. Data Management in the Secure Database

ETS currently maintains a secure database to house all student demographic data and assessment results. Information associated with each student has a database relationship to the LEA, school, and grade codes as data is collected during operational testing. Only individuals with the appropriate credentials can access the data. ETS builds all interfaces with the most stringent security considerations, including interfaces with data encryption for databases that store test items and student data. ETS applies best and up-to-date security practices, including system-to-system authentication and authorization, in all solution designs.

All stored test content and student data is encrypted. ETS complies with the Family Educational Rights and Privacy Act (20 *United States Code [USC]* § 1232g; 34 *Code of Federal Regulations* Part 99) and the Children's Online Privacy Protection Act (15 *USC* §§ 6501–6506, P.L. No. 105–277, 112 Stat. 2681–1728).

In TOMS, staff at LEAs and test sites have different levels of access appropriate to the role assigned to them.

5.2.6. Statistical Analysis on Secure Servers

During CAASPP testing, ETS information technology staff members retrieve data files from CAI and load those files into a database. The ETS Data Quality Services staff extracts the data from the database and performs quality control procedures (e.g., the values of all variables are as expected) before passing files to the ETS statistical analysis group (refer to section *9.4 Quality Control of Psychometric Processes* for data validation processes undertaken by ETS Data Quality Services) The statistical analysis staff stores the files on secure servers. All staff involved with the data adhere to the ETS Code of Ethics and the ETS Information Protection Policies to prevent any unauthorized access to the data.

5.2.7. Student Confidentiality

To meet the requirements of the Every Student Succeeds Act as well as state requirements, LEAs must collect demographic data about students' ethnicity, disabilities, parent/guardian education, and so forth during the school year. ETS takes every precaution to prevent any of this information from becoming public or being used for anything other than testing and score-reporting purposes. These procedures are applied to all documents in which student demographic data appears, including reports and the Pre-ID files and response booklets used in paper–pencil testing.

5.2.8. Student Test Results

5.2.8.1. Types of Results

The following deliverables are produced for reporting of the CAASPP Smarter Balanced Summative Assessments:

- Preliminary student reports for online assessments in the Online Reporting System (ORS)
- Preliminary student reports for paper–pencil tests in the ORS
- Individual Student Score Reports (SSRs) (electronic)
- Internet reports aggregated by content area and state, county, LEA, or test site

5.2.8.2. Security of Results Files

ETS takes measures to protect files and reports that show students' scores and achievement levels. ETS is committed to safeguarding all secure information in its possession from unauthorized access, disclosure, modification, or destruction. ETS has strict information security policies in place to protect the confidentiality of both student and client data. ETS staff access to production databases is limited to personnel with a business need to access the data. User IDs for production systems must be person-specific or for systems use only.

ETS has implemented network controls for routers, gateways, switches, firewalls, network tier management, and network connectivity. Routers, gateways, and switches represent points of access between networks. However, these do not contain mass storage or represent points of vulnerability, particularly for unauthorized access or denial of service.

ETS has many facilities, policies, and procedures to protect computer files. Software and procedures such as firewalls, intrusion detection, and virus control are in place to provide for physical security, data security, and disaster recovery. ETS is certified in the BS 25999-2 standard for business continuity and conducts disaster recovery exercises annually. ETS routinely backs up all data to either disks through deduplication or to tapes, all of which are stored off site.

Access to the ETS Computer Processing Center is controlled by employee and visitor identification badges. The Center is secured by doors that can only be unlocked by the badges of personnel who have functional responsibilities within its secure perimeter. Authorized personnel accompany visitors to the ETS Computer Processing Center at all times. Extensive smoke detection and alarm systems, as well as a preaction fire-control system, are installed in the Center.

5.2.8.3. Security of Individual Results

ETS protects individual students' results during the following events:

- Scoring
- Transfer of scores by means of secure data exchange
- Reporting
- Analysis and reporting of erasure marks
- Posting of aggregate data
- Storage

In addition to protecting the confidentiality of testing materials, ETS' Code of Ethics further prohibits ETS employees from financial misuse, conflicts of interest, and unauthorized appropriation of ETS property and resources. Specific rules are also given to ETS employees and their immediate families who may take a test developed by ETS (e.g., a CAASPP assessment). The ETS OTI verifies that these standards are followed throughout ETS. This verification is conducted, in part, by periodic on-site security audits of departments, with follow-up reports containing recommendations for improvement.

5.2.9. Security and Test Administration Incident Reporting System Process

Test security incidents, such as improprieties, irregularities, and breaches, are prohibited behaviors that give a student an unfair advantage or compromise the secure administration of the tests, which, in turn, compromises the reliability and validity of test results (CDE, 2020e). Whether intentional or unintentional, failure by staff or students to comply with security rules constitutes a test security incident. Test security incidents have impacts on scoring and affect students' performance on the test.

LEA CAASPP coordinators and CAASPP test site coordinators ensured that all test security and summative administration incidents were documented by following the prompts in TOMS that guided coordinators in their submittal. An Appeal is a request to reset, restore, re-open, invalidate, or grant a grace period extension to a student's test. If an Appeal to a student's test was warranted, TOMS provided additional prompts to file the Appeal.

After a case was submitted, an email containing a case number and next steps was sent to the submitter (and to the LEA CAASPP coordinator, if the case was submitted by the CAASPP test site coordinator). The STAIRS case in TOMS provided the LEA CAASPP coordinator, the CDE, and the California Technical Assistance Center (CalTAC) with the opportunity to interact and communicate regarding the STAIRS process (CDE, 2020e).

Prior to the assessment administration, ETS and the CDE agreed that the following types of STAIRS cases were also forwarded to the CDE:

- Student cheating or accessing unauthorized devices
- Security breach (where a student exposed secure materials)
- Student unable to review previous answers (20-minute pause rule for the CAT was exceeded)

Appeals requests were reviewed by the CDE. When a request to submit an Appeal was approved, the coordinator received a system-generated email with the Appeal type that was approved (CDE, 2020e).

Types of Appeals available during the 2019–2020 CAASPP administration are described in table 5.1.

Table 5.1 Types of Appeals

Type of Appeal	Description
Reset	Resetting a student's summative assessment removes that assessment from the system and enables the student to start a new assessment from the beginning.
Invalidate	Invalidated summative assessments will be scored, and scores will be provided on the SSR with a note that an irregularity occurred. The student(s) will be counted as participating in the calculation of the school's participation rate for accountability purposes. The score will be counted as "not proficient" for aggregation into the CAASPP results.
Re-open	Reopening a summative assessment allows a student to access an assessment that has already been submitted or has expired.

Table 5.1 *(continuation)*

Type of Appeal	Description
Restore	Restoring a summative assessment returns an assessment from the Reset status to its prior status. This action can only be performed on tests that have been previously reset.
Grace Period Extension	Permitting a grace period extension allows the student to review previously answered questions upon logging back on to the assessment after expiration of the pause rule. Note that for a PT, having the test administrator open a new testing session may be all that is needed to continue testing. A grace period extension will only be granted in cases where there was a disruption to a test session, such as a technical difficulty, fire drill, schoolwide power outage, earthquake, or other act beyond the control of the test administrator.

5.2.9.1. Impropriety

A testing impropriety 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. An impropriety can be corrected and contained at a local level. An impropriety should be reported to the LEA CAASPP coordinator and CAASPP test site coordinator immediately. The coordinator should report the incident within 24 hours, using the STAIRS/Appeals process in TOMS.

5.2.9.2. Irregularity

A testing irregularity is an unusual circumstance that impacts an individual or a group of students who are testing and may potentially affect student performance on the test or impact test security or test validity. These circumstances can be corrected and contained at the local level and submitted using the STAIRS/Appeals process in TOMS. An irregularity must be reported to the LEA CAASPP coordinator and CAASPP test site coordinator immediately. The coordinator must report the irregularity within 24 hours, using the online STAIRS/Appeals process in TOMS.

5.2.9.3. Breach

A testing breach is an event that poses a threat to the validity of the test. Breaches require immediate attention and escalation to CalTAC (for social media breaches) or the CDE (for all other breaches) via telephone. Following the call, the CAASPP test site coordinator or LEA CAASPP coordinator must report the incident using the online STAIRS/Appeals process in TOMS within 24 hours. Examples may include such situations as a release of secure materials or a security or system risk. These circumstances have external implications for the Smarter Balanced Assessment Consortium and may result in a Consortium decision to remove the test item(s) from the available secure bank.

5.2.10. Appeals

For incidents that resulted in a need to reset, re-open, invalidate, or restore individual online student assessments, the request was approved by the CDE. In most instances, an Appeal was submitted to address a test security breach or irregularity. The LEA CAASPP coordinator or CAASPP test site coordinator submitted Appeals in TOMS. All submitted Appeals were available for retrieval and review by the appropriate credentialed users within a given organization. However, the view of Appeals is restricted according to the user role as established in TOMS. An Appeal could be requested only by the LEA CAASPP coordinator or CAASPP test site coordinator if prompted while filing a STAIRS case in TOMS (CDE, 2020e). Types of Appeals available during the 2019–2020 CAASPP administration are described in table 5.1.

Table 5.2 and table 5.3 present the number of Appeals in STAIRS in the 2019–2020 administration for ELA and mathematics, respectively, as well as the number of Statewide Student Identifiers (SSIDs) submitted and approved.

Table 5.2 Number and Types of Incidents Submitted in STAIRS in the 2019–2020 Administration—ELA

Description	Appeal Type	Number of Incidents	Total Number of SSID(s) Submitted	Appeals SSID(s) Approved
Accessibility Issue	Reset	3	3	3
Accidental Summative Access	Reset or Re-open or No Appeal	21	286	187
Administered Incorrect Assessment	Reset	0	0	0
Administration Error	No Appeal	0	0	0
Data Entry Issue	Reset or Re-open	0	0	0
Expired or Accidentally Submitted Test	Re-open	3	3	3
Exposing Secure Materials	Invalidate or No Appeal	0	0	0
Incorrect SSID Used	Reset or No Appeal	2	3	0
Restore from Reset	Restore	0	0	0
Student Cheating or Accessing Unauthorized Devices	Invalidate	1	1	1
Student Disruption	No Appeal	1	0	0
Technical Issues	Grace Period Extension or No Appeal	0	0	0
Validity Issue	Invalidate or Reset	2	2	2

Table 5.3 Number and Types of Incidents Submitted in STAIRS in the 2019–2020 Administration—Mathematics

Description	Appeal Type	Number of Incidents	Total Number of SSID(s) Submitted	Appeals SSID(s) Approved
Accessibility Issue	Reset	4	10	3
Accidental Summative Access	Reset or Re-open or No Appeal	9	84	36
Administered Incorrect Assessment	Reset	0	0	0
Administration Error	No Appeal	0	0	0
Data Entry Issue	Reset or Re-open	0	0	0
Expired or Accidentally Submitted Test	Re-open	1	1	1
Exposing Secure Materials	Invalidate or No Appeal	1	0	0
Incorrect SSID Used	Reset or No Appeal	0	0	0
Restore from Reset	Restore	0	0	0
Student Cheating or Accessing Unauthorized Devices	Invalidate	0	0	0
Student Disruption	No Appeal	0	0	0
Technical Issues	Grace Period Extension or No Appeal	0	0	0
Validity Issue	Invalidate or Reset	4	4	3

Table 5.4 and table 5.5 present the number of Appeals approved and rejected in ELA and mathematics, respectively, by Appeal type in STAIRS.

Table 5.4 Number of Appeals Approved in STAIRS in the 2019–2020 Administration

Appeal Type	Number of Appeals Approved in ELA	Number of Appeals Approved in Mathematics
Reset	154	39
Re-open	21	1
Invalidate	3	3
Grace Period Extension	0	0
Restore	0	0

Table 5.5 Number of Appeals Rejected in STAIRS in the 2019–2020 Administration

Appeal Type	Number of Appeals Rejected in ELA	Number of Appeals Rejected in Mathematics
Reset	99	56
Re-open	0	0
Invalidate	0	0
Grace Period Extension	0	0
Restore	0	0

5.3. Processing and Scoring

The constructed-response (CR) data and the TDS-scored data for tests completed by students in a given day flow from the TDS to ETS. The TDS is capable of scoring a variety of item types, referred to as “machine-scored” items, which are described in section *7.1 Approach to Scoring Item Responses*. Outcomes of CR items are scored by artificial intelligence or by human scoring.

Targeted efforts are made to recruit California educators for participation as raters in the human scoring portion of the Smarter Balanced assessments. Raters are certified based on their ability to use a rubric and accurately score sample responses. Once approved, raters are trained to access the Measurement Incorporated and ETS scoring interfaces and Smarter Balanced–specific scoring policies and procedures and are provided interactive training to practice scoring sample responses with feedback from the scoring leader.

Raters work in shifts and are supervised by a scoring leader who has received special training in scoring and monitoring. Raters are provided Smarter Balanced materials to aid scoring; these materials include anchor sets, scoring rubrics, validity samples, qualifying sets, and condition codes. (Refer to section *7.3 Rater Training* for the definitions of these materials.) A scoring leader gives direct feedback to raters for additional content support. Scoring of California student responses is given priority routing to raters who are California-based educators.

5.4. Procedures to Maintain Standardization

The test administration procedures are designed so that the tests are administered in a standardized manner. ETS takes all necessary measures to ensure the standardization of test administration, as described in this section.

5.4.1. Local Educational Agency CAASPP Coordinator

An LEA CAASPP coordinator was designated by the district superintendent at the beginning of the 2019–2020 school year. LEAs include public school districts, statewide benefit charter schools, State Board of Education–authorized charter schools, county office of education programs, and direct funded charter schools.

LEA CAASPP coordinators are responsible for ensuring the proper and consistent administration of the CAASPP assessments. In addition to the responsibilities set forth in 5 *CCR* Section 857, their responsibilities include

- adding CAASPP test site coordinators and test administrators into TOMS;
- training CAASPP test site coordinators and test administrators regarding the state and Smarter Balanced assessment administration as well as security policies and procedures;
- reporting test security incidents (including testing irregularities) to the CDE;
- overseeing test administration activities;
- printing out checklists for CAASPP test site coordinators and test administrators to review in preparation for administering the summative assessments;
- distributing and collecting scorable and nonscorable materials for students who take paper–pencil tests;
- filing a report of a testing incident in STAIRS; and
- requesting an Appeal (if indicated by TOMS prompts while reporting an incident using the STAIRS/Appeal process).

5.4.2. CAASPP Test Site Coordinator

A CAASPP test site coordinator is trained by the LEA CAASPP coordinator for each test site (5 *CCR* Section 857[f]). A test site coordinator must be an employee of the LEA and must sign a security agreement (5 *CCR* Section 859[a]).

A test site coordinator is responsible for identifying test administrators and ensuring that they have signed CAASPP Test Security Affidavits (5 *CCR* Section 859[d]). CAASPP test site coordinators' duties may include

- adding test administrators into TOMS;
- entering test settings for students;
- creating testing schedules and procedures for a school consistent with state and LEA policies;
- working with technology staff to ensure secure browsers are installed and any technical issues are resolved;
- monitoring testing progress during the testing window and ensuring all students take the test, as appropriate;
- coordinating and verifying the correction of student data errors in the California Longitudinal Pupil Achievement Data System;
- ensuring a student's test session is rescheduled, if necessary;
- addressing testing problems;
- reporting security incidents;
- overseeing administration activities at a school site;
- filing a report of a testing incident in STAIRS; and
- requesting an Appeal (if indicated by TOMS prompts while reporting an incident using the STAIRS/Appeals process).

5.4.3. Test Administrators

Test administrators are identified by CAASPP test site coordinators as individuals who will administer the Smarter Balanced Summative Assessments.

A test administrator must sign a security affidavit (5 *CCR* Section 850[ae]). A test administrator's duties may include

- ensuring the physical conditions of the testing room meet the criteria for a secure test environment;
- administering the CAASPP assessments;
- reporting all test security incidents to the test site coordinator and LEA CAASPP coordinator in a manner consistent with Smarter Balanced, state, and LEA policies;
- viewing student information prior to testing to ensure that the correct student receives the proper test with appropriate resources and reporting potential data errors to test site coordinators and LEA CAASPP coordinators;
- monitoring student progress throughout the test session using the Test Administrator Interface; and
- fully complying with all directions provided in the *Directions for Administration* for the Smarter Balanced Online Summative Assessments (CDE, 2020d).

5.4.4. Instructions for Test Administrators

5.4.4.1. Test Administrator *Directions for Administration*

The *Directions for Administration* of the Smarter Balanced Summative Assessments used by test administrators to administer the Smarter Balanced assessments to students are included in the *CAASPP Online Test Administration Manual* (CDE, 2020d). Test administrators must follow all directions and guidelines and read, word-for-word, the instructions to students in the "SAY" boxes to ensure standardization of test administration. Additionally, the *CAASPP Online Test Administration Manual* provides information to test administrators regarding the systems involved in testing, including sections on the TDS, so they may become familiar with the testing application used by their students (CDE, 2020d).

5.4.4.2. CAASPP Online Test Administration Manual

The *CAASPP Online Test Administration Manual* (CDE, 2020d) contains information and instructions on overall procedures and guidelines for all LEA and test site staff involved in the administration of online assessments. Sections include the following topics:

- Roles and responsibilities of those involved with CAASPP testing
- Test administration resources
- Test security
- Administration preparation and planning
- General test administration
- Test administration directions and scripts for test administrators
- Overview of the student testing application
- Instructions for steps to take before, during, and after testing

Appendices include definitions of common terms, descriptions of different aspects of the test and systems associated with the test, and checklists of activities for LEA CAASPP coordinators, CAASPP test site coordinators, and test administrators.

5.4.4.3. CAASPP and ELPAC Test Operations Management System User Guide

TOMS is a web-based application that allows LEA CAASPP coordinators to set up test administrations, add and manage users, submit online student test settings, and order paper–pencil tests.

TOMS modules described in the *TOMS User Guide* included the following (CDE, 2020c):

- **Test Administration Setup—**This module allows LEAs to determine and calculate dates for the LEA's 2019–2020 administration of the CAASPP, including the Smarter Balanced assessments.
- **Adding and Managing Users—**This module allows LEA CAASPP coordinators to add CAASPP test site coordinators and test administrators to TOMS so that the designated user can administer, monitor, and manage the CAASPP Smarter Balanced assessments.
- **Student Test Assignment—**This module allows LEA CAASPP coordinators to designate students to take the alternate assessment.
- **Online Student Test Settings—**This module allows LEA CAASPP coordinators and CAASPP test site coordinators to configure online test settings so students receive the assigned accessibility resources for the online assessments.

5.4.4.4. Other System Manuals

Other manuals were created to assist LEA CAASPP coordinators and others with the technological components of the CAASPP System and are listed next:

- ***CAASPP and ELPAC Technical Specifications and Configuration Guide*—**This manual provides information, tools, and recommended configuration details to help technology staff prepare computers and install the secure browser to be used for the online CAASPP assessments (CDE, 2020b).
- ***CAASPP Security Incidents and Appeals Procedure Guide*—**This manual provides information on how to report a testing incident and submit an Appeal to reset, reopen, invalidate, or restore individual online student assessments (CDE, 2020e).
- ***CAASPP and ELPAC Accessibility Guide Online Testing*—**This manual provides descriptions of the accessibility resources for online tests as well as information about supported hardware and software requirements for administering tests to students using accessibility resources, including those with a braille accommodation using Job Access With Speech (JAWS®) (software) or a braille embosser (hardware). Students with a braille accommodation are able to take advantage of the adaptive algorithm using the TDS's Enhanced Accessibility Mode and JAWS (CDE, 2020a).

5.5. LEA Training

Each year, ETS, in collaboration with the CDE and their Assessment Validity and Outreach contractor, the Sacramento County Office of Education (SCOE), establishes and implements a comprehensive training plan for LEA assessment staff and educators on all aspects of the assessment program. The ETS and SCOE annual training plans are developed with stakeholder feedback and specify the audience, topics, frequency, and mode (in-person, webcast, videos, modules, etc.) of the training, including such elements as format, participants, and logistics.

In 2019−2020, ETS and SCOE increased their collaboration efforts to provide a more streamlined training experience for LEA and school staff. ETS and SCOE began coordinating training plans and posting all training opportunities in one centralized location on the CAASPP website. LEA staff were able to register for training opportunities, across both CDE contractors' offerings, in one place; and access all archived training materials on the 2019−2020 Training Opportunities web page at https://bit.ly/3elzm20. This new streamlined and coordinated process provided easy access to all the trainings that were offered.

5.5.1. Workshops

The in-person training series began in July and continued through August 2019, with eight two-day Summer Institutes that helped classroom teachers and other educators understand the purpose of the different types of Smarter Balanced assessments and how they work together to support learning. The workshop helped educators understand the design of Smarter Balanced Interim Assessment items aligned with college- and career-ready standards and used evidence-based scoring to analyze student responses. Information about how to use interim assessment and Digital Library systems, including accessibility features, to support teaching and learning was included in the content provided to LEA staff.

ETS also conducted eight in-person pretest workshops and a pretest webcast in January 2020 for the 2019–2020 administration, which focused on training LEA CAASPP coordinators on how to prepare for administering all aspects of the CAASPP online assessments.

Training was also provided to focus on interpreting and using results. ETS typically provides eight in-person "CAASPP Results Are In—Now What?" workshops. However, because of the COVID-19 pandemic and its impact on statewide testing, this workshop was converted to a virtual training. With the cancellation of statewide testing and limited results being released to LEAs, the title of the training was changed to "CAASPP: Using Assessment Data for Decision-Making." This training was made available to LEAs as four stand-alone modules that focused on what data can tell educators about current student learning, how to interpret data, how to communicate data to local stakeholders, and making sense of Smarter Balanced data.

In addition to the in-person training opportunities offered by ETS, SCOE held the first California Assessment Conference (CAC) in Oakland, California, in October 2019. This conference focused on building connections between assessments and the classroom by providing classroom educators with information about using statewide assessment data to improve teaching and learning. The conference included a session on the three parts of a balanced assessment system: formative assessments, interim assessments, and summative assessments. The conference also offered sessions that provided information about accessibility resources.

5.5.2. Virtual Training and Webcasts

ETS provided a series of virtual trainings and live webcasts throughout the school year that were archived and made available for training LEA and test site staff and test administrators. Webcast viewers were provided with a method of electronically submitting questions to the presenters during the webcast. The webcasts were recorded and archived for on-demand viewing on the 2019−2020 CAASPP Archived Webcasts web page at https://bit.ly/3xZRZjF. CAASPP webcasts were available to everyone and required neither preregistration nor a logon account.

In addition to the webcasts provided by ETS, SCOE hosted a number of virtual trainings intended to support new LEA CAASPP coordinators throughout all aspects of administration. This training series provided opportunities for new LEA CAASPP coordinators to receive timely training nearly every month.

SCOE also offered assessment update meetings intended to provide LEA CAASPP coordinators with timely updates about California’s assessment system. The meetings were recorded and archived for on-demand viewing on the 2019−2020 Training Opportunities web page at https://bit.ly/3elzm20.

5.5.3. Videos

To supplement the virtual trainings, webcasts, and in-person workshops, ETS also produced short demonstration videos on various aspects of administering the CAASPP, which were available on the CAASPP Quick Reference Guides and Videos web page at https://www.caaspp.org/administration/instructions/qrgs-and-videos/index.html. SCOE produced quick reference guides to accompany many of the video resources, providing multiple avenues of support for educators administering the assessments.

5.5.4. Training for Proper Identification of Designated Supports and Accommodations

ETS produced short demonstration videos for every embedded accessibility resource that demonstrated how to use the resource for educators, students, and parents. The videos were available in both English and Spanish on the Accessibility Resources Demonstration Videos web page at https://www.caaspp.org/training/caaspp/uaag.html.

In addition, ETS developed a video with LEA staff about the importance of implementing CAASPP accessibility resources to help California educators learn more about the importance of accessibility resources and best practices used by educators in the field. The “Importance of Implementing CAASPP and ELPAC Accessibility Resources: Voices from Educators” video was available on the CAASPP Training Video web page at ~~https://www.caaspp.org/rsc/videos/archived-training_importance-of-implementing-accessibility.html.~~ (This link is not active)

A video on how to use the Individual Student Assessment Accessibility Profile (ISAAP) Tool was also available to support educators in the process of creating an individual student profile and matching accessibility resources to student needs to ensure a fair and valid testing experience.

At the CAC, SCOE offered three sessions on accessibility. A Plenary Accessibility 101 Session was presented to all conference attendees and was intended to build a shared understanding of basic accessibility-related terms and considerations. The Creating an Equitable Process breakout session focused on developing an equitable and systematic process for matching students with appropriate accessibility resources. Matching Resources to Student Needs was another breakout session focused on providing an opportunity to practice appropriately matching student needs to the various accessibility resources.

5.6. Monitoring Assessment of Students

The U.S. Department of Education's peer-review process includes several critical elements that address the need to monitor testing resources for students with disabilities, English learner (EL) students, and EL students with disabilities. The Every Student Succeeds Act reaffirms the importance of ensuring that assessments are accessible to special populations, and the Individuals with Disabilities Education Act has monitoring requirements for students with disabilities. This section describes the accessibility resources used to support students in the Smarter Balanced assessment, as well as the procedures to identify and assign students with accommodations and designated supports. Finally, the number of students who were assigned accessibility resources was reported based on available data.

5.6.1. Universal Tools, Designated Supports, and Accommodations for Students with Disabilities

The purpose of universal tools, designated supports, and accommodations in testing is to allow *all* students the opportunity to demonstrate what they know and what they are able to do, rather than giving students who use these resources an advantage over other students or artificially inflating their scores. Universal tools, designated supports, and accommodations minimize or remove barriers that could otherwise prevent students from demonstrating their knowledge, skills, and achievement in a specific content area.

