California Department of Education Assessment Development & Administration Division

California Assessment of
Student Performance and Progress

California Assessment of Student Performance and Progress Smarter Balanced Technical Report 2017–18 Administration

Final Submitted August 2, 2019

Educational Testing Service

Contract No. CN150012

Table of Contents

List of Appendices

List of Tables

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

Term	Definition
2PL	two-parameter logistic
AD	Assessment Development
AERA	American Educational Research Association
AI	artificial intelligence
AIR	American Institutes for Research
AYP	adequate yearly progress
CAASPP	California Assessment of Student Performance and Progress
CAPA	California Alternate Performance Assessment
CAT	computer-adaptive test
CCR	*California Code of Regulations*
CCSS	Common Core State Standards
CDE	California Department of Education
CDS	county/district/school
CI	confidence interval
CMA	California Modified Assessment
CR	constructed response
CRESST	Center for Research on Evaluation, Standards, & Student Testing
CSEMs	conditional standard errors of measurement
CSTs	California Standards Tests
CSU	California State University
DFA	*Directions for Administration*
DIF	differential item functioning
EAP	Early Assessment Program
EC	*Education Code*
EL	English learner
ELA	English language arts/literacy
eSKM	Enterprise Score Key Management
ETS	Educational Testing Service
GPCM	generalized partial credit model
HOSS	highest obtainable scale score
HOT	highest obtainable theta
HumRRO	Human Resources Research Organization
ICC	item characteristic curve *or* intraclass correlation
IEP	individualized education program
IFEP	initially fluent English proficient
IRT	item response theory
JAWS®	Job Access With Speech
ISAAP	Individual Student Assessment Accessibility Profile
LEA	local educational agency

Term	Definition
LOT	lowest obtainable theta
LOSS	lowest obtainable scale score
MI	Measurement Incorporated
MLE	maximum likelihood estimation
NAEP	National Assessment of Educational Progress
NCME	National Council on Measurement in Education
ORS	Online Reporting System
OTI	Office of Testing Integrity
PAR	Psychometrics, Analysis, and Research
PIN	problem item notification
PISA	Program for International Student Assessment
PT	performance task
RFEP	reclassified fluent English proficient
SBE	State Board of Education
SEM	standard error of measurement
SFTP	secure file transfer protocol
SGID	School and Grade Identification sheet
SS	scale score
STS	Standards-based Tests in Spanish
TAG	Technical Advisory Group
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
VSC	Virtual Scoring System
wABC	weighted Area Between the Curves
WCAG	Web Content Accessibility Guidelines
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 2017–18 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 2017–18, 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—Tools and practices designed to help teachers utilize 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, eight, and high school (grades ten, eleven, or twelve; these are the California Science Test and the CAA for Science)
- A primary language assessment, the Standards-based Tests in Spanish for Reading/Language Arts in grades two through eleven (optional for eligible Spanish-speaking English learners)
- A new primary language assessment, the California Spanish Assessment, delivered in pilot form at selected local educational agencies (LEAs), to students in grades three through eight and high school who are Spanish-speaking English learners or students seeking a measure that recognizes their Spanish reading, writing mechanics, and listening skills

The CAASPP Smarter Balanced tests are presented as online assessments. Paper-pencil and braille versions of the Smarter Balanced assessments are made available to local educational agencies (LEAs) that do not have the necessary computer network infrastructure to administer the online tests; these are available with prior permission from 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 http://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 (CAT)

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 a wrong answer, the TDS will follow up with an easier question; if the student answers correctly, the next question will be slightly more difficult. Because the answers on items used to estimate the student's ability are machine-scored, the student's performance on the items already administered is known immediately, and the successive items are selected to adapt to the estimated ability of the student. The CAT selects questions based on a student's responses, scores the responses, and revises its estimate of the student's ability. This process continues until the test content outlined in the test's blueprint is covered.

The CAT requires 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 2013–14 field test. Each year, new items are added to the Smarter Balanced item pools.

1.3.2. Performance Tasks (PTs)

The PT is a nonadaptive test 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

Each grade-level, content area Smarter Balanced Summative Assessment was administered to approximately 435,000 to 483,000 students during the 2017–18 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 learners (ELs) who are in their first 12 months of attending school in the United States are exempt from taking the ELA portion of the assessment. ELs 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 testing, but their scores are excluded from all aggregate calculations.

For students with significant cognitive disabilities, the decision to administer the Smarter Balanced Summative Assessments or the CAAs is made by their individualized education program team. Parents/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. Only students whose parents/guardians submit a written request may be exempted from taking the tests (*Education Code [EC]* Section 60615).

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 http://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

[1] "English Learner (EL) Students (Formerly Known as Limited-English-Proficient or LEP)," from the CDE Glossary of Terms web page at http://www.cde.ca.gov/ds/sd/cb/glossary.asp.

> 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 or 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.”

Sections 60602.5(c) and (d) provide additional information regarding intent and context 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

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 2017–18 CAASPP administration, the window started on January 9 and ended on July 16, 2018. The 12-week window for each LEA begins on the day of completion in which 66 percent of the instructional year is completed.

1.7. Significant CAASPP Developments in 2017–18

1.7.1. Updated Accessibility Resources

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

- *Amplification,* a non-embedded designated support that permits volume control beyond a device’s built-in settings using headphones or other non-embedded devices
- *Audio Transcript,* an embedded accommodation for the ELA assessment that displays a transcript of the closed captioning created for the listening packages; this includes braille transcript
- *Equation editor for braille,* a universal tool allowing for accessible mathematics input and output and full scoring
- *Hybrid Adaptive Test (HAT)* for braille readers, an online, multistage form for students using refreshable braille to access the Smarter Balanced for Mathematics assessments, featuring an adaptive section with items that do not require a supplemental graphics package, followed by a fixed-form section requiring a tactile graphics package for the student

- *Line reader*, a universal tool that permits a student to move an on-screen horizontal line that surrounds each line of text with shading
- *Mouse pointer*, a non-embedded accommodation that permits the selection of size and color options for the student's mouse pointer
- *Word prediction*, a non-embedded accommodation that is accessed using a physically separate device, allowing students to begin writing a word and choose from a list of words that have been predicted from word frequency and syntax rules

1.7.2. Historical Comparisons

Trends in examinee performance and test characteristics over time, which include cross-sectional and longitudinal comparisons, now include data from three operational administrations (2015–16, 2016–17, and 2017–18).

1.7.3. New Paper-Pencil Form

New paper-pencil forms in ELA and mathematics were released and used during the 2017–18 administration.

1.8. Groups and Organizations Involved with the CAASPP System

1.8.1. State Board of Education (SBE)

The 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, 2017).

1.8.2. California Department of Education (CDE)

The CDE oversees California's public school system, which is responsible for the education of more than 6,200,000 children and young adults in more than 10,450 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 Performance, Planning, and Technology Branch that oversees programs promoting innovation and improved student achievement. Programs include oversight of statewide assessments and the collection and reporting of educational data (CDE, 2017b).

[2] Retrieved from the CDE Fingertip Facts on Education in California – *CalEdFacts* web page at https://www.cde.ca.gov/ds/sd/cb/ceffingertipfacts.asp

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, including defining the purpose and scopes, test design, item development, standard setting, score reporting, and scoring the constructed-response items of the Smarter Balanced Summative Assessment.

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. The Smarter Balanced Assessment Consortium offers year-end summative assessments, optional interim assessments, and the Digital Library, an online collection of resources to help teachers improve classroom-based assessment practices. The roles of Smarter Balanced in the CAASPP System are to provide the collection of test items in the item bank and to provide access to the Digital Library (Smarter Balanced, n.d.).

1.8.5. Contractors

1.8.5.1 Educational Testing Service (ETS)

The CDE and the SBE contract with Educational Testing Service (ETS) to administer and report the CAASPP Smarter Balanced assessments. As the prime contractor, ETS has overall responsibility for working with the CDE to implement and maintain an effective assessment system and to coordinate the work of ETS with its subcontractors. Activities directly conducted 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
- Developing, hosting, and providing support for TOMS
- Processing student test assignments
- Processing orders and shipment of test materials and pre-identification services
- 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 American Institutes for Research (AIR)

ETS also monitors and manages the work of AIR, subcontractor to ETS for the CAASPP System of online assessments. Activities conducted by AIR include

- providing the AIR 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), a component of the overall CAASPP Assessment Delivery System;
- scoring machine-scorable items; and
- providing Level 3 technology help desk support to LEAs.

1.8.5.3 Measurement Incorporated (MI)

ETS monitors and manages the work of Measurement Incorporated (MI), a subcontractor to ETS for the CAASPP System. MI uses its artificial intelligence (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)

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 supports
- Ordering test materials and pre-identification services
- Viewing and downloading reports
- 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* form and the Appeals module

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.”[3] 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

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

user roles is explained more extensively in the *2017–18 CAASPP Online Test Administration Manual* (CDE, 2018b).

1.9.2. Test Delivery System (TDS)

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 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 purpose of the training tests is 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 the experience of a grade-level assessment, grade-specific items and difficulty levels, performance tasks, and the format and structure of an operational assessment.

1.9.4. Online Reporting System (ORS)

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.

1.9.5. Constructed-Response (CR) Scoring Systems for Educational Testing Service (ETS) and Measurement Incorporated (MI)

CRs 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 2018. The technical report contains 10 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 distributions of scores aggregated by student groups based on demographics and the use 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 2017–18 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 spring 2017–18 administration. These include
  - item response theory parameters,
  - omission and completion analyses,
  - conditional exposure analyses,
  - reliability analyses that include assessments of the reliability of test scores and claim scores for the population as a whole and for selected student groups,
  - consistency and accuracy of the performance-level classifications,
  - interrater reliability statistics for the human-scoring items and statistics showing the agreement of artificial intelligence scoring with human scoring, and
  - procedures designed to ensure the validity of score uses and interpretations.
- Chapter 9 highlights the quality control processes used at various stages of administration of the Smarter Balanced assessments.
- Chapter 10 presents cross-sectional and longitudinal historical comparisons of the overall tests and claims for all students and selected student groups. Descriptions and data are provided on the basis of student performances and test characteristics.

- Chapter 11 provides a summary of test assembly, test administration, calibration, and scaling procedures that are specifically applied to the paper-pencil tests; and the results of the analyses performed on the data for students who took paper-pencil tests instead of the online assessments. Analyses include
  - score distributions,
  - item response theory parameter values,
  - reliability analyses,
  - conditional standard error of measurement,
  - correlations between claims and between content areas, and
  - the use of designated supports and accommodations.
- Chapter 12 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. Retrieved from 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. Retrieved from https://bit.ly/2zNHQO4

California Department of Education. (2017, October). *State Board of Education responsibilities.* Retrieved from http://www.cde.ca.gov/be/ms/po/sberesponsibilities.asp

California Department of Education. (2018b). *CAASPP online test administration manual, 2017–18 administration.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.online_tam.2017-18.pdf

California Department of Education. (2018a, August). *Organization.* Retrieved from http://www.cde.ca.gov/re/di/or/

Smarter Balanced Assessment Consortium. (n.d.). *Smarter assessments*. Retrieved from http://www.smarterbalanced.org/assessments/

Chapter 2: Overview of CAASPP Smarter Balanced Processes

This chapter overviews the processes conducted by Smarter Balanced to develop the summative assessments. It also describes the processes implemented by Educational Testing Service (ETS) to administer the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced assessments.

Also described here are each process and a summary of the associated specifications. More details about the specifications and the analyses associated with each process are described in other chapters that are referenced in the subsections that follow.

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 specifications provide guidance on how to translate the Smarter Balanced content specifications into actual assessment items (Smarter Balanced, 2016, 2017a, 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, school districts, higher education professionals, and other stakeholders (Smarter Balanced, 2016). *For more information regarding the item response theory 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 section *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, 2017a, 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 computer adaptive test (CAT) and the performance task (PT). The number of PT items that a student is administered depends on the particular PT a student is assigned. Refer to Table 5.A.1 and Table 5.A.2 in appendix 5.A for the number of items in each PT. Refer to Table 5.B.1 through Table 5.B.3 in appendix 5.B for the distributions of number of items presented to students in the total test, PT, and CAT components respectively.

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], 2015).

2.2.1.2 Field Testing

Field test PTs have been embedded into the Smarter Balanced operational tests since the 2016-17 administration. Students who were assigned an embedded field test PT were not assigned an operational performance task. 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 2017–18 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, 2017c).

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 2017–18, the blueprints for the field test are provided specifically along with the blueprints for the operational tests provided in appendix 2.A, in subsection *Enhanced Computer Adaptive Testing (CAT) Blueprints for Students Participating in the 2017–18 Smarter Balanced Embedded Field Test of Performance Tasks (PTs)*. 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. Refer to Table 7.12, Table 7.13, Figure 7.7, and Figure 7.8 for the summary statistics associated with the test performances of the students assigned the field test PTs.

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 2015–16 testing is discussed in the report *Smarter Balanced Summative Assessments Testing Procedures for Adaptive Item Selection Algorithm* (AIR, 2015).

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, 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, 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 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 these subsections.

2.3.2.1 Test Administrators

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

Staff at local educational agencies (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], 2018a).

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*****—**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, 2018a). (Refer to *5.4.4.2 CAASPP Online Test Administration Manual* in chapter 5 for more information.)
- ***Test Operations Management System (TOMS) Pre-Administration Guide for CAASPP Testing*****—**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, 2017b). (Refer to *5.4.4.3 TOMS Pre-Administration Guide for CAASPP Testing* in chapter 5 for more information.)

2.4. Participation

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 Assessments (CAAs) for Mathematics based on alternate achievement standards (approximately one percent or fewer of the student population). The decision to assign a student to take an alternate assessment is made by his or her 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:

- Students with the most significant cognitive disabilities who meet the criteria for the CAA for ELA alternate assessment based on alternate achievement standards (approximately one percent or fewer of the student population). The decision to assign a student to take an alternate assessment is made by his or her IEP team.
- English learners 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 participation situations is illustrated in Table 7.9 in subsection *7.4.1.3 Scoring of Incomplete Cases*. Refer to appendix 7.A regarding the number of participants and the percent of participation of all students and selected demographic groups for each test.

2.5. Universal Tools, Designated Supports, and Accommodations

All public school students participate in the CAASPP System of assessments, including students with disabilities and English learners. Additional resources are sometimes needed for these students. The CDE provides a full range of assessment resources for all students, including those who are English learners 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: Universal Tools, Designated Supports, and Accommodations for the CAASPP System” (CDE, 2018c). [4]

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 test

[4] This technical report is based on the version of Matrix One that was available during the 2017–18 CAASPP administration.

delivery system at the time of testing. Refer to subsection *1.9 Systems Overview and Functionality* in *Chapter 1: Introduction* for more details regarding these systems.

Appendix 2.B presents counts and percentages of students assigned designated supports, accommodations, or unlisted resources for PTs and CAT respectively during the 2017–18 CAASPP Smarter Balanced administration. The majority of students do not use any designated supports, accommodations, or unlisted resources. The tables in appendix 2.B were created using student demographic data that is in version 2 of the production data file ("P2") which was updated on August 31, 2018.