5.6.2. Identification

All public school students participate in the CAASPP System, including students with disabilities and EL students. The Smarter Balanced Assessment Consortium's *Usability, Accessibility, and Accommodations Guidelines* (Smarter Balanced, 2020) and the CDE's Matrix One (CDE, 2019) are intended for school-level personnel and individualized education program (IEP) and Section 504 plan teams to select and administer the appropriate universal tools, designated supports, and accommodations as deemed necessary for individual students.[43]

The *Guidelines* apply to all students and promote an individualized approach to the implementation of assessment practices. Another web document, the *Smarter Balanced Resources and Practices Comparison Crosswalk* (Smarter Balanced, 2018), connects the assessment resources described in the *Guidelines* with associated classroom practices.

Another manual, the *Smarter Balanced Usability, Accessibility, and Accommodations Implementation Guide* (Smarter Balanced, 2014), provides suggestions for implementation of these resources. Test administrators are given the opportunity to participate in the Smarter Balanced practice and training tests so that students have the opportunity to familiarize themselves with a designated support or accommodation prior to testing.

[43] This technical report is based on the version of Matrix One that was available during the 2019–2020 CAASPP administration. Note that Matrix One has since been combined with the English Language Proficiency Assessments for California Matrix Four to form a single accessibility resources matrix, the California Assessment Accessibility Resources Matrix (CDE, 2020f).

5.6.3. Assignment

Once the student’s IEP or Section 504 plan team decided which accessibility resource(s) the student should use, LEA CAASPP coordinators and CAASPP test site coordinators used TOMS to assign designated supports and accommodations to students prior to the start of a test session.

There were three ways the student’s accessibility resource(s) could be assigned:

1. Using the ISAAP Tool to identify the accessibility resource(s) and then uploading the spreadsheet it creates into TOMS (This process is discussed in more detail in subsection *2.5.1 Resources for Selection of Accessibility Resources*.)
2. Using the Online Student Test Settings template to enter students’ assignments and then uploading the spreadsheet into TOMS
3. Entering assignments for each student individually in TOMS

If a student’s IEP or Section 504 plan team identified and designated a resource not identified in an accessibility matrix, the LEA CAASPP coordinator or CAASPP test site coordinator needed to submit a request for an unlisted resource to be approved by the CDE. The CDE then determined whether the requested unlisted resource changed the construct being measured after all testing was completed.

5.6.4. Usage of Accessibility Resources

After schools and LEAs assigned eligible students to accommodations or designated supports, CAI’s TDS provided and captured whether a certain accommodation or designated support (or multiple accommodations or designated supports) were used by a student as the student progressed through the test.

Table 5.6 and table 5.7 report the number of students who, based on the availability of data, were assigned to a certain accommodation or designated support and actually used this accommodation or designated support at least once in ELA and mathematics, respectively. Embedded accessibility resources are those that are part of the online TDS, whereas non-embedded accessibility resources are provided outside of that system.

Types of accommodations and designated supports—labeled “ACC” and “DS” in the *Resource Type* column—included in table 5.6 and table 5.7 are as follows:

- **Text-To-Speech:** Text is read aloud to the student via embedded text-to-speech technology.
- **American Sign Language (ASL):** ASL videos are available for any item that has a listening component. The ASL human signer and the signed test content are viewed on the same screen.
- **Print on Demand:** Paper copies of passages and stimuli, items, or all of these are printed for students.
- **Masking:** This resource involves blocking off content that is not of immediate need or that may be distracting to the student.
- **Audio Transcript:** This resource allows students to view a transcript of the audio content for the current test page. This is useful for students with visual impairment who are accustomed to accessing information presented via audio in the form of braille.

Table 5.6 Summary of Accommodations and Designated Supports Used by Students—ELA

Content Area	Grade	Opportunity	Accessibility Resource	Resource Type	Students Assigned	Students Used
ELA	3	CAT	Embedded American Sign Language	ACC	0	0
ELA	3	PT	Embedded American Sign Language	ACC	0	0
ELA	3	CAT	Embedded Audio Transcript	ACC	0	0
ELA	3	PT	Embedded Audio Transcript	ACC	0	0
ELA	3	CAT	Embedded Text-to-Speech Passages	ACC	24	13
ELA	3	PT	Embedded Text-to-Speech Passages	ACC	18	15
ELA	3	CAT	Non-Embedded Print on Demand	ACC	0	0
ELA	3	PT	Non-Embedded Print on Demand	ACC	0	0
ELA	3	CAT	Embedded Masking	DS	2	0
ELA	3	PT	Embedded Masking	DS	0	0
ELA	3	CAT	Embedded Text-to-Speech Items	DS	65	42
ELA	3	PT	Embedded Text-to-Speech Items	DS	52	27
ELA	4	CAT	Embedded American Sign Language	ACC	0	0
ELA	4	PT	Embedded American Sign Language	ACC	0	0
ELA	4	CAT	Embedded Audio Transcript	ACC	0	0
ELA	4	PT	Embedded Audio Transcript	ACC	0	0
ELA	4	CAT	Embedded Text-to-Speech Passages	ACC	23	19
ELA	4	PT	Embedded Text-to-Speech Passages	ACC	13	12
ELA	4	CAT	Non-Embedded Print on Demand	ACC	1	0
ELA	4	PT	Non-Embedded Print on Demand	ACC	1	0
ELA	4	CAT	Embedded Masking	DS	1	0
ELA	4	PT	Embedded Masking	DS	0	0
ELA	4	CAT	Embedded Text-to-Speech Items	DS	35	22
ELA	4	PT	Embedded Text-to-Speech Items	DS	24	9

Table 5.6 *(continuation one)*

Content Area	Grade	Opportunity	Accessibility Resource	Resource Type	Students Assigned	Students Used
ELA	5	CAT	Embedded American Sign Language	ACC	0	0
ELA	5	PT	Embedded American Sign Language	ACC	0	0
ELA	5	CAT	Embedded Audio Transcript	ACC	0	0
ELA	5	PT	Embedded Audio Transcript	ACC	0	0
ELA	5	CAT	Embedded Text-to-Speech Passages	ACC	29	22
ELA	5	PT	Embedded Text-to-Speech Passages	ACC	22	18
ELA	5	CAT	Non-Embedded Print on Demand	ACC	0	0
ELA	5	PT	Non-Embedded Print on Demand	ACC	0	0
ELA	5	CAT	Embedded Masking	DS	2	0
ELA	5	PT	Embedded Masking	DS	2	0
ELA	5	CAT	Embedded Text-to-Speech Items	DS	80	44
ELA	5	PT	Embedded Text-to-Speech Items	DS	56	21
ELA	6	CAT	Embedded American Sign Language	ACC	1	1
ELA	6	PT	Embedded American Sign Language	ACC	0	0
ELA	6	CAT	Embedded Audio Transcript	ACC	2	0
ELA	6	PT	Embedded Audio Transcript	ACC	0	0
ELA	6	CAT	Embedded Text-to-Speech Passages	ACC	24	20
ELA	6	PT	Embedded Text-to-Speech Passages	ACC	31	24
ELA	6	CAT	Non-Embedded Print on Demand	ACC	1	0
ELA	6	PT	Non-Embedded Print on Demand	ACC	0	0
ELA	6	CAT	Embedded Masking	DS	3	0
ELA	6	PT	Embedded Masking	DS	5	1
ELA	6	CAT	Embedded Text-to-Speech Items	DS	24	19
ELA	6	PT	Embedded Text-to-Speech Items	DS	48	15

Table 5.6 *(continuation two)*

Content Area	Grade	Opportunity	Accessibility Resource	Resource Type	Students Assigned	Students Used
ELA	7	CAT	Embedded American Sign Language	ACC	0	0
ELA	7	PT	Embedded American Sign Language	ACC	0	0
ELA	7	CAT	Embedded Audio Transcript	ACC	0	0
ELA	7	PT	Embedded Audio Transcript	ACC	0	0
ELA	7	CAT	Embedded Text-to-Speech Passages	ACC	62	35
ELA	7	PT	Embedded Text-to-Speech Passages	ACC	60	46
ELA	7	CAT	Non-Embedded Print on Demand	ACC	0	0
ELA	7	PT	Non-Embedded Print on Demand	ACC	0	0
ELA	7	CAT	Embedded Masking	DS	21	2
ELA	7	PT	Embedded Masking	DS	25	3
ELA	7	CAT	Embedded Text-to-Speech Items	DS	76	38
ELA	7	PT	Embedded Text-to-Speech Items	DS	86	32
ELA	8	CAT	Embedded American Sign Language	ACC	0	0
ELA	8	PT	Embedded American Sign Language	ACC	0	0
ELA	8	CAT	Embedded Audio Transcript	ACC	0	0
ELA	8	PT	Embedded Audio Transcript	ACC	0	0
ELA	8	CAT	Embedded Text-to-Speech Passages	ACC	71	42
ELA	8	PT	Embedded Text-to-Speech Passages	ACC	42	33
ELA	8	CAT	Non-Embedded Print on Demand	ACC	0	0
ELA	8	PT	Non-Embedded Print on Demand	ACC	0	0
ELA	8	CAT	Embedded Masking	DS	31	3
ELA	8	PT	Embedded Masking	DS	28	1
ELA	8	CAT	Embedded Text-to-Speech Items	DS	104	58
ELA	8	PT	Embedded Text-to-Speech Items	DS	67	20

Table 5.6 *(continuation three)*

Content Area	Grade	Opportunity	Accessibility Resource	Resource Type	Students Assigned	Students Used
ELA	11	CAT	Embedded American Sign Language	ACC	7	2
ELA	11	PT	Embedded American Sign Language	ACC	1	0
ELA	11	CAT	Embedded Audio Transcript	ACC	4	0
ELA	11	PT	Embedded Audio Transcript	ACC	0	0
ELA	11	CAT	Embedded Text-to-Speech Passages	ACC	316	150
ELA	11	PT	Embedded Text-to-Speech Passages	ACC	189	108
ELA	11	CAT	Non-Embedded Print on Demand	ACC	1	0
ELA	11	PT	Non-Embedded Print on Demand	ACC	1	0
ELA	11	CAT	Embedded Masking	DS	47	1
ELA	11	PT	Embedded Masking	DS	28	2
ELA	11	CAT	Embedded Text-to-Speech Items	DS	442	182
ELA	11	PT	Embedded Text-to-Speech Items	DS	259	66

Table 5.7 Summary of Accommodations and Designated Supports Used by Students—Mathematics

Content Area	Grade	Opportunity	Accessibility Resource	Resource Type	Students Assigned	Students Used
Mathematics	3	CAT	Embedded American Sign Language	ACC	0	0
Mathematics	3	PT	Embedded American Sign Language	ACC	0	0
Mathematics	3	CAT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	3	PT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	3	CAT	Embedded Masking	DS	1	0
Mathematics	3	PT	Embedded Masking	DS	0	0
Mathematics	3	CAT	Embedded Text-to-Speech	DS	43	28
Mathematics	3	PT	Embedded Text-to-Speech	DS	44	28
Mathematics	4	CAT	Embedded American Sign Language	ACC	0	0
Mathematics	4	PT	Embedded American Sign Language	ACC	0	0
Mathematics	4	CAT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	4	PT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	4	CAT	Embedded Masking	DS	0	0
Mathematics	4	PT	Embedded Masking	DS	0	0
Mathematics	4	CAT	Embedded Text-to-Speech	DS	21	19
Mathematics	4	PT	Embedded Text-to-Speech	DS	18	16
Mathematics	5	CAT	Embedded American Sign Language	ACC	0	0
Mathematics	5	PT	Embedded American Sign Language	ACC	0	0
Mathematics	5	CAT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	5	PT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	5	CAT	Embedded Masking	DS	0	0
Mathematics	5	PT	Embedded Masking	DS	0	0
Mathematics	5	CAT	Embedded Text-to-Speech	DS	38	25
Mathematics	5	PT	Embedded Text-to-Speech	DS	32	14

Table 5.7 *(continuation one)*

Content Area	Grade	Opportunity	Accessibility Resource	Resource Type	Students Assigned	Students Used
Mathematics	6	CAT	Embedded American Sign Language	ACC	1	1
Mathematics	6	PT	Embedded American Sign Language	ACC	0	0
Mathematics	6	CAT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	6	PT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	6	CAT	Embedded Masking	DS	1	0
Mathematics	6	PT	Embedded Masking	DS	2	0
Mathematics	6	CAT	Embedded Text-to-Speech	DS	16	13
Mathematics	6	PT	Embedded Text-to-Speech	DS	17	7
Mathematics	7	CAT	Embedded American Sign Language	ACC	0	0
Mathematics	7	PT	Embedded American Sign Language	ACC	0	0
Mathematics	7	CAT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	7	PT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	7	CAT	Embedded Masking	DS	20	3
Mathematics	7	PT	Embedded Masking	DS	12	2
Mathematics	7	CAT	Embedded Text-to-Speech	DS	36	20
Mathematics	7	PT	Embedded Text-to-Speech	DS	23	17
Mathematics	8	CAT	Embedded American Sign Language	ACC	0	0
Mathematics	8	PT	Embedded American Sign Language	ACC	0	0
Mathematics	8	CAT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	8	PT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	8	CAT	Embedded Masking	DS	25	0
Mathematics	8	PT	Embedded Masking	DS	17	1
Mathematics	8	CAT	Embedded Text-to-Speech	DS	75	36
Mathematics	8	PT	Embedded Text-to-Speech	DS	70	35

Table 5.7 *(continuation two)*

Content Area	Grade	Opportunity	Accessibility Resource	Resource Type	Students Assigned	Students Used
Mathematics	11	CAT	Embedded American Sign Language	ACC	3	1
Mathematics	11	PT	Embedded American Sign Language	ACC	2	0
Mathematics	11	CAT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	11	PT	Non-Embedded Print on Demand	ACC	0	0
Mathematics	11	CAT	Embedded Masking	DS	295	17
Mathematics	11	PT	Embedded Masking	DS	374	19
Mathematics	11	CAT	Embedded Text-to-Speech	DS	473	49
Mathematics	11	PT	Embedded Text-to-Speech	DS	499	53

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.

California Code of Regulations, Title 5, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, Section 855 (n.d.) ~~https://govt.westlaw.com/calregs/Document/I2DB6A0BAA54F41B69BAF5553FABBE5EF?viewType=FullText&originationContext=documenttoc&transitionType=CategoryPageItem&contextData=(sc.Default) (This link is not active)~~

California Code of Regulations, Title 5, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, Section 859 (n.d.) ~~https://govt.westlaw.com/calregs/Document/I2DB6A0BAA54F41B69BAF5553FABBE5EF?viewType=FullText&originationContext=documenttoc&transitionType=CategoryPageItem&contextData=(sc.Default) (this link is not active)~~

California Department of Education. (2019). *Matrix one: Universal tools, designated supports, and accommodations for the California Assessment of Student Performance and Progress for 2019–20*. Sacramento, CA: California Department of Education. https://bit.ly/3h0EkD2

California Department of Education. (2020a). *CAASPP and ELPAC accessibility guide for online testing*. Sacramento, CA: California Department of Education. https://bit.ly/3usOzny

California Department of Education. (2020b). *CAASPP and ELPAC technical specifications and configuration guide*. Sacramento, CA: California Department of Education. https://bit.ly/3upqiyD

California Department of Education. (2020c). *CAASPP and ELPAC Test Operations Management System user guide*. Sacramento, CA: California Department of Education. https://bit.ly/2QKKJaV

California Department of Education. (2020d). *CAASPP online test administration manual, 2019–2020 test administration.* Sacramento, CA: California Department of Education. https://bit.ly/3eXYBXc

California Department of Education. (2020e). *CAASPP security incidents and appeals procedure guide, 2019–2020 test administration.* Sacramento, CA: California Department of Education. https://bit.ly/3b5saoH

California Department of Education. (2020f). *California assessment accessibility resources matrix*. Sacramento, CA: California Department of Education. https://bit.ly/3vHmKbm

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service. https://www.ets.org/s/about/pdf/standards.pdf

Khattri, N., Reeve, A., & Kane, M. (1998). Principles and so practices of performance assessment. Mahwah, NJ: Routledge.

Office of Governor Gavin Newsom. (2020). *Governor Newsom issues executive order to suspend standardized testing for students in response to COVID-19 outbreak* [Press release]. https://www.gov.ca.gov/2020/03/18/governor-newsom-issues-executive-order-to-suspend-standardized-testing-for-students-in-response-to-covid-19-outbreak/

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations implementation guide.* Los Angeles: Smarter Balanced Assessment Consortium. https://bit.ly/3xN8U9b

Smarter Balanced Assessment Consortium. (2018). *Smarter Balanced resources and practices comparison crosswalk.* Los Angeles: Smarter Balanced Assessment Consortium. https://bit.ly/3h5xBHS

Smarter Balanced Assessment Consortium. (2020). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines.* Los Angeles: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

Sympson, J., & Hetter, R. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings from the 27th Annual Meeting of the Military Testing Association* (pp. 973–77). San Diego, CA: Navy Personnel Research and Development Center.

Chapter 6: Standard Setting

This chapter briefly discusses the standard setting process outlined by Smarter Balanced.

6.1. Description

Standard setting, which also is referred to as achievement level setting, refers to a class of methodologies by which one or more thresholds are used to determine achievement levels. The Smarter Balanced Assessment Consortium set four achievement levels—*Standard Not Met, Standard Nearly Met*, *Standard Met,* and *Standard Exceeded*—with three threshold cuts for each grade and content area.

In coordination with its member states, the Smarter Balanced Assessment Consortium implemented an extensive achievement-level-setting process involving software development, item mapping, review panels, committees, workshops, and extensive validity research to set the final thresholds and achievement level descriptors. For detailed information regarding this process, refer to Chapter 10 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016).

Reference

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: 2013–14 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf

Chapter 7: Scoring and Reporting

To determine individual students' scores for the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced Summative Assessments, student item responses were scored and individual student scores were calculated based on the item responses. In addition, student test scores were aggregated to produce information for schools and local educational agencies (LEAs).

This chapter summarizes the types of scores and score reports that are produced at the end of each administration of the Smarter Balanced Summative Assessments for English language arts/literacy (ELA) and mathematics.

7.1. Approach to Scoring Item Responses

7.1.1. Structure of the Assessments

To understand the basis of the scoring approach, an understanding of the structure of the CAASPP Smarter Balanced online summative assessments is necessary. These assessments are designed to gather evidence that can be used to make inferences about student mastery of the Common Core State Standards (CCSS). The assessments are based on claims and targets.

Claims are inferences made about a student based on the student's test score. They are broad statements about learning outcomes. These statements require evidence that articulates the types of data and observations that support interpretations of progress toward the achievement of the claim. Claims identify the set of knowledge and skills being measured. The following is an example of a mathematics claim:

> **Claim 1: Concepts and Procedures—**Students can explain and apply mathematical concepts and carry out mathematical procedures with precision and fluency.

Targets describe the evidence that can be used to support a claim about a student. Targets are specific to claims. The following is a target associated with the previous claim:

> **Target C—**Understand the connections between proportional relationships, lines, and linear equations.

The items are designed based on a variety of task models that define item characteristics such as item type, allowable stimuli, prompt feature, and item interactions.

7.1.2. Certification of the Scoring System

ETS staff from the Assessment and Learning Technology Research & Development, Enterprise Score Key Management (eSKM), Psychometric Analysis & Research (PAR), Constructed Response Scoring, Systems & Capabilities, and Information Technology divisions participated in the certification of the scoring system. Each team followed procedures required by the ETS Office of Quality for operational readiness and Standard 7.8 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

ETS staff reviewed operational answer keys and scoring rubrics provided by Smarter Balanced staff. In addition, item parameter estimates for items were loaded into the ETS operational scoring system. Central aspects of the validity of the CAASPP online summative test scores are the degree to which scoring rubrics are related to the appropriate assessment targets and claims based on Smarter Balanced assessments. A key facet of validity is the degree to which scoring rules are applied accurately throughout the scoring sessions.

7.1.3. Types of Item Responses

In accordance with the Smarter Balanced Online Summative Assessment specifications, students are administered a computer adaptive test (CAT) component and a selected performance task (PT) (Smarter Balanced, 2017a through 2017i [ELA]; and 2018a through 2018k [mathematics]). The combination of the CAT and the PT components fulfills the content requirements for the test blueprint (refer to appendix 2.A).

CAASPP Smarter Balanced online summative assessments include traditional selected response items, short constructed-response (CR) items, writing extended response (WER) items, and technology-enhanced items. Some items are machine-scored, which means that they can be scored by the test delivery system (TDS). Other items are scored with the artificial intelligence (AI) scoring engine; still others are human scored by a trained rater. The scoring approach used depends on the item type and scoring requirements provided by the Smarter Balanced item specifications. Table 7.1 lists the types of items that are machine-scored.

Table 7.1 Machine-Scored Online Item Types

Item Type	Description	Content Area
Equation	Students enter an equation or numeric response using an on-screen panel containing mathematical characters.	Mathematics only
Evidence-based selected response	This is a two-part item: the student responds to a multiple-choice item and then responds to a multiple-select item.	ELA only
Grid item—Drag and drop	Students respond by dragging and dropping a single choice (“source”) into the appropriate location (“target”). The scoring key is a set of numeric identifiers that specifies which source needs to be placed in which target to answer the item correctly.	Mathematics only
Grid item—Graphing	Students plot points, lines, and multisegmented lines on a graph. Items can be answered by looking at a graph. For some items, students must manipulate the elements in the graph to respond.	Mathematics only
Hot text	Students are presented with a stem that contains multiple underlined words or phrases from which students select sections of text or drag-and-drop sections of text.	ELA only
Multiple choice	Three to five answer choices are provided, and students can select only one choice to respond.	ELA and mathematics

Table 7.1 *(continuation)*

Item Type	Description	Content Area
Multiple select	Five to eight answer choices are provided, and students are instructed to select one or more choices to respond. These item types can have multiple keys; students may be awarded partial credit for partially correct answers or may need to select all correct answers to receive credit.	ELA and mathematics
Table interaction	Students are required to respond by making a keyboard entry into one or more cells in a table grid. The response can be restricted to one selection of row, column, or table, or no restrictions.	Mathematics only

Item types that require students to provide a response by writing words or numbers are called "constructed-response" items. Both the CAT and the PT include CR items. The CAT section contains both machine-scored items worth 0–1 or 0–2 points, as well as short-text items worth 0–2 points. The PT section contains machine-scored items, short-text items worth 0–2 points, and WER items worth 0–6 points.[44] A small number of mathematics PTs include CR items with a 0–4 point range. CR items for CAASPP Smarter Balanced assessments include the following item types:

- ***Short-answer text response items*** require students to respond with words, phrases, short sentences, or mathematical expressions. These items have a value of 0–2 points, with a small number of mathematics short-answer items having values ranging from 0–4 points. These items are scored holistically based on a rubric. Holistic scoring gives students a single, overall assessment score for the response as a whole.
- ***WER items (full-write response)*** require students to write one or more paragraphs. The WER is scored for three dimensions of writing (purpose, focus, and organization; evidence and elaboration; and conventions). These items are scored analytically based on rubrics; readers assign a score based on each dimension.

7.1.4. Scoring the Item Types

The specifications regarding which CR items are eligible for machine scoring are described in an ETS memorandum (ETS, 2015a).

ETS staff review operational answer keys and scoring rubrics provided by the Smarter Balanced Assessment Consortium and follow scoring specifications to enter scores into the ETS operational scoring system. The target of the scoring specifications is to optimize the validity, reliability, and efficiency of scoring. A central aspect of the validity of the test scores is the degree to which scoring rubrics are related to the appropriate assessment targets, depth of knowledge, and claims based on Smarter Balanced assessments. A key facet of

[44] Smarter Balanced blueprints describe that three WER items are worth 0–10 points, including one item with 2 points and two items with 4 points each. The scoring specifications from Smarter Balanced instruct combining the two 4-point items to take the average of the two for scoring. As a result, the total WER items are worth 0–6 points.

validity is whether the scoring rules are applied accurately during the scoring sessions. The validity and reliability of the scoring of CR items are evaluated in *Chapter 8: Analyses*.

The scoring specifications include details on the type of training provided to raters, the rater screening and qualification process, and the metrics used to evaluate rater accuracy that apply to the human scoring of CR items. ETS' subcontractor, Measurement Incorporated (MI), scores the machine-scorable CR items utilizing AI scoring engines.

The scoring rubrics for the short-answer items are holistic, with the exception of the rubrics used to score the ELA PT full-write response, which are analytic. The full-write response item is also referred to as a WER item. An example of scoring rubrics of the WER items is available in the *Smarter Balanced Hand-scoring Rules* (Smarter Balanced, 2014c).

7.2. Quality Control of Scoring

7.2.1. Human Scoring

7.2.1.1. Quality Control in the Scoring Process

In general, the scoring model is based on scoring one item at a time (i.e., raters score responses to a single prompt until there are no more responses to that prompt during the shift). However, some mathematics PT items have scoring dependencies, which means that students base their calculations and responses on the answers to previous items associated with the PT. When these items are human scored, all the items in the PT, along with the student responses, are provided to the rater. This allows the rater to evaluate dependent items based on the previous items that serve as the basis for the dependent item.

The three traits measured by the extended writing tasks (full-write responses)—Organization and Purpose, Evidence and Elaboration or Development and Elaboration, and Conventions—are evaluated together by a single rater. The rater assigns a separate trait score for each of the three traits.

Items are scored by a team of 5 to 10 raters under the direction of a scoring leader. Scoring leaders are supervised by chief scoring leaders. Each chief scoring leader is responsible for multiple teams in a specific content area and grade band. Responses to individual prompts are assigned to teams of no fewer than three raters. If there is not a sufficient number of responses during a shift to occupy at least three raters, the responses are held until a sufficient number to occupy at least three raters is reached. Each rater works individually on the rater's own device to read each student response and enter a score for each item.

7.2.1.2. Quality Control Related to Raters

ETS has developed a variety of procedures to control the quality of ratings and monitor the consistency of scores provided by raters. These procedures specify rater qualifications and procedures for rater certification and daily rater calibration. Raters are required to demonstrate their accuracy by passing a certification test before ETS assigns them to score a specific assessment and by passing a shorter, more focused calibration test before each scheduled scoring session. Rater certification and calibration are key components in maintaining quality and consistency.

Scoring leaders monitor raters' performance by reading their scored responses to determine whether the rater assigned the correct rating. Some scoring leaders choose to read the response before finding out what score the rater has assigned; others choose to know what score the rater has assigned before reading the response. Refer to the *Monitoring Raters* subsection for more information on this process.

7.2.1.3. Rater Qualification

Raters should meet the following requirements prior to being hired:

- All candidates must have a bachelor's degree and be eligible to work in the United States (and are e-verified prior to hire).
- Teaching experience is strongly preferred.
- Graduate students and substitute teachers are encouraged to apply.
- Bilingual English and Spanish speakers are encouraged to apply.
- Candidates must complete training and achieve qualifications through the certification process.

Table 7.2 through table 7.4 summarize the overall active human raters who were trained and prepared for the 2019–2020 scoring for ETS (table 7.2), MI (table 7.3), and combined (table 7.4) across both organizations. Due to the suspension of the testing window in the 2019–2020 administration, a small number of raters actually scored the Smarter Balanced assessments.

Table 7.2 Summary of Characteristics of ETS Human Raters Scoring CAASPP Smarter Balanced Assessments

Characteristic	N	Percent
Total raters scoring in 2019–2020	**265**	**100%**
Fluent in Spanish and expressed interest in scoring assessments in Spanish	4	2%
Experience teaching in a kindergarten through grade twelve (K–12) school	83	31%
Currently works in a K–12 school in California	45	17%
Others—Not meeting any of the previous criteria	133	50%

Table 7.3 Summary of Characteristics of MI Human Raters Scoring CAASPP Smarter Balanced Assessments

Characteristic	N	Percent
Total raters scoring in 2019–2020	**194**	**100%**
Fluent in Spanish and expressed interest in scoring assessments in Spanish	0	0%
Experience teaching in a K–12 school	2	1%
Currently works in a K–12 school in California	0	0%
Others—Not meeting any of the previous criteria	192	99%

Table 7.4 Summary of Characteristics of ETS and MI Human Raters Scoring CAASPP Smarter Balanced Assessments

Characteristic	N	Percent
Total raters scoring in 2019–2020	**2,520**	**100%**
Fluent in Spanish and expressed interest in scoring assessments in Spanish	58	2%
Experience teaching in a K–12 school	905	33%
Currently works in a K–12 school in California	510	19%
Others—Not meeting any of the previous criteria	1,241	46%

California educators should meet the following qualifications:

- Must have a current California teaching credential (although California charter school teachers may or may not have a teaching credential for the scoring participation requirement)
- May be retired educators and other administrative staff with a teaching credential who are not current classroom teachers
- Must have achieved, at minimum, a bachelor’s degree

All team leaders and raters are required to qualify before scoring and are informed of what they are expected to achieve to qualify. Refer to *7.3 Rater Training* for a more complete description of this training.

ETS makes a distinction between training sets and calibration (qualification) sets. Training sets are nonconsequential, as the sets provide the raters the opportunity to score sample papers and receive feedback, including the correct score point and rationale associated with that score point and the sample paper. Training sets are a learning tool that the raters are required to complete. Nonadjacent scores may occur in the training sets, as minimum agreement standards are not part of training sets.

Upon completion of the required training sets, raters move on to a consequential calibration set that will determine rater eligibility for operational scoring of a particular item type. Calibration (qualification) sets have minimum agreement levels that are enforced, and nonadjacent scores are not allowed. All 0–4 and 0–3 point items adhere to the Smarter Balanced recommendation of a 70 percent exact and 0 percent discrepant (nonadjacent) agreement rate to score.

The standards, provided in table 7.5, are minimum qualification expectations for the various score point ranges and the qualification standard in terms of the percent of exact agreement. This qualification set, like the validity papers discussed in the next subsection (*Monitoring Raters*), has been previously scored by scoring experts. Raters must score the papers in the same manner according to the percentage of agreements listed in table 7.5.

Table 7.5 Rater Qualification Standard for Agreement with Correct Scores

Score Point Range	Qualification Standard (Exact Agreement)
0–1	90%
0–2	80%
0–3	70%
0–4	60%

The qualification process is conducted through an online system that captures the results electronically for each individual trainee.

7.2.1.3.1. Monitoring Raters

ETS staff created performance scoring reports so that scoring leaders can monitor the daily human-scoring process and plan any retraining activities, if needed. For monitoring interrater reliability, 10 percent of the student responses that have already been scored by the raters are randomly selected for a second scoring and assigned to raters by the scoring system; this process is referred to as back-reading. The second rater is unaware of the first rater's score. The evaluation of the response from the second rater is compared to that of the first rater. Scoring leaders and chief scoring leaders provide second reads during their shifts for additional quality review.

Validity papers, carefully selected and prescored by scoring experts, also are used to monitor rater performance. They are inserted randomly into each rater's scoring queue at a rate of 9 percent of the total papers scored by a rater during a rater's shift. Validity papers serve as another real-time evaluation of rater accuracy.

Real-time management tools allow everyone, from scoring leaders to content specialists, access to

- the overall interrater reliability rate, which measures the percentage of agreement when the scores assigned by raters are compared to the scores assigned by other raters, including scoring managers;
- the read rate, which is defined as the number of responses read per hour;
- the individual and overall percentage of agreement for validity paper ratings; and
- the projected date for completion of the scoring for a specific prompt or task.

7.2.2. Quality Control of Artificial Intelligence Scoring

The responses to some of the short-answer items on the CAASPP Smarter Balanced Online Summative Assessments are scored by MI's AI scoring engine. MI's AI scoring engine analyzes a training set of papers and calculates features that pertain to the content in question for each individual item. The scoring engine then sends the features to dozens of different models that compete to determine which ones can best associate the features with the corresponding human-assigned scores. The strongest models then are blended automatically to create a final model that retains the best elements from the various algorithms. After the model is built, the model elements are selected to maximize scoring accuracy for the response data.

The goal of MI's AI scoring is to provide scores that are statistically comparable to those obtained from human raters. To ensure this continues to be true after the initial model development, MI conducts ongoing quality checks to ensure that the scoring models consistently perform as expected. Statistics such as perfect or adjacent agreement, the Pearson product-moment correction coefficient, or the quadratic weighted kappa (QWK) are used for comparing the agreement between AI scoring and human scoring. MI meets with the California Department of Education (CDE) to specify the evaluation metric and expected level of accuracy for AI scoring. If an analysis of the human and AI agreement for an item indicates that the scoring engine needs to be adjusted, MI recalibrates the scoring model for that item. Using a new set of training papers (500–1,000, depending on the item type and complexity), MI retrains and recalibrates the scoring model until it meets or exceeds the agreement level established by the CDE, using agreed-upon evaluation metrics.

ETS and MI have developed and documented a proprietary standardized system for addressing the complexities inherent in monitoring and maintaining quality throughout largescale, human-scoring projects. ETS processes ensure that both organizations maintain a quality assurance system through 10 percent of AI-scored items being scored by a human rater and used for agreement sample analysis.

7.2.3. Score Verification Process

Various measures are taken to ascertain that the scoring keys are applied to the student responses as intended and the student overall and claim scores are computed accurately. ETS' eSKM system uses scoring specifications provided by psychometricians to derive all types of scores, such as theta scores, overall scale scores, claim scale scores, achievement levels, etc., from individual item scores. A series of quality control checks are carried out by ETS psychometricians to ensure the accuracy of each score. The details are described in *9.4 Quality Control of Psychometric Processes*.

7.2.4. Interrater Reliability Results

At least 10 percent of the test responses of CR items in ELA and mathematics were scored independently by a second reader. ETS and MI used at least 30 validity papers that covered the full range of scores. Validity sets were monitored throughout the administration and postadministration periods for performance. Supplemental samples were added as needed. The statistics for interrater reliability for all items at all grades are usually presented, except for this year due to extremely insufficient data. These statistics include the percentage of perfect agreement and adjacent agreement between the two readers and the QWK statistic. QWK is a statistic used to measure the degree of association between two ratings with values ranging from 0.0 (indicating no agreement) to 1.0 (indicating perfect agreement). Refer to subsection *8.5.6.3 Quadratic Weighted Kappa (QWK)* for detailed information on QWK.

Smarter Balanced provided flagging criteria (Smarter Balanced, 2016a) based on the statistics that follow for identifying items to be reviewed for potential elimination after scoring was completed. Polytomous items are flagged if any of the following conditions occur:

- Adjacent plus exact agreement < 0.80
- Exact agreement < 0.60
- QWK < 0.20

Dichotomous items are flagged if either of the following conditions occur:

- Exact agreement < 0.80
- QWK < 0.20

ETS follows the Smarter Balanced recommended exact and adjacent agreement rates criteria. However, ETS uses a more stringent agreement criterion for QWK; that is, dichotomous and polytomous items are flagged if their QWK is below 0.70.

7.3. Rater Training

7.3.1. Training Overview

7.3.1.1. English Language Arts/Literacy

To score ELA items, raters received training based on the task model used to design a group of items with similar characteristics. Raters were first trained by grade band, claim, and target and then applied generic rubrics to score the responses. For example, raters were trained to score Claim 1 Target 5 responses for grade band three through five. The training was further focused based on the item type—short answer or WER—as well as the grade band (grades three through five, six through eight, or grade eleven).

“Baseline” training sets of papers, also called anchors, as well as scoring rubrics, were provided to raters based on writing purpose (e.g., informational or explanatory writing) for the WER items. Baseline anchor and training sets of papers consisted of student responses that have been scored, reviewed by scoring experts, and selected to be exemplars of each score point. Often, these were annotated to provide a specific explanation of how the paper exemplified a response that should earn that particular score. Raters could refer to these sets to increase their understanding of how to accurately apply the scoring rubric.

Additional anchor and training sets were created for periodic qualification, a process in which raters engaged in a brief training and then scored a prescored set of papers to ensure they were scoring accurately before their shift begins.

Qualification and validity sets were provided for each WER essay type. Anchor and training sets were also provided for the task models associated with the ELA short-answer items in the CAT and PT sections. For the ELA short-answer items in the CAT and the PT sections, raters received training for a grade band—grades three through five, six through eight, or grade eleven—instead of a grade level.

Although training was provided at the task-model level, rater qualification occurred on an item type and grade-span basis for all ELA human-scored items. Qualification and validity papers were provided for each ELA CR item. Raters qualified for each item type within a specific grade band before being assigned to score that item type (American Institutes for Research [AIR], 2015).

7.3.1.2. Mathematics

To score mathematics items, raters receive training and must qualify on all task models before scoring items on any task model. Similar to the training procedures for ELA, for mathematics, the Smarter Balanced Assessment Consortium provides anchor papers, the baseline paper, and training sets for the task models. The consortium also provides item-specific rubrics and item-specific validation sets for all mathematics items (AIR, 2015).

7.3.2. Training Process: English Language Arts/Literacy Performance Task Extended Writing Tasks

Baseline anchor sets for each writing purpose (e.g., informational writing or explanatory writing) were used to train raters on each of the writing traits—Organization and Purpose, Evidence and Elaboration or Development and Elaboration, and Conventions—within a particular grade band. The writing purposes are narrative, informational, and opinion at grades three through five; narrative, informational, and argumentative at grades six through eight; and explanatory and argumentative at grade eleven.

For all writing purposes, Organization and Purpose is the first trait and Conventions is the third trait. Evidence and Elaboration is the second trait for the opinion, argumentative, informational, and explanatory writing purposes. Development and Elaboration is the second trait for the narrative writing purpose.

Writing traits for opinion, argumentative, informational, or explanatory writing are

- Organization and Purpose,
- Evidence and Elaboration, and
- Conventions.

Writing traits for narrative writing are

- Organization and Purpose,
- Development and Elaboration, and
- Conventions.

A chart that presents the traits to their purposes is shown in figure 7.1.

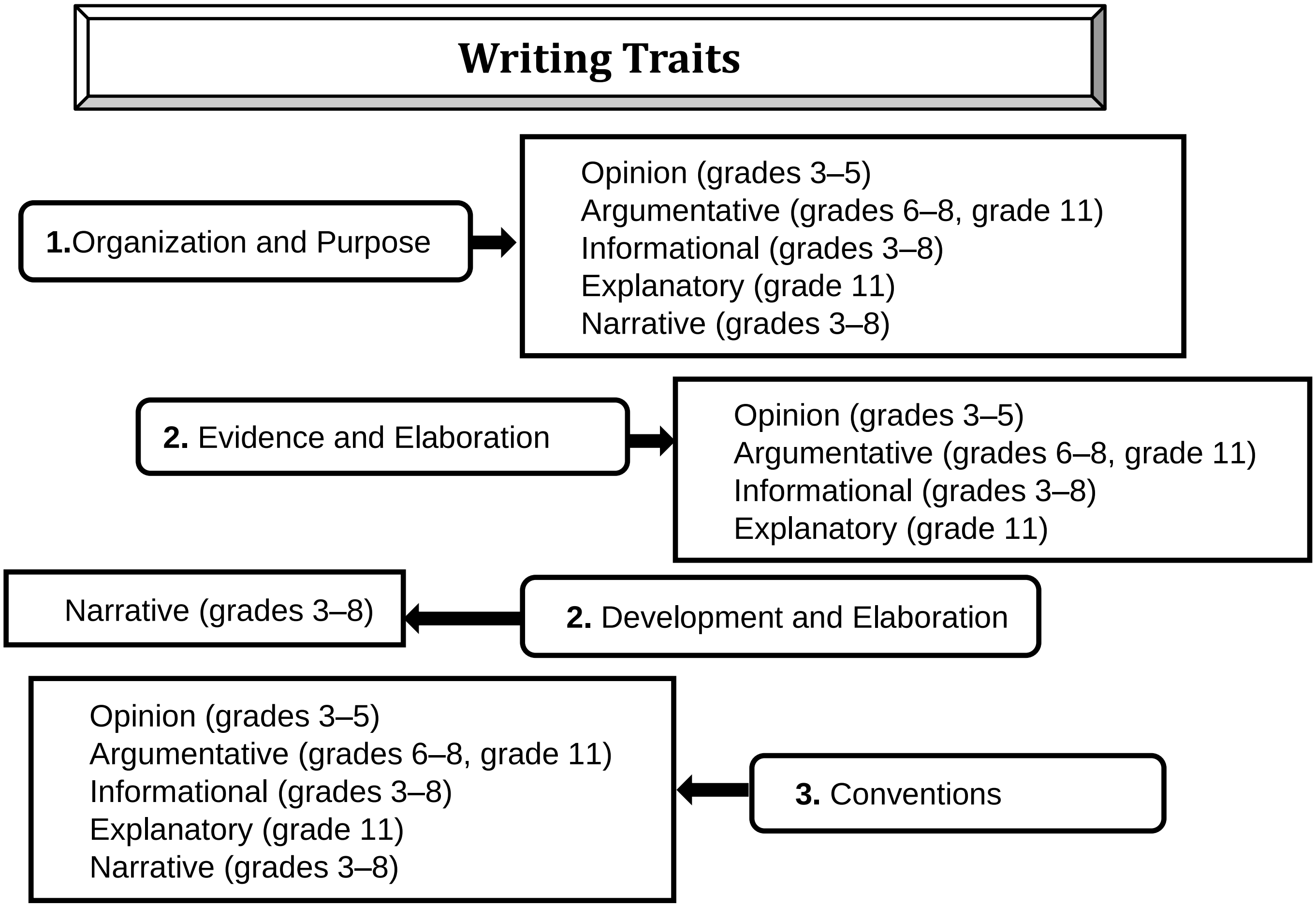
Writing Traits
1.Organization and Purpose
Opinion (grades 3–5)
Argumentative (grades 6–8, grade 11)
Informational (grades 3–8)
Explanatory (grade 11)
Narrative (grades 3–8)
2. Evidence and Elaboration
Opinion (grades 3–5)
Argumentative (grades 6–8, grade 11)
Informational (grades 3–8)
Explanatory (grade 11)
2. Development and Elaboration
Narrative (grades 3–8)
3. Conventions
Opinion (grades 3–5)
Argumentative (grades 6–8, grade 11)
Informational (grades 3–8)
Explanatory (grade 11)
Narrative (grades 3–8)

Figure 7.1 Writing traits

The training process for extended writing tasks is described next. The training steps are described in the following list.

Training steps:

1. Trainees read the task, rubrics, and source materials for the WER items in a particular grade band and writing purpose (for example, grades three through five informational). Trainees read sample responses and annotations.
2. Trainees read a training set of five responses to the same item (Essay 1) and score those responses for Conventions.
3. Trainees review the correct scores and the scoring rationale for the Conventions scores for those responses.
4. Trainees read another training set of five responses to that item (Essay 1) and score those responses for Organization and Purpose. They then review the correct scores and the scoring rationale for the Organization and Purpose scores for those responses.
5. Trainees read another training set of five responses to that item (Essay 1) and score those responses for Evidence and Elaboration. They then review the correct scores and the scoring rationale for the Evidence and Elaboration scores for those responses.
6. Trainees read another training set of five responses to that item (Essay 1) and score each of those responses for all three traits.
7. Trainees review the scoring rationale for the training responses and answer training questions.
8. Trainees score a qualification round (10 papers) for all three traits for Essay 1.
9. Qualified raters—those who meet the standard in the qualification round—begin scoring.
10. Trainees who do not meet the qualification standard on their first attempt have an opportunity to review correct scores and the scoring rationale with a scoring leader before making a second attempt.

The training materials are described in the following list.

Materials for training raters of WER items, at each grade level:

1. Baseline anchor sets approved during Smarter Balanced pre–range finding (Range finding activities include the review of student responses against item rubrics, the validation of rubric effectiveness, and the selection of anchor papers used by human scoring for the larger population of responses.)
2. Field test prompt and stimulus materials
3. Purpose- and task-specific rubrics
4. Conventions charts approved by the Smarter Balanced Assessment Consortium
5. Supplemental scoring guidelines approved by the Smarter Balanced Assessment Consortium
6. Training sets specific to the first WER task for each grade and purpose

7. Qualification sets generally administered in two rounds of approximately 10 responses per WER task

7.3.3. Training Process: English Language Arts/Literacy Short-Answer Items

The process for training raters to score short-answer items is also organized by grade band (grades three through five, six through eight, or eleven). These training steps are described in the following list.

Training steps:

1. Trainees read the rubrics and scoring notes for the short-answer items in a particular grade band and purpose category (for example, grades three through five evidence). Trainees read sample responses to a prompt and the associated annotations.
2. Trainees review the scoring rationale for each of the anchors (i.e., anchor sets for the claim, target, and subclaim).
3. Trainees score the training set (5–10 papers) for the short-answer claim, target, and subclaim.
4. Trainees review the correct scores and scoring rationale for the training set.
5. Trainees read the prompt, source materials, or stimuli for the first short-answer item in the claim, target, and subclaim (e.g., Grade 6, Claim 1, Reading Item 1).
6. Trainees score a qualification round.
7. Qualified raters begin scoring.
8. Trainees who do not meet the qualification standard on their first attempt have an opportunity to review correct scores and the scoring rationale with a scoring leader before making a second attempt.

The training materials are described in the following list.

Materials for short-answer item training:

1. Anchors and training sets by grade band, claim, target, and subcategory
2. Prompts and source materials or stimuli
3. Item-specific rubrics
4. One qualification set with 10 responses per item

7.3.4. Training Process: Mathematics Items

The training process for mathematics items is described next. The training steps for scoring mathematics items are described in the following list.

Training steps:

1. Trainees review the items that are represented in the anchor and training sets, any associated source materials or stimuli, and the item-specific rubrics.
2. Trainees read the associated source materials or stimuli, as appropriate.
3. Trainees score the training set for the item category.
4. Trainees review the correct scores and scoring rationale for the training set.
5. Trainees score a qualification round.

6. Trainees who do not meet the qualification standard on their first attempt have an opportunity to review correct scores and the scoring rationale with a scoring leader before making a second attempt.
7. Qualified raters begin scoring.

The training materials are described in the following list.

Materials for mathematics training:

1. Anchors and training sets by PT grade, family, and item category or by CAT item
2. Prompts and source materials or stimuli
3. Item-specific rubrics
4. One or two qualification rounds per item category, depending on item complexity, with 10 responses per round

Unlike ELA PTs, mathematics PTs may contain interdependencies among the items within a task. Each mathematics PT is made up of four to six items. Items may be dependent on any of the previous items within the PT. For example, if item 6 is dependent on items 3 and 5, the rubric for item 6 specifies the correct response based on prior correct responses to items 3 and 5. Raters are responsible for determining the appropriate response to item 6 and awarding credit accordingly, even when the student’s responses to items 3 and 5 are incorrect. It is also possible for the first two of the six items to be AI-scored while two or more of the other four are human scored.

The proper handling of tasks with dependencies is addressed in the training process. Raters have practice working through PT responses and recognizing correct work that is based on previous incorrect values. PTs are composed of items based on several different task models. In general, training materials are organized so raters train on a task model rather than on a complete PT. However, when PT items that are dependent on previous items in the set are presented in training, the entire set of items and responses is included. This allows raters to identify the previous responses that serve as the basis for the item that is being scored.

7.3.5. Supplemental Training for Scoring Supervisors

Scoring condition codes allow raters to categorize certain responses as unscorable. The code indicates the reason that the response cannot be scored. Responses with condition codes were routed to scoring supervisors for final code assignment. Supervisors required detailed training on the Smarter Balanced condition codes and definitions (Smarter Balanced, 2014a).

Table 7.6 presents the valid condition codes used for scoring, along with descriptions of the responses that would warrant the assignment of the different codes.

Table 7.6 Scoring Condition Codes

Condition Code	Reason	Use
B	**Blank**	No response
I	**Insufficient**	a. Use the “I” code when a student has not provided a meaningful response; this may include the following: • Random keystrokes • Undecipherable text • “I hate this test” • “I don’t know, IDK” • “I don’t care” • “I like pizza!” (in response to a reading passage about helicopters) • Response consisting entirely of profanity b. For ELA WER items, use the “I” code for responses described previously and also if • the student’s original work is insufficient for the rater to determine whether the student is able to organize, cite evidence and elaborate, and use conventions as defined in the rubrics; or • the response is too brief to make a determination regarding whether it is on purpose or on topic.
L	**Nonscorable Language**	• ELA: Language other than English • Mathematics: Language other than English or Spanish
T	**Off-Topic for ELA WER Items Only**	• The response is unrelated to the task or sources, or shows no evidence that the student has read the task or the sources (especially for informational or explanatory and opinion or argumentative); or • “Off topic” responses are generally substantial responses.

Table 7.6 *(continuation)*

Condition Code	Reason	Use
M	**Off-Purpose for ELA WER Items Only**	The student has clearly not written to the purpose designated in the task: • An off-purpose response addresses the topic of the task but not the purpose of the task. • Students may use narrative techniques in an explanatory essay or use argumentative or persuasive techniques to explain, for example, and still be on purpose. • Off-purpose responses are generally developed responses (essays, poems, etc.) clearly not written to the designated purpose.

7.3.6. Human-Scoring Alerts

Raters were also trained to watch for indications of a “crisis paper” and cheating. Such information can require urgent attention. Any student response of a sensitive nature to any human-scored test item was assigned a score and identified as an “alert.” Raters received a process document as part of their training materials that described the steps to follow should they determine that a response should be classified as an alert response. The different types of crisis paper alerts are as follows:

- Suicide
- Criminal activity
- Alcohol or drug use
- Extreme depression
- Violence
- Rape, sexual, or physical abuse
- Self-harm or intent to harm others
- Neglect

For crisis paper alerts, the local educational agency’s (LEA’s) superintendent and LEA CAASPP coordinator in the LEA for the flagged student were sent a copy of the response and the student’s Statewide Student Identifier via tracked delivery.

7.4. Student Test Scores

ETS developed two parallel scoring systems to produce students’ scores: the eSKM scoring system, which scores and delivers individual students’ scores to the ETS reporting system; and the parallel scoring system developed by ETS Technology and Information Processing Services (TIPS), which computes individual students’ scores. The two scoring systems independently applied the same scoring algorithms and specifications. ETS psychometricians verified the eSKM scoring by comparing all individual student scores from TIPS and resolving any discrepancies. This process redundancy is an internal quality control step that is in place to verify the accuracy of scoring. Students’ scores were reported only when the two parallel systems produced identical results with acceptable tolerance.

Were scores not to match, the mismatch would be investigated by ETS' PAR and eSKM teams and resolved. (For example, the mismatch could be a result of a Smarter Balanced and CDE decision to not score an item as a problem was identified in a particular item or rubric.) ETS would apply a problem item notification (PIN) not to score the item through the systematic process in eSKM, which might result in a mismatch if TIPS were still in the process of applying the PIN in the parallel system when the student score was being compared. This real-time scoring check is designed to detect mismatches and track remediation.