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 is documented in Matrix One (CDE, 2018c). Part 1 of Matrix One lists the embedded universal tools, designated supports, and accommodations available for CAASPP Smarter Balanced online testing. Parts 2 and 3 of Matrix One include the non-embedded universal tools, designated supports, accommodations, and unlisted resources that are available. School-level personnel, IEP teams, and Section 504 teams use 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, 2018d) aids in the selection of universal tools, designated supports, and accommodations deemed necessary for individual students. 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 supports.

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*.* LEA personnel, including IEP and Section 504 plan teams, used the CAASPP 2017–18 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.

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 includes an inventory of unlisted resources that have already been identified and are preapproved (CDE, 2018c). During the 2017–18 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 may be assigned with the CDE's approval.

For an unlisted resource to be approved, it must not change the construct of what is being tested. If it does, test results for a student using an unlisted resource that was approved but changes the construct of what is being tested will not be considered valid for accountability purposes. The student receives a score with 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 as PDF files and aggregated results for the LEA.

Another means of viewing CAASPP scores is 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 student score report, 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.

2.6.2. Aggregation Procedures

In order 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 CAASPP Results web page at http://caaspp.cde.ca.gov/. 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 subsection *7.5 Overview of Score Aggregation Procedures* of this report. In Table 7.E.1 through Table 7.E.56 in appendix 7.E, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, special education service status, and economic status, as well as crosstab analysis for ethnicity and economic status. The tables show the numbers of students with valid scores in each group, scale score means and standard deviations, and the percentage in each achievement level. To protect student privacy, statistics are presented in the tables as “NA” when the number of students in the sample is fewer than 11.

Table 7.16 in subsection *7.5.1 Score Distributions and Summary Statistics* provides definitions for the demographic student groups included in the tables.

2.7. Calibration and Scaling

Item response theory (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 2013–14 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 2017–18 administration of the Smarter Balanced assessments were expressed on the baseline scale. The baseline scale was defined following the 2013–14 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 from the field-test calibration and scaling that comprised the 2017–18 administration item pool are available in appendix 8.A.

2.7.2. Horizontal Scaling

Item parameters derived for the Smarter Balanced assessment were linked during the Smarter Balanced field test administration by concurrently calibrating items within 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 is 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 four items was administered to some students in grade five.

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.)

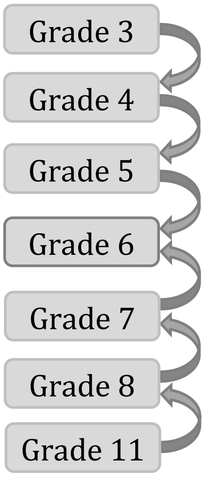

Figure 2.1 Vertical scaling

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/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. Retrieved from 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. Retrieved from 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. (2018a). *CAASPP online test administration manual, 2017–18 administration.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.online_tam.2017-18.pdf

California Department of Education. (2018c). *Matrix one: Universal tools, designated supports, and accommodations for the California Assessment of Student Performance and Progress for 2018–19*. Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ai/caasppmatrix1.asp

California Department of Education. (2017b). *TOMS pre-administration guide for CAASPP testing*. Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.TOMS-pre-admin-guide.2017-18.pdf

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

Smarter Balanced Assessment Consortium. (2017a). *ELA computer adaptive test (CAT) and performance task (PT) item specifications*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017c). *Enhanced CAT blueprints for students participating in the 2016-17 Smarter Balanced embedded field test of performance tasks*. Retrieved from https://web.archive.org/web/20170705015948/https://portal.smarterbalanced.org/library/en/v1.0/enhanced-cat-blueprints-for-students-participating-in-the-2016-17-embedded-field-test-of-performance-tasks.pdf

Smarter Balanced Assessment Consortium. (2018b). *Mathematics CAT and PT item specifications*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

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

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: 2014–15 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf

Smarter Balanced Assessment Consortium. (2017). *Smarter Balanced Assessment Consortium: 2015–16 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/2015-16-summative-technical-report.pdf

Smarter Balanced Assessment Consortium. (2018). *Smarter Balanced Assessment Consortium: 2016–17 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://portal.smarterbalanced.org/library/en/2016-17-summative-assessment-technical-report.pdf

Smarter Balanced Assessment Consortium. (2018d). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines.* Los Angeles, CA: Smarter Balanced Assessment Consortium and National Center on Educational Outcomes. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

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. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-implementation-guide.pdf

Stocking, M. L., & Lord, F. M. (1983). *Developing a common metric in item response theory*. *Applied Psychological Measurement, 7*, 201–210

Chapter 3: Item Development

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–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 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. Retrieved from 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 2014–15 school year. The summative assessments each consist of two parts: a computer adaptive test (CAT) and performance tasks (PTs). 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.

4.1. Smarter Balanced Adaptive Item Selection Algorithm

This subsection describes the algorithm and the design for implementation of adaptive item selection for the Smarter Balanced test delivery system. The implementation builds extensively on the algorithm implemented in American Institutes for Research's (AIR's) test delivery system.

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* (AIR, 2014).

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

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

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 assessment target. Each test is described by a single blueprint for each claim of 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 operational testing window, AIR 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.

In *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 (subclaims) 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 AIR'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 AIR 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	NA
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

AIR 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. Retrieved from 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. Retrieved from https://portal.smarterbalanced.org/library/en/testing-procedures-for-adaptive-item-selection-algorithm.pdf

Chapter 5: Test Administration

This chapter provides an overview of the Smarter Balanced California Assessment of Student Performance and Progress (CAASPP) test administration and describes the measures to ensure test security, procedures to maintain standardization, and procedures for implementation of test accommodations based on Standard 7.8 of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

5.1. Test Administration

The testing window for 2017–18 administration of the CAASPP Smarter Balanced assessments was January 9 through July 16, 2018. 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).

Educational Testing Service (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, and general testing rules. These resources are discussed in more detail in the subsection *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 seen by 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 (refer to Table 5.A.1 and Table 5.A.2 in appendix 5.A for the numbers of items in each operational PT). During the 2017–18 administration, PT tasks were randomly assigned at the student level.

The distributions of the number of items presented to students for the total test and the CAT and the PT components are presented in Table 5.B.1 through Table 5.B.3 in appendix 5.B. Table 5.B.4 presents the counts and percentages of students administered items who meet the criteria specified in the operational blueprints, students who do not meet the criteria, and students who exceed the criteria. Table 5.B.5 presents the counts and percentages of students administered items who meet the criteria specified in the embedded field test blueprints, students who do not meet the criteria, and students who exceed the criteria. 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 (CAT) 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 (PT) 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], 2018a).

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. Table 5.A.1 and Table 5.A.2 in appendix 5.A list the performance tasks given to students and the number of items in each PT.

Note that, during the 2017–18 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 are 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* (AERA, APA, & NCME, 2014), "The documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session" (p. 128).

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 (OTI)

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* (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.

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 test delivery system (TDS) (CDE, 2018a and 2018b).

The following measures ensured the security of CAASPP System assessments administered in 2017–18:

- 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 to the California Technical Assistance Center (CalTAC) before ETS granted the coordinators access to the Test Operations Management System (TOMS). (5 *CCR*, Section 859[a])
- Anyone having access to the testing materials must have 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. Due to 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 the subsection *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

Due to 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, the information technology staff at ETS retrieves data files from the American Institutes for Research and loads them into a database. The ETS Data Quality Services staff extract the data from the database and perform quality control procedures before passing files to the ETS statistical analysis group. The statistical analysis staff store the files on secure servers. All staff members involved with the data adhere to the ETS Code of Ethics and the ETS Information Protection Policies to prevent any unauthorized access to 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 (printed and 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 on both electronic files and paper reports 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 onsite security audits of departments, with follow-up reports containing recommendations for improvement.

5.2.9. Security and Test Administration Incident Reporting System (STAIRS) 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, 2018b). 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 filling out the secure STAIRS form for reporting, which contains selectable options to guide coordinators in their submittal. After the form was submitted, an email containing a case number and next steps was sent to the submitter (and to the LEA CAASPP coordinator, if the form was submitted by the CAASPP test site coordinator). Coordinators could not file an appeal without the case number that is created by submitting the *CAASPP STAIRS* form. The *CAASPP STAIRS* form provided the LEA CAASPP coordinator, the CDE, and CalTAC with the opportunity to interact and communicate regarding the STAIRS process. (CDE, 2018b)

Incidents were then resolved when the LEA CAASPP coordinator or CAASPP test site coordinator either filed an appeal to reset, re-open, invalidate, restore, or grant a grace period extension to a student's test, or by following other instructions in a system-generated email in response to the *STAIRS* form submittal.

The following types of STAIRS reports were also forwarded to the CDE:

- Student cheating
- Security breach (where either a student or an adult exposed secure materials)
- Accidental access to a summative assessment
- Incorrect SSID used (intentionally switched)
- 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 has been approved. The coordinator then returned to TOMS to access the Appeals System, where the appeal was filed (CDE, 2018b).

Types of appeals available during the 2017–18 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.
Invalidation	Invalidated summative assessments will be scored and scores will be provided on the Student Score Report 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.

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.

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 reported the incident within 24 hours, using the online *CAASPP STAIRS* form.

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 in the online Appeals System for resolution. An irregularity must be reported to the LEA CAASPP coordinator and CAASPP test site coordinator immediately. The coordinator reported the irregularity within 24 hours, using the online *CAASPP STAIRS* form.

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 complete the online *CAASPP STAIRS* form 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 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 directed in the email response to the *STAIRS* form (CDE, 2017d).

Types of appeals available during the 2017–18 CAASPP administration are described in Table 5.1.

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 the subsection *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 MI 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 subsection *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 subsection. Refer also to subsection *11.4 Test Administration* for additional information about administration of the CAASPP Smarter Balanced paper-pencil tests.

5.4.1. LEA CAASPP Coordinator

An LEA CAASPP coordinator was designated by the district superintendent at the beginning of the 2017–18 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 the STAIRS response email indicates that an appeal is warranted).

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 ensure all students participate, 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 the STAIRS response email indicates that an appeal is warranted).

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, 2018a).

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 Assessment used by test administrators to administer the Smarter Balanced assessments to students are included in the *CAASPP Online Test Administration Manual* (CDE, 2018a). 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, 2018a).

5.4.4.2 CAASPP Online Test Administration Manual

The *CAASPP Online Test Administration Manual* (CDE, 2018a) 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 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 TOMS Pre-Administration Guide for CAASPP Testing

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 include the following (CDE, 2017c):

- **Test Administration Setup—**This module allows LEAs to determine and calculate dates for the LEA’s 2017–18 administration of the CAA 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 assessments.
- **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.

- ***Technical Specifications and Configuration Guide for CAASPP Online Testing*—**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, 2017d).
- ***Security Incidents and Appeals Procedure Guide*—**This manual provides information on how to report and submit an appeal to the CDE to reset, reopen, invalidate, or restore individual online student assessments (CDE, 2018b).
- ***Accessibility Guide for CAASPP Online Testing*—**This manual provides descriptions of the accessibility features 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 the software Job Access With Speech (JAWS®) tool 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, 2018e).

5.5. LEA Training

ETS established and implemented a training plan for LEA assessment staff on all aspects of the assessment program. The CDE and ETS, in collaboration with the CDE Senior Assessment Fellows and other stakeholders as needed, determined the audience, topics, frequency, and mode (in-person, webcast, videos, modules, etc.) of the training, including such elements as format, participants, and logistics.

ETS conducted 16 in-person pretest workshops and presented four webcasts for the 2017–18 administration.

Following approval by the CDE, the ancillary materials were posted for each webcast on the CAASPP website at http://www.caaspp.org/training/caaspp/ so the LEAs could download the training materials.

5.5.1. In-person Training

ETS also provided a series of in-person trainings. Beginning in January 2018, the first in-person trainings provided were the pretest CAASPP workshops, which focused on training LEA CAASPP coordinators on how to prepare for administering the CAASPP online assessments. Additionally, a two-session Post-Test Workshop was offered in May and June

2018 with the sessions “Principles of Scoring and Reporting” and “The Results Are In—Now What?”

5.5.2. Webcasts

ETS provided a series of live webcasts throughout the school year that were archived and made available for training LEA and test site staff as well as 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 CAASPP Summative Assessments Videos and Archived Webcasts web page at http://www.caaspp.org/training/caaspp/. CAASPP webcasts are available to everyone and require neither preregistration nor a logon account.

5.5.3. Videos and Narrated PowerPoint Presentations

To supplement the live webcasts and in-person workshops, ETS also produced short “how-to” videos and narrated PowerPoint presentations that were available on the CAASPP Summative Assessments Videos and Archived Webcasts web page. In total, 20 recorded webcasts and tutorials were produced for the 2017–18 administration year.

5.6. 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.

5.6.1. Identification

All public school students participate in the CAASPP System, including students with disabilities and English learners. The Smarter Balanced Assessment Consortium’s *Usability, Accessibility, and Accommodations Guidelines* (Smarter Balanced, 2018) and the CDE’s Matrix One (CDE, 2018f) 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.

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, 2018a), 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.

5.6.2. Assignment

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

There are three ways the student's accessibility resource(s) can be assigned:

1. Using the Individual Student Assessment Accessibility Profile Tool (ISAAP) 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 identifies and designates a resource not identified in Matrix One, the LEA CAASPP coordinator or CAASPP test site coordinator needs to submit a request for an unlisted resource to be approved by the CDE. The CDE then determines if the requested unlisted resource changes the construct being measured after all testing has been completed.

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 Department of Education. (2018e). *Accessibility guide for CAASPP online testing*. Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.accessibility-guide.2017-18.pdf

California Department of Education. (2018a). *CAASPP online test administration manual, 2017–18 administration.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.online_tam.2017-18.pdf

California Department of Education. (2018f). *Matrix one: Universal tools, designated supports, and accommodations for the California Assessment of Student Performance and Progress (CAASPP) system*. Sacramento, CA: California Department of Education. Retrieved from https://web.archive.org/web/20190801214751/https://www.cde.ca.gov/ta/tg/ai/documents/caasppmatrixone0918.docx

California Department of Education. (2018b). *Security incidents and appeals procedure guide, 2017–18 administration.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.stairs-appeals-guide.2017-18.pdf

California Department of Education. (2017d). *Technical specifications and configuration guide for CAASPP online testing*. Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.tech-specs-and-config-guide.2017-18.pdf

California Department of Education. (2017c). *TOMS pre-administration guide for CAASPP testing*. Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.TOMS-pre-admin-guide.2017-18.pdf

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service. Retrieved from 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.

Smarter Balanced Assessment Consortium. (2018). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines.* Los Angeles: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations implementation guide.* Los Angeles: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-implementation-guide.pdf

Smarter Balanced Assessment Consortium. (2018a). *Smarter Balanced Resources and Practices Comparison Crosswalk.* Los Angeles: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/uaag-resources-and-practices-comparison-crosswalk.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. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

Chapter 6: Standard Setting

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 cut scores 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 cut scores 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. Retrieved from 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 Online Summative Assessments in English language arts/literacy (ELA) and mathematics, student item responses are scored and individual student scores—overall scale scores and claims and subscores—are calculated based on the item responses. In addition, student test scores must be aggregated to produce information for schools and local educational agencies (LEAs). This chapter describes how various types of student responses are scored for the CAASPP online assessments, as well as the various types of scores that are generated. This chapter also presents information on the concept of measurement error and how measurement error should be considered when interpreting student test scores.