All scores must comply with the ETS scoring specifications and the parallel scoring process to ensure the quality and accuracy of scoring and to support the transfer of scores into the database of the student records scoring system, the Test Operations Management System (TOMS).

7.4.1. Total Test Scores

7.4.1.1. Theta Scores

For all of the tests, theta scores (item response theory [IRT] ability estimates) are obtained through maximum likelihood estimation (MLE) applied to item scores (Birnbaum, 1968). Items scored as one (correct) or zero (incorrect) are referred to as dichotomous items. Items scored from zero to some number of points greater than one are called polytomous items. The generalized partial credit model (GPCM) is applied to both types of items. The GPCM (Muraki, 1992) is:

$$P_{ih}(\theta_j) = \begin{cases} \dfrac{\exp[\sum_{v=1}^{h} Da_i(\theta_j - b_i + d_{iv})]}{1 + \sum_{c=1}^{n_i} \exp[\sum_{v=1}^{c} Da_i(\theta_j - b_i + d_{iv})]}, & \text{if score } h = 1, 2, \ldots., n_i \\[2ex] \dfrac{1}{1 + \sum_{c=1}^{n_i} \exp[\sum_{v=1}^{c} Da_i(\theta_j - b_i + d_{iv})]}, & \text{if score } h = 0 \end{cases} \tag{7.1}$$

Refer to the Alternative Text for Equation 7.1 for a description of this equation.

where,

$P_{ih}(\theta_j)$ is the probability of student with proficiency θ_j obtaining score h on item i,

n_i is the maximum number of score points for item i,

a_i is the discrimination parameter for item i,

b_i is the location parameter for item i,

d_{iv} is the category parameter for item i on score v, and

D is a scaling constant of 1.7 that makes the logistic model approximate the normal ogive model.

When $n_i = 1$, equation 7.1 becomes an expression of the two-parameter logistic model for dichotomous items.

The log-likelihood of a student with proficiency θ_j, given the observed response vector v, is:

$$L(\theta_j \,|U\,) = \ln(\prod_{i=1}^{I}\prod_{v=0}^{n_i} P_{ih}(\theta_j)^{u_{iv}}) \tag{7.2}$$

$$u_{iv} = \begin{cases} 1, \text{ if the score } h \text{ on polytomous item } i \text{ is equal to } v, \\ 0, \text{ otherwise} \end{cases}$$

Refer to the Alternative Text for Equation 7.2 for a description of this equation.

where,

I is the total number of items in the response vector,

n_i is the maximum number of score points for item i, and

P_{ih} is the probability of the score h observed on item i, as expressed in equation 7.1.

The theta that is associated with the largest log-likelihood for a particular pattern of scores is the maximum likelihood theta estimate. The equation for the MLE cannot generally be solved explicitly as it is nonlinear in nature (Hambleton & Swaminathan, 1985, p. 79). As a result, an iterative process such as the Newton-Raphson procedure is employed. At iteration t, a student's estimated ability θ is:

$$\theta_t = \theta_{t-1} - \frac{L_{t-1}^{'}}{L_{t-1}^{''}} \tag{7.3}$$

Refer to the Alternative Text for Equation 7.3 for a description of this equation.

where,

L'_{t-1} is the first derivative of the log-likelihood at iteration $t-1$, and

L''_{t-1} is the second derivative.

When the difference between the estimates in successive iterations becomes acceptably small (i.e., difference is less than .0001), the process is said to converge. The convergence criterion determines the level of accuracy of estimation, provided that the process converges. Theta scores are the basis for scale scores but are not reported. Scale scores and the transformation from theta scores to scale scores are described in the *Scale Scores for the Total Assessment* subsection.

7.4.1.2. Inverse Test Characteristic Curve Method

There are some special cases in which the score reported for a student is not based on the MLE approach:

- The student got the lowest possible score on the total test, which would lead to an MLE of -∞.
- The student got the highest possible score on the total test, which would lead to an MLE of +∞.
- The student's response pattern did not lead to a single most likely MLE of the student's ability, or the likelihood function was so flat that its maximum was not much greater than the likelihood over a wide range of theta values.

In these cases, the student's score is computed by the inverse test characteristic curve (TCC) method (Stocking, 1996). This method transforms the sum of the student's item scores into an ability estimate. That estimate is the ability level at which the sum of the expected scores on the items the student took is equal to the sum of the scores that the student actually earned on those items.

The item characteristic curve for an item shows the probability of a correct answer to the item (in the case of dichotomous items) or the probability of responding in a score category (in the case of polytomous items) as a function of the student's ability. The TCC for a set of items shows the expected total score on those items as a function of the student's ability. Because information is lost by not utilizing each student's unique pattern of responses, this method is used only when the response pattern does not lead to one clear MLE of the student's ability or the likelihood function is so flat that although it has a maximum, there is a wide range of theta values at which the likelihood is only slightly less than the maximum.

The lowest obtainable theta (LOT) and the highest obtainable theta (HOT) defined by the Smarter Balanced Assessment Consortium are presented in table 7.7 for each grade and content area (Smarter Balanced Assessment Consortium, 2016b). The theta scores for grades three through eight and grade eleven are on a common vertical scale.

Table 7.7 Theta of Lowest and Highest Obtainable Scores

Content Area and Grade	LOT	HOT
ELA 3	-4.5941	1.3374
ELA 4	-4.3962	1.8014
ELA 5	-3.5763	2.2498
ELA 6	-3.4785	2.5140
ELA 7	-2.9114	2.7547
ELA 8	-2.5677	3.0430
ELA 11	-2.4375	3.3392
Mathematics 3	-4.1132	1.3335
Mathematics 4	-3.9204	1.8191
Mathematics 5	-3.7276	2.3290
Mathematics 6	-3.5348	2.9455
Mathematics 7	-3.3420	3.3238
Mathematics 8	-3.1492	3.6254
Mathematics 11	-2.9564	4.3804

7.4.1.3. Scoring of Incomplete Cases

Sometimes students fail to complete their tests. Depending on the nature of the missing data, different actions are taken. This subsection covers the following three situations:

1. Attemptedness/Test-Taking rules that describe when a test is considered attempted or participated
2. When a test is scored
3. How and when incomplete tests are scored

As defined in the Smarter Balanced scoring specifications, tests are considered “complete” if students respond to at least the minimum number of operational items specified in the blueprint. Otherwise, the tests are “incomplete.” (Refer to table 8.1 and table 8.2 for the minimum number of operational items in each claim for students who are assigned only operational items and for students who are assigned items for embedded field test PTs, respectively.) In a fixed-form (i.e., not CAT) assessment, unanswered items are treated as incorrect. However, in a CAT environment, all but one of the specific unanswered items are unknown, because the test administration terminates when a student stops responding to items. ETS implemented several procedures that score an incomplete test in a CAT environment; these procedures are presented in table 7.8.

Table 7.8 Treatment of Incomplete Tests

If the student. . .	Classify the student as taking the test?	Include the data in the student file?	Score the responses for the student?	Classify the student as attempting the test?	Report a score for the student?
Logged on to both the CAT and PT but answered no items	Yes	Yes	No	No	No
Logged on to both the CAT and PT and answered at least one item for only CAT or PT	Yes	Yes	Lowest obtainable score for the test	No	No
Logged on to both the CAT and PT and answered at least one PT item but fewer than 10 CAT items	Yes	Yes	Lowest obtainable score for the test	Yes	No
Logged on to both the CAT and PT, answered at least one PT item and at least 10 CAT items, but did not answer a specified minimum number of items for a complete test	Yes	Yes	MLE (unanswered items in the middle of the test scored as incorrect), or for an incomplete test, estimate from equation 7.4	Yes	Yes

The number and percent of students who took the tests are presented in the tables of appendix 7.A for all students in each test by a variety of incomplete conditions. Sometimes a student stops answering items before the TDS has administered all the items the student is supposed to answer. When that happens, the student's test is considered complete if the student has answered at least a specified minimum number of items (less than the number of items in the full test). Otherwise, the student's score is based on an adjusted ability estimate calculated by the formula in equation 7.4.

$$\theta_{Adj.} = \theta_{\min} + (\theta_{achieved} - \theta_{\min}) * PropAdj \qquad (7.4)$$

Refer to the Alternative Text for Equation 7.4 for a description of this equation.

where,

θ_{adj} is the student's adjusted ability estimate,

$\theta_{achieved}$ is the theta estimate based on the incomplete test,

θ_{min} is a predetermined theta estimate equal to -3.5, which is the average of the LOT values across all tests (on the vertical theta scale), and

$PropAdj$ is the proportion of the test completed by the student.

7.4.1.4. Scale Scores for the Total Assessment

After MLE scoring is performed on the theta scale and the scoring rules are implemented, the scaling constants are applied. Scale scores (SS) are on the Smarter Balanced vertical scale and are formed by linking across grades using common items in adjacent grades. The vertical scale score is the linear transformation of the post–vertically scaled IRT ability estimate (refer to subsection *2.7.3 Vertical Scaling* for the procedure). The student's estimated theta score is converted to a scale score by the following formulas:

For ELA: $\text{SS} = 85.8\,\theta + 2508.2$ (7.5)

Refer to the Alternative Text for Equation 7.5 for a description of this equation.

For mathematics: $\text{SS} = 79.3\,\theta + 2514.9$ (7.6)

Refer to the Alternative Text for Equation 7.6 for a description of this equation.

There is a restriction that the scale score cannot be higher or lower than the specified highest and lowest possible scores for that content area and grade level. The lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS) for each test are displayed in table 7.9. Scale scores are rounded to the nearest integer.

Detailed information regarding the establishment of scale scores for the Smarter Balanced Summative Assessments can be found in chapter 10 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a) and the *Smarter Balanced Scoring Specification: 2014–2015 Administration* (AIR, 2015).

Table 7.9 Lowest and Highest Obtainable Scale Scores

Content Area and Grade	LOSS	HOSS
ELA 3	2114	2623
ELA 4	2131	2663
ELA 5	2201	2701
ELA 6	2210	2724
ELA 7	2258	2745
ELA 8	2288	2769
ELA 11	2299	2795
Mathematics 3	2189	2621
Mathematics 4	2204	2659
Mathematics 5	2219	2700
Mathematics 6	2235	2748
Mathematics 7	2250	2778
Mathematics 8	2265	2802
Mathematics 11	2280	2862

7.4.1.5. Achievement Levels

Standard settings were performed by the Smarter Balanced Assessment Consortium, which defined four achievement levels based on overall scale scores. These achievement level categories were labeled “Standard Not Met,” “Standard Nearly Met,” “Standard Met,” and “Standard Exceeded.” The combined categories of “Standard Met” and “Standard Exceeded” are used to define students meeting the proficiency criterion for accountability purposes. Refer to *Chapter 10 Achievement Level Setting* of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a) for details related to the standard setting procedure; *Reporting Achievement Level Descriptors* (Smarter Balanced, 2015) for the descriptors used to describe Smarter Balanced achievement levels; and *Interpretation and Use of Scores and Achievement Levels* (Smarter Balanced, 2014b) for more information about using achievement levels.

- **Level 1—Standard Not Met:** Student demonstrates minimal understanding of ELA and mathematics and the ability to apply the knowledge and skills for his or her grade level that are associated with college and career readiness.
- **Level 2—Standard Nearly Met:** Student demonstrates partial understanding of ELA and mathematics and the ability to apply the knowledge and skills for his or her grade level that are associated with college and career readiness.
- **Level 3—Standard Met:** Student demonstrates adequate understanding of ELA and mathematics and the ability to apply the knowledge and skills for his or her grade level that are associated with college and career readiness.
- **Level 4—Standard Exceeded:** Student demonstrates thorough understanding of ELA and mathematics and the ability to apply the knowledge and skills for his or her grade level that are associated with college and career readiness.

The thresholds for the achievement levels vary by grade and content area. Table 7.10 provides the theta thresholds for Standard Nearly Met, Met, and Exceeded at each grade level. For example, the threshold of −0.888 for “Standard Met” in grade three ELA means that a student must earn a theta score (θ) of −0.888 or higher to achieve that classification.

Table 7.10 Theta Thresholds for Achievement Levels

Content Area and Grade	Standard Nearly Met	Standard Met	Standard Exceeded
ELA 3	-1.646	-0.888	-0.212
ELA 4	-1.075	-0.410	0.289
ELA 5	-0.772	-0.072	0.860
ELA 6	-0.597	0.266	1.280
ELA 7	-0.340	0.510	1.641
ELA 8	-0.247	0.685	1.862
ELA 11	-0.177	0.872	2.026
Mathematics 3	-1.689	-0.995	-0.175
Mathematics 4	-1.310	-0.377	0.430
Mathematics 5	-0.755	0.165	0.808
Mathematics 6	-0.528	0.468	1.199
Mathematics 7	-0.390	0.657	1.515
Mathematics 8	-0.137	0.897	1.741
Mathematics 11	0.354	1.426	2.561

Table 7.11 shows the scale score range of each achievement level for the ELA and mathematics assessments, respectively.

Table 7.11 Scale Score Ranges for Achievement Levels

Content Area and Grade	Standard Not Met	Standard Nearly Met	Standard Met	Standard Exceeded
ELA 3	2114–2366	2367–2431	2432–2489	2490–2623
ELA 4	2131–2415	2416–2472	2473–2532	2533–2663
ELA 5	2201–2441	2442–2501	2502–2581	2582–2701
ELA 6	2210–2456	2457–2530	2531–2617	2618–2724
ELA 7	2258–2478	2479–2551	2552–2648	2649–2745
ELA 8	2288–2486	2487–2566	2567–2667	2668–2769
ELA 11	2299–2492	2493–2582	2583–2681	2682–2795
Mathematics 3	2189–2380	2381–2435	2436–2500	2501–2621
Mathematics 4	2204–2410	2411–2484	2485–2548	2549–2659
Mathematics 5	2219–2454	2455–2527	2528–2578	2579–2700
Mathematics 6	2235–2472	2473–2551	2552–2609	2610–2748
Mathematics 7	2250–2483	2484–2566	2567–2634	2635–2778
Mathematics 8	2265–2503	2504–2585	2586–2652	2653–2802
Mathematics 11	2280–2542	2543–2627	2628–2717	2718–2862

7.4.2. Claim Scores (Subscores)

Claims identify knowledge and skills being measured through a set of items. Groups of items in each combination of grade and content area are formed based on related content standards; outcomes for these groups of items are called claim scores. A claim score is a measure of a student's performance on the items in that claim.

There are four claims for ELA assessments and three claims for mathematics assessments. Claims 2 and 4 of mathematics scores are combined because of content similarity and to provide flexibility for item development. Consequently, only three claim scores are reported with the overall mathematics score.

Like the overall test, results of each claim are reported as a theta score, a scale score, and a claim strength or weakness. The claims for ELA are identified in table 7.12 and are also available in the blueprints, which are provided in appendix 2.A.

Table 7.12 Claims Identified for ELA

Claim	Description
1. Reading	Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
2. Writing	Students can produce effective and well-grounded writing for a range of purposes and audiences.
3. Listening/ Speaking	Students can employ effective listening skills for a range of purposes and audiences.
4. Research	Students can engage in research and inquiry to investigate topics and to analyze, integrate, and present information.

The claims for mathematics are identified in table 7.13 and are also available in the blueprints, which are provided in appendix 2.A. Note that for mathematics, claims 2 and 4 are reported together as defined by the Smarter Balanced Assessment Consortium, so there are only three reporting categories with four claims.

Table 7.13 Claims Identified for Mathematics

Claim	Description
1. Concepts and Procedures	Students can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
2. Problem Solving	Students can solve a range of complex, well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies.
3. Model and Data Analysis	Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.
4. Communicating/ Reasoning	Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.

7.4.2.1. Scale Scores for Claims

Claim scores are calculated by applying the MLE approach to the items contained in a particular claim. The resulting ability estimates are converted to claim scale scores by applying equation 7.5 for ELA assessments and equation 7.6 for mathematics assessments. ELA scores are computed for each claim. Mathematics scores are computed for Claim 1, claims 2 and 4 combined, and Claim 3.

Claim scores are based on fewer items than total test scores. As a result, the number of students whose claim scores cannot be estimated by the MLE approach is larger than for the total score. ETS uses the inverse TCC approach when MLE-derived theta estimates are not available for a claim.

7.4.2.2. Performance Levels for Claims

The relative strengths and weaknesses for each student are reported for each claim. The three performance levels for each claim are as follows:

- **Above Standard**—Student clearly understands and can successfully apply his or her knowledge to the standards tested in this content area for his or her grade.
- **Near Standard**—Student shows understanding and can apply his or her knowledge to the standards tested in this content area for his or her grade.
- **Below Standard**—Student has limited understanding and difficulty applying his or her knowledge to the standards tested in this content area for his or her grade.

Because claim scores are based on fewer items than overall test scores, the standard error of the claim scale scores is included in the determination of the student's performance level on a claim. SS_{claim} is a student's estimated scale score on a claim. A range of possible student scale scores is calculated for each student from $SS_{Claim} - 1.5 \times SE_{SS_{Claim}}$ to $SS_{Claim} + 1.5 \times SE_{SS_{Claim}}$, each of which is converted to a scale score and rounded to an integer.

If the value at the high end of the score range is less than the minimum scale score associated with the overall "Standard Met" achievement classification, the claim performance level is reported as "Below Standard." This achievement classification is also assigned when all student responses to items associated with a claim are incorrect.

If the value at the low end of the range is greater than the minimum scale score associated with the overall "Standard Met" achievement classification, the claim performance level is reported as "Above Standard." This claim performance level is also reported when all student responses are correct.

Scale score ranges that do not meet either of these classifications are reported as "At/Near Standard."

7.4.3. Theta Scores Standard Error

A student's true ability level or theta score and standard error of theta are not known. The standard error of measurement (SEM) is the standard deviation of the distribution of theta scores that the student would earn under different testing conditions. In IRT, the only differences taken into account in the SEM are those associated with different sets of items that could be presented to the student. An error band can be calculated from the student's theta score minus one SEM to the student's theta score plus one SEM. Over a large number of replications of this procedure, the error band will contain the student’s true score approximately 68 percent of the time. The error band is transformed to the scale score metric and reported for the CAASPP Smarter Balanced assessments. It is useful to take into account the size of measurement errors because no assessment measures student ability with perfect accuracy or consistency. (Error bands are also discussed in subsection *7.4.5 Error Band*.)

In the framework of IRT, the SEM is the reciprocal of the square root of the test information function (TIF) based on the items taken by each student. It is also the estimate of standard error for the estimate of theta. The TIF is the sum of information from each item on the test. With MLE, the SEM for a student with proficiency θ_j is:

$$SEM(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}} \tag{7.7}$$

Refer to the Alternative Text for Equation 7.7 for a description of this equation.

where,

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

$$I(\theta_j) = \sum_{i=1}^{n} I_i(\theta_j) \tag{7.8}$$

Refer to the Alternative Text for Equation 7.8 for a description of this equation.

and $I_i(\theta_j)$ is the item information of item i for student j.

When item information is based on the GPCM for both dichotomous and polytomous items, it is calculated as:

$$I_i(\theta_j) = (Da_i)^2[s_{i2}(\theta_j) - s_i^2(\theta_j)] \tag{7.9}$$

Refer to the Alternative Text for Equation 7.9 for a description of this equation.

where,

$S_i(\theta_j)$ is the expected item score for item i on a theta scale score θ_j, calculated as

$$s_i(\theta_j) = \sum_{h=0}^{n_i} hp_{ih}(\theta_j) \tag{7.10}$$

Refer to the Alternative Text for Equation 7.10 for a description of this equation.

and

$$s_{i2}(\theta_j) = \sum_{h=0}^{n_i} h^2 p_{ih}(\theta_j) \tag{7.11}$$

Refer to the Alternative Text for Equation 7.11 for a description of this equation.

where,

$P_{ih}(\theta_j)$ is the probability of an examinee with θ_j getting score h on item i, the computation of which is shown in equation 7.1, and

n_i is the maximum number of score points for item i.

The SEM is calculated based only on the answered item(s) for both complete and incomplete tests. The upper bound of the SEM is set to 2.5 on the theta metric, and any value larger than 2.5 is truncated at 2.5, as is required by the Smarter Balanced Assessment Consortium (AIR, 2015).

7.4.4. Scale Score Standard Errors

Standard errors of the maximum likelihood theta estimates are also transformed onto the reporting scale. This transformation is

$$SE_{scaled} = a * SE_{\theta_j} \tag{7.12}$$

Refer to the Alternative Text for Equation 7.12 for a description of this equation.

where,

SE_θ is the standard error of the ability estimate on the θ scale, and

a is the slope of the scaling constants that transform θ to the reporting scale.

The value of a is 85.8 for ELA and 79.3 for mathematics.

7.4.5. Error Band

A band of scale scores showing the measurement error associated with each scale score is reported. It is generated by developing a band of indeterminacy surrounding the scale score:

$$\text{error band} = (SS - SE_{scaled}, SS + SE_{scaled}) \tag{7.13}$$

Refer to the Alternative Text for Equation 7.13 for a description of this equation.

where,

SS is the scale score,

SE_{scaled} is the SEM associated with this scale score,

$SS - SE_{scaled}$ is the lower boundary of the error band, and

$SS + SE_{scaled}$ is the upper boundary of the error band.

7.4.6. Assessment Target Reports

7.4.6.1. Overview of Assessment Target Reports

Assessment target standards are specific to each content domain and linked to the CCSS associated with claim areas. For Smarter Balanced tests, assessment targets are intended to support the development of high-quality items and tasks that contribute evidence to the claims. The relationship between assessment targets and CCSS elements is made explicit in the Smarter Balanced content specifications (ETS, 2015a; 2015b).

Assessment target scores, which are reported only at the group level, provide insight into strengths and weaknesses for a group of students relative to their performance on the test as a whole. For a selected group of students (for example, a classroom), if their performance on an assessment target is better than their performance on the test as a whole, the assessment target is an area of relative strength. Conversely, if the group of students did not perform as well on an assessment target in relation to the test as a whole, it would be an area of relative weakness.

Assessment target scores are derived from item *residuals*, which are the differences between a student's observed score and expected score for a particular item. For the selected group of students, the assessment target scores for each student are calculated by summing the differences between the observed and expected scores for each student for all items that the student attempted within a particular assessment target. The sum of these differences is then divided by the total number of points possible for items within a particular target. Next, the mean assessment target scores, as well as the standard error for all students in the selected student group, are calculated. Finally, strengths and weaknesses thresholds are established after the values for each assessment target are calculated. More details on the calculation of the assessment targets and the establishment of the strengths and weaknesses thresholds are described in an ETS memorandum, *Target Score Reporting* (ETS, 2015b).

Note, however, that while assessment targets are based on target standards, not all claim areas support assessment target reporting. For example, assessment targets are reported for all claims in ELA but only for Claim 1 in mathematics.

7.4.6.2. Limitations

Caution should be used when reporting or interpreting assessment targets. First, assessment targets can only be meaningfully reported at the group level because they are neither reliable nor generalizable enough to support inferences for individual students. Second, because residuals are sensitive to model fit, student strengths and weaknesses evaluated this way are sometimes the result of a misfit in item calibration. Therefore, it is necessary to compute the average residuals of each item across all students within each assessment target to determine whether the average residuals across all students are uniformly close to zero. Finally, assessment targets that are based on 10 or fewer items in the item bank are not reported, except the WER items.

The extent to which the scores are *generalizable* depends on the total number of items administered from that domain across all students. A small number of items is not sufficient to broadly represent the target domain or to support the general conclusions required of actionable information.

7.4.6.3. Reporting

The distribution of the average assessment target scores depends both on the number of students in the defined group and on the number of items that these students answered in a target. As both numbers grow large, the average residuals increasingly cluster symmetrically around zero. To support California schools in making valid inferences based on the assessment target information, the number of items per target standard is considered when reporting the assessment target. A criterion that there are at least 10 items within the item pool for a target standard is recommended. No target score reports were conducted and reported because of the impact of the novel coronavirus disease 2019 pandemic.

7.5. Overview of Score Aggregation Procedures

To provide meaningful results to the stakeholders, test scores for a given grade and content area are aggregated at the school, LEA or direct funded charter school, county, and state levels. The aggregated scores are generated both for selected groups and for the population. The next subsection contains a description of the types of aggregation performed on CAASPP Smarter Balanced online summary assessment scores. Score aggregation includes only students with valid scores, refer to subsection *7.6.2 Special Cases* for more information.

7.5.1. Score Distributions and Summary Statistics

Summary statistics that describe student performance on each assessment that contains only operational items are presented in table 7.14. Included in the tables are the number of students for each assessment and the mean and standard deviation of student scores expressed in terms of both scale score and theta score.

Table 7.14 Mean and Standard Deviation of Theta and Scale Scores[45]

Content Area and Grade	Number of Students	Mean Scale Score	Scale Score SD	Mean Theta Score	Theta Score SD
ELA 3	137	2387	86	-1.41	1.00
ELA 4	143	2447	110	-0.72	1.28
ELA 5	163	2462	92	-0.54	1.07
ELA 6	123	2480	110	-0.33	1.28
ELA 7	488	2496	109	-0.15	1.27
ELA 8	203	2448	97	-0.70	1.13
ELA 11	13,975	2623	121	1.33	1.42
Mathematics 3	101	2393	96	-1.53	1.21
Mathematics 4	69	2385	104	-1.64	1.31
Mathematics 5	110	2425	89	-1.14	1.12
Mathematics 6	47	2387	121	-1.61	1.53
Mathematics 7	65	2410	104	-1.32	1.31
Mathematics 8	560	2460	117	-0.69	1.47
Mathematics 11	10,522	2604	144	1.12	1.82

The number and the percentage of students in each achievement level and the number and the percentage who meet or exceed the standard are shown in table 7.15.

[45] Results of table 7.14 and table 7.15 include students who took regular online tests and students who were assigned to the embedded field test PTs.

Table 7.15 Count and Percentage in Achievement Levels for CAASPP Online Summative Assessments

Content Area and Grade	Standard Not Met N	Standard Not Met %	Standard Nearly Met N	Standard Nearly Met %	Standard Met N	Standard Met %	Standard Exceeded N	Standard Exceeded %	Standard Met/Exceeded* N	Standard Met/Exceeded* %
ELA 3	55	40%	38	28%	30	22%	14	10%	44	32%
ELA 4	48	34%	28	20%	38	27%	29	20%	67	47%
ELA 5	62	38%	42	26%	47	29%	12	7%	59	36%
ELA 6	51	41%	27	22%	32	26%	13	11%	45	37%
ELA 7	219	45%	104	21%	120	25%	45	9%	165	34%
ELA 8	145	71%	26	13%	26	13%	6	3%	32	16%
ELA 11	2,370	17%	2,357	17%	4,070	29%	5,178	37%	9,248	66%
Mathematics 3	41	41%	19	19%	27	27%	14	14%	41	41%
Mathematics 4	43	62%	14	20%	7	10%	5	7%	12	17%
Mathematics 5	68	62%	26	24%	13	12%	3	3%	16	15%
Mathematics 6	37	79%	6	13%	1	2%	3	6%	4	9%
Mathematics 7	54	83%	4	6%	3	5%	4	6%	7	11%
Mathematics 8	374	67%	106	19%	45	8%	35	6%	80	14%
Mathematics 11	3,638	35%	2,023	19%	2,356	22%	2,505	24%	4,861	46%

* May not exactly match the sum of Level 3 and Level 4 percentages, because of rounding

Figure 7.2 presents a graphical representation of the percentage of students at each ELA achievement level by grade. These are the achievement levels for ELA shown in table 7.15.

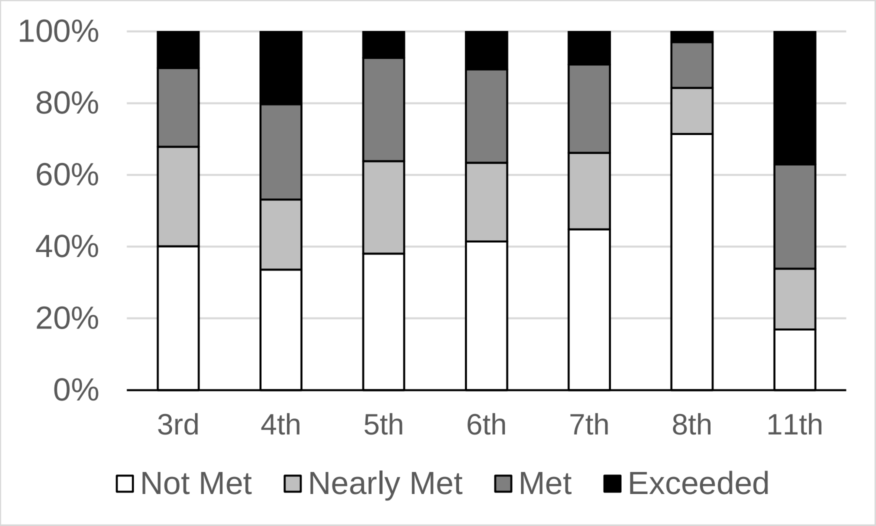

Figure 7.2 Percentage of achievement levels in ELA

Figure 7.3 presents a graphical representation of the percentage of students at each mathematics achievement level by grade. These are the achievement levels for mathematics shown in table 7.15.

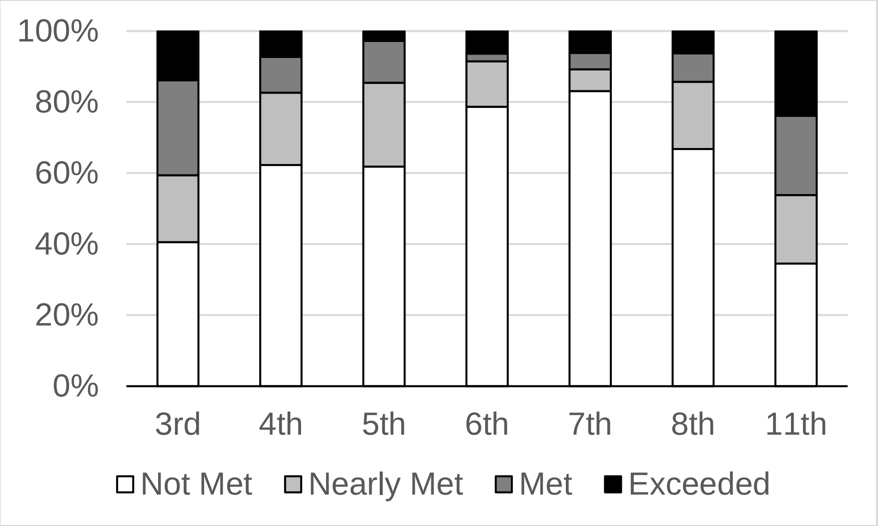

Figure 7.3 Percentage of achievement levels in mathematics

7.5.2. Group Scores

Because of the extremely small number of students who completed tests this year, group score analyses are not reported.

7.6. Reports Produced and Scores for Each Report

The tests that make up the CAASPP online summative assessments provide results or score summaries that are reported for different purposes. The four major purposes are to

1. help facilitate conversations between parents/guardians and teachers about student performance,
2. serve as a tool to help parents/guardians and teachers work together to improve student learning,
3. help schools and LEAs identify strengths and areas that need improvement in their educational programs, and
4. provide the public and policymakers with information about student achievement.

This section provides detailed descriptions of the uses and applications of CAASPP reporting for students.

7.6.1. Online Reporting

TOMS is a secure website hosted by ETS that permits LEA users to manage the CAASPP online summative assessments and to inform the TDS. This system uses a role-specific design to restrict access to certain tools and applications based on the user's designated role. Specific functions of TOMS include the following:

- Manage user access privileges
- Manage test administration calendars and testing windows
- Manage student test assignments
- Manage and confirm the accuracy of students' test settings (i.e., designated supports and accommodations) prior to testing
- Run and download various reports

In addition to TOMS, there were two California online reporting systems used during the 2019–2020 administration: the Online Reporting System (ORS) and the California Educator Reporting System (CERS).

TOMS communicated with the ORS, which provided authorized users with interactive and cumulative online reports for ELA and mathematics at the student, school, and LEA levels. The ORS provided access to two CAASPP reports: Score Reports, which provided preliminary score data for each administered test available in the reporting system; and the Completion Status Reports, which provided completion data in the reporting system for students taking an assessment.

TOMS also communicated with the CERS, which, starting in January 2021, is the primary source for LEA staff to analyze CSA results at the LEA, school, grade, classroom, or customized group level. CERS provides these reports, which can be downloaded to plan instruction.

LEA staff who have TOMS logon credentials can enter CERS through the CAASPP website (https://www.caaspp.org/) to access student assessment results. CERS allowed educators to view their students' assessment results. For example, educators could create customized groups from assigned student groups; for interim assessments, they could access specific assessment items with student responses; and they could use the distractor analysis feature to identify student strengths and needs.

7.6.2. Special Cases

Student scores are not reported for the following cases:

- Student had a medical emergency during testing
- Student's parent/guardian requested exemption from testing
- Student was tested but marked no answers
- Student did not log on to both CAT and PT portions
- Student logged on to two parts (PT and CAT) without any recorded answers
- Student logged on to one part (PT or CAT) but not both parts, and had no recorded answers
- Student attempted fewer than 10 CAT items and fewer than 1 PT item
- Because of the suspension of testing in 2019–2020, individual Student Score Reports (SSRs) were not made available for students who did not complete and submit at least one content area (ELA or mathematics) of the Smarter Balanced assessment

7.6.3. Types of Score Reports

There are two categories of CAASPP reports. The specific reports within each category are presented in this subsection.

7.6.3.1. Student Score Report

The CAASPP SSR is the official score report for parents or guardians and describes the student's results. Because testing was suspended, only students who completed and submitted at least one Smarter Balanced content area received an SSR in 2019–2020. Results presented for the CAASPP online summative assessments include the following metrics:

- Scale score for each content area assessment reported (The ranges of scale scores for both ELA and mathematics are provided in table 7.9.)
- Achievement level for each content area assessment reported (Smarter Balanced achievement levels for both ELA and mathematics are "Standard Exceeded," "Standard Met," "Standard Nearly Met," and "Standard Not Met.")
- Performance levels for all claims in each content area assessment reported (Smarter Balanced performance levels for claims are "Above Standard," "Near Standard," and "Below Standard.")

Scores for students who were assigned accommodations or designated supports are reported in the same way as for students without accommodations or designated supports. (Refer to section *2.5 Universal Tools, Designated Supports, and Accommodations* for more information about accessibility resources.)

In all, LEAs had four options for accessing and distributing SSRs to parents/guardians:

1. Accessing electronic SSR PDFs using a locally provided parent/guardian or student portal

2. Downloading SSR PDFs from TOMS and making them available electronically using a secure local method
3. Downloading SSR PDFs from TOMS, printing them, and making them available locally
4. Purchasing paper SSRs from ETS (This option was not available in 2019–2020 because of the suspension of testing.)

The LEA CAASPP coordinator could forward the appropriate reports to test sites. In the case of a locally printed CAASPP SSR, the LEA sent the printed report(s) to the child's parent or guardian. Downloaded SSRs were forwarded to the test site. CAASPP SSRs that include individual student results were not distributed beyond the student's school.

Further information about the SSR and its interpretation is provided on the Starting Smarter web page at https://ca.startingsmarter.org/.

7.6.3.2. Student Data Files and Aggregations

The CAASPP student data files for the LEA were available for the LEA CAASPP coordinator and CAASPP test site coordinator to download from TOMS.

Preliminary student scores and aggregations were also available to LEAs using the ORS and CERS. These applications permitted LEAs to view preliminary results data for all tests taken.

Official aggregations of Smarter Balanced results were not performed in 2019–2020 because of the suspension of testing. Historical online results are accessible to the public on the Test Results for California's Assessments website at https://caaspp-elpac.cde.ca.gov/.

7.6.4. Score Report Applications

CAASPP online summative assessment results provided parents/guardians with information about their child's progress. The results were a tool for increasing communication and collaboration between parents/guardians and teachers. Along with the results from the Smarter Balanced Interim Assessments, the SSR could be used by parents/guardians while talking with teachers about ways to improve their child's achievement of the CCSS.

Schools could use the CAASPP online summative assessment results to help make decisions about how best to support student achievement. CAASPP online summative assessment results, however, should never be used as the only source of information to make important decisions about a child's education.

CAASPP online summative assessment results help schools and LEAs identify strengths and weaknesses in their instructional programs. Each year, staff from schools and LEAs examine CAASPP test results at each grade level and content area tested. Their findings are used to help determine

- the extent to which students are learning the academic standards,
- instructional areas that can be improved,
- teaching strategies that can be developed to address the needs of students, and
- decisions about how to use funds to ensure that students achieve the standards.

CAASPP online summative assessments results were used to rank the academic performance of schools, compare schools with similar characteristics (e.g., size and ethnic composition), identify low-performing and high-performing schools, and set yearly targets for academic progress.

7.6.5. Criteria for Interpreting Test Scores

An LEA may use CAASPP online summative assessment results to help make decisions about student placement, promotion, retention, or other considerations related to student achievement. However, it is important to remember that a single test can provide only limited information. Other relevant information should be considered as well. It is advisable for parents and guardians to evaluate their child's strengths and weaknesses in the relevant topics by reviewing classroom work and progress reports in addition to the child's CAASPP online summative assessment results. It is also important to note that a student's score in a content area could vary somewhat if the student were retested.

7.6.6. Criteria for Interpreting Score Reports

The information presented in various reports must be interpreted with caution when making performance comparisons. When comparing scale score and achievement-level results, the user is limited to comparisons within a content area. The scale scores are on a vertical scale across grades for each content area (ELA or mathematics), but the score scales for ELA and mathematics are not comparable to each other. The user may compare scale scores for the same content area and grade, within a school, between schools, or between a school and its LEA, its county, or the state.

The user can also make comparisons within the same grade and content area across years. Caution should be taken when comparing scale scores from different grades within a content area, because the curricula are different across grade levels. Comparing scores obtained in different content areas should be avoided because the results are not on the same scale. Finally, note that for 2019–2020, individual student results for students who were unable to complete testing prior to the suspension of testing may not accurately represent student abilities.

For more details on the criteria for interpreting information provided on the score reports, refer to https://ca.startingsmarter.org/. Refer also to *2018–19 CAASPP Post-Test Guide* (CDE, 2019), which was applicable for the 2019–2020 CAASPP administration.

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.

American Institutes for Research. (2015). *Smarter Balanced scoring specification, 2014–2015 administration: Summative and interim assessments: ELA grades 3–8, 11 and mathematics grades 3–8, 11, version 7.* Washington, DC: American Institutes for Research. http://www.smarterapp.org/documents/TestScoringSpecs2014-2015.pdf

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

California Department of Education (2019). *2018–19 CAASPP post-test guide: Technical information for student score reports for CAASPP LEA and test site coordinators and research specialists.* Sacramento, CA: California Department of Education. https://bit.ly/3b3l4ko

Educational Testing Service. (2015a). *Selection of Smarter Balanced field trial items for operational scoring.* [Memorandum]. Sacramento, CA: Educational Testing Service.

Educational Testing Service. (2015b). *Target score reporting.* [Memorandum]. Sacramento, CA: Educational Testing Service.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

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

Smarter Balanced Assessment Consortium. (2014a). *Hand-scoring rules.* Los Angeles, CA: Smarter Balanced Assessment Consortium. http://www.smarterapp.org/documents/Smarter_Balanced_Hand_Scoring_Rules.pdf

Smarter Balanced Assessment Consortium. (2014b). *Interpretation and use of scores and achievement levels.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/interpretation-and-use-of-scores-and-achievement-levels.pdf

Smarter Balanced Assessment Consortium. (2014c). *Smarter Balanced scoring guide for grades three, six, and eleven ELA PT full-write baseline sets.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/scoring-guide-for-ela-full-writes.pdf

Smarter Balanced Assessment Consortium. (2015). *Reporting achievement level descriptors.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/achievement-level-descriptors.pdf

Smarter Balanced Assessment Consortium. (2016a). *Smarter Balanced Assessment Consortium: 2013–14 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf

Smarter Balanced Assessment Consortium. (2016b). Smarter Balanced Assessment Consortium: *2014-15 technical report*. https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf/

Smarter Balanced Assessment Consortium. (2017a). *ELA CAT item specifications, high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017b). *ELA CAT item specifications, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017c). *ELA CAT item specifications, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017d). *ELA PT item specifications, argumentative, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017e). *ELA PT item specifications, explanatory, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017f). *ELA PT item specifications, , informative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017g). *ELA PT item specifications, narrative, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017h). *ELA PT item specifications, narrative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017i). *ELA PT item specifications, opinion, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018a). *Mathematics CAT and performance task (PT) item specifications*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018b). *Mathematics CAT item specifications, Claim 1, grade eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018c). *Mathematics CAT item specifications, Claim 1, grade five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018d). *Mathematics CAT item specifications, Claim 1, grade four*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018e). *Mathematics CAT item specifications, Claim 1, grade seven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018f). *Mathematics CAT item specifications, Claim 1, grade six*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018g). *Mathematics CAT item specifications, Claim 1, grade three*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018h). *Mathematics CAT item specifications, Claim 1, high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018i). *Mathematics CAT item specifications, Claim 2, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018j). *Mathematics CAT item specifications, Claim 3, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018k). *Mathematics CAT item specifications, Claim 4, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21*, 365–89.

Accessibility Information

Alternative Text for Equation 7.1

P sub I h of theta sub j equals the numerator exp open parenthesis the sum from v equals 1 to h of D times a sub i of the quantity open parenthesis theta sub j minus b sub I plus d sub iv close parenthesis close parenthesis and denominator 1 plus the sum from c equals 1 to n sub I exp open parenthesis the sum from v equals 1 to c D times a sub i of the quantity open parenthesis theta sub j minus b sub I plus d sub iv close parenthesis close parenthesis, if score h equals 1, 2, …., n sub i.

P sub I h of theta sub j equals 1 divided by denominator 1 plus the sum from c equals 1 to n sub I exp open parenthesis the sum from v equals 1 to c D times a sub i of the quantity open parenthesis theta sub j minus b sub I plus d sub iv close parenthesis close parenthesis, if score h equals 0.

Alternative Text for Equation 7.2

L of the union of theta sub j, U equals the natural logarithm of open parenthesis the product from I equals 1to I times the product from v equals 0 to n sub I of P sub ih of theta sub j to the u sub iv power.

u sub i v equals 1, if the score h on polytomous item i is equal to v, 0, otherwise.

Alternative Text for Equation 7.3

Theta sub t equals theta sub t minus 1 minus the quantity with the numerator L prime sub t minus 1 and the denominator L double prime sub t minus 1.

Alternative Text for Equation 7.4

Theta sub adj equals theta sub min plus open parenthesis theta sub achieved minus theta sub min close parenthesis times PropAdj.

Alternative Text for Equation 7.5

ELA scale score is the sum of 2508.2 and 85.8 times theta.

Alternative Text for Equation 7.6

Mathematics scale score is the sum of 2514.9 and 79.3 times theta.

Alternative Text for Equation 7.7

SEM of Theta sub j equals 1 divided by the square root of I of theta sub j.

Alternative Text for Equation 7.8

I of Theta sub j equals the sum from I equals 1 to n of I sub I of Theta sub j.

Alternative Text for Equation 7.9

I sub i of Theta sub j equals open parenthesis D times a sub I close parenthesis squared times open bracket s sub i2 of theta sub j minus s squared sub I of theta sub j close bracket.

Alternative Text for Equation 7.10

S sub i of Theta sub j equals the sum from h equals 0 to n sub i of h times p sub ih of theta sub j.

Alternative Text for Equation 7.11

S sub i2 of Theta sub j equals the sum from h equals 0 to n sub i of h squared times p sub ih of theta sub j.

Alternative Text for Equation 7.12

Scale score standard error (SE sub scaled) equals a times SE sub theta sub j.

Alternative Text for Equation 7.13

Error band equals open parenthesis SS minus SE sub scaled comma SS plus SE sub scaled).

Appendix 7.A: Student Completion Conditions

Table 7.A.1 CAASPP Smarter Balanced Student Completion Conditions, ELA

If the Student:	Grade 3 N	Grade 4 N	Grade 5 N	Grade 6 N	Grade 7 N	Grade 8 N	Grade 11 N
Logged on to both CAT and PT, but answered no items	0	0	0	0	0	0	0
Logged on to both CAT and PT and answered at least one item for only CAT or PT	0	0	0	0	0	0	0
Logged on to both CAT and PT and answered at least 1 CAT and 1 PT item but fewer than 10 CAT items	0	0	0	0	0	0	0
Logged on to both CAT and PT and answered at least 1 PT item and at least 10 CAT items but did not answer specified minimum number of items	0	0	0	0	0	0	0
Completed both CAT and PT	119	117	142	111	474	198	12,716
Did not log on to both CAT and PT—not tested medical emergency (NTE)	2	0	2	2	6	6	27
Did not log on to both CAT and PT—parent/guardian exemption (PGE)	149	143	145	109	121	135	292
Did not log on to both CAT and PT—less than 12 months in U.S. exemption, not tested English learner (NEL)	7,077	6,426	5,590	6,002	5,431	4,694	4,314
Did not log on to both CAT and PT—other reasons	445,954	446,949	439,608	456,505	462,360	479,660	452,806
Logged on to both CAT and PT and answered questions, but at least one of the tests was expired	26	26	22	16	23	10	1,594

Table 7.A.2 CAASPP Smarter Balanced Student Completion Conditions, Mathematics

If the Student:	Grade 3 N	Grade 4 N	Grade 5 N	Grade 6 N	Grade 7 N	Grade 8 N	Grade 11 N
Logged on to both CAT and PT, but answered no items	0	0	0	0	0	0	0
Logged on to both CAT and PT and answered at least one item for only CAT or PT	0	0	0	0	0	0	0
Logged on to both CAT and PT and answered at least 1 CAT and 1 PT item but fewer than 10 CAT items	0	0	0	0	0	0	0
Logged on to both CAT and PT and answered at least 1 PT item and at least 10 CAT items but did not answer specified minimum number of items	0	0	0	0	0	0	0
Completed both CAT and PT	99	66	101	46	62	516	10,005
Did not log on to both CAT and PT—NTE	2	0	2	2	6	5	26
Did not log on to both CAT and PT—PGE	146	140	143	108	121	134	296
Did not log on to both CAT and PT—less than 12 months in U.S. exemption, NEL	0	0	0	0	0	0	0
Did not log on to both CAT and PT—other reasons	453,077	453,450	445,254	462,587	468,221	483,994	460,753
Logged on to both CAT and PT and answered questions, but at least one of the tests was expired	3	5	9	2	5	54	667

Chapter 8: Analyses

This chapter summarizes the results of the analyses performed on the data resulting from the 2019–2020 administration. These include item response theory (IRT) parameters from Smarter Balanced item pools; omission, expiration, and completion analyses; overall testing summaries as the means and standard deviations of scale scores and theta values; and the percentages of students at each performance level for each test. Other psychometric analyses, such as item exposure analyses, reliability analysis, consistency and accuracy of the performance level classifications, correlation analysis, student group reliability, interrater reliability for the human-scoring items, the agreement between human scoring and artificial intelligence (AI) scoring, claim scoring analysis, and student group analyses were not conducted and reported, because of the constraints of sample size. All analysis procedures are described in this chapter, although some of analyses could not be conducted for the 2019–2020 administration.

8.1. Background

There are five primary statistical analysis procedures presented in this chapter:

1. IRT parameters
2. Omission and completion analyses
3. Conditional exposure analyses
4. Reliability analyses
5. Analyses in support of validity evidence

8.1.1. Summary of the Analyses

Only analyses that were conducted for the 2019–2020 administration are presented in the body of the text of this subsection and in the listed appendices. Please note that classical item analyses and differential item functioning (DIF) analysis are not presented because these analyses were performed by the Smarter Balanced Assessment Consortium during the 2013–2014 field test administration and the embedded field test administration in each year from 2014–2015 (Smarter Balanced, 2016b; Smarter Balanced, 2019).

1. **IRT Parameters—**Appendix 8.A presents summaries of item difficulty parameter estimates (*b*-values) and item discrimination parameter estimates (*a*-values) for all of the items in each assessment and separate summaries for each claim.

2. **Omission and Completion Analysis—**A student’s record is considered *complete* when the student answers at least one operational performance task (PT) and at least 10 computer adaptive test (CAT) items. Table 8.1 and table 8.2 present the minimum number of operational items in each claim for students who are assigned only operational items and for students who are assigned embedded field test PTs, respectively.

8.1.2. Sample Used for the Analyses

Analyses were conducted based on version 2 of the production data file (“P2”) received on July 16, 2020. The P2 file comprised the full California Assessment of Student Performance and Progress (CAASPP) online summative assessments’ data for the majority of tests. All valid student records were used for the technical report analyses. Students whose records were flagged as “not scored” and students who were enrolled in a different grade than the one in which they were tested were not included.

Items for the embedded field test PTs were embedded into the 2019–2020 operational tests. However, because the field test data was not provided to ETS, none of the PT field test items were analyzed in this chapter.

8.2. Item Response Theory Parameter Values

The purpose of the IRT calibration and scaling is to place item difficulty and student ability estimates onto a common theta scale in each content area. The Common Core State Standards (CCSS) provide a foundation for developing Smarter Balanced assessments that support inferences concerning student changes in achievement (i.e., progress). One approach to modeling student progress across grades is to report scores on a vertical scale, which is a single scale for reporting scores on tests at different grade levels of the same content area. Its purpose is to report scores in a way that shows a student’s progress in a content area, from one grade level to the next. One key assumption with vertical scaling is that it is possible to make meaningful comparisons between scores on tests in the same content area at different grade levels.

Item parameters used in the CAASPP online summative assessments were estimated and scales were constructed during the Smarter Balanced field test administration. Item parameter calibration software, model fit, and evaluation of vertical scale anchor items are not described in the current technical report. For more detailed information on these and other psychometric topics, refer to chapter 6 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a) and subsequent Smarter Balanced technical reports (Smarter Balanced, 2016b, 2017j, 2018k, 2019).

Unidimensional IRT models were used to calibrate items within each content area. Based on the results from the psychometric analyses occurring during the pilot and field test administrations, the Smarter Balanced Assessment Consortium chose the two-parameter logistic model (Birnbaum,1968) for calibration of the dichotomous items and the generalized partial credit model (Muraki, 1992) for calibration of polytomous items. The formula associated with these models is provided in equation 7.1 in subsection *7.4.1.1 Theta Scores*.

Chapter 9 of the *2013–14 Smarter Balanced Technical Report* provides more detailed information about how Smarter Balanced assessments were calibrated and scaled both horizontally and vertically through IRT processes (Smarter Balanced, 2016a).

8.2.1. Summary Information

Parameter estimates for the 2019–2020 operational items were obtained mainly from the 2013–2014 Smarter Balanced field test analyses, but also from the subsequent Smarter Balanced embedded field test analyses after the 2013–2014 administration. Summary statistics of these parameter estimates are calculated to show the difficulty and discrimination of the overall test, as well as the difficulty and discrimination of claims; distributions of *b*-value and *a*-value parameter estimates are created to provide more detail. The step parameters for all polytomous items are also provided.