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 his or her 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. Here 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. Here 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

Educational Testing Service (ETS) staff from the Assessment Development, Enterprise Score Key Management, Psychometrics Analysis and 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 [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 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 2017h [ELA]; and 2018a through 2018j [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 multisegment 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

Item Type	Description	Content Area
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
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.[5] A small number of mathematics PTs include CR items with a 0–4 point range. CR items for CAASPP Smarter Balanced 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 to 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.

[5] 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.

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 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 Scoring Guide* (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 of 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 his or her 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 human rater pool for ETS (Table 7.2), MI (Table 7.3), and combined (Table 7.4) across both organizations.

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

Characteristic	N	Percent
Total raters scoring in 2017–18	**4,147**	**NA**
Fluent in Spanish and expressed interest in scoring assessments in Spanish	84	2%
Experience teaching in a kindergarten (K)–12 school	969	23%
Currently works in a K–12 school in California	733	18%
Others—Not meeting any of the previous criteria	2,361	57%

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

Characteristic	N	Percent
Total raters scoring in 2017–18	**1,776**	**NA**
Fluent in Spanish and expressed interest in scoring assessments in Spanish	NA[6]	NA
Experience teaching in a kindergarten (K)–12 school	285	16%
Currently works in a K–12 school in California	285	16%
Others—Not meeting any of the previous criteria	1,206	68%

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

Characteristic	N	Percent
Total raters scoring in 2017–18	**5,923**	**NA**
Fluent in Spanish and expressed interest in scoring assessments in Spanish	84	NA
Experience teaching in a kindergarten (K)–12 school	1,254	21%
Currently works in a K–12 school in California	1,018	17%
Others—Not meeting any of the previous criteria	3,567	61%

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)
- 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 in order 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 which 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

[6] MI does not hire raters specifically for CAASPP, so the counts presented are specifically for California educators and residents. Distinct counts of some groups are not available.

Balanced recommendation of a 70 percent exact and 0 percent discrepant (nonadjacent) agreement rate to score.

The standards, provided in Table 7.5, are 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 nine percent of the total papers scored by a rater during his or her 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 (SA) 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 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 large-scale, 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. The results of the agreement analysis are presented in *8.6.4.8 Interrater Agreement*.

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' Enterprise Score Key Management (eSKM) system utilizes 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 use at least 30 validity papers that cover the full range of scores. Validity sets are monitored throughout the administration and post-administration periods for performance. Supplemental samples are added as needed. The statistics for interrater reliability for all items at all grades are presented in appendix 8.G. These statistics include the percentage of perfect agreement and adjacent agreement between the two readers and the quadratic-weighted kappa statistic.

Smarter Balanced provides flagging criteria (Smarter Balanced, 2016) based on the statistics that follow for identifying items to be reviewed for potential elimination after scoring is completed. ETS uses the Smarter Balanced flagging criteria and reports items flagged in the technical documentation. Polytomous items are flagged if any of the following conditions occur:

- Adjacent agreement < 0.80
- Exact agreement < 0.60
- Quadratic-weighted kappa < 0.20

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

- Exact agreement < 0.80
- Quadratic-weighted kappa < 0.20

Table 7.6 shows the number of items flagged by content area, grade, and scoring method. There were 158 items flagged among 1,253 scored items across all grades in ELA and mathematics.

Table 7.6 Number of Constructed-Response Items Flagged, by Content Area and Grade, 2017–18

Scoring Method	Content Area	Grade	Flagged Polytomous Items	Flagged Dichotomous Items	Total Flagged Items	Total Number of Scored Items	Percentage Flagged
Human to Human SA	ELA	3	1	5	6	18	33
Human to Human SA	ELA	4	2	7	9	20	45
Human to Human SA	ELA	5	7	2	9	21	43
Human to Human SA	ELA	6	0	0	0	14	0
Human to Human SA	ELA	7	0	0	0	15	0
Human to Human SA	ELA	8	0	0	0	19	0
Human to Human SA	ELA	11	9	0	9	22	41
Human to Human SA	Mathematics	3	0	0	0	28	0
Human to Human SA	Mathematics	4	0	0	0	39	0
Human to Human SA	Mathematics	5	0	0	0	33	0
Human to Human SA	Mathematics	6	0	0	0	38	0
Human to Human SA	Mathematics	7	0	0	0	30	0
Human to Human SA	Mathematics	8	0	0	0	30	0
Human to Human SA	Mathematics	11	0	2	2	31	6
Human to AI SA	ELA	3	3	0	3	38	8
Human to AI SA	ELA	4	3	0	3	58	5
Human to AI SA	ELA	5	11	0	11	58	19

Scoring Method	Content Area	Grade	Flagged Polytomous Items	Flagged Dichotomous Items	Total Flagged Items	Total Number of Scored Items	Percentage Flagged
Human to AI SA	ELA	6	6	0	6	45	13
Human to AI SA	ELA	7	1	0	1	53	2
Human to AI SA	ELA	8	5	0	5	55	9
Human to AI SA	ELA	11	14	0	14	87	16
Human to AI SA	Mathematics	3	0	0	0	24	0
Human to AI SA	Mathematics	4	0	1	1	14	7
Human to AI SA	Mathematics	5	0	0	0	23	0
Human to AI SA	Mathematics	6	0	0	0	9	0
Human to AI SA	Mathematics	7	2	2	4	6	67
Human to AI SA	Mathematics	8	0	0	0	12	0
Human to AI SA	Mathematics	11	0	0	0	29	0
Human to Human WER	ELA	3	7	NA	7	21	33
Human to Human WER	ELA	4	18	NA	18	54	33
Human to Human WER	ELA	5	20	NA	20	60	33
Human to Human WER	ELA	6	0	NA	0	24	0
Human to Human WER	ELA	7	1	NA	1	57	2
Human to Human WER	ELA	8	0	NA	0	60	0
Human to AI WER	ELA	3	6	NA	6	18	33
Human to AI WER	ELA	6	4	NA	4	18	22
Human to AI WER	ELA	11	19	NA	19	72	26
Overall	**-**	**-**	**139**	**19**	**158**	**1,253**	**13**

7.3. Rater Training

7.3.1. Training Overview

7.3.1.1 ELA

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

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

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

Qualification and validity sets are provided for each WER essay type. Anchor and training sets are 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 receive training for a grade span (grades three through five, six through eight, or grade eleven) instead of a grade level.

Although training is provided at the task-model level, rater qualification occurs on an item-type and grade-span basis for all ELA human-scored items. Qualification and validity papers are provided for each ELA CR item. Raters must qualify for each item type within a specific grade span before being assigned to score that item type (AIR, 2014).

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, 2014).

7.3.2. Training Process: ELA PT Extended Writing Tasks

Baseline anchor sets for each writing purpose (e.g., informational writing or explanatory writing) are 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 span. 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.

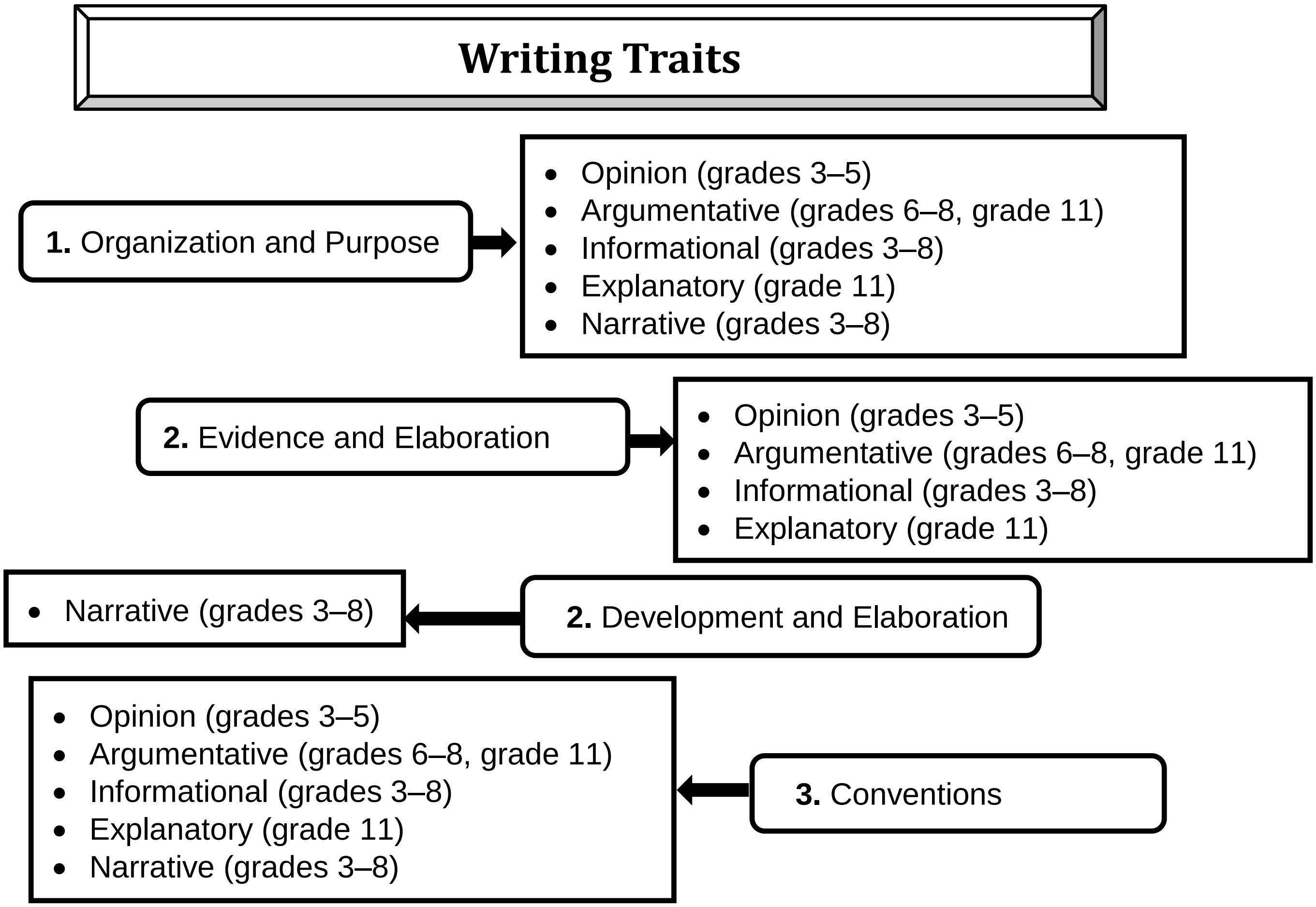
Writing Traits
1. Organization and Purpose
Opinion (grades 3–5)
Argumentative (grades 6–8, grade 11)
Informational (grades 3–8)
Explanatory (grade 11)
Narrative (grades 3–8)
2. Evidence and Elaboration
Opinion (grades 3–5)
Argumentative (grades 6–8, grade 11)
Informational (grades 3–8)
Explanatory (grade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

Figure 7.1 Writing Traits

The training steps are described in the top panel of Figure 7.2; the training materials are described in the bottom panel.

Training steps:

1. Trainees read the task, rubrics, and source materials for the WER items in a particular grade span 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.

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

1. Baseline anchor sets approved during Smarter Balanced Pre-Range Finding*
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

* 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.

Figure 7.2 Training Process for Extended Writing Tasks

7.3.3. Training Process: ELA 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 grade eleven). These training steps are described in the top panel of Figure 7.3, and the training materials are described in the bottom panel.

Training steps:

1. Trainees read the rubrics and scoring notes for the short-answer items in a particular grade span 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.

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

Figure 7.3 Training Process for ELA Short Answer Items

7.3.4. Training Process: Mathematics Items

The training steps for scoring mathematics items are described in the top panel of Figure 7.4, and the training materials are described in the bottom panel.

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.

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

Figure 7.4 Training Process for Mathematics Items

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 are routed to scoring supervisors for final code assignment. Supervisors require detailed training on the Smarter Balanced condition codes and definitions (Smarter Balanced, 2014a).

Table 7.7 presents the valid condition codes used for scoring, along with descriptions of the responses that would warrant the assignment of the different codes.

Table 7.7 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; for example: • 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.

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 are 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 is assigned a score and identified as an “alert.” Raters receive a process document as part of their training materials that describes 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 LEA’s superintendent and LEA CAASPP coordinator in the LEA for the flagged student are sent a copy of the response and the student 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 apply the same scoring algorithms and specifications. ETS psychometricians verify 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 are reported only when the two parallel systems produce identical results with acceptable tolerance.

When scores do not match, the mismatch is 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 applies a problem item notification (PIN) not to score the item through the systematic process in eSKM, which might result in a mismatch if TIPS is still in the process of applying the PIN in the parallel system when the student score is 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 (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 (GPC) model is applied to both types of items. The GPC model (Muraki, 1992) is:

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

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 2-parameter logistic model for dichotomous items.

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

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

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

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, student's estimated ability θ is:

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

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 is so flat that its maximum is 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 test characteristic curve 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 7.8 for each grade and content area (Smarter Balanced Assessment Consortium, 2016). The theta scores for grades three through eight and grade eleven are on a common vertical scale.

Table 7.8 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 following three situations:

1. Attemptedness/Participation 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 7.9.

Table 7.9 Treatment of Incomplete Tests

If the student. . .	Classify the student as participating?	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 treated as incorrect), or for an incomplete test, estimate from equation 7.4	Yes	Yes

The number and percent of students who participated in the tests are presented in the tables of appendix 7.A for all students in each test and for the selected demographic student groups by grade and content area. In addition, the numbers of students in the selected demographic student groups with different test completion conditions are presented in the tables of appendix 7.F. Note that in appendix 7.A and appendix 7.F, all students are counted, including the students assigned to take embedded field test PTs.

Sometimes a student stops answering items before the test delivery system 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.

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

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 lowest obtainable theta 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 item response theory (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: $SS = 85.8\,\theta + 2508.2$ (7.5)

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

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 7.10.

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, 2016) and the *Smarter Balanced Scoring Specification: 2014–2015 Administration* (AIR, 2015b).

Table 7.10 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

Content Area and Grade	LOSS	HOSS
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” or “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, 2016) for details related to the standard setting procedure; *Reporting Achievement Level Descriptors* (Smarter Balanced, 2015c) 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 cut scores for the achievement levels vary by grade and content area. Table 7.11 provides the theta cut scores for Standard Nearly Met, Met, and Exceeded at each grade level. For example, the cut score 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 7.11 Theta Cut Scores

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

Table 7.12 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 7.13 and are also available in the blueprints, which are provided in appendix 2.A.

Table 7.13 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 7.14 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 7.14 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.
- **At/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:

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

where,

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

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

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

When item information is based on the generalized partial credit model for both dichotomous and polytomous items, it is calculated as:

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

where,

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

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

and

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

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, 2015a).