Appendix 8.A provides summary statistics describing the distributions of item difficulty and discrimination parameter estimates at each test level from the field test calibration and scaling. Note that only operational items from the item pool administered as part of the CAASPP administration are included in this analysis.

For more information regarding the IRT methodology used by Smarter Balanced to form the basis for new item development, test equating, and computer-adaptive testing, refer to chapter 9 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a).

8.2.1.1. All Items

Table 8.A.1 through table 8.A.14 in appendix 8.A present univariate statistics (mean, standard deviation, minimum, and maximum) of the scaled IRT *a*-values. These statistics for each test are presented for all items in the test and for the items in each claim. Table 8.A.15 through table 8.A.28 present the univariate statistics of the IRT *b*-values for all items in the test and for the items in each claim.

8.3. Omission and Completion Analyses

8.3.1. Omit Rates

If a student views an item, leaves it unanswered, and goes on to view and answer another item, the missing response is classified as an “omit.” If the student omits an item—that is, leaves the item unanswered—and does not view additional items, the responses for the successive items are classified as “not seen.”

The percentage of students leaving an item blank can indicate a problem with the time allowed for the test or with some feature of the item. If students are given an adequate amount of testing time, at least 95 percent of the students should attempt to answer each item. The CAASPP online summative assessments are designed to be untimed, allowing all students to respond to all of the items. Because there is no time limit for the test, a percentage of blank responses that is greater than 5 percent for any single item may be an indication of a problem with an item.

8.3.2. Completion Rates

Completion rates indicate the proportion of students who failed to complete a certain number of items in either the CAT or PT portion of the test. Regardless of whether or not the test contains only operational items or also includes embedded field test PTs, a student’s record for the CAT portion is considered incomplete if the student completed fewer than 10 CAT items. For tests that contain only operational items, a student’s record is considered incomplete if the student did not complete at least one operational PT item and at least 10 CAT items. A student’s record is considered *complete* when the student answers at least one operational PT and at least 10 CAT items. However, for tests with embedded field test

PTs, there is no requirement for a student to complete any PT items, so a student's record is considered complete if the student completed at least 10 CAT items.

A student's record for a claim is not considered complete unless the student completed at least the specified minimum number of items for that claim—refer to table 8.1 and table 8.2 for the minimum number of operational items in each claim, for both English language arts/literacy (ELA) and mathematics, for students who are assigned only operational items and for students who are assigned embedded field test PTs, respectively.

Table 8.1 Minimum Number of Items for a Complete Claim Score If No Field Test PT Items

Content Area and Claim	Grades 3–5	Grades 6–8	Grade 11
ELA Claim 1	14	14	15
ELA Claim 2	8	8	8
ELA Claim 3	8	8	8
ELA Claim 4	9	9	9
Mathematics Claim 1	17	16	19
Mathematics Claim 2	8	8	8
Mathematics Claim 3	8	8	8

Table 8.2 Minimum Number of Items for a Complete Claim Score If Test Includes Field Test PT Items

Content Area and Claim	Grades 3–5	Grades 6–8	Grade 11
ELA Claim 1	14	14	15
ELA Claim 2	9	9	9
ELA Claim 3	8	8	8
ELA Claim 4	9	9	9
Mathematics Claim 1	17	16	19
Mathematics Claim 2	8	8	8
Mathematics Claim 3	8	8	8

8.4. Conditional Exposure Rates of Items

Item exposure refers to the frequency of item administration in the student population. Items that are selected too frequently may become known to students in advance of the test administration and, as a result, fail to perform as expected.

Conditional exposure control refers to the establishment of exposure controls to be applied to the items at a specified level of difficulty (*b*-value). These controls become necessary when items at a particular level of difficulty are especially likely to be used too often. For example, it may be necessary to limit item exposure for very difficult items. Because of extremely small samples, this analysis cannot be conducted and reported in the 2019–2020 administration.

8.5. Reliability Analyses

There are many definitions of reliability (Haertel, 2006) that have their genesis in classical test theory and a variety of methods that can be used to estimate reliability.

The general concept of reliability concerns the extent to which the test scores measure *a particular construct* consistently. The variance in the distribution of test scores—essentially, the differences among individuals—is partly due to factors that are consistent over permissible differences in the testing process (e.g., different items or tasks or different raters) and partly due to factors that are not consistent. The measure of variation associated with the first kind of differences—consistent differences—is called "true variance"; the measure of variation associated with the remaining differences—those that operate essentially at random—is called "error variance." Reliability is the proportion of total variance that is due to true variance. The standard error of measurement (SEM) is a statistic that characterizes the error variance. Although no reliability analyses were conducted and reported for the 2019–2020 test administration because of extremely small testing samples, discussion of the concepts and calculation of reliability remains in this chapter.

8.5.1. Sample for Reliability Analyses

The reliability analyses performed for CAASPP require that the sample be screened beyond the requirements listed in subsection *8.1.2 Samples for the Analyses*. When students' ability estimates on the overall test or a claim are lower than the lowest obtainable theta (LOT) for that test, they are assigned the lowest obtainable scale score (LOSS) for that test. When students' ability estimates on the overall test or a claim are higher than the highest obtainable theta (HOT) for that test, they are assigned the highest obtainable scale score (HOSS) for that test. When a student is assigned either the LOSS or HOSS, a measure of the student's true performance is not known, as it would be lower than LOSS or higher than HOSS, which ultimately impacts any reliability analyses. Because of this, the reliability analyses in this section further exclude students assigned the LOSS or HOSS from the student data used for general analyses that was described at the beginning of this chapter. (Refer to subsection *7.4.1.4 Scale Scores for the Total Assessment* for the definitions of LOSS–LOT and HOSS–HOT.)

8.5.2. Marginal Reliability

In a specified population of students, the reliability of test scores, X, is defined as the proportion of the test score variance that is attributable to true differences in student abilities and is sometimes operationalized as the correlation between scores on two replications of the same testing procedure, $\rho_{XX'}$.

Reliability coefficients may range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely students would be to obtain very similar scores if they were retested. In applied settings, the requirement of repeated administrations is impractical, and methodologies estimating reliability from relationships among student performances on items within a single test form are often used. Coefficient alpha (Cronbach, 1951) is among the most common of these methodologies. These reliability indices are not directly applicable to a CAT because each student takes a different test form.

An IRT-based approach called marginal reliability (Green, Bock, Humphreys, Linn, & Reckase, 1984) can be used to estimate the reliability of CAT scores. The estimates of reliability coefficients reported here are for item response model-based ability estimates.

This reliability coefficient for theta estimates, $\rho_{\theta\theta'}$, is defined based on a single test administration, as shown in equation 8.1:

$$\rho_{\theta\theta'} = 1 - \frac{M_{SEM_\theta^2}}{s_\theta^2} \qquad (8.1)$$

Refer to the Alternative Text for Equation 8.1 for a description of this equation.

where,

s_θ^2 is the measure of variance in ability estimates,

θ is an ability estimate, and

$M_{SEM_\theta^2}$ is an average of the squared conditional standard error of measurement (CSEM) (i.e., error variances) at each value of the ability estimate.

8.5.3. Standard Error of Measurement

The SEM provides a measure of score instability in the scale score metric. The SEM is the square root of the error variance in the scores (i.e., the standard deviation of the distribution of the differences between students' observed scores and their true scores). The SEM is calculated by:

$$SEM_{Scaled} = a \times s_\theta \sqrt{1 - \rho_{\theta\theta'}} \qquad (8.2)$$

Refer to the Alternative Text for Equation 8.2 for a description of this equation.

where,

$\rho_{\theta\theta'}$ is the reliability estimated in equation 8.1,

S_θ is the standard deviation of the total test θ score, and

a is the slope of the scaling transformation of θ to the reporting scale.

The SEM is useful in determining the confidence interval (CI) that likely captures a student's true score. A student's true score can be thought of as the mean of observed scores a student would earn over an infinite number of independent administrations of the test. Across those administrations, approximately 95 percent of the CIs from the student's observed score -1.96 SEMs to the student's observed score +1.96 SEMs would contain that student's true score (Crocker & Algina, 1986). Therefore, this interval is called a 95 percent CI for the student's true score. For example, if a student's observed score on a given test equals 2440 points, and the SEM equals 23, one can be 95 percent confident that the student's true score lies between 2395 and 2485 points (2440 ± 45).

8.5.4. Conditional Standard Errors of Measurement

CSEMs are estimated as part of the IRT-based scoring procedure. CSEMs for scale scores are based on IRT and are estimated as a function of measured ability. The CSEMs of theta scores (or of linearly transformed theta scores) are typically smaller in scale score units toward the center of the scale in the test metric where more items are located. The CSEMs are usually larger at the extreme ends of the scale, because there is no way to know how much better than that a student really is in the case of an extremely high score, or how much worse than that a student really is in the case of an extremely low score, given the difficulty of content administered to the student. A student's CSEM under the IRT framework is equal to the reciprocal of the square root of the test information function (TIF):

$$\mathrm{CSEM(SS)} = a \times \frac{1}{\sqrt{\mathrm{I}(\theta)}} \tag{8.3}$$

Refer to the Alternative Text for Equation 8.3 for a description of this equation.

where,

$SS = a \times \theta + b$,

CSEM (SS) is the conditional SEM on scale score scale, and

$I(\theta)$ is the TIF at ability level θ, as is shown in equations 7.8 to 7.11, which are in subsection *7.4.3 Theta Scores Standard Error*.

The statistic is multiplied by a, where a is the scaling factor needed to transform theta to the scale score metric. The intercept to transform theta to the scale score is denoted as b. The values of a and b vary by content area and are shown in equations 7.5 and 7.6 for ELA and mathematics, respectively. (These equations are in subsection *7.4.1.4 Scale Scores for the Total Assessment*.)

Because the Smarter Balanced assessments use item pattern scoring, each response pattern can have a unique ability estimate and CSEM. Some response patterns have more uncertainty or random error associated with their ability estimates at the upper or lower ends of the reporting scale, where items administered to students may not be well aligned to a student's true ability level. For example, if there are not enough difficult items in the item pool, a high-ability student may not be presented with difficult items on every replication of the CAT. Under these circumstances, while the student's scale score will be high, the student's CSEM may not be well estimated.

To reduce the level of uncertainty, the CSEMs were averaged at each scale score point. In addition, the uncertainty associated with CSEMs across the entire ability continuum, including the extreme ends, was further reduced by loglinear smoothing. Loglinear smoothing is implemented by using loglinear models to replace a discrete empirical dataset with a discrete dataset that preserves some features of the observed data without the irregularities that are attributable to sampling. Loglinear models can preserve a variety of different features in observed data with a relatively small number of parameters (Moses, von Davier, & Casabianca, 2004). Loglinear smoothing is implemented through LOGLIN, which is a function of an open-source software *KE* (ETS, 2011).

The average CSEMs at each scale score point are estimated from the 2017–2018 Smarter Balanced Summative Assessment data for all students. Given the stability across the 2017–2018 through 2019–2020 California student populations and the stability of the item pools, the relationship between the reporting scale and CSEMs should remain stable across administration years. The stability of this relationship helps facilitate the estimation of CSEMs prior to the test administration instead of after the completion of all testing windows.

CSEMs vary across the θ scale. When a test has thresholds, it is important to estimate CSEMs at those thresholds. Table 8.3 presents the scale score CSEMs at the lowest score required for a student to be classified in the *Standard Nearly Met*, *Standard Met*, and *Standard Exceeded* achievement levels for each test.

Table 8.3 Scale Score CSEM at Performance-Level Thresholds

Content Area and Grade	Standard Nearly Met Minimum SS	Standard Nearly Met CSEM	Standard Met Minimum SS	Standard Met CSEM	Standard Exceeded Minimum SS	Standard Exceeded CSEM
ELA 3	2367	24.00	2432	22.00	2490	23.00
ELA 4	2416	26.00	2473	25.00	2533	25.00
ELA 5	2442	24.00	2502	24.00	2582	25.00
ELA 6	2457	25.00	2531	24.00	2618	25.00
ELA 7	2479	27.00	2552	25.00	2649	26.00
ELA 8	2487	26.00	2567	25.00	2668	27.00
ELA 11	2493	31.00	2583	28.00	2682	29.00
Mathematics 3	2381	19.00	2436	17.00	2501	16.00
Mathematics 4	2411	19.00	2485	17.00	2549	17.00
Mathematics 5	2455	23.00	2528	19.00	2579	18.00
Mathematics 6	2473	24.00	2552	21.00	2610	20.00
Mathematics 7	2484	27.00	2567	23.00	2635	20.00
Mathematics 8	2504	28.00	2586	26.00	2653	21.00
Mathematics 11	2543	30.00	2628	25.00	2718	22.00

Table 8.4 presents the average CSEMs in each achievement level by content area and grade level. The CSEMs tended to be smaller in the achievement levels of *Standard Nearly Met*, *Standard Met,* and *Standard Exceeded* than *Standard Not Met* for all tests. The pattern of average CSEMs is similar for the tests in each content area.

Table 8.4 Mean CSEM for Each Achievement Level

Content Area and Grade	Standard Not Met	Standard Nearly Met	Standard Met	Standard Exceeded
ELA 3	27.46	22.58	22.22	23.52
ELA 4	28.74	25.91	25.00	26.03
ELA 5	26.99	24.00	24.37	26.74
ELA 6	28.62	24.20	24.28	26.05
ELA 7	30.14	25.55	25.29	28.08
ELA 8	29.87	25.77	25.44	28.22
ELA 11	36.22	29.51	28.03	30.25
Mathematics 3	21.78	17.84	16.66	17.32
Mathematics 4	23.70	18.38	17.00	17.81
Mathematics 5	28.58	20.72	18.13	17.85
Mathematics 6	31.21	22.02	20.23	20.08
Mathematics 7	34.95	25.14	21.63	20.31
Mathematics 8	36.65	26.92	23.42	21.47
Mathematics 11	41.01	27.71	23.50	21.45

8.5.5. Decision Classification Analyses

When an assessment uses achievement levels as the primary method to report test results, accuracy and consistency of decisions become key indicators of the quality of the assessment.

Decision accuracy is the extent to which students are classified in the same way as they would be if each student’s score were the average over all possible forms of the test (the student’s true score). Decision accuracy answers the following question: How closely does the actual classification of test takers, based on their single-form scores, agree with the classification that would be made on the basis of their true scores, if their true scores could somehow be known?

Decision consistency is the extent to which students are classified in the same way as they would be on the basis of a single form of a test other than the one for which data is available. Decision consistency answers the following question: What is the agreement between the classifications based on two nonoverlapping, equally difficult forms of the test?

The methodology used for estimating the reliability of classification decisions is described in Livingston and Lewis (1995). The necessary input information includes only the maximum and minimum possible scores on the test and the observed score distribution and the reliability coefficient for the group of students that the estimates will refer to. The method was implemented by the ETS proprietary computer program RELCLASS-COMP (Version 4.14).

Reliability of classification at a threshold is estimated by combining the achievement levels above a particular threshold and combining the achievement levels below that threshold. The result is a two-by-two table indicating whether the students are above or below the threshold. The sum of the entries in the main diagonal is the number of students accurately (or consistently) classified as above or below that threshold. Table 8.5 and table 8.6 illustrate these two-by-two contingency tables.

Table 8.5 Decision Accuracy for Reaching an Achievement Level

Achievement Level Status	Does Not Reach an Achievement Level Based on True Score	Reaches an Achievement Level Based on True Score
Does not reach an achievement level	Consistent classification	Inconsistent classification
Reaches an achievement level	Consistent classification	Consistent classification

Table 8.6 Decision Consistency for Reaching an Achievement Level

Achievement Level Status	Does Not Reach an Achievement Level Based on an Alternate Form	Reaches an Achievement Level Based on an Alternate Form
Does not reach an achievement level	Consistent classification	Inconsistent classification
Reaches an achievement level	Inconsistent classification	Consistent classification

8.5.6. Interrater Agreement

To monitor the consistency of ratings assigned to students’ responses by raters, approximately 10 percent of the constructed-response (CR) items received a second rating. The two sets of ratings are used to compute statistics describing the consistency (or reliability) of the ratings. This interrater consistency is described in three ways:

1. Percentage agreement between two raters
2. Cohen’s Kappa
3. Quadratic weighted kappa (QWK) coefficient

8.5.6.1. Percentage Agreement

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

8.5.6.2. Kappa

Interrater reliability or consistency is an indicator of homogeneity and is most frequently measured using an intraclass correlation (ICC) which incorporates the exact agreement between raters over and above that expected by chance. The index is defined as the following:

$$\mathrm{ICC} = r_I = (ms_{between} - ms_{within})/(ms_{between} + [k - 1]ms_{within}) \qquad (8.4)$$

Refer to the Alternative Text for Equation 8.4 for a description of this equation.

where,

$ms_{between}$ is the mean-square estimate of between-subjects variance, and

ms_{within} is the mean-square estimate of within-subjects variance.

For categorical ratings, Cohen's Kappa statistic (1960) has the properties of an ICC and can be used for interrater reliability. Cohen's Kappa is therefore used as a primary indicator of the interrater reliability of the human-scored items. In addition, the percentage of ratings on which the raters are in exact agreement or differ by just one point are computed.

8.5.6.3. Quadratic Weighted Kappa

QWK is used because kappa does not take into account the degree of disagreement between raters. It is a generalization of the simple kappa coefficient using weights to quantify the relative difference between categories. The range of the QWK is from 0.0 to 1.0, with perfect agreement being equal to 1.0.

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

$$\kappa_{ij} = \frac{\left(\sum_{i=0}^{m}\sum_{j=0}^{m} w_{ij}\frac{n_{ij}}{n_{++}}\right) - \left(\sum_{i=0}^{m}\sum_{j=0}^{m} w_{ij}\frac{n_{i+}n_{+j}}{n_{++}^2}\right)}{1-\left(\sum_{i=0}^{m}\sum_{j=0}^{m} w_{ij}\frac{n_{i+}n_{+j}}{n_{++}^2}\right)} \tag{8.5}$$

Refer to the Alternative Text for Equation 8.5 for a description of this equation.

For QWK, the weights are:

$$w_{ij} = 1 - \frac{(i-j)^2}{m^2} \tag{8.6}$$

Refer to the Alternative Text for Equation 8.6 for a description of this equation.

The interrater reliability analyses are performed on approximately 10 percent of the overall testing population, randomly selected from the total population; those students' responses are scored by two raters. In some scoring rubrics, zero is a valid score for the responses but is not provided by a rater. Instead, a score of zero is assigned when the student attempted the writing task but did not provide a response. Responses with zero scores should not be included in the calculation of the agreement statistics for these items.

Refer to *Chapter 7: Scoring and Reporting* of this report and the *Smarter Balanced Scoring Guide for Grades Three, Six, and Eleven: English/Language Arts PT Full-Write Baseline Sets* (Smarter Balanced, 2014) for scoring dimensions.

8.5.7. Agreement Between Artificial Intelligence and Human Scoring

To ensure that the AI scoring engine awards scores that are consistent with the scores assigned by qualified human raters, Measurement Incorporated, the CAASPP subcontractor scoring some of the CR items, conducts ongoing quality checks to ensure that the scoring models perform consistently. A description of these quality checks is provided in subsection *7.2.2. Quality Control of Artificial Intelligence Scoring*.

8.6. Validity Evidence

Validity refers to the degree to which each interpretation or use of a test score is supported by the accumulated evidence (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; ETS, 2014). It constitutes the central notion underlying the development, administration, and scoring of a test and the uses and interpretations of test scores.

Validation is the process of accumulating evidence to support each proposed score interpretation or use. This validation process does not rely on a single study or gathering only one type of evidence. Rather, validation involves multiple investigations and different kinds of supporting evidence (AERA, APA, & NCME, 2014; Cronbach, 1971; ETS, 2014; Kane, 2006). It begins with the test design and is implicit throughout the entire assessment process, which includes item development and field testing, analyses of items, test scaling and linking, scoring, reporting, and score usage.

In this section, the evidence gathered is presented to support the intended uses and interpretations of scores for the CAASPP online summative assessments. This section is organized primarily around the principles prescribed by AERA, APA, and NCME's *Standards for Educational and Psychological Testing* (2014). These *Standards* require a clear definition of the purpose of the test, a description of the constructs to be assessed, and the population to be assessed, as well as how the scores are to be interpreted and used. Since many aspects of the CAASPP System are still under development at the time of this report, additional research to further support the Smarter Balanced goals is mentioned as appropriate throughout this section.

The *Standards* identify five kinds of evidence that can provide support for score interpretations and uses:

1. Evidence based on test content
2. Evidence based on relations to other variables
3. Evidence based on response processes
4. Evidence based on internal structure
5. Evidence based on the consequences of testing

The next subsection defines the purpose of the CAASPP online summative assessments, followed by a description and discussion of the kinds of validity evidence that have been gathered. For general test validity evidence collected by the Smarter Balanced Assessment Consortium, refer to chapter 1 of the *2014–15 Smarter Balanced Technical Report* (Smarter Balanced, 2016b). The validity evidence presented in chapter 1 of that report was collected from the results of a pilot test and a field test prior to the operational administration of the nationwide Smarter Balanced Online Summative Assessments.

8.6.1. Evidence in the Design of CAASPP

8.6.1.1. Purpose

The purpose of the CAASPP assessment system is to provide school staff and teachers with information and tools they need to improve teaching and learning so as to prepare all students for college and career readiness.

8.6.1.2. Constructs to Be Measured

The CAASPP online summative assessments are designed to show how well students perform relative to the Smarter Balanced Assessment Consortium content standards, which are aligned to the CCSS. These standards describe what students should know and be able to do at each grade level.

Test blueprints define the procedures used to measure the claims and standards. These blueprints, for ELA and mathematics, are provided in appendix 2.A. They also provide an operational definition of the construct to which each set of standards refers. That is, they define, for each content area, the subject to be assessed, the tasks to be presented, the administration instructions to be given, and the rules used to score student responses. The test blueprints control as many aspects of the measurement procedure as possible so that the testing conditions will remain the same over test administrations (Cronbach, 1971) to minimize construct-irrelevant score variance (Messick, 1989).

The Smarter Balanced Assessment Consortium also created the content specifications used to create the CAASPP online summative assessments (Smarter Balanced, 2015a and 2015b).

8.6.1.3. Interpretations and Uses of the Scores

Overall student performance is expressed as scale scores and achievement levels, which are generated for both ELA and mathematics assessments, as are strength and weakness levels for each claim. An inference is drawn about how much knowledge and skill in the content area the student has, on the basis of a student’s total score. The total score is also used to classify students in terms of their level of knowledge and skill in the content area. These levels are called achievement levels and are labeled *Standard Exceeded*, *Standard Met*, *Standard Nearly Met*, and *Standard Not Met*.

The strength and weakness levels are used to draw inferences about a student’s achievement in each of the claims for each test. A detailed description of the uses and applications of the CAASPP online summative assessment scores is presented in chapter 7. Parents/Guardians have access to the Starting Smarter website, which describes CAASPP Student Score Reports and how parents/guardians can use the reports to communicate with teachers about a child’s learning (The Regents of the University of California & CDE, 2020). The information provided is available in English and Spanish. Finally, additional information can be found in the *2018–19 CAASPP Post-Test Guide* (CDE, 2019), which also was applicable for the 2019–2020 CAASPP administration.

The results for tests within the CAASPP System have four primary purposes:

1. Help facilitate conversations between parents/guardians and teachers about student performance
2. Serve as a tool to help parents/guardians and teachers work together to improve student learning
3. Help staff from schools and local educational agencies identify strengths and areas that need improvement in their educational programs
4. Provide the public and policymakers with information about student achievement

More detailed descriptions regarding score use can be found in the *Education Code* Section 60602 web page at https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=EDC&division=4.&title=2.&part=33.&chapter=5.&article=1 (outside source).

8.6.1.4. Intended Test Population

Students enrolled in grades three through eight and grade eleven are required to take part in the Smarter Balanced Summative Assessments, unless they are eligible to participate in the alternate assessments. English learners who were in their first 12 months of attending school in the United States were exempt from taking the ELA portion of the assessments.

8.6.2. Evidence Based on Test Content

Evidence based on test content refers to traditional forms of content validity evidence, such as the rating of test specifications and test items (Crocker, Miller, & Franks, 1989; Sireci, 1998), as well as alignment methods for educational tests that evaluate the interactions between curriculum frameworks, testing, and instruction (Rothman, Slattery, Vranek, & Resnick, 2002; Bhola, Impara & Buckendahl, 2003; Martone & Sireci, 2009).

The degree to which the Smarter Balanced test specifications captured the CCSS, and the items adequately represent the domains delineated in the test specifications, were demonstrated in the *Alignment Study Report* (Smarter Balanced, 2016c). The major finding presented here is that the knowledge, skills, and abilities measured by the Smarter Balanced assessments are consistent with the ones specified in the CCSS. With computer-adaptive testing, an extra dimension of content validity evidence is to ensure that the item-selection algorithm produces forms for individual students that conform to the test blueprint. It was found that across content areas and grade levels, 98 percent or more of the simulated tests covered the test blueprint (American Institutes for Research [AIR], 2015).

8.6.2.1. Description of the State Standards

As noted in section *1.1 Background*, the Smarter Balanced Summative Assessments are aligned with the CCSS for ELA and mathematics. The purpose of the CCSS is to provide school staff and teachers with the information and tools they need to improve teaching and learning so as to prepare all students for college and career readiness. These content standards describe what students should know and be able to do at each grade level (Smarter Balanced, 2015a and 2015b).

8.6.2.2. Item Specifications

Item specifications describe the characteristics of items that are written to measure each content standard. Specifications were developed for each target, within each claim, and at each grade level, and are published by the Smarter Balanced Assessment Consortium for ELA (Smarter Balanced, 2017a through 2017i) and mathematics (Smarter Balanced, 2018a through 2018j).

8.6.2.3. Item Selection Algorithm

The item selection algorithm is designed to cover a standards-based blueprint in the assembly of CAT forms. The general item selection approach is based on an item selection algorithm (refer to *Chapter 4: Test Assembly*) that evaluates an item's contribution to each of the following measures:

1. A measure of content match to the blueprint
2. A measure of overall test information
3. Measures of test information for each reporting category on the test

Details can be found in the *Smarter Balanced Adaptive Item Selection Algorithm Design Report (*AIR, 2014).

8.6.2.4. Assessment Blueprints

The Smarter Balanced summative test blueprints provided in appendix 2.A describe the content of the ELA and mathematics summative assessments for all grades tested and how that content is assessed. The summative online test blueprints reflect the depth and breadth of the performance expectations of the CCSS. The test blueprints have information about the number of items and depth of knowledge for items associated with each assessment target. Each test is described by a single blueprint for each segment of the test and identifies the order in which the segments appear.

8.6.2.5. Item Development Process

A detailed description of the content and psychometric criteria applicable to the construction of the Smarter Balanced item pool is included in *Chapter 4: Test Design*, for overall content validity, and *Chapter 3: Item Development*, for item development, of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a).

8.6.2.6. Alignment Study

A strong alignment between the CCSS and assessments is fundamental to the meaningful measurement of student achievement and instructional effectiveness. Alignment results demonstrate that the assessments represent the full range of the content standards and that these assessments measure student knowledge in the same manner and at the same level of complexity as expected in the content standards. For example, across all grades, 64.7 percent of the items are identified in alignment with the ELA grade-level CCSS and 76.7 percent of the items are identified in alignment with the mathematics grade-level CCSS by at least 50 percent of the reviewers (Smarter Balanced, 2016c).

8.6.2.7. Form Assembly Process

The content standards, blueprints, and item-selection algorithm are the basis for choosing items for each assessment. Additional item difficulty and discrimination targets are defined in light of what are desirable statistical characteristics in test items and statistical evaluations. Refer to *Chapter 4: Test Assembly* for additional information.

8.6.2.8. Simulation Study

Simulations are conducted to evaluate and ensure the implementation and quality of the adaptive item-selection algorithm and the scoring algorithm. The simulation tool allows for the manipulation of key blueprint and configuration settings to match the blueprint and minimize measurement error. The report *Smarter Balanced Summative Assessments Testing Procedures for Adaptive Item-Selection Algorithm* contains more information about the algorithms used (AIR, 2015).

The findings from the 2016–2017 simulation study demonstrate that the Smarter Balanced adaptive test delivery system administers assessments with items representing the breadth and depth identified in the test specifications and content standards, and that scores are comparable with respect to the targeted content and are measured with good precision across the range of proficiency. Refer to *Simulation Results, 2016–17 Test Administrations English Language Arts/Literacy grades 3–8, 11, and Mathematics Grades 3–8, 11* for detailed information (AIR, 2016)*.*

8.6.3. Evidence Based on Response Processes

Validity evidence based on response processes refers to “evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by test takers” (AERA et al., 2014, p. 15). This type of evidence generally includes documentation of activities such as

- interviews with students concerning their responses to test items (i.e., think alouds);
- systematic observations of test response behavior;
- evaluation of the criteria used by judges when scoring PTs, analysis of student item response-time data, and features scored by automated algorithms; and
- evaluation of the reasoning processes students employ when solving test items (Embretson, 1983; Messick, 1989; Mislevy, 2009).

This type of evidence is used to confirm that the Smarter Balanced assessments are measuring the cognitive skills that are intended to be the objects of measurement and that students are using these targeted skills to respond to the items.

8.6.3.1. Think Alouds

One way to evaluate response process is through think-aloud protocols (Lewis, 1982). Think-aloud protocols were conducted early in the development of the Smarter Balanced assessments and were described by Smarter Balanced (2015a) in the following way:

> “Using the revised item and task specifications, a small set of items was developed and administered in fall 2012 during a small-scale trial. This provided the Consortium with their first opportunity to administer and score the new item types. During the small-scale trials, the Consortium also conducted cognitive laboratories to better understand how students respond to various types of items. The cognitive laboratories used a think-aloud methodology in which students speak their thoughts while working on a test item. The item and task specifications were again revised based on the findings of the cognitive laboratories and the small-scale trial. These revised specifications were used to develop items for the 2013 pilot test, and they were again revised based on 2013 pilot test results and subsequent review by content experts.”

8.6.3.2. Analysis of Testing Time

Testing times for each administration can be evaluated for consistency, with the expected response processes for the tasks presented to students. The length of time it takes students to take a test is recorded and analyzed to build a profile describing what a typical testing event looks like for each content area and grade. In addition, variability in testing time is investigated to determine whether a student’s testing time should be viewed as unusual or irregular. It should be noted that the Smarter Balanced assessments are untimed tests.

In these analyses, only students who completed at least 10 CAT items and 1 PT item and had timing records are included. The students having the shortest testing time in the PT portion—1 percent of all the students taking the test—and the students with the shortest testing time in the CAT portion—also 1 percent of all the students taking the test—are removed from the analysis. The remaining testing population is partitioned into quartiles based on scale scores on the total test. These groupings are not the same as the achievement levels.

The descriptive statistics—e.g., the number of students, mean, standard deviation, minimum and maximum, and percentiles—of the following time variables are computed for each of the four quartile groups derived from the scale scores for each content area:

- Time required to complete the total test
- Time required to complete the CAT section of each test
- Time required to complete the PT section of each test

Some cases of extremely long testing time may be attributed to students with special needs taking longer to complete the tests, or the test not being closed down properly. Therefore, mean testing times may be misleading. The medians (50th percentile) are more meaningful in the interpretation of the time comparisons because medians are less impacted by the extreme values than means. The removal of the 1 percent of the student data with the shortest testing time is a modest exclusion that leaves some students with very short durations in the results for each of the tests. Similarly, some very long durations are present in the data, which may indicate errors such as the failure to close a testing session. Therefore, the median is a better statistic than the mean for evaluating testing time information. Because of the extremely small testing sample, time analysis was not conducted and reported for the 2019–2020 administration.

8.6.4. Evidence Based on Internal Structure

Validity evidence based on *internal structure* refers to the statistical analysis of item and score subdomains to investigate the primary and secondary (if any) dimensions measured by an assessment. Procedures for gathering such evidence include factor analysis—both exploratory and confirmatory—or multidimensional IRT scaling. With a vertical scale, a consistent primary dimension across the levels of the test should be maintained.

8.6.4.1. Dimensionality

A dimensionality study was conducted during the pilot test phase to determine the factor structure of the assessments and the types of scales developed, as well as the associated IRT models used to calibrate them. In part, that study used the Akaike Information Criterion (Akaike, 1973) to evaluate the fit of potential multidimensional models relative to the unidimensional model. The results suggested that the unidimensional model fit better than the multidimensional model, once model complexity was taken into account. More detailed results for the Smarter Balanced pilot test are available in the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a).

8.6.4.2. Differential Item Functioning

Analysis of item functioning using IRT and DIF falls under the internal structure category. For Smarter Balanced, DIF analyses were conducted to assess differences in the item performance of groups of students who differ in their demographic characteristics. DIF analyses were implemented during the pilot test and field test phases when the tests were delivered in linear fixed-length forms (Smarter Balanced, 2016a, chapter 6; and Smarter Balanced, 2016b, chapter 8). For both ELA and mathematics, few items were identified as having significant levels of DIF. In the operational assessment, by virtue of the CAT delivery, non-embedded field test items are not amenable to DIF analyses.

8.6.4.3. Overall Reliability Estimates

The results of reliability analyses on the total test theta scores on each summative test are presented for previous tests but not for the 2019–2020 administration because of the impacts of the novel coronavirus disease 2019 (COVID-19) pandemic. Previous results indicated that the reliability estimates for all summative test total scores were high, ranging from 0.92 to 0.95. Theta score standard deviations and SEMs increased with grade level; this is often an artifact of vertical scaling.

8.6.4.4. Claim Reliability Estimates

For each CAASPP online summative assessment, theta scores are computed for claims. The reliability estimates of these scores are usually conducted and reported, except this year because of the impacts of the COVID-19 pandemic. Previous reliability estimates of claims were invariably lower than those for the total tests because they were based on fewer items. Because the reliabilities of scores at the claim level are lower than for total scores, and because each claim contains a different number of items, educators should supplement the score results with other information when interpreting claim scores.

8.6.4.5. Student Group Reliability Estimates

The reliabilities are usually examined for various student groups within the student population that differ in their demographic characteristics; they were not this year, because of the impacts of the COVID-19 pandemic. The characteristics considered are gender, ethnicity, economic status, special education services status, migrant status, English language fluency, military status, homeless status, and ethnicity by economic status.

8.6.4.6. Reliability of Performance Classifications

The methodology used for estimating the reliability of classification decisions is described with the decision classification analyses in subsection *8.5.5 Decision Classification Analyses*.

8.6.4.7. Interrater Reliability

Cohen’s Kappa statistics provide evidence of the degree to which a student’s score is consistent from one rater to another. Research has shown values of kappa between 0.41 and 0.60 exhibit moderate levels of agreement between the two ratings (Landis & Koch, 1977; Flack, Afifi, Lachenbruch, & Schouten, 1988) and that values of QWK greater than 0.70 indicate excellent agreement (Williamson, Xi, & Breyer, 2012).

8.6.5. Evidence Based on Relations to Other Variables

Evidence based on *relations to other variables* refers to traditional forms of criterion-related validity evidence such as concurrent and predictive validity, as well as more comprehensive investigations of the relationships among test scores and other variables such as multitrait-multimethod studies (Campbell & Fiske, 1959). External variables can be used to evaluate hypothesized relationships between test scores and other measures of student achievement (e.g., test scores) to evaluate the degree to which different tests actually measure different skills and the utility of test scores for predicting specific criteria (e.g., college grades). This type of evidence is essential for supporting the validity of certain inferences based on scores from the Smarter Balanced assessments for certifying college and career readiness, which are the primary test purposes.

A subset of students who took National Assessment of Educational Progress (NAEP) and Program for International Student Assessment (PISA) items also took Smarter Balanced CAT items and PTs. A summary of the resulting item performance for NAEP, PISA, and all Smarter Balanced items can be found in chapters 7 and 8 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a). That study found item-level performance to be similar for NAEP and Smarter Balanced populations. A study taking the next step of relating Smarter Balanced scales to NAEP or PISA scales has not yet been completed.

Another study established the relationship between Smarter Balanced field test scores and the likelihood of achieving “Conditionally Exempt” status based on achieving the required minimum scores for the California State University Early Assessment Program (EAP). During the 2013–2014 administration, students in grade eleven took the EAP for ELA, mathematics, or both. The comparison showed a correlation of 0.68 between Smarter Balanced ELA and EAP ELA assessments and correlations from 0.49 to 0.61 between Smarter Balanced mathematics and EAP mathematics tests (ETS, 2015a, 2015b, and 2015c). These correlations indicate that Smarter Balanced Summative Assessments might be measuring different aspects of college readiness than the EAP assessments, which previously provided insight into the readiness of California students in grade eleven for college-level mathematics and ELA courses. Other predictive validity research is being pursued by the Smarter Balanced Assessment Consortium as part of their research agenda.

8.6.6. Evidence Based on Consequences of Testing

Evidence based on *consequences of testing* refers to the evaluation of the intended and unintended consequences associated with a testing program. Examples of evidence based on testing consequences include investigations of adverse impact, evaluation of the effects of testing on instruction, and evaluation of the effects of testing on issues such as high school dropout rates. With respect to educational tests, the *Standards* stress the importance of evaluating test consequences. For example, they state the following:

> “When educational testing programs are mandated . . . the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the use of the test, both intended and unintended, should also be examined by the test user.” (AERA et al., 1999, p. 145)

Investigations of testing consequences relevant to the Smarter Balanced goals include analyses of students’ opportunity to learn the CCSS and analyses of changes in textbooks and instructional approaches. Unintended consequences, such as changes in instruction, diminished morale among teachers and students, increased pressure on students leading to increased dropout rates, or the pursuit of college majors and careers that are less challenging can be evaluated. These sorts of investigations require information beyond what has been available to the CAASPP program to date. Refer to the *Smarter Balanced Assessment Consortium: 2017–18 Technical Report* (Smarter Balanced, 2019) for more validity evidence.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), Proceedings from *2nd International Symposium Information Theory* (pp. 267–81). Budapest, Hungary: Akademia Kiado.

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.

American Institutes for Research. (2014). *Smarter Balanced adaptive item selection algorithm design report.* Washington, DC: American Institutes for Research. http://www.smarterapp.org/documents/AdaptiveAlgorithm-Preview-v3.pdf

American Institutes for Research. (2015). *Smarter Balanced summative assessments testing procedures for adaptive item-selection algorithm, 2014–2015 test administrations, English language arts/literacy (ELA), grades 3–8 and 1, and mathematics, grades 3–8 and1.* Washington, DC: American Institutes for Research. https://portal.smarterbalanced.org/library/en/testing-procedures-for-adaptive-item-selection-algorithm.pdf

American Institutes for Research. (2016). *Smarter Balanced Summative Assessments simulation results, 2016–17 test administrations English language arts/literacy grades 3-8,11, mathematics grades 3-8, 11.* Washington, DC: American Institutes for Research. https://portal.smarterbalanced.org/library/en/2016-17-summative-assessments-simulation-results.pdf

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*, 21–29.

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

California Department of Education. (2019). *2018–19 CAASPP post-test guide: Technical information for student score reports for CAASPP LEA and test site coordinators and research specialists.* Sacramento, CA: California Department of Education. https://bit.ly/3b3l4ko

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York, NY: Holt.

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*, 179–94.

Cronbach L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

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

Educational Testing Service. (2011). KE (Version 3) [Software]. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2014). *ETS standards for quality and fairness.* Princeton, NJ: Educational Testing Service.

Educational Testing Service (2015a). *Linking study between Smarter Balanced ELA field test and California State University (CSU) English Placement Test.* [Memorandum]. Sacramento, CA: Educational Testing Service.

Educational Testing Service (2015b). *Linking study between Smarter Balanced ELA field test and CSU entry-level mathematics test.* [Memorandum]. Sacramento, CA: Educational Testing Service.

Educational Testing Service. (2015c). *Study of the relationship between the Early Assessment Program and the Smarter Balanced field tests.* Sacramento, CA: Educational Testing Service. https://www.cde.ca.gov/ta/tg/ca/documents/eapstudy.pdf

Embretson (Whitley), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Flack, V. F., Afifi, A. A., Lachenbruch, P. A., & Schouten, H. J. A. (1988). Sample size determinations for the two rater Kappa statistics. *Psychometrika, 53*(3), 321–325.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347–360.

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

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

Landis, J. R., & Koch, G. G. (1977). The measurement of interrater agreement for categorical data. *Biometrics*, *33*, 159–74.

Lewis, C. H. (1982). *Using the “thinking aloud” method in cognitive interface design* [Technical report]. IBM. RC-9265.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement*, *32,* 179–97.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction. *Review of Educational Research, 4*, 1332–61.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. *CRESST Report 752*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Moses, T., von Davier, A. A., & Casabianca, J. (2004). *Loglinear smoothing: An alternative numerical approach using SAS*. Princeton, NJ: Educational Testing Service. https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2004.tb01954.x/pdf

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2): 159–176.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing.* [Technical Report 566]. Washington, DC: Center for the Study of Evaluation.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*, 299–321.

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced scoring guide for grades 3, 6, and 11 English/language arts PT full-write baseline sets*. Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/scoring-guide-for-ela-full-writes.pdf

Smarter Balanced Assessment Consortium. (2015a). *Content specifications for the summative assessment of the Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/english-language-artsliteracy-content-specifications.pdf

Smarter Balanced Assessment Consortium. (2015b). *Content specifications for the summative assessment of the Common Core State Standards for mathematics*. Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/mathematics-content-specifications.pdf

Smarter Balanced Assessment Consortium. (2016a). *Smarter Balanced Assessment Consortium: 2013–14 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf

Smarter Balanced Assessment Consortium. (2016b). *Smarter Balanced Assessment Consortium: 2014–15 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf/

Smarter Balanced Assessment Consortium. (2016c). *Smarter Balanced Assessment Consortium: Alignment study report.* Alexandria, VA: Human Resource Research Organization. http://www.smarterapp.org/documents/AlignmentStudyReport.pdf

Smarter Balanced Assessment Consortium. (2017a). *ELA CAT item specifications, grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017b). *ELA CAT item specifications, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017c). *ELA CAT item specifications, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017d). *ELA PT item specifications, argumentative, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017e). *ELA PT item specifications, explanatory, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017f). *ELA PT item specifications, informative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017g). *ELA PT item specifications; narrative, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017h). *ELA PT item specifications, narrative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017i). *ELA PT item specifications, opinion, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2017j). *Smarter Balanced Assessment Consortium: 2015–16 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/2015-16-summative-technical-report.pdf

Smarter Balanced Assessment Consortium. (2018a). *Mathematics CAT item specifications, Claim 1, grade eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018b). *Mathematics CAT item specifications, Claim 1, grade five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018c). *Mathematics CAT item specifications, Claim 1, grade four*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018d). *Mathematics CAT item specifications, Claim 1, grade seven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018e). *Mathematics CAT item specifications, Claim 1, grade six*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018f). *Mathematics CAT item specifications, Claim 1, grade three*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018g). *Mathematics CAT item specifications, Claim 1, high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018h). *Mathematics CAT item specifications, Claim 2, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018i). *Mathematics CAT item specifications, Claim 3, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018j). *Mathematics CAT item specifications, Claim 4, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium. ~~https://case.smarterbalanced.org/cfdoc/~~

Smarter Balanced Assessment Consortium. (2018k). *Smarter Balanced Assessment Consortium: 2016–17 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. http://portal.smarterbalanced.org/library/en/2016-17-summative-assessment-technical-report.pdf

Smarter Balanced Assessment Consortium. (2019). *Smarter Balanced Assessment Consortium: 2017–18 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. http://portal.smarterbalanced.org/library/en/2017-18-summative-assessment-technical-report.pdf

The Regents of the University of California & California Department of Education. (2020). *Starting smarter*. [Website]. https://ca.startingsmarter.org/

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46,* 247–72.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*, 2–13.

Accessibility Information

Alternative Text for Equation 8.1

Rho sub theta theta prime equals 1 minus M sub SEM squared sub theta divided by s squared sub theta.

Alternative Text for Equation 8.2

SEM sub scaled equals a times s sub theta times the square root of 1 minus rho sub theta theta prime.

Alternative Text for Equation 8.3

CSEM of SS equals 1 times a divided by the square root of I of theta hat.

Alternative Text for Equation 8.4

ICC is interrater reliability, which is equal to the difference between mean square between variance and mean square within variance divided by the sum of mean square between variance and k minus 1 of mean square within variance.

Alternative Text for Equation 8.5

K sub ij equals open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub ij divided by n sub plus plus close parenthesis minus open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub iplus times n sub plusj divided by n squared sub plusplus close parenthesis divided open parenthesis 1 minus open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub iplus times n sub plusj divided by n squared sub plusplus close parenthesis close parenthesis, K sub ij equals open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub ij divided by n sub plus plus close parenthesis minus open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub iplus times n sub plusj divided by n squared sub plusplus close parenthesis divided open parenthesis 1 minus open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub iplus times n sub plusj divided by n squared sub plusplus close parenthesis close parenthesis.

Alternative Text for Equation 8.6

W sub ij equals 1 minus open parenthesis I minus j close parenthesis squared divided by m squared.

Appendix 8.A: Item Response Theory Parameter Estimates

IRT parameter estimates used in the 2019–2020 administration of the Smarter Balanced Summative Assessments were derived from the 2013–2014 field tests and the operational assessments from the 2014–2015 through the 2018–2019 assessments with embedded field tests. The Smarter Balanced Assessment Consortium conducted calibration and equating for all item parameter estimates.

Table 8.A.1 IRT *a*-values for Grade Three—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	317	0.70	0.25	0.19	1.52
Claim 2	262	0.67	0.20	0.18	1.21
Claim 3	183	0.54	0.17	0.21	1.01
Claim 4	147	0.67	0.24	0.18	1.29
All Items	909	0.66	0.23	0.18	1.52

Table 8.A.2 IRT *a*-values for Grade Four—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	279	0.64	0.23	0.15	1.41
Claim 2	300	0.60	0.22	0.17	1.24
Claim 3	201	0.55	0.17	0.19	1.14
Claim 4	172	0.58	0.21	0.15	1.29
All Items	952	0.60	0.21	0.15	1.41

Table 8.A.3 IRT *a*-values for Grade Five—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	318	0.63	0.23	0.18	1.38
Claim 2	313	0.62	0.22	0.17	1.25
Claim 3	168	0.53	0.15	0.21	0.99
Claim 4	165	0.66	0.19	0.23	1.18
All Items	964	0.62	0.21	0.17	1.38

Table 8.A.4 IRT *a*-values for Grade Six—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	276	0.62	0.21	0.20	1.18
Claim 2	312	0.59	0.22	0.17	1.34
Claim 3	178	0.51	0.18	0.19	0.95
Claim 4	174	0.58	0.20	0.17	1.05
All Items	940	0.58	0.21	0.17	1.34

Table 8.A.5 IRT *a*-values for Grade Seven—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	277	0.59	0.20	0.19	1.30
Claim 2	309	0.58	0.22	0.17	1.65
Claim 3	176	0.51	0.16	0.20	1.00
Claim 4	144	0.56	0.20	0.16	1.00
All Items	906	0.57	0.20	0.16	1.65

Table 8.A.6 IRT *a*-values for Grade Eight—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	284	0.59	0.20	0.15	1.12
Claim 2	341	0.54	0.18	0.15	1.07
Claim 3	212	0.50	0.16	0.13	0.91
Claim 4	143	0.57	0.20	0.20	1.19
All Items	980	0.55	0.19	0.13	1.19

Table 8.A.7 IRT *a*-values for Grade Eleven—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	905	0.55	0.18	0.13	1.16
Claim 2	815	0.48	0.15	0.14	1.07
Claim 3	604	0.45	0.16	0.10	0.98
Claim 4	403	0.48	0.18	0.12	1.19
All Items	2,727	0.49	0.17	0.10	1.19

Table 8.A.8 IRT *a*-values for Grade Three—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	772	0.84	0.29	0.16	1.59
Claim 2	280	0.89	0.27	0.20	1.51
Claim 3	239	0.73	0.31	0.13	1.53
All Items	1,291	0.83	0.29	0.13	1.59

Table 8.A.9 IRT *a*-values for Grade Four—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	853	0.84	0.28	0.17	1.64
Claim 2	316	0.80	0.30	0.20	1.63
Claim 3	249	0.75	0.28	0.25	1.51
All Items	1,418	0.82	0.29	0.17	1.64

Table 8.A.10 IRT *a*-values for Grade Five—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	831	0.77	0.29	0.14	1.62
Claim 2	301	0.80	0.30	0.16	1.56
Claim 3	249	0.67	0.30	0.16	1.77
All Items	1,381	0.76	0.30	0.14	1.77

Table 8.A.11 IRT *a*-values for Grade Six—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	758	0.69	0.26	0.13	1.40
Claim 2	235	0.78	0.26	0.15	1.71
Claim 3	220	0.59	0.25	0.13	1.51
All Items	1,213	0.69	0.26	0.13	1.71

Table 8.A.12 IRT *a*-values for Grade Seven—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	693	0.73	0.28	0.10	1.49
Claim 2	240	0.79	0.29	0.11	1.43
Claim 3	171	0.61	0.33	0.12	1.68
All Items	1,104	0.73	0.30	0.10	1.68

Table 8.A.13 IRT *a*-values for Grade Eight—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	636	0.58	0.26	0.09	1.33
Claim 2	185	0.71	0.28	0.16	1.40
Claim 3	160	0.50	0.24	0.14	1.36
All Items	981	0.59	0.27	0.09	1.40

Table 8.A.14 IRT *a*-values for Grade Eleven—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	1,820	0.60	0.27	0.09	1.57
Claim 2	390	0.58	0.27	0.10	1.40
Claim 3	425	0.46	0.24	0.09	1.34
All Items	2,635	0.58	0.27	0.09	1.57

Table 8.A.15 IRT *b*-values for Grade Three—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	317	-0.54	1.10	-2.72	4.69
Claim 2	262	-0.79	1.04	-2.90	4.12
Claim 3	183	-0.17	1.21	-2.92	3.82
Claim 4	147	-0.21	0.96	-2.22	1.86
All Items	909	-0.49	1.11	-2.92	4.69

Table 8.A.16 IRT *b*-values for Grade Four—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	279	0.16	1.30	-2.60	6.23
Claim 2	300	-0.47	1.21	-3.25	2.94
Claim 3	201	0.01	1.31	-2.82	4.25
Claim 4	172	0.34	1.24	-2.03	3.73
All Items	952	-0.04	1.30	-3.25	6.23

Table 8.A.17 IRT *b*-values for Grade Five—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	318	0.44	1.42	-2.60	5.65
Claim 2	313	-0.21	1.31	-3.25	3.29
Claim 3	168	0.31	1.32	-2.82	3.48
Claim 4	165	0.36	1.17	-2.03	3.83
All Items	964	0.19	1.35	-3.25	5.65