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

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

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:

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

where,

SS is the scale score,

SE_{scaled} is the standard error of measurement 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 he or she 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.

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. Table 7.15 summarizes the number of reportable assessment targets for the 2017–18 CAASPP Smarter Balanced administration.

Table 7.15 Number of Targets with 10 Items or More

Grade	ELA Claim 1	ELA Claim 2	ELA Claim 3	ELA Claim 4	ELA Total	Mathematics Claim 1
3	13	7	1	3	24	11
4	14	7	1	3	25	11
5	14	7	1	3	25	11
6	14	6	1	3	24	10
7	14	7	1	3	25	9
8	14	7	1	3	25	10
11	14	7	1	3	25	16

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.

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 7.16. Summary statistics on assessment where the field-test PT items are embedded are presented in Table 7.17. 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. The mean thetas and corresponding scale scores increase as expected over increasing grade levels across the vertical scale. [7]

Table 7.16 Operational Mean and Standard Deviation of Theta and Scale Scores

Content Area and Grade	Number of Students	Mean Scale Score	Scale Score SD	Mean Theta Score	Theta Score SD
ELA 3	425,170	2424	92	-0.98	1.08
ELA 4	444,084	2464	99	-0.52	1.15
ELA 5	449,833	2496	101	-0.14	1.18
ELA 6	462,575	2519	100	0.13	1.17
ELA 7	451,848	2544	104	0.42	1.21
ELA 8	449,042	2559	104	0.59	1.21
ELA 11	430,454	2593	119	0.98	1.38
Mathematics 3	427,118	2431	85	-1.06	1.07
Mathematics 4	445,894	2468	87	-0.60	1.10
Mathematics 5	451,321	2490	96	-0.31	1.20
Mathematics 6	463,902	2511	110	-0.05	1.39
Mathematics 7	453,151	2524	117	0.12	1.48
Mathematics 8	449,516	2541	127	0.32	1.61
Mathematics 11	429,299	2562	130	0.59	1.64

Table 7.17 Embedded Field Test PTs Mean and Standard Deviation of Theta and Scale Scores

Content Area and Grade	Number of Students	Mean Scale Score	Scale Score SD	Mean Theta Score	Theta Score SD
ELA 3	8,611	2419	92	-1.04	1.08
ELA 4	9,002	2459	101	-0.57	1.18
ELA 5	9,087	2489	102	-0.23	1.19
ELA 6	9,439	2507	103	-0.02	1.19
ELA 7	9,175	2532	105	0.28	1.22
ELA 8	9,100	2549	105	0.47	1.22
ELA 11	8,640	2581	117	0.85	1.36

[7] Note that this information in this technical report may differ slightly from information found on the CDE CAASPP Results website at http://caaspp.cde.ca.gov/ due to different dates on which the data was accessed.

Content Area and Grade	Number of Students	Mean Scale Score	Scale Score SD	Mean Theta Score	Theta Score SD
Mathematics 3	8,677	2430	85	-1.07	1.07
Mathematics 4	9,054	2467	88	-0.60	1.11
Mathematics 5	9,154	2487	97	-0.35	1.22
Mathematics 6	9,443	2507	112	-0.10	1.41
Mathematics 7	9,216	2522	116	0.09	1.46
Mathematics 8	9,113	2539	128	0.31	1.62
Mathematics 11	8,562	2557	130	0.54	1.64

For students who took only operational items, 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 7.18.

Table 7.18 Percentages and Counts of Operational-only Students 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	120,052	28%	99,941	24%	93,937	22%	111,240	26%	205,177	48%
ELA 4	142,263	32%	85,550	19%	99,337	22%	116,934	26%	216,271	49%
ELA 5	137,323	31%	89,908	20%	124,393	28%	98,209	22%	222,602	49%
ELA 6	125,697	27%	115,156	25%	141,704	31%	80,018	17%	221,722	48%
ELA 7	120,250	27%	104,426	23%	153,265	34%	73,907	16%	227,172	50%
ELA 8	115,645	26%	112,467	25%	149,062	33%	71,868	16%	220,930	49%
ELA 11	93,871	22%	95,330	22%	130,784	30%	110,469	26%	241,253	56%
Mathematics 3	117,609	28%	100,497	24%	119,012	28%	90,000	21%	209,012	49%
Mathematics 4	117,131	26%	137,391	31%	109,028	24%	82,344	18%	191,372	43%
Mathematics 5	167,283	37%	121,593	27%	73,907	16%	88,538	20%	162,445	36%
Mathematics 6	160,809	35%	129,065	28%	86,496	19%	87,532	19%	174,028	38%
Mathematics 7	165,785	37%	118,285	26%	84,320	19%	84,761	19%	169,081	37%
Mathematics 8	180,508	40%	103,232	23%	72,443	16%	93,333	21%	165,776	37%
Mathematics 11	196,417	46%	98,110	23%	79,353	18%	55,419	13%	134,772	31%

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

Figure 7.5 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 7.18.

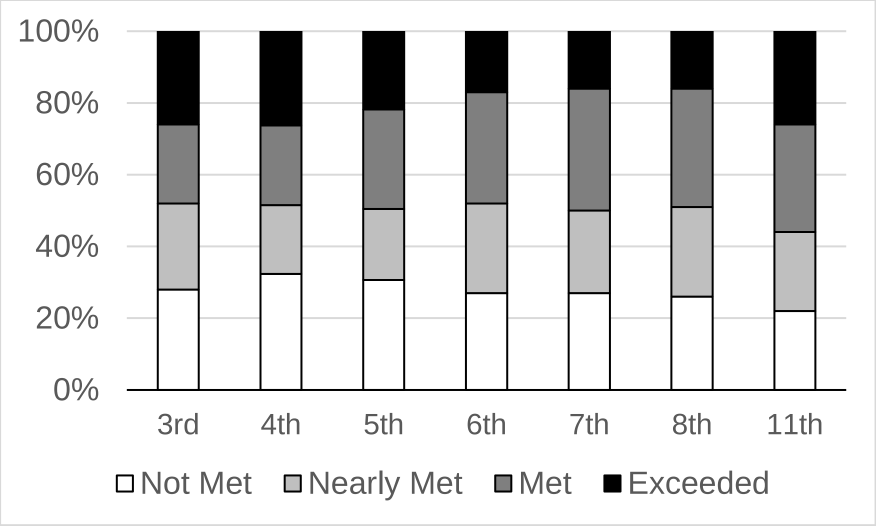

Figure 7.5 Percentages of Achievement Levels in ELA, Operational Assessments

Figure 7.6 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 7.18.

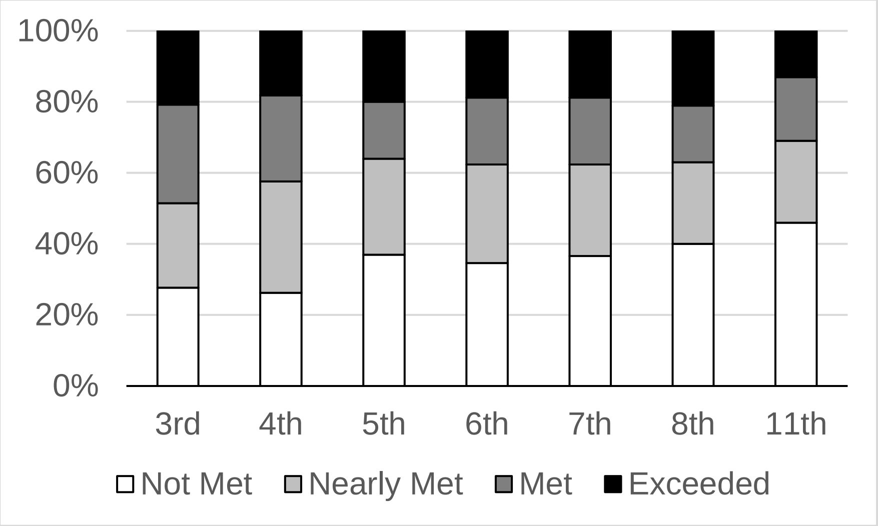

Figure 7.6 Percentages of Achievement Levels in Mathematics, Operational Assessments

For students who took an embedded field test PT, 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 7.19.

Table 7.19 Percentages and Counts of Embedded Field Test–only PTs Students 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	2,653	31%	1,995	23%	1,875	22%	2,088	24%	3,963	46%
ELA 4	3,129	35%	1,676	19%	1,937	22%	2,260	25%	4,197	47%
ELA 5	3,063	34%	1,830	20%	2,382	26%	1,812	20%	4,194	46%
ELA 6	3,043	32%	2,355	25%	2,613	28%	1,428	15%	4,041	43%
ELA 7	2,828	31%	2,324	25%	2,705	29%	1,318	14%	4,023	44%
ELA 8	2,763	30%	2,246	25%	2,812	31%	1,279	14%	4,091	45%
ELA 11	2,098	24%	2,071	24%	2,585	30%	1,886	22%	4,471	52%
Mathematics 3	2,422	28%	2,161	25%	2,250	26%	1,844	21%	4,094	47%
Mathematics 4	2,426	27%	2,750	30%	2,220	25%	1,658	18%	3,878	43%
Mathematics 5	3,529	39%	2,438	27%	1,426	16%	1,761	19%	3,187	35%
Mathematics 6	3,438	36%	2,567	27%	1,710	18%	1,728	18%	3,438	36%
Mathematics 7	3,470	38%	2,383	26%	1,711	19%	1,652	18%	3,363	36%
Mathematics 8	3,758	41%	1,981	22%	1,492	16%	1,882	21%	3,374	37%
Mathematics 11	4,055	47%	1,900	22%	1,569	18%	1,038	12%	2,607	30%

* May not exactly match the sum of Level 3 and Level 4 percentages due to rounding.

Figure 7.7 presents a graphical representation of the percentage of students who took the embedded field test PTs at each ELA achievement level by grade. These are the achievement levels for ELA shown in Table 7.19.

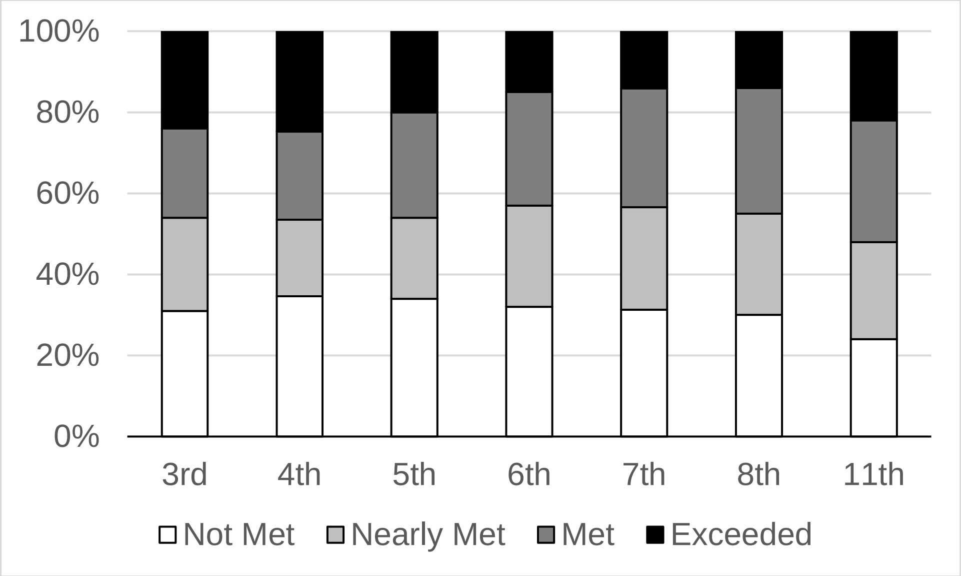

Figure 7.7 Percentages of Achievement Levels in ELA, Embedded Field Test PTs

Figure 7.8 presents a graphical representation of the percentage of students who took the embedded field test PTs at each mathematics achievement level by grade. These are the achievement levels for mathematics shown in Table 7.19.

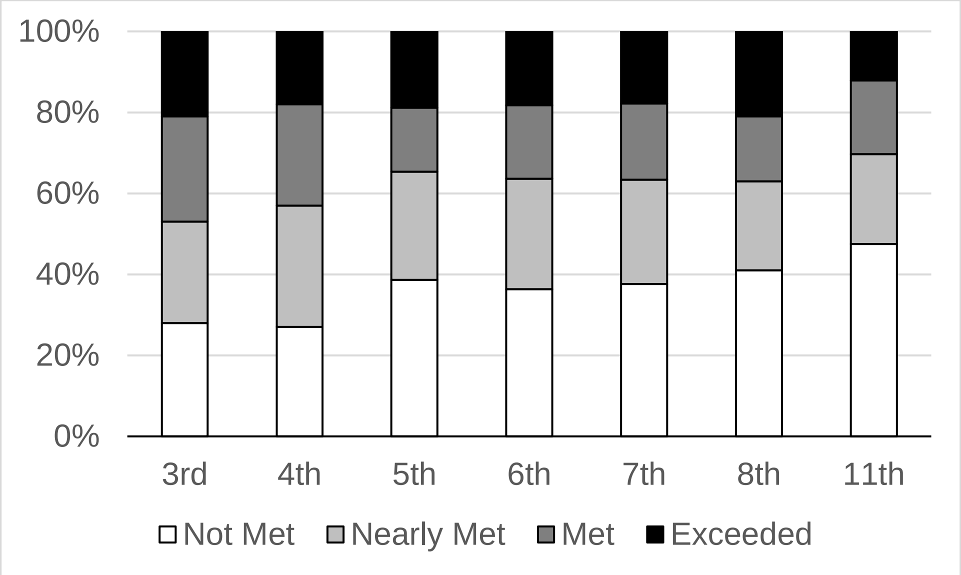

Figure 7.8 Percentages of Achievement Levels in Mathematics, Embedded Field Test PTs

Detailed score distribution information is available in the appendices. Table 7.B.1 and Table 7.B.2 in appendix 7.B show the estimated distributions of theta scores for each test. Table 7.C.1 and Table 7.C.2 in appendix 7.C present the selected percentiles of the scale score distributions. Table 7.C.3 through Table 7.C.16 present the frequency distributions of scale scores for each assessment.

Table 7.B.3 through Table 7.B.16 contain the distributions of theta scores for each claim. Table 7.D.1 through Table 7.D.4 in appendix 7.D show the number of items presented within each test, number of students with valid score in each claim, and the mean and standard deviation of student scores expressed in terms of both scale score and theta score. “Valid score” means the student records were not flagged as “not scored” or the students were enrolled in the grade for which they were tested. The number of students in each claim performance level are reported in Table 7.D.5 through Table 7.D.8. For frequency distributions in appendix 7.B, appendix 7.C, and appendix 7.D, all students are counted, including the students assigned the embedded field test PTs.

7.5.2. Group Scores

Statistics summarizing student performance by content area and grade for selected groups of the students who took only operational items are provided in appendix 7.E: for each test in Table 7.E.1 through Table 7.E.14, and for each test claim in Table 7.E.29 through Table 7.E.42. The summary statistics of student performance by content area and grade for selected groups of students who were assigned to take embedded field test PTs are presented in Table 7.E.15 through Table 7.E.28 and for each test claim in Table 7.E.43 through Table 7.E.56.

In the tables, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status (disadvantaged or not), special education services status, migrant status, and ethnicity by economic status. The tables show, for each demographic group, the number of students with a valid scale score, scale score mean and standard deviation, and the percentage of students in each achievement level and claim performance level.

Table 7.20 lists the demographic student groups included in the tables. Students’ economic status was determined by the education level of their parents and whether or not the student participated in the National School Lunch Program. To protect privacy when the number of students in a student group is 10 or fewer, the summary statistics at the achievement and claim level are not reported, but are replaced by “NA.”

Table 7.20 Demographic Student Groups to Be Reported

Value	Student Groups
Gender	• Male • Female
Ethnicity	• American Indian or Alaska Native • Asian • Black or African American • Filipino • Hispanic or Latino • Native Hawaiian or Other Pacific Islander • White • Two or more races

Value	Student Groups
English Language Fluency	• English only • Initial fluent English proficient • English learner • Reclassified fluent English proficient • To be determined • English proficiency unknown
Economic Status	• Not economically disadvantaged • Economically disadvantaged
Special Education Services Status	• No special education services • Special education services
Migrant Status	• Eligible for the Title I Part C Migrant Program • Not eligible for the Title I Part C Migrant Program

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 school districts identify strengths and areas that need improvement in their educational programs, and
4. provide the public and policymakers with information about student achievement.

This subsection 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 test delivery system. 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, TOMS communicates with the Online Reporting System (ORS) that provides authorized users with interactive and cumulative online reports for ELA and mathematics at the student, school, and LEA levels. The ORS provides access to two CAASPP functions:

Score Reports, which provide preliminary score data for each administered test available in the reporting system; and the Completion Status Reports, which provide completion data in the reporting system for students taking an assessment.

Based on the Smarter Balanced reporting requirements for ELA and mathematics, the ORS provides the preliminary summative reports containing information outlining student knowledge and skills, as well as performance levels aligned to the assessment-specific claims. The online aggregate reports provide functionality at the student, classroom, school, and LEA levels. The online aggregate reports are available to be downloaded in PDF, Excel, and comma-separated value formats.

7.6.2. Special Cases

Student scores are not reported for the following cases:

- Student was absent from the test
- Student's answer document was blank or student moved or had a medical emergency
- 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
- Student was invalidated in the system

7.6.3. Types of Score Reports

There are three categories of CAASPP reports. The categories and the specific reports within each category are as follows:

- Student Score Report
  - The Student Score Report is the official score report for the parents or guardians and describes the student's results.
  - 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 7.8.)
    - 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 use accommodations or designated supports are reported in the same way as for students without accommodations or designated

  supports. (Refer to subsection *2.5 Universal Tools, Designated Supports, and Accommodations* for more information about accessibility resources.)
  - LEAs receive paper Student Score Reports to distribute to parents/guardians and students' schools. This report is also provided as a printable PDF that the LEA CAASPP coordinator may download from TOMS.
  - Further information about the CAASPP online summative assessments Student Score Report and the other reports is provided at http://caaspp.cde.ca.gov/.
- School Reports
  - The school performance report provides group information by content area, including the school average scale score and percentage of students at or above "Standard Met."
  - This report provides a list of students' scale scores, achievement levels, and performance levels for claims.
  - The school scale score report is presented as a dashboard to provide group information by content area. It includes a histogram showing the distribution of students' scale scores.
- District Reports
  - The district performance report provides school-level information by content area, including the school average scale score and percentage of students at or above "Standard Met."
  - This report lists all the proficiency information for each school, including the testing status, number of students who completed testing, average scale score, and percentage of students in each achievement level.
  - The district scale score report is presented as a dashboard to provide cumulative information. The histogram shows the frequency of schools with mean scores in each score interval.

The CAASPP aggregate reports and student data files for the LEA are available for the LEA CAASPP coordinator to download from TOMS. The LEA CAASPP coordinator forwards the appropriate reports to test sites. In the case of the CAASPP Student Score Report, the LEA sends the printed report(s) to the child's parent or guardian and forwards a copy to the student's school or test site. Downloaded Student Score Reports are forwarded to the test site. CAASPP Student Score Reports that include individual student results are not distributed beyond the student's school.

Internet reports are described on the CDE website and are accessible to the public online at http://caaspp.cde.ca.gov/.

Preliminary individual student scores are also available to LEAs prior to the release of final reports via electronic reporting, accessed using the ORS. This application permits LEAs to view preliminary results data for all tests taken.

7.6.4. Score Report Applications

CAASPP online summative assessment results provide parents and guardians with information about their child's progress. The results are a tool for increasing communication and collaboration between parents or guardians and teachers. Along with the results from the Smarter Balanced Interim Assessments, the Student Score Report can be used by parents and guardians while talking with teachers about ways to improve their child's achievement of the CCSS.

Schools may 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 needs of students, and
- decisions about how to use funds to ensure that students achieve the standards.

CAASPP online summative assessments results are 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 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 performance-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.

For more details on the criteria for interpreting information provided on the score reports, refer to the *2017–18 CAASPP Post-Test Guide* (CDE, 2018).

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. Retrieved from 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 (2018). *2017–18 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. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.post-test_guide.2017-18.pdf

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. (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. Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/08/ELA_Content_Specs.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. Retrieved from https://portal.smarterbalanced.org/library/en/mathematics-content-specifications.pdf

Smarter Balanced Assessment Consortium. (2016). Smarter Balanced Assessment Consortium: 2014-15 Technical Report. Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf/

Smarter Balanced Assessment Consortium. (2017a). *ELA CAT item specifications, grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017b). *ELA CAT item specifications, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017c). *ELA CAT item specifications, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017d). *ELA PT item specifications, argumentative, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017e). *ELA PT item specifications, explanatory, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017e). *ELA PT item specifications, informative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017f). *ELA PT item specifications, narrative, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017g). *ELA PT item specifications, narrative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017h). *ELA PT item specifications, opinion, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2014a). *Hand-scoring rules.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from 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. Retrieved from https://portal.smarterbalanced.org/library/en/interpretation-and-use-of-scores-and-achievement-levels.pdf

Smarter Balanced Assessment Consortium. (2018a). *Mathematics computer adaptive test (CAT) and performance task (PT) item specifications*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018b). *Mathematics CAT item specifications, Claim 1, grade eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018c). *Mathematics CAT item specifications, Claim 1, grade five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018d). *Mathematics CAT item specifications, Claim 1, grade four*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018e). *Mathematics CAT item specifications, Claim 1, grade seven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018f). *Mathematics CAT item specifications, Claim 1, grade six*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018g). *Mathematics CAT item specifications, Claim 1, grade three*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018p). *Mathematics CAT item specifications, Claim 1, high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

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. Retrieved from http://www.smarterbalanced.org/assessments/development/

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. Retrieved from http://www.smarterbalanced.org/assessments/development/

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. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2015c). *Reporting achievement level descriptors.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/achievement-level-descriptors.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. Retrieved from https://portal.smarterbalanced.org/library/en/scoring-guide-for-ela-full-writes.pdf

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

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

Chapter 8: Analyses

This chapter summarizes the item- and test-level statistics calculated for the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced Online Summative Assessments administered during the 2017–18 administration.

8.1. Background

There are five primary statistical analyses presented in this chapter:

1. Item Response Theory (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

Each of these sets of analyses is presented in the body of the text and in the listed appendixes. 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 2013–14 field test administration (Smarter Balanced, 2016b).

1. **Item Response Theory (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. Also presented for each test are conditional distributions of *a*-values and *b*-values for students at specified ability levels (scale-score intervals) and the *a*-values and *b*-values of all performance task (PT) items. For polytomous items, partial credit step values (*d*-values) are included.

2. **Omission and Completion Analyses.** Appendix 8.B shows item parameter estimate summaries for items with different omit rates. Statistics are shown for the PTs and computer adaptive test (CAT) items in each test. The item parameter estimates are from the field-test calibrations. The purpose of these analyses is to examine whether the items with high omit rates are systematically more difficult or more discriminating than items with low omit rates. Appendix 8.B also shows the completion rates for each test.

3. **Conditional Exposure Analyses.** Appendix 8.C shows, for each assessment, distributions (in intervals) of item exposure frequency for all items in that test, for the items in each claim, and for items at different difficulty levels.

4. **Reliability Analyses.** The following results of the analyses are presented:

    - Appendix 8.D presents results of the reliability analyses of test scores and claim scores for the population as a whole and for selected student groups.

    - Table 8.3 presents the reliability results for the population as a whole.

    - Table 8.4 shows the conditional standard errors of measurement (CSEMs) at achievement-level scale score cuts.

    - Tables in Appendix 8.E present CSEM distributions for the total test scores.

- Figure 8.E.1 through Figure 8.E.14 in appendix 8.E present plots of CSEMs conditional on scale scores.
- Table 8.5 presents the mean CSEM for each achievement level.
- Tables in appendix 8.F present statistics describing the accuracy and consistency of the performance classifications.
- Appendix 8.G shows interrater reliability statistics for the human-scored items and statistics showing the agreement of artificial intelligence (AI) scoring with human scoring for the constructed-response (CR) items.

5. **Analyses in Support of Validity Evidence.** Validity evidence related to the CAASPP online summative assessments is discussed in subsection *8.6 Validity Evidence*. Appendix 8.H presents distributions of the time required to complete the total test for each content area, including both the PT and CAT portions. Table 8.6 and the tables in appendix 8.I present correlations between English language arts/ literacy (ELA) and mathematics scores calculated for all students and for demographic student groups of interest.

8.1.2. Samples for the Analyses

Analyses were conducted based on version 5 of the production data file (“P5”) received in October 2018. The P5 file comprised the full 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 2017–18 operational tests. However, because the field test data was not provided to Educational Testing Service (ETS), none of the PT field-test items were analyzed in this chapter.

8.2. IRT 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, 2016b) and subsequent Smarter Balanced technical reports (Smarter Balanced, 2016c, 2017i, 2018s).

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 (2PL) model (Birnbaum,1968) for calibration of the dichotomous items and the generalized partial credit model (GPCM; 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, 2016b).

8.2.1. Summary Information

Parameter estimates for the 2017–18 operational items were obtained mainly from the 2013–14 Smarter Balanced field-test analyses, but also from the subsequent Smarter Balanced embedded field-test analyses after the 2013–14 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, 2016b).

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.

8.2.1.2 CAT Items

Table 8.A.29 through Table 8.A.42 in appendix 8.A show the distributions of CAT item *a*-values across 10 intervals of the ability scale, conditional on 6 intervals of student ability indicated by ranges of the overall test scale score. Table 8.A.43 through Table 8.A.56 present the distributions of CAT items across 16 intervals of *b*-values conditional on 6 intervals of overall test scale scores. The mode of each distribution is in bold text and indicated with an asterisk.

8.2.1.3 Performance Task Items

Table 8.A.57 through Table 8.A.70 in appendix 8.A show the conditional distribution of *a*-values for the PT items. Table 8.A.71 through Table 8.A.84 show the conditional distribution of *b*-values for the PT items. Parameter values of all PT items are presented in Table 8.A.85 through Table 8.A.98.

For Table 8.A.29 through Table 8.A.84, the scale score intervals included in the table range from the lowest 100 scale scores containing the lowest obtainable scale score (LOSS) to the highest obtainable scale score (HOSS) with increments of 100 scale score points. For example, “2100–2199” to “2600–2699” for ELA in grade three includes the LOSS of 2114 and the HOSS of 2623.

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 five percent for any single item may be an indication of a problem with an item.

Table 8.B.1 and Table 8.B.2 in appendix 8.B present the summary of omit rates, including the number of items in each omit rate interval, for the PT and CAT items respectively. The tables also contain the average difficulty and discrimination for these items. As shown, the overall omit rates for CAT items across contents and grades are very low, and no items have omit rates higher than five percent.

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 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 students who are assigned only operational items and for students who are assigned embedded field-test PTs, respectively. The percentages of students completing each test, each claim on the test, and each of the two parts of the test are presented in Table 8.B.3 and Table 8.B.4 in appendix 8.B. Note that all students are counted in these tables, including the students assigned with embedded field-test PTs.

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	13	15
ELA Claim 2	12	12	12
ELA Claim 3	8	8	8
ELA Claim 4	8	8	8
Mathematics Claim 1	17	16	19
Mathematics Claim 2	8	8	8
Mathematics Claim 3	8	8	8

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

Content Area and Claim	Grades 3–5	Grade 6	Grade 7	Grade 8	Grade 11
ELA Claim 1	14	14	14	16	15
ELA Claim 2	10	10	10	10	10
ELA Claim 3	8	8	8	8	8
ELA Claim 4	9	9	9	9	9
Mathematics Claim 1	20	19	20	20	22
Mathematics Claim 2	8	8	8	8	8
Mathematics Claim 3	8	8	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. Table 8.C.1 and Table 8.C.2 in appendix 8.C show, for each test and for each claim, the numbers of items in five intervals of exposure, with the lowest being 1 to 100 student testing events and the highest being greater than or equal to 3,000 student testing events. These tables also show how many items were not administered.

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. Table 8.C.3 through Table 8.C.16 present the same information as Table 8.C.1 and Table 8.C.2, computed separately for items in several intervals of difficulty.

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.

This subsection documents the reliability and SEM statistics that are used for the CAASPP Smarter Balanced Summative Assessments.

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 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 HOSS for that test. When a student is assigned either the LOSS or HOSS, a measure of his or her 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:

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

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 CSEM (i.e., error variances) at each value of the ability estimate.

8.5.3. Standard Error of Measurement (SEM)

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:

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

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).

Table 8.3 gives the total score reliability for theta as well as the mean, standard deviation, and SEM of both thetas and scale scores for each of the 14 tests, along with the number of student results upon which those analyses were performed. Note that in the case of the total test reliability, the reliability is for the whole test on the theta score scale; it is calculated using the total test theta scale score of individual students.

In Table 8.3, only students who finished at least 10 CAT items and 1 PT item are included in the analysis.

Table 8.3 Summary Statistics for Scale Scores and Theta Scores, Reliability, and SEMs

Content Area/Grade	Number of Students	Reliability	Scale Score Mean	Scale Score SD	Scale Score SEM	Theta Score Mean	Theta Score SD	Theta Score SEM
ELA 3	429,403	0.93	2422	91	24.43	-1.00	1.05	0.28
ELA 4	444,957	0.92	2460	96	26.82	-0.56	1.12	0.31
ELA 5	450,541	0.93	2493	98	25.89	-0.18	1.14	0.30
ELA 6	465,310	0.93	2516	98	26.06	0.09	1.14	0.30
ELA 7	452,366	0.93	2541	100	27.33	0.38	1.17	0.32
ELA 8	450,172	0.93	2557	100	27.27	0.56	1.17	0.32
ELA 11	420,726	0.92	2587	112	31.04	0.91	1.31	0.36
Mathematics 3	428,570	0.95	2429	82	18.78	-1.08	1.03	0.24
Mathematics 4	446,767	0.94	2465	84	19.74	-0.62	1.05	0.25
Mathematics 5	453,641	0.94	2488	92	23.43	-0.34	1.17	0.30
Mathematics 6	461,950	0.94	2511	104	25.51	-0.05	1.31	0.32
Mathematics 7	452,386	0.94	2524	112	28.40	0.11	1.41	0.36
Mathematics 8	442,651	0.93	2537	119	31.33	0.28	1.50	0.40
Mathematics11	427,941	0.93	2561	124	33.95	0.59	1.57	0.43

8.5.4. Intercorrelations, Reliabilities, and SEMs for Claims Scores

For each test, theta scores and scale scores are computed for claims. As is described in subsection *7.1.1 Structure of the Assessments* in *Chapter 7: Scoring and Reporting*, claims identify the set of knowledge and skills being measured. Claim scores are scores on the set of items that form the basis for a claim.

Intercorrelations, reliability estimates, and theta-based SEMs for the claims are presented in Table 8.D.1 through Table 8.D.14 in appendix 8.D. The reliability estimates vary significantly across claims according to both the number of items and the types of content standards that are included in each claim. The standards of claims can be found in the Smarter Balanced blueprints that are provided in appendix 2.A.

8.5.5. Student Group Reliabilities and SEMs

The reliabilities of the total test scores and the claim scores are examined for various student groups within the student population. The student groups included in these analyses are defined by gender, economic status, special education services status, accommodations for students with special education services, English language fluency, primary ethnicity, and migrant status. The reliability analyses are also presented by primary ethnicity within economic status.

Reliabilities and theta-based SEMs for the total test scores and the claim scores are reported for each student group analysis. Table 8.D.15 through Table 8.D.24 in appendix 8.D present the overall test reliabilities for student groups defined by student gender, economic status, special education services status, English language fluency, primary ethnicity, and migrant status. Table 8.D.25 through Table 8.D.30 present the reliabilities for the student groups based on primary ethnicity within economic status.

The next set of tables, Table 8.D.31 through Table 8.D.100, present the claim-level reliabilities for the student groups. Table 8.D.31 through Table 8.D.44 present the claim-level reliabilities for the student groups based on gender, economic status, and migrant status. Table 8.D.45 through Table 8.D.58 show the same analyses for the student groups based on special education services status and English language fluency. Table 8.D.59 through Table 8.D.72 present results for the student groups based on primary ethnicity of the students. The last set of tables, Table 8.D.73 through Table 8.D.100, present the claim-level reliabilities for the student groups based on primary ethnicity within economic status.

Note that the reliabilities are reported only for samples that are comprised of 11 or more students. Also, in some cases, score reliabilities are not estimable and are presented in the tables as “NA.” The reliability estimates for some of the student groups are negative due to small variation in scale scores and large CSEMs for extreme score values. These negative reliabilities and their associated SEMs also are presented as “NA.”

8.5.6. Conditional Standard Errors of Measurement (CSEMs)

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):

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

where,

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

CSEM (SS) is the conditional standard error of measurement on scale score scale, and

$I(\theta)$ is the test information function 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 2014–15 Smarter Balanced Summative Assessment data for all students (Smarter Balanced, 2016c). Given the stability across the 2014–15 through 2017–18 California student populations and the stability of the item pool, 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 cut scores, it is important to estimate CSEMs at those cut scores. Table 8.4 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 8.4 Scale Score CSEM at Performance-level Cut Points

Content Area/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	2432	22	2490	23
ELA 4	2416	25	2473	24	2533	25
ELA 5	2442	24	2502	24	2582	25
ELA 6	2457	27	2531	25	2618	26
ELA 7	2479	27	2552	26	2649	26
ELA 8	2487	27	2567	26	2668	27
ELA 11	2493	32	2583	29	2682	28

Content Area/Grade	Standard Nearly Met Minimum SS	Standard Nearly Met CSEM	Standard Met Minimum SS	Standard Met CSEM	Standard Exceeded Minimum SS	Standard Exceeded CSEM
Mathematics 3	2381	19	2436	17	2501	17
Mathematics 4	2411	20	2485	17	2549	17
Mathematics 5	2455	23	2528	19	2579	18
Mathematics 6	2473	25	2552	21	2610	20
Mathematics 7	2484	30	2567	23	2635	20
Mathematics 8	2504	32	2586	26	2653	22
Mathematics 11	2543	35	2628	27	2718	22

Table 8.5 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 8.5 Mean CSEM for Each Achievement Level

Content Area/Grade	Standard Not Met	Standard Nearly Met	Standard Met	Standard Exceeded
ELA 3	27.84	22.77	22.00	23.74
ELA 4	28.88	25.00	24.47	25.89
ELA 5	28.07	24.00	24.59	26.49
ELA 6	31.99	26.01	25.45	27.32
ELA 7	31.91	26.15	25.62	28.13
ELA 8	30.88	26.28	26.07	27.89
ELA 11	36.86	30.05	28.06	29.97
Mathematics 3	22.37	17.95	17.00	17.81
Mathematics 4	23.64	18.11	17.00	17.68
Mathematics 5	29.41	20.81	18.23	17.99
Mathematics 6	33.62	23.10	20.58	20.91
Mathematics 7	41.07	26.23	21.41	20.67
Mathematics 8	40.56	29.02	24.04	21.97
Mathematics 11	49.00	30.97	24.79	22.52

Scale score CSEM distributions are shown in Table 8.E.1 through Table 8.E.14 in appendix 8.E. The plots of the CSEMs conditional for scale scores are also presented in Figure 8.E.1 through Figure 8.E.14. In the figures, the vertical axis is defined as the CSEMs and the horizontal axis is designated as scale scores, which is a common metric for tests within the same content area. Each data point represents an individual student. Typically, for fixed-form tests, the pattern of the CSEMs tends to be U–shaped, such that the plotted values of CSEMs for the middle scale scores tend to be lower than those for extreme scale scores. Table 8.4 and Table 8.5 and Figure 8.E.1 through Figure 8.E.14 in appendix 8.E show CSEMs are smallest in the upper-middle portion of the score range, slightly larger for high scores, and much larger for low scores, getting larger as the score gets lower. This is partially due to the impact of the CAT and vertical scales, which, in relation to a fixed-form test, is the attenuation of the U–shaped relationship between CSEMs and scale scores.

8.5.7. 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).

The results of these analyses are presented in Table 8.F.1 through Table 8.F.28 in appendix 8.F. Included are the contingency tables for both accuracy and consistency of the various achievement-level classifications. The proportion of students accurately classified is determined by summing the main diagonal of the upper table. The proportion of students consistently classified is determined by summing the main diagonal of the lower table. The classifications are collapsed to *Standard Not Met* and *Standard Nearly Met* versus *Standard Met* and *Standard Exceeded*, which are the critical categories for accountability. In each case, the estimated proportion of classifications with exact agreement is the sum of the entries in the diagonal of the contingency table of the achievement level placements.

Reliability of classification at a cut score is estimated by combining the achievement levels above a particular cut score and combining the achievement levels below that cut. The result is a two-by-two table indicating whether the students are above or below the cut score. The sum of the entries in the main diagonal is the number of students accurately (or consistently) classified as above or below that cut score. Figure 8.1 and Figure 8.2 illustrate these 2 × 2 contingency tables.

True status on all-forms average	**Does not reach an achievement level**	**Reaches an achievement level**
Does not reach an achievement level	Correct classification	Misclassification
Reaches an achievement level	Misclassification	Correct classification

Figure 8.1 Decision Accuracy for Reaching an Achievement Level

Decision made on the form taken	**Does not reach an achievement level**	**Reaches an achievement level**
Does not reach an achievement level	Correct classification	Misclassification
Reaches an achievement level	Misclassification	Correct classification

Figure 8.2 Decision Consistency for Reaching an Achievement Level

8.5.8. Interrater Agreement

To monitor the consistency of ratings assigned to students’ responses by raters, approximately 10 percent of the CRs 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 coefficient

8.5.8.1 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 increasing numbers of item score points. The fewer the item score points, the fewer degrees of freedom on which two raters can vary, and the higher the percentage of agreement.

8.5.8.2 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:

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

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 percentages of ratings on which the raters are in exact agreement or differ by just one point are computed.

8.5.8.3 Quadratic-Weighted Kappa

Quadratic-weighted kappa 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 quadratic weighted kappa 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:

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

For quadratic-weighted kappa, the weights are:

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

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.

Table 8.G.1 through Table 8.G.14 in appendix 8.G present the results of the interrater analyses and descriptive statistics of the ratings by the two raters on short-answer items, including the following:

- Number of score points in each item
- Number of raters for each round of rating
- Kappa
- Quadratic-weighted kappa
- Percent of exact agreement
- Percent of adjacent agreement
- Mean of the item score
- Standard deviation of the item score

Table 8.G.15 through Table 8.G.20 present the results of the interrater analyses on writing extended-response (WER) items. The number of items that did not meet the interrater agreement standards by Smarter Balanced were flagged and presented in Table 7.6. In addition to the statistics described previously, the dimension name is also identified.

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.

8.5.9. Agreement Between AI 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*.

Two sets of ratings for the same item, one set from the AI scoring engine and the other set from human raters, are evaluated and compared. Table 8.G.21 through Table 8.G.34 in appendix 8.G present the agreement statistics between AI and human scoring for short answer items for ELA and mathematics. Table 8.G.35 through Table 8.G.37 present the agreement statistics between AI and human scoring for WER items. The dimension name is identified in the case of WER items. These tables include the following:

- Number of score points in each item
- Number of raters for each round of rating
- Kappa
- Quadratic-weighted kappa
- Percent of exact agreement
- Percent of adjacent agreement

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 subsection, 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, 2016c). 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) in order 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. The CDE also publishes *The Guide to Your CAASPP Student Score Report* for parents/guardians of students in grades three (CDE, 2018a); four, six, and seven (CDE, 2018b); eleven (CDE, 2018c); and five and eight (CDE, 2018d). The guides are published in English and Spanish.

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 http://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. Validity 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, 2014). 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 subsection *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 2017h) and mathematics (Smarter Balanced, 2018a through 2018k).

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 these measures:

1. a measure of content match to the blueprint,
2. a measure of overall test information, and
3. measures of test information for each reporting category on the test.

Details can be found in 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.

The degree to which test forms administered in 2014–15 met the blueprint is provided in *Chapter 5: Test Administration* and in Table 5.B.4 in appendix 5.B.

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, 2016b).

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, 2014).

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 2016–17 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. Validity 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—one percent of all the students taking the test—and the students with the shortest testing time in the CAT portion—also one 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, 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 one 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.

Table 8.H.1 and Table 8.H.2 in appendix 8.H provide descriptive statistics for ELA and mathematics testing time, respectively. These tables include total testing time and percentile information at each ability level. The unit of testing time is minutes; for example, in Table 8.H.1, the median of the testing time for the ELA grade three Q1 group is 165 minutes. At every grade level, in both ELA and mathematics, students in the lowest ability level (1st quartile, Q1) have shorter median testing times than students in the other groups. The median of total testing time generally increases with ability level from Q1 to Q4. Students at the 50th percentile within each ability quartile spent 103 to 258 minutes on ELA assessments across all grades and 63 to 164 minutes on mathematics assessments across all grades.

Table 8.H.3 (for ELA) and Table 8.H.4 (for mathematics) provide the descriptive statistics of testing time for the CAT portion and the percentile information at each ability level. The number of CAT items presented to each student is reported in Table 5.B.2 in appendix 5.B. Similar to total testing time, the median of testing time in the CAT portion generally increases with ability level from Q1 to Q4 in mathematics. For ELA, median testing times also increase with ability level, although there are no substantial differences in testing times between the Q3 and Q4 groups for ELA. Students at the 50th percentile within each ability quartile spent 60 to 132 minutes on the CAT portion of ELA assessments across all grades and 46 to 119 minutes on the CAT portion of mathematics tests across all grades.

After testing time distributions for CAT were reviewed, testing times for the PTs were investigated. During testing, each student was presented with a few items (one to six) that were randomly assigned in each grade. (More details on assignment of PTs can be found in *Chapter 5: Test Administration*.) Table 8.H.5 and Table 8.H.6 in appendix 8.H provide the descriptive statistics for ELA and mathematics testing times for each PT and the percentile information at each ability level, respectively. Overall, students in the lowest ability level (1st quartile, Q1) have shorter testing times than students in the other groups. For ELA, the median of the PT testing time increases with ability level from Q1 to Q4. Students at the 50th percentile within each ability quartile spent 37 to 130 minutes on the PT portion of ELA assessments across PTs and all grades and 14 to 68 minutes on the PT portion of mathematics tests across PTs and all grades. For mathematics, there are no significant differences in PT testing time from Q2 to Q4 groups.

For the CAT administrations, results are consistent with past studies suggesting that testing time for items increases with more difficult items (van der Linden, 2009).

8.6.4. Validity 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, 2016b).

8.6.4.2 Differential Item Functioning (DIF)

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, 2016b, chapter 6; and Smarter Balanced, 2016c, 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 in Table 8.3. The results indicate that the reliability estimates for all summative test total scores are high, ranging from 0.92 to 0.94. Theta score standard deviations and SEMs are increasing 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 presented in Table 8.D.1 through Table 8.D.14 in appendix 8.D. The reliability estimates of claims are invariably lower than those for the total tests because they are 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 also are examined for various student groups within the student population that differ in their demographic characteristics. The characteristics considered are gender, ethnicity, economic status, special education services status, migrant status, English language fluency, and ethnicity-by-economic status (refer to Table 7.20 for the demographic groups reported). Reliability estimates and SEM information for the total test theta scores and the claim theta scores are reported for each student group. Table 8.D.15 through Table 8.D.30 in appendix 8.D present the reliabilities and SEMs on the overall test theta scores for the various student groups. Table 8.D.31 through Table 8.D.100 present the reliabilities and SEMs of theta scores for the claims.

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 on page 102. The results of these analyses are presented in Table 8.F.1 through Table 8.F.28 in appendix 8.F. When the classifications are collapsed to below *Standard Met* versus *Standard Met* and above, which are the critical categories for accountability analyses, the estimated proportion of students who are classified accurately ranges from 0.70 to 0.81 across all tests. Similarly, the estimated proportion of students who are classified consistently ranges from 0.88 to 0.93 for students classified into below *Standard Met* versus *Standard Met* and above. These are considered high levels of accuracy and consistency.

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 quadratic-weighted kappa greater than 0.70 indicate excellent agreement (Williamson, Xi, & Breyer, 2012).

The results in Table 8.G.1 through Table 8.G.14 in appendix 8.G show at least moderate levels of agreement between raters who scored students’ responses for 47 percent of the human-scored, short-answer items in ELA and 23 percent of the human-scored items in mathematics. The rater agreement is at least high, with kappa over 0.60 for 8 percent of the ELA human-scored items and 75 percent of the mathematics human-scored items. The rater agreement is excellent, with the quadratic-weighted kappa over 0.7 for 12 percent of the ELA human-scored items and 80 percent of the mathematics human-scored items.

The results in Table 8.G.15 through Table 8.G.20 show at least moderate levels of agreement between raters who scored students' responses for 42 percent of the human-scored WER items, and high levels of agreement for 4 percent of the human-scored WER items in ELA for grades three through eight. The rater agreement is excellent, with the quadratic-weighted kappa over 0.7 for 23 percent of the human-scored WER items.

Table 8.G.21 through Table 8.G.34 present the results for AI machine-scored items for ELA and mathematics. The results show at least moderate levels of agreement between human raters and AI engines that scored students' responses for 70 percent of the AI machine-scored short-answer items in ELA and 39 percent of the AI machine-scored short-answer items in mathematics. The agreement is high, with Kappa over 0.6 for 10 percent of ELA AI machine-scored short-answer items and 55 percent of mathematics AI machine-scored short-answer items. The rater agreement is excellent, with the quadratic-weighted Kappa over 0.7 for 37 percent of the ELA and 73 percent of the mathematics AI machine-scored items.

Table 8.G.35 through Table 8.G.37 presents the results for AI machine-scored WER items for ELA in grades three, six, and eleven. The results show at least moderate levels of agreement between human raters and AI engines for 47 percent of the AI machine-scored WER items. The rater agreement is excellent, with the quadratic-weighted kappa over 0.7 for 46 percent of the AI machine-scored WER items.

8.6.4.8 Interrater Agreement

As shown in Table 8.G.1 through Table 8.G.14 in appendix 8.G, all human-scored items in ELA assessments can be awarded a maximum of two points (0, 1, or 2) for short-text items and a maximum of four points for WER items. In mathematics, human-scored items can be awarded between one (0, 1) and four (0, 1, 2, 3, 4) points. Approximately 10 percent of the test population's responses to the human-scored items are scored by two raters. The percentage of students for whom the raters are in exact agreement ranges from 52 percent to 90 percent for ELA and 63 percent to 100 percent for mathematics. The percentage of students for whom the raters are in exact or adjacent agreement ranges from 92 percent to 100 percent for ELA and 95 percent to 100 percent for mathematics.

As is reported in Table 8.G.15 through Table 8.G.20, WER items have two points for convention dimension and four points for Organization and Purpose, Development and Elaboration, or Evidence and Elaboration scoring dimensions. The percentage of students for whom the raters are in exact agreement ranges from 51 percent to 86 percent; the percentage of students for whom the raters are in exact or adjacent agreement ranges from 94 percent to 100 percent in ELA tests for grades three through eight.

As presented in Table 8.G.21 through Table 8.G.34, 10 percent of the students' responses that are scored by the AI engine are also scored by human raters. The percentages of students for whom the AI engine and human raters are in exact agreement range from 46 percent to 92 percent for ELA across the grades and from 48 percent to 99 percent for mathematics across the grades. The percentages of students for whom the AI engine and human raters are in exact or adjacent agreement range from 85 percent to 100 percent for the ELA tests and 84 percent to 100 percent for the mathematics tests.

Table 8.G.35 through Table 8.G.37 present the interrater agreement between the AI engine and human raters for ELA WER items in grades three, six, and eleven; only these three tests contain AI-scored WER items. The percentages of students for whom the AI engine and human raters are in exact agreement range from 43 percent to 69 percent. The percentages of students for whom the AI engine and human raters are in exact or adjacent agreement range from 93 percent to 99 percent.

8.6.4.9 Correlations Between the Claims Within Content Areas

The distinctiveness and reliability of the claim theta scores in each content area are important because CAASPP strength and weakness levels are reported based on claim scores. The interrelationships of claim scores should be shown to be consistent with the construct being assessed. Table 8.D.1 through Table 8.D.14 in appendix 8.D provide the intercorrelations between claim scores within each test in the two content areas (i.e., ELA and mathematics). Results show that the correlations between claim scores are consistent across the grades. Correlations range from 0.60 to 0.78 for ELA and from 0.68 to 0.84 for mathematics.

8.6.4.10 Correlations Between Content Area Test Scores

The degree to which students' content area test scores correlate as expected provides evidence of those scores as measures of the intended constructs. Table 8.6 provides the correlations between scores on the 2017–18 CAASPP ELA and mathematics assessments and the numbers of students on which these correlations are based. Sample sizes for individual assessments are shown in bold and indicated with an asterisk; the numbers of students on which the correlations are based are shown on the lower left without bolding. The correlations are provided in the upper right. Results are based on all students with valid scale scores and are provided by grade.

Table 8.6 Correlations for All Students

Content Area and Grade	Sample Size	R and Sample Size
ELA 3	***433,781**	0.82
Mathematics 3	432,870	***435,795**
ELA 4	***453,086**	0.81
Mathematics 4	452,157	***454,948**
ELA 5	***458,920**	0.81
Mathematics 5	457,870	***460,475**
ELA 6	***472,014**	0.84
Mathematics 6	470,599	***473,345**
ELA 7	***461,023**	0.82
Mathematics 7	459,303	***462,367**
ELA 8	***458,142**	0.82
Mathematics 8	455,941	***458,629**
ELA 11	***439,094**	0.79
Mathematics 11	433,609	***437,861**

Notes:

- Sample sizes of the individual assessments are in **bold** font and indicated with an asterisk.
- Numbers that are not in bold font are the sample sizes to calculate the correlations.

- R denotes the correlation coefficient; these are decimals that begin with “0” (zero).

Results for these students appear to be consistent with expectations. In general, students’ ELA scores correlated moderately with their mathematics scores. They are correlated slightly more highly among students in grades three through eight than in grade eleven.

Table 8.I.1 through Table 8.I.8 in appendix 8.I provide the content area test score correlations by gender, ethnicity, English language fluency, economic status, migrant status, and special education services status. The correlation between students’ ELA and mathematics scores was approximately .80 at all grade levels, for nearly all the student groups. One exception was English learners, who showed lower correlations at all grades.

Note that the correlations are reported only for groups of more than 10 students. Correlations between scores on any two content area tests where 10 or fewer students took the tests are expressed as “NA.”

8.6.5. Validity 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, 2016b). 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 (CSU) Early Assessment Program (EAP). During the 2013–14 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. Validity 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,

> “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.

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. Retrieved from 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. Retrieved from 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. Retrieved from 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. (2018c). *Understanding your student score report, grade eleven.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.Smarter-Balanced-SSR-guides-English-gr11.2017-18.pdf

California Department of Education. (2018a). *Understanding your student score report, grade three.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.Smarter-Balanced-SSR-guides-English-gr3.2017-18.pdf

California Department of Education. (2018d). *Understanding your student score report, grades five and eight.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.Smarter-Balanced-SSR-guides-English-gr5-8.2017-18.pdf

California Department of Education. (2018b). *Understanding your student score report, grades four, six, and seven.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.Smarter-Balanced-SSR-guides-English-gr4-6-7.2017-18.pdf

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. (2014). *ETS standards for quality and fairness.* Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2011). KE (Version 3) [Software]. 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. Retrieved from http://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. Retrieved from http://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. (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. Retrieved from https://www.smarterbalanced.org/wp-content/uploads/2015/08/ELA_Content_Specs.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. Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/08/Mathematics-Content-Specifications.pdf

Smarter Balanced Assessment Consortium. (2016a). *Smarter Balanced Assessment Consortium: Alignment study report.* Alexandria, VA: Human Resource Research Organization. Retrieved from http://www.smarterapp.org/documents/AlignmentStudyReport.pdf

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

Smarter Balanced Assessment Consortium. (2016c). *Smarter Balanced Assessment Consortium: 2014–15 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf/

Smarter Balanced Assessment Consortium. (2017a). *ELA CAT item specifications, grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017b). *ELA CAT item specifications, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017c). *ELA CAT item specifications, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017a). *ELA PT item specifications, argumentative, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017d). *ELA PT item specifications, explanatory, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017e). *ELA PT item specifications, informative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017f). *ELA PT item specifications; narrative, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017g). *ELA PT item specifications, narrative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017h). *ELA PT item specifications, opinion, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2017i). *Smarter Balanced Assessment Consortium: 2015–16 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/2015-16-summative-technical-report.pdf

Smarter Balanced Assessment Consortium. (2018a). *Mathematics computer adaptive test (CAT) item specifications, Claim 1, grade eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018b). *Mathematics CAT item specifications, Claim 1, grade five*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018c). *Mathematics CAT item specifications, Claim 1, grade four*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018d). *Mathematics CAT item specifications, Claim 1, grade six*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018e). *Mathematics CAT item specifications, Claim 1, grade seven*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018f). *Mathematics CAT item specifications, Claim 1, grade three*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018g). *Mathematics CAT item specifications, Claim 1, high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/assessments/development/

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. Retrieved from http://www.smarterbalanced.org/assessments/development/

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. Retrieved from http://www.smarterbalanced.org/assessments/development/

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. Retrieved from http://www.smarterbalanced.org/assessments/development/

Smarter Balanced Assessment Consortium. (2018k). *Smarter Balanced Assessment Consortium: 2016–17 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://portal.smarterbalanced.org/library/en/2016-17-summative-assessment-technical-report.pdf

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. Retrieved from https://portal.smarterbalanced.org/library/en/scoring-guide-for-ela-full-writes.pdf

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.

Chapter 9: Quality Control Procedures

The California Department of Education (CDE), Smarter Balanced Assessment Consortium, and Educational Testing Service (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 evaluate 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 for accessibility, bias and sensitivity, as well as content prior to their administration in the 2013–14 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 performance tasks, and made recommendations for item development and item-quality criteria.

After the 2012–13 pilot test, staff from the Smarter Balanced Assessment Consortium used statistical criteria to flag items that were potentially problematic due to 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 simulations studies were conducted:

1. the simulation study conducted prior to the 2013–14 Smarter Balanced field test that is described in chapter 4 of the *2013–14 Technical Smarter Balanced Report* (Smarter Balanced, 2016); and
2. the simulation study conducted prior to the 2016–17 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 subsection *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 in addition to a place for a student barcode label.

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 California Assessment of Student Performance and Progress (CAASPP) System specifications. Each document must be approved by the CDE before it can be published to the CAASPP Portal at http://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 2017–18 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 scorable and nonscorable materials within five working days after the last day of each test administration period. The freight-return kits provided to LEAs contain color-coded labels identifying scorable and nonscorable materials and labels with bar-code information identifying the school and district. The LEAs packed all materials into cartons, applied the appropriate labels, and then numbered the cartons prior to returning the materials to the processing center by means of their assigned carrier. The use of the color-coded labels streamlines the return process.

9.3.3. Processing Test Materials

9.3.3.1 Online Assessments

Online tests that were submitted by students were transmitted from the American Institutes for Research (AIR) 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 performance task [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 physical damage, shipping errors, and omission of materials. The number of students recorded on the student 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; AIR, 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 system (eSKM) 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 (eSKM) Processing

ETS developed two independent and parallel scoring structures to produce students' scores: the eSKM[8] 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' responses. The two scoring systems independently apply the same scoring algorithms and specifications. ETS psychometricians verify 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 are reported only when the two parallel systems produce identical results.

When scores do not match, the mismatch is 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 applies the problem item notification (PIN) not to score the item through the systematic process in eSKM and a mismatch is possible, if TIPS is still in the process of applying the PIN in the parallel system when the student score is 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 collect and parse .xml files that contain student response data from AIR and send constructed-response (CR) item responses to ETS and MI for human scoring. After receiving the results of human scoring, eSKM merges student scores from the CAT and PT test components, calculates individual student scores, and generates student scores in the approved statistical extract format on a daily basis. These data extracts are sent to ETS' Data Quality Services for data validation. Following validation, the student response statistical extracts are made available to the psychometricians.

9.4.2.2 Psychometric Processing

Psychometricians verify 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 check variables such as total scale scores, achievement levels, number of scored items, and performance levels of claims. To investigate discrepancies, theta scores and completeness are also checked; refer to *7.4 Student Test Scores* for definitions of these scores. Refer also to subsection *12.4 Psychometric Analysis* for more information on psychometric quality control.

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 before student reports are generated. In addition to parallel scoring for both online and paper-pencil assessments, ETS provides verification of answer keys and item analysis for paper-pencil assessments.

[8] The eSKM system produces the ETS scores of record.

9.5. Quality Control of Constructed-Response (CR) 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 are led through a training period by trained assessment 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 are trained to appropriately apply the rubrics by using the Smarter Balanced–provided benchmark sample papers.

Trained raters are scheduled to score in four or eight hour shifts. Prior to starting a shift, a rater must take and pass a calibration test that demonstrates sufficient training in Smarter Balanced scoring criteria and ability to score accurately.

Scoring leaders are qualified raters who have the responsibility of providing feedback to raters in order to provide additional content support and offer corrective mentoring for struggling raters.

Each rater is assigned a secure user ID and password to log on to the scoring system and is required to sign a confidentiality agreement. System access for the rater is restricted to the hours that he or she is scheduled to work.

9.5.2. Hand Scoring Verification

9.5.2.1 Criteria for Read-Behinds

Ten percent of responses are scored twice (i.e., “read behind”) to check agreement among raters, although the percentage can vary, depending on item type and reader performance. Scoring leaders read behind raters throughout a shift and enter their own scores on responses that raters have read. Both first and second readings are eligible for read-behind. Results of interrater reliability are shown in appendix 8.G.

A scoring leader reviews the randomly selected responses after raters submit scores. Leaders review 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 may be rescored by leadership and used to inform and instruct raters as a performance-improvement strategy.

When a response is selected for a second reading, the corrected score is used for interrater reliability calculation. The original rater’s score is not be used for any calculation.

9.5.2.2 Validity Responses

Validity responses are provided randomly as part of the set of “live” responses being scored, so a rater does not know that the response being scored is for validity. These responses are selected from “live” responses by scoring leaders prior to the scoring of the item. Leadership staff identifies the response to be used for validity and the system adds the response to the validity pool for use during scoring.

All staffing levels are eligible to score second readings. Ten percent of responses are assigned to be read a second time. Second readings are scored independently from the first reading.

Only scorable responses are selected for second readings. Nonscorable (i.e., condition code) responses are not eligible for second readings and so are not included in the calculation of interrater reliability.

The second reading sample is not a stratified random sample. The selection of a second reading response is also not based on the first reading score or any demographic information associated with the response. Instead, responses flagged for second reading are flagged at random by the scoring system for each item identification number.

Second reading scores are used only for statistical analysis to obtain interrater reliability. They are not included in the calculation of the final item score.

9.5.3. AI Scoring Verification

To ensure the quality of machine scoring with artificial intelligence (AI), ETS and MI maintain a quality assurance system where 10 percent of AI-scored items being scored by a human rater and used for agreement sample analysis. The results of the agreement analysis are presented in section *8.6.4.8 Interrater Agreement*. Also, refer to subsection *7.2 Quality Control of Scoring* and subsection *12.3 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, the completed student answer documents were routed for scoring. Quality control of paper-pencil tests is ensured by an independent group that signs into eSKM and checks scoring keys. This group must sign off and approve the keys in order for scoring to commence for the administration. This team also creates scoring stencils to be used during the administration to overlay on top of a student's answer document to verify the score computed by eSKM is accurate.

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, four 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 through ETS' patented QC Integrator software;
3. Evaluation of the production of all Student Score Reports reports—available in paper and electronic versions—by verifying the print quality, comparing number of report copies, sequence of report order, and offset characteristics to the CDE requirements; and
4. Proofreading of the pilot and production reports by the CDE and ETS prior to any LEA mailings or file availability.

All reports are required to include a single, accurate LEA code, a charter school number (if applicable), a school district name, and a school name. All elements conform 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 is used to verify and confirm accurate codes and names. CDE provides a revised LEA Master File to ETS throughout the year as updates become available.

After the reports are validated against the CDE's requirements, a set of reports for pilot LEAs are provided to the CDE and ETS for review and approval. Paper reports are sent on the actual report forms, organized as they are expected to look in production. The CDE and ETS review and approve the report package after a thorough examination.

Upon the CDE's approval of the reports generated for the pilot districts, ETS proceeds with the first batch of report production. The first production batch is selected to validate a subset of LEAs that contain key reporting characteristics (e.g., academic achievement) and demographics of the state. The first production batch incorporates CDE-selected LEAs and provides the final check prior to generating all reports and mailing them to the LEAs as well as making them available for the LEA to 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 include the logic for handling submitted assessments and Answer Books that, for example, indicate the student tested but responded to no items, was absent, was not tested due to parent/guardian request, or did not complete the assessment due to 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 conducts 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 are 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 is allotted for user acceptance testing to confirm that the systems meet requirements and to make identified corrections before final deployment. To accomplish system acceptance and sign-off, ETS deploys 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

To begin the quality control process for paper-pencil test administration, the ETS resolutions team members complete response documents by marking responses on response booklets for fictitious students in selected schools and across several LEAs. They mark response booklets with answers that are all correct, all incorrect, and other test response combinations. These response combinations are the expected results across achievement levels and score ranges. The response booklets are sent for processing, batching, and data entry. Once released from scanning, the test results are sent through the system for scoring and reporting. Student Score Reports are created along with data files for subject matter experts in the teams to review and verify.

Individual student score reports were generated based on the fictitious students and 100 percent quality control was demonstrated by ETS' Resolution staff.

References

American Institutes for Research. (2015). *Smarter Balanced scoring specification: 2014–2015 administration, version 7*. Retrieved from http://www.smarterapp.org/documents/TestScoringSpecs2014-2015.pdf

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

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

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

Chapter 10: Historical Comparisons

Historical comparisons are performed to identify the trends in student performance and test characteristics over time. Such comparisons were performed for the three most recent administration years of California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced test administration—2017–18, 2016–17, and 2015–16. The comparisons include both cross-sectional comparisons for the same grades in different years (with different students) and longitudinal comparisons for the same students in different years (in different grades).

The indicators of student performance include the mean and standard deviation of scale scores and the percentage of students classified into achievement levels for an overall test and into performance levels for claims. Test characteristics are compared by examining the reliability and standard error of measurement (SEM) for each test.

10.1. Student Performances

10.1.1. Cross-Sectional Comparisons on the Overall Tests

In cross-sectional comparisons, cohorts of students from the 2015–16 CAASPP administration are compared to students in the same grades from the 2016–17 and 2017–18 CAASPP administrations. For example, students enrolled in grade three for the 2015–16 CAASPP administration are compared with students enrolled in grade three for the 2016–17 and 2017–18 CAASPP administrations.

As noted in Table 7.10 in *Chapter 7: Scoring and Reporting*, the reporting scale ranges from 2114 to 2795 for English language arts/literacy (ELA) and from 2189 to 2862 for mathematics. The difference between the two adjacent years in average scale score and percentage of students meeting or exceeding standards is the later year's values minus the previous year's values for the same grade. For example, in comparing the values from the 2016–2017 and 2017–2018 administrations, a positive value indicates an increase from 2017–17 to 2017–18 and a negative value indicates a decrease. The achievement level percentages may not sum to exactly 100 or to exactly the combined achievement level percentage due to rounding.

10.1.1.1 Summary Statistics

Table 10.A.1 in appendix 10.A contains the number of students assessed, the number of students with valid scores, the means and standard deviations of students' scale scores in 2015–16, 2016–17, and 2017–18 for each test, as well as the differences in scale scores between 2015–16 and 2016–17 and between 2017–18 and 2016-2017.

10.1.1.2 Achievement Levels of Overall Students

Scale cut scores are used to classify each student into one of four achievement levels: *Standard Not Me*t, *Standard Nearly Met, Standard Met,* or *Standard Exceeded*. Refer to Table 7.12 in *Chapter 7: Scoring and Reporting* for the achievement level scale score ranges for each test. The percentages of students for each achievement level and qualifying for the *Standard Met* and *Standard Exceeded* levels, as well as the differences in the percentages of the students in *Standard Met* and *Standard Exceeded* between 2015–16 and 2016–17 and between 2016–17 and 2017–18, are presented in Table 10.A.2 in appendix 10.A. Note that this information may differ slightly from information found on the California Department of Education (CDE) CAASPP Results website at http://caaspp.cde.ca.gov/ due to different dates on which the data was accessed.

10.1.1.3 Scale Score Distributions

Table 10.A.3 through Table 10.A.6 in appendix 10.A show the distribution of scale scores observed in 2015–16, 2016–17, and 2017–18 for each grade and content area. Frequency counts are provided for each scale score interval of 30. "NA" indicates that there is no obtainable scale score in the interval. The scale score ranges for each grade on the vertical scale are those defined by the Smarter Balanced Assessment Consortium. Refer to Table 7.10 in chapter 7 for the scale score ranges.

10.1.1.4 Achievement Levels of Selected Student Groups

Table 10.A.7 through Table 10.A.20 in appendix 10.A provide statistics summarizing student achievement by content area and grade for selected student groups. In the tables, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status (disadvantaged or not), need for special education services, migrant status, the use of designated supports (using designated supports or not), and the use of accommodations (using accommodations or not). The tables show, for each demographic student grouping, the numbers of students with a valid scale score, scale score means and standard deviations, and the percentages of students in each achievement level, for 2015–16, 2016–17, and 2017–18, as well as the differences in the percentages of *Standard Met* or *Standard Exceeded* between 2015–16 and 2016–17, and between 2016–17 and 2017–18.

10.1.2. Cross-Sectional Comparisons on Claims

10.1.2.1 Summary Statistics

Table 10.B.1 through Table 10.B.4 in appendix 10.B contain the number of students assessed, the number of students with valid scores, the means and standard deviations of students' scale scores in 2015–16, 2016–17, and 2017–18 on each claim by grade and content area. Also presented are the year-to-year differences in average scale scores. The claim scores are on the same scale as the scores on the total test in which the claims are included. Refer to the score ranges of the total test (Table 7.10) for the score ranges of claims.

10.1.2.2 Performance Levels of Overall Students

Table 10.B.5 through Table 10.B.8 in appendix 10.B present the percentages of students in each performance level of each claim in 2015–16, 2016–17, and 2017–18. Student results on each claim are classified into three performance levels: *Below Standard*, *Near Standard*, and *Above Standard*. The year-to-year differences in the percentages of students *Near Standard* or *Above Standard* are also presented. Refer to Table 7.12 in chapter 7 for the achievement level scale score ranges for each test. Refer to subsection *7.4.2.2 Performance Levels for Claims* for the details regarding the classification of performance levels on claims.

10.1.2.3 Performance Levels of Selected Student Groups

Table 10.B.9 through Table 10.B.57 in appendix 10.B show the statistics summarizing performance by content area and grade for selected student groups. Table 10.B.9 through Table 10.B.36 show the statistics for the ELA assessments; Table 10.B.37 through Table 10.B.57 show the statistics for mathematics. In these tables, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status (disadvantaged or not), need for special education services, migrant status, the use of designated supports (using designated supports or not), and the use of accommodations (using accommodations or not).

The tables show, for each demographic student grouping, the number of students with a valid scale score, the scale score mean and standard deviations, and the percentage of students in each claim performance level, for 2015–16, 2016–17, and 2017–18, as well as the year-to-year differences in the average scale scores and the percentages of *Near Standard* or *Above Standard*.

10.1.3. Longitudinal Comparisons on the Overall Tests

For longitudinal comparisons, the data is gathered and compared for the same students in 2017–18, 2016–17, and 2015–16. Through vertical scaling, scores on tests at different grade levels of the same content area were placed on a common scale. For Smarter Balanced Summative Assessments, reporting scores on a vertical scale allows student progress to be tracked for a particular content area across grade levels.

The difference in average scale scores or in the percentages of students meeting or exceeding standards is the later year's (e.g., 2017–18) values minus the previous year's (2016–17) values for the same students. Therefore, a positive value indicates an increase in the later year (e.g., 2017–18) and a negative value indicates a decrease in the later year (e.g., 2017–18). Individual achievement level percentages may not sum to exactly 100 or the combined achievement level percentage due to rounding.

For year-to-year comparisons, only the differences between 2017–18 and 2016–17 and the differences between 2016–17 and 2015–16 are presented. The statistics in these tables include only those students who advanced one grade each year and whose scores were available in all three years.

Refer to the *2016–17 CAASPP Smarter Balanced Technical Report* (CDE, 2017) for the comparison between data from the 2016–17 and 2015–16 administrations.

10.1.3.1 Summary Statistics

Table 10.C.1 in appendix 10.C shows the number of students assessed, the number of students with valid scores, the means and standard deviations of students' scale scores in 2016–17 and 2017–18 for each test, as well as the differences in scale scores between 2016–17 and 2017–18. Table 10.C.2 presents the same set of statistics as in Table 10.C.1, but for all three administration years (2015–16, 2016–17, and 2017–18), as well as the year-to-year differences in scale scores.

10.1.3.2 Achievement Levels of Overall Students

The percentages of students of each achievement level and qualifying for the *Standard Met* and *Standard Exceeded* levels, as well as the differences in the percentages of the students in *Standard Met* and *Standard Exceeded* between 2016–17 and 2017–18 are presented in Table 10.C.3 in appendix 10.C. The same information is presented in Table 10.C.4 but for all three administration years (2015–16, 2016–17, and 2017–18).

10.1.3.3 Scale Score Distributions

Table 10.C.5 and Table 10.C.7 in appendix 10.C show the distribution of scale scores observed in 2016–17 and 2017–18 on the same students per each grade and content area. Frequency counts are provided for each scale score interval of 30. The scale score distributions for 2015–16, 2016–17, and 2017–18 are presented in Table 10.C.6 and Table 10.C.8.

10.1.3.4 Achievement Levels of Selected Groups

Table 10.C.9 through Table 10.C.18 in appendix 10.C provide statistics summarizing student performance by content area and grade for selected groups of students. In the tables, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status (disadvantaged or not), need for special education services, migrant status, the use of designated supports (using designated supports or not), and the use of accommodations (using accommodations or not).

The tables show, for each demographic group, the numbers of students with valid scale scores in 2016–17 and 2017–18 as well as the scale score means and standard deviations, and the percentage of students in each achievement level, for these students. Additionally, the differences in the percentages of *Standard Met* and *Standard Exceeded* between 2016–17 and 2017–18 are shown. The statistics for three years 2015–16, 2016–17, and 2017–18 are presented in Table 10.C.19 through Table 10.C.26.

10.1.4. Longitudinal Comparisons on Claims

10.1.4.1 Summary Statistics

Table 10.D.1 through Table 10.D.4 in appendix 10.D contain the number of students assessed, the number of students with valid scores, the means and standard deviations of students' scale scores in 2016–17 and 2017–18 on each claim by grade and content area, as well as the differences in the scale scores between 2016–17 and 2017–18.

The statistics for each claim in 2015–16, 2016–17, and 2017–18 are presented in Table 10.D.5 through Table 10.D.8. The claims are on the same scale as the total test in which the claims are included. Refer to the score ranges of the total test (Table 7.10) for the score ranges of claims.

10.1.4.2 Performance Levels of Overall Students

Table 10.D.9 and Table 10.D.10 in appendix 10.D present the percentages of students in each performance level of each claim in 2016–17 and 2017–18. Student results on each claim are classified into three achievement levels: *Below Standard*, *Near Standard*, and *Above Standard*. Refer to Table 7.12 in chapter 7 for the achievement level scale score ranges for each test. The percentages of students of each performance level, as well as the differences in the percentages of *Near Standard* or *Above Standard* between 2017–18 and 2015–16. Refer to *7.4.2.2 Performance Levels for Claims* in chapter 7 for the details regarding the classification of achievement levels on claims. Table 10.D.11 through Table 10.D.14 present the percentages of each performance level of each claim in 2015–16, 2016–17, and 2017–18.

10.1.4.3 Performance Levels of Selected Student Groups

Table 10.D.15 through Table 10.D.49 in appendix 10.D show the statistics summarizing student performance by content area and grade for selected student groups. Data in Table 10.D.15 through Table 10.D.34 is calculated from the data for the ELA assessments; data in Table 10.D.35 through Table 10.D.49 is calculated from the data for mathematics.

In these tables, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status (disadvantaged or not), need for special education services, migrant status, the use of designated supports (using designated supports or not), and the use of accommodations (using accommodations or not).

The tables show, for each demographic student grouping, the number of students with a valid scale score, scale score means and standard deviations, and the percentage of students in each performance level, for 2016–17 and 2017–18 respectively, as well as the differences in the percentages of *Near Standard* or *Above Standard* between 2016–17 and

2017–18. Table 10.D.50 through Table 10.D.77 present the percentages of each performance level of each claim in 2015–16, 2016–17, and 2017–18.

10.2. Test Characteristics

The marginal reliabilities and SEMs expressed in theta score units for each test are presented in Table 10.E.1 in appendix 10.E. The same statistics as in Table 10.E.1 for claims 1 and 2 appear in Table 10.E.2. Those for claims 3 and 4 are presented in Table 10.E.3.

Reliabilities are affected by both item characteristics and student characteristics. Refer to subsections *8.5.2 Marginal Reliability* and *8.5.3 Standard Error of Measurement* for the methods used to calculate marginal reliability and SEM, respectively.

Reference

California Department of Education. (2017). *2016–17 California Assessment of Student Performance and Progress Smarter Balanced technical report.* Retrieved from https://www.cde.ca.gov/ta/tg/ca/documents/sbac17techrpt.pdf

Chapter 11: Paper-Pencil Versions of Smarter Balanced Summative Assessments

11.1. Background

Paper-pencil versions of the Smarter Balanced Summative Assessments are made available to local educational agencies (LEAs) that either do not have the necessary computer network infrastructure to administer the online tests or do not include computers as a part of their curricula. The paper-pencil versions contain a fixed set of questions that includes components of the online assessment such as multiple-choice items, constructed-response (CR) items, and performance tasks (PTs).

Paper-pencil versions exist for all grade levels and content areas assessed by Smarter Balanced and were administered to nearly 1,300 students across California in 2017–18. There were approximately 400 students who took the English language arts/literacy (ELA) and mathematics paper-pencil tests in grades three and four during the 2017–18 administration. For all other tests, there were fewer.

Paper-pencil versions were available only with prior permission from the California Department of Education (CDE).

11.2. Testing Window

The window for 2017–18 Paper-Pencil testing was