Table 8.A.18 IRT *b*-values for Grade Six—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	276	0.94	1.33	-2.10	4.78
Claim 2	312	0.57	1.48	-3.25	5.54
Claim 3	178	0.64	1.56	-2.82	7.38
Claim 4	174	0.75	1.15	-1.76	3.61
All Items	940	0.72	1.40	-3.25	7.38

Table 8.A.19 IRT *b*-values for Grade Seven—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	277	1.24	1.46	-1.84	6.63
Claim 2	309	0.83	1.46	-2.72	5.31
Claim 3	176	0.72	1.37	-1.71	5.88
Claim 4	144	1.36	1.47	-1.49	5.61
All Items	906	1.02	1.46	-2.72	6.63

Table 8.A.20 IRT *b*-values for Grade Eight—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	284	1.49	1.41	-1.84	6.42
Claim 2	341	0.83	1.38	-3.01	4.56
Claim 3	212	0.78	1.30	-2.12	3.87
Claim 4	143	1.46	1.39	-1.79	5.19
All Items	980	1.10	1.41	-3.01	6.42

Table 8.A.21 IRT *b*-values for Grade Eleven—ELA

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	905	1.91	1.45	-2.09	9.10
Claim 2	815	1.65	1.46	-1.98	9.15
Claim 3	604	1.33	1.43	-1.65	6.62
Claim 4	403	2.02	1.42	-1.20	8.94
All Items	2,727	1.72	1.46	-2.09	9.15

Table 8.A.22 IRT *b*-values for Grade Three—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	772	-1.13	1.08	-4.34	4.16
Claim 2	280	-0.35	0.91	-2.68	3.20
Claim 3	239	-0.13	1.05	-2.42	5.12
All Items	1,291	-0.77	1.13	-4.34	5.12

Table 8.A.23 IRT *b*-values for Grade Four—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	853	-0.35	1.15	-3.38	4.48
Claim 2	316	0.09	1.04	-2.68	3.28
Claim 3	249	0.31	1.03	-2.08	5.18
All Items	1,418	-0.14	1.13	-3.38	5.18

Table 8.A.24 IRT *b*-values for Grade Five—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	831	0.17	1.17	-3.38	6.20
Claim 2	301	0.99	1.08	-2.68	4.63
Claim 3	249	0.89	1.19	-2.01	5.98
All Items	1,381	0.48	1.21	-3.38	6.20

Table 8.A.25 IRT *b*-values for Grade Six—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	758	0.78	1.41	-3.93	9.16
Claim 2	235	1.39	1.30	-2.98	6.44
Claim 3	220	1.79	1.60	-2.16	8.75
All Items	1,213	1.08	1.48	-3.93	9.16

Table 8.A.26 IRT *b*-values for Grade Seven—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	693	1.74	1.26	-1.79	7.80
Claim 2	240	1.90	1.05	-1.09	5.07
Claim 3	171	2.22	1.56	-1.65	6.59
All Items	1,104	1.85	1.28	-1.79	7.80

Table 8.A.27 IRT *b*-values for Grade Eight—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	636	1.98	1.68	-1.87	7.75
Claim 2	185	2.30	1.43	-2.84	6.48
Claim 3	160	2.80	1.82	-1.65	9.02
All Items	981	2.18	1.69	-2.84	9.02

Table 8.A.28 IRT *b*-values for Grade Eleven—Mathematics

Claim	Number of Items	Mean	Standard Deviation	Minimum	Maximum
Claim 1	1,820	2.37	1.50	-4.43	8.72
Claim 2	390	3.11	1.31	-1.10	6.68
Claim 3	425	3.04	1.54	-1.05	9.25
All Items	2,635	2.59	1.51	-4.43	9.25

Chapter 9: Quality Control Procedures

The California Department of Education (CDE), Smarter Balanced Assessment Consortium, and ETS implemented rigorous quality control procedures throughout the test development, administration, scoring, and analyses processes. As part of this effort, ETS staff worked with its Office of Professional Standards Compliance, which publishes and maintains the *ETS Standards for Quality and Fairness* (ETS, 2014). These standards support the goal of delivering technically sound, fair, and useful products and services while assisting the public and auditors in evaluating those products and services. Quality control procedures are outlined in this chapter.

9.1. Quality Control of Item Development

Item writers hired to develop Smarter Balanced assessment items were trained in Smarter Balanced policies on sensitivity and bias guidelines, as well as guidelines for accessibility, to ensure that the items allow the widest possible range of students to demonstrate their content knowledge (Smarter Balanced, 2016). A group of educators reviewed the items and performance tasks (PTs) for accessibility, bias and sensitivity, as well as content prior to their administration in the 2013–2014 field test.

To further ensure the quality of Smarter Balanced assessment items, in early May 2013, Smarter Balanced recruited a panel of English language arts/literacy (ELA) and mathematics content experts and decision-makers with expertise in the needs of students with disabilities and students who were English learners. This panel reviewed item specifications, item types, items, and PTs and made recommendations for item development and item-quality criteria.

After the 2012–2013 pilot test, staff from the Smarter Balanced Assessment Consortium used statistical criteria to flag items that were potentially problematic because of content, bias, or accessibility issues.

For more information regarding the steps taken by the Smarter Balanced Assessment Consortium to ensure quality during item development, refer to chapter 3 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016).

9.2. Quality Control of Test Assembly and Delivery

The assembly of all test forms must conform to blueprints that represent a set of constraints and specifications. There were separate specifications for the ELA and mathematics assessments. These blueprints are critical to the formation of valid assessments and can be found in appendix 2.A.

The Smarter Balanced Assessment Consortium conducted computer simulations to evaluate the test delivery system and the adaptive testing algorithm. Two sets of simulation studies were conducted:

1. The simulation study conducted prior to the 2013–2014 Smarter Balanced field test that is described in chapter 4 of the *Smarter Balanced Assessment Consortium: 2013–14 Technical Report* (Smarter Balanced, 2016).
2. The simulation study conducted prior to the 2016–2017 California Assessment of Student Performance and Progress (CAASPP) operational administration that is described in *Chapter 4: Test Assembly* in this current technical report.

9.3. Quality Control of Test Materials

9.3.1. Developing Assessments

9.3.1.1. Online Assessments

The steps taken to develop and ensure the quality of the online assessments is described in section *5.1 Test Administration*.

9.3.1.2. Paper–Pencil Forms

Test forms and response booklets received from the Smarter Balanced Assessment Consortium are carefully reviewed by ETS staff to ensure that they meet quality standards. Each document is reviewed for accuracy, completeness, and alignment with supporting materials.

Print-ready PDFs received for the paper versions of the Smarter Balanced Summative Assessments undergo a stringent quality control process to ensure that there is adequate space for student identification and demographic information.

9.3.1.3. Test Administration Manuals

ETS staff consult with internal subject matter experts and conduct validation checks to verify that test instruction manuals accurately match the test booklets and testing processes. Copy editors and content editors review each document for spelling, grammar, accuracy, and adherence to CDE style. Manuals received from Smarter Balanced are customized to fit the CAASPP System specifications. Each document must be approved by the CDE before it can be published to the CAASPP website at https://www.caaspp.org/. Only nonsecure documents are posted to this website.

9.3.2. Collecting Test Materials

9.3.2.1. Online Assessments

During the 2019–2020 CAASPP administration, there were no collectable materials associated with online testing.

9.3.2.2. Paper–Pencil Forms

Once the paper–pencil tests are administered at test sites whose local educational agencies (LEAs) had received prior approval from the CDE, LEAs returned all scorable and nonscorable materials within five working days after the last day of each test administration period. The LEAs packed all materials into cartons, applied the labels, and then numbered the cartons prior to returning the materials to the processing center by means of their assigned carrier.

9.3.3. Processing Test Materials

9.3.3.1. Online Assessments

Online tests that were submitted by students were transmitted from Cambium Assessment, Inc. (CAI) to ETS each day. Each system checked for the completeness of the student record and stopped records that were identified as having an error. (For example, the system would identify a test part that was missing a content registration ID, a unique identifier that matches the student's opportunities—computer-adaptive testing [CAT] and PT—in final scoring.)

Test responses were separated for human scoring between ETS and Measurement Incorporated (MI), and the reader's ratings were delivered to ETS scoring systems for merging with machine-scored items, final scoring, and scoring quality checks.

9.3.3.2. Paper–Pencil Forms

Upon receipt of the test materials, ETS personnel examined each shipment for a number of conditions, including shipping errors and omission of materials. The number of students recorded on the School and Grade Identification (SGID) sheet was compared to the number of answer documents returned to ETS.

ETS' staff compared scorable material quantities reported on the SGIDs to actual documents received. LEAs were contacted by phone if there were any missing shipments or the quantity of materials returned appeared to be less than expected.

9.4. Quality Control of Psychometric Processes

9.4.1. Development of Scoring Specifications

A number of measures are taken to ascertain that the scoring keys are applied to the student responses as intended and the student scores are computed accurately. ETS builds and reviews the scoring system models based on the Smarter Balanced Assessment Consortium scoring specifications and CDE requirements (Smarter Balanced, 2014; American Institutes for Research, 2015). Machine-scored item responses and demographic information are collected and provided electronically to ETS in a master student data file. Human-scored item responses are sent electronically to the ETS Online Network for Evaluation or MI scoring centers for scoring by trained, qualified raters. Record counts are verified against the counts obtained during security check-in from the document processing staff to ensure all students are accounted for in the file.

Once the record counts are reviewed, the machine-scored item responses are scored against the appropriate answer key provided by the Smarter Balanced Assessment Consortium. In addition, the student's original response string is stored for data verification and auditing purposes.

The Smarter Balanced Assessment Consortium provided the specifications for scoring the assessments well in advance of the receipt of student response data. These specifications contain detailed scoring procedures, along with the procedures for determining whether a student had attempted a test and whether that student response data should be included in the statistical analyses and calculations for computing summary data. Standard quality inspections are performed on all data files, including the evaluation of each student data record for correctness and completeness. Student results are kept confidential and secure at all times.

9.4.2. Development of Scoring Procedures

ETS' Enterprise Score Key Management (eSKM) system uses scoring procedures specified by psychometricians and provides scoring services. Following scoring, a series of quality control checks are carried out by ETS psychometricians to ensure the accuracy of each score.

9.4.2.1. Enterprise Score Key Management System Processing

ETS developed two independent and parallel scoring structures to produce students' scores: the eSKM[46] scoring system, which collects, scores, and delivers individual students' scores to the ETS reporting system; and the parallel scoring system developed by ETS Technology and Information Processing Services (TIPS), which scores individual students'

[46] The eSKM system produces the ETS scores of record.

responses. The two scoring systems independently applied the same scoring algorithms and specifications. ETS psychometricians verified the eSKM scoring by comparing all individual student scores from TIPS and resolving any discrepancies. This process redundancy is an internal quality control step and is in place to verify the accuracy of scoring. Students' scores were reported only when the two parallel systems produce identical results.

If scores did not match, the mismatch would be investigated by ETS' Psychometrics, Statistics, and Data Science and eSKM teams and resolved. The mismatch could be a result of a Smarter Balanced and CDE decision not to score an item because a problem was identified in a particular item or rubric. ETS applied the problem item notification (PIN) not to score the item through the systematic process in eSKM and a mismatch could be possible, if TIPS were still in the process of applying the PIN in the parallel system when the student score was being compared. This real-time scoring check is designed to continually detect mismatches and track remediation.

ETS' Centralized Repository Distribution System and Enterprise Service Bus departments collected and parsed .xml files that contain student response data from CAI and sent constructed-response (CR) item responses to ETS and MI for human scoring. After receiving the results of human scoring, eSKM merged student scores from the CAT and PT test components, calculated individual student scores, and generated student scores in the approved statistical extract format on a daily basis. These data extracts were sent to ETS' Data Quality Services for data validation. Following validation, the student response statistical extracts were made available to the psychometricians.

9.4.2.2. Psychometric Processing

Psychometricians verified the eSKM scoring by comparing the parallel scoring programs, conducting extensive analyses to resolve any discrepancies, and verifying the accuracy of all student scores and reported results. In particular, psychometricians checked variables such as total scale scores, achievement levels, number of scored items, and performance levels of claims. To investigate discrepancies, theta scores and completeness were also checked; refer to *7.4 Student Test Scores* for definitions of these scores. Refer also to section *10.3 Psychometric Analyses* for more information on psychometric quality control.

All scores complied with the ETS scoring specifications and the parallel scoring process to ensure the quality and accuracy of scoring and to support the transfer of scores into the database of the student records scoring system before student reports were generated. In addition to parallel scoring for both online and paper–pencil assessments, ETS provided verification of answer keys and item analysis for paper–pencil assessments.

9.5. Quality Control of Constructed-Response Scoring

9.5.1. Team Training and Calibration

Rater qualifications, rater certifications, and daily rater calibrations are all processes used to control the reliability of CR scoring. Raters were led through a training period by trained Assessment and Learning Technology Research & Development staff, content scoring leaders, group scoring leaders, and scoring leaders for an assigned grade level and specific prompt types prior to the annual scoring period. In the training period, raters were trained to appropriately apply the rubrics by using the Smarter Balanced–provided benchmark sample papers.

Trained raters were scheduled to score in four- or eight-hour shifts. Prior to starting a shift, a rater was required to take and pass a calibration test that demonstrated sufficient training in Smarter Balanced scoring criteria and ability to score accurately.

Scoring leaders were qualified raters with the responsibility of providing feedback to raters to provide additional content support and offer corrective mentoring for struggling raters.

Each rater was assigned a secure user ID and password to log on to the scoring system and was required to sign a confidentiality agreement. System access for the rater was restricted to the hours that the rater was scheduled to work.

9.5.2. Hand Scoring Verification

9.5.2.1. Criteria for Read-Behinds

Ten percent of responses were scored twice (i.e., “read behind”) to check agreement among raters, although the percentage could vary, depending on item type and reader performance. Scoring leaders read behind raters throughout a shift and entered their own scores on responses that raters had read. Both first and second readings were eligible for read-behind.

A scoring leader reviewed the randomly selected responses after raters submitted scores. Leaders reviewed rater scoring statistics (i.e., interrater reliability, score point distributions, and validity performance) to determine the need for monitoring via read-behinds or additional training. Responses determined to be scored incorrectly during read-behind review could be rescored by leadership and used to inform and instruct raters as a performance-improvement strategy.

When a response was selected for a second reading, the corrected score was used for interrater reliability calculation. The original rater’s score was not used for any calculation.

9.5.2.2. Validity Responses

Validity responses were provided randomly as part of the set of “live” responses being scored, so a rater did not know that the response being scored was for validity. These responses were selected from “live” responses by scoring leaders prior to the scoring of the item. Leadership staff identified the response to be used for validity and the system added the response to the validity pool for use during scoring.

All staffing levels were eligible to score second readings. Ten percent of responses were assigned to be read a second time. Second readings were scored independently from the first reading.

Only scorable responses were selected for second readings. Nonscorable (i.e., condition code) responses were not eligible for second readings and so were not included in the calculation of interrater reliability.

The second reading sample was not a stratified random sample. The selection of a second reading response was also not based on the first reading score or any demographic information associated with the response. Instead, responses flagged for second reading were flagged at random by the scoring system for each item identification number.

Second reading scores were used only for statistical analysis to obtain interrater reliability. They were not included in the calculation of the final item score.

9.5.3. Artificial Intelligence Scoring Verification

To ensure the quality of machine scoring with artificial intelligence (AI), ETS and MI maintained a quality assurance system where 10 percent of AI-scored items were scored by a human rater and used for agreement sample analysis. More details are presented in subsection *8.5.6 Interrater Agreement*. Also, refer to section *7.2 Quality Control of Scoring* and section *10.2 Hand Scoring* for more information.

9.6. Quality Control of Paper–Pencil Scoring

If an LEA was approved to administer the paper–pencil version of the Smarter Balanced Summative Assessments, student responses were entered into the Data Entry Interface and scored electronically and by a rater, depending on the item type.

9.7. Quality Control of Reporting

To ensure the quality of CAASPP Smarter Balanced Online Summative Assessment results, for both individual student and summary reports, three general areas are evaluated:

1. Comparison of report formats with input sources from the CDE-approved samples
2. Validation of the report data through quality control checks performed by ETS' Data Quality Services and Resolutions teams, as well as running of all Student Score Reports (SSRs) through ETS' patented QC Interrogator software
3. Proofreading of the pilot and production reports by the CDE and ETS prior to making reports available to LEAs

All reports were required to include a single, accurate LEA code, a charter school number (if applicable), an LEA name, and a school name. All elements conformed to the CDE's official county/district/school (CDS) code and naming records. From the start of processing through scoring and reporting, the CDS Master File was used to verify and confirm accurate codes and names. The CDE provided a revised LEA Master File to ETS throughout the year as updates became available.

After the reports were validated against the CDE's requirements, a set of reports for pilot LEAs were provided to the CDE and ETS for review and approval. Electronic reports were sent on the actual report template, organized as they were expected to look in production. The CDE and ETS reviewed and approved the reports after a thorough examination.

Upon the CDE's approval of the reports generated for the pilot districts, ETS proceeded with the first batch of report production. The first production batch was selected to validate a subset of LEAs that contained key reporting characteristics (e.g., academic achievement) and demographics of the state. The first production batch incorporated CDE-selected LEAs and provided the final check prior to generating all reports and making them available electronically for download in the Test Operations Management System.

9.7.1. Exclusion of Student Scores from Summary Reports

ETS provides specifications to the CDE that document when to exclude student scores from summary reports. These specifications included the logic for handling submitted assessments that, for example, indicated the student tested but responded to no items, was absent, was not tested because of parent/guardian request, or did not complete the assessment because of illness. The methods for handling other anomalies are also covered in the specifications. These anomalies are described in more detail in *7.6.2 Special Cases*.

9.8. End-to-End Operational Tests

ETS conducted end-to-end testing prior to the start of the test administration. The purpose of this testing is to verify that all systems, processes, and resources were ready for the operational administration.

9.8.1. Online Assessments

ETS employs a number of strategies to verify ongoing systems performance, including monitoring of system availability and online system usage. Time was allotted for user acceptance testing to confirm that the systems met requirements and to make identified corrections before final deployment. To accomplish system acceptance and sign-off, ETS deployed systems to a staging area, which mirrors the final production environment, for operational and user acceptance testing. Final approval by the CDE triggers final deployment of the system.

9.8.2. Paper–Pencil Tests

The Data Entry Interface (DEI) underwent user acceptance testing to ensure that the correct test items were available for a grade-level assessment in the DEI. Then, during testing, information technology personnel monitor daily feeds to ensure the completeness and timeliness of records sent for hand scoring.

References

American Institutes for Research. (2015). *Smarter Balanced scoring specification: 2014–2015 administration, version 7*. http://www.smarterapp.org/documents/TestScoringSpecs2014-2015.pdf

Educational Testing Service. (2014). *ETS standards for quality and fairness.* Princeton, NJ: Author. https://www.ets.org/s/about/pdf/standards.pdf

Smarter Balanced Assessment Consortium. (2014). *Hand-scoring rules.* Los Angeles, CA: Smarter Balanced Assessment Consortium. http://www.smarterapp.org/documents/Smarter_Balanced_Hand_Scoring_Rules.pdf

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: 2013–14 technical report.* https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf

Chapter 10: Continuous Improvement

The sixth operational administration of the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced Summative Assessments for English language arts/literacy (ELA) and mathematics occurred in 2019–2020. Throughout the past six years, continuous efforts have been made to improve the assessments in various ways. This chapter summarizes accomplishments and ongoing improvements for the Smarter Balanced assessments in test delivery and administration, hand scoring, psychometric analyses, and accessibility.

Because the Smarter Balanced Assessment Consortium owns the test design and item development of these assessments, the focus of ETS' continuous improvement is limited to test administration, scoring and reporting, and analyses.

10.1. ETS Administration and Delivery

10.1.1. Post-Test Survey

The CAASPP program annually solicits feedback from CAASPP stakeholders through the CAASPP Post-Test Survey. Local educational agency and test site staff, as well as test administrators and test examiners, were invited to participate in the 2019–2020 CAASPP Post-Test Survey. More than 3,000 California educators provided specific, actionable insights about their testing experience. This survey gathered information and data from educators who were part of the administration of the CAASPP and English Language Proficiency Assessments for California. Its goal was to highlight successes and identify areas for improvement, both immediate and long term.

Overall, the state's educators felt that the resources and training materials they were given were useful in preparing them and their students in the 2019–2020 administration. Their feedback generally described smooth preparation, training, support, and assessment administration experiences. Also, educators provided valuable feedback for potential improvements to the 2020–2021 administration.

In the area of preparation and training, educators reported in 2019–2020 that they had received adequate preparation and training for a successful administration but could use additional support in specific areas such as interim assessments and accessibility resources.

Many improvements were made to the Test Operations Management System (TOMS) in the 2019–2020 administration. Most users (83%) reported that the improvements made the application easier to use. They provided various suggestions for improvements to TOMS that would make managing test operations more efficient.

10.2. Hand Scoring

10.2.1. Document Summative Assessment Scoring Activities

Continuous improvements that occurred in 2019–2020 included that rater agreement and validity statistics for rater agreement were monitored each week. Scoring leaders provided feedback to ETS Assessment and Learning Technology Research & Development to determine what adjustments to training or samples were to be made.

Because of a shortened scoring season in 2019–2020, the following scheduled improvements were unable to be fully implemented and evaluated. Instead, full implementation and evaluation will take place during the 2020–2021 administration.

Planned improvements for 2020–2021 will include the following:

- ETS will use standardized training to assist the scoring leader in utilizing the performance indicator panels, which allows easier access to quantitative feedback regarding individual raters.
  - Improved training will be conducted via online learning courses, as opposed to the WebEx sessions used previously. Online learning courses provide the following expected benefits:
    - Standardized explanation on what information is available within the performance indicator panel, as well as how to use the information
    - Standardized format for providing feedback to individual raters to ensure the area of improvement needed is clear and consistent regardless of which score leader may be monitoring an individual rater on any given day
    - Automatic restriction of scoring leaders from monitoring raters until training requirements have been satisfied (previously done manually)

10.2.2. Documenting and Revisiting Summative Assessment Item Flagging Criteria

At least 10 percent of the ELA and mathematics responses are scored independently by a second reader each year. Of these, the statistics for the interrater reliability were calculated for all items at all grades. To determine the reliability of scoring, ETS examined the percentage of perfect agreement and adjacent agreement between the two readers. The item-level quadratic weighted kappa (QWK) statistic was calculated to reflect the level of improvement beyond the chance level in the consistency of scoring in chapter 8. In the 2019–2020 administration, the item flagging criterion was raised from the QWK at 0.2 established by the Smarter Balanced Assessment Consortium (Smarter Balanced, 2016) to 0.7 as suggested by Williamson, Xi, and Breyer (2012). Refer to subsection 7.2.4 for detailed information on flagging criteria.

10.2.3. Monitoring, Documenting, and Evaluating Rater Qualifications to Industry Standards

Starting with this current technical report, ETS documents rater qualification in subsection *7.2.1.2 Quality Control Related to Raters* and shares reports with the California Department of Education on the counts of California educators and California residents participating in both the existing rater pool and as potential raters in the recruitment pipeline. Systemically, ETS Human Resources analyzes the data received in its application process and uses the answers to these questions to support the development of a strong, qualified workforce.

In 2019–2020, ETS updated and documented the qualifications of the rater pools in subsection *7.2.1.3 Rater Qualification*. Documentation of the qualifications of the rater pool will be produced annually.

10.3. Psychometric Analyses

10.3.1. Scoring Verification Process

Improvements for item flagging for the quality assurance procedures were planned in a discussion between ETS and Smarter Balanced in 2018–2019 and implemented in the 2019–2020 administration. As a new part of the score verification effort, the following improvements were made:

- The *p*-value for item flagging was changed to a range of [0.03, 0.97], from [.05, 0.95], to avoid more type I errors. In addition, polytomous items will be flagged if any one of the scoring categories has less than 3 percent of responses.
- An average theta was added to a flagged item when results were sent to the Smarter Balanced Assessment Consortium. When a polytomous item was flagged, its historical performance in the previous year was included for a comparison. The corresponding average theta for the group in the previous year was included.
- ETS notified the Smarter Balanced psychometric team two weeks before ETS sent the list of flagged items that required Smarter Balance’s review, giving Smarter Balanced sufficient time to identify appropriate assessment development staff who could be made available to perform timely content reviews.

In the 2019–2020 administration, the ETS psychometric team delivered flagged items and comprehensive information on the flagged items to Smarter Balanced for verification in a timely manner. Before Smarter Balanced test scores were released, ETS and Smarter Balanced ensured all scores reported were based on quality items.

10.4. Accessibility

Like all CAASPP assessments, the Smarter Balanced Summative Assessments are administered using the test delivery system created by Cambium Assessment, Inc., for the Smarter Balanced assessments. As such, implementation of new online universal tools, designated supports, and accommodations are provided by Smarter Balanced (Smarter Balanced, 2020) and aligned with the test delivery system.

The following changes will be implemented during the 2019–2020 Smarter Balanced administration:

- Illustration glossaries for mathematics items will be available for selected construct-irrelevant terms. This resource, which is a type of translation glossary, is available as an embedded designated support for online assessments and as a nonembedded designated support for paper–pencil tests. Students who are assigned the illustration glossaries resource will be given a fixed-form assessment.
- Somali and Hmong will be offered as translation glossaries for mathematics items.
- Unified English Braille Technical will be available for the mathematics assessment.

References

Williamson, D.M., Xi, X., & Breyer, F.J. (2012), A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31: 2–13.

Smarter Balanced Assessment Consortium. (2020). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines.* Los Angeles: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: 2013–14 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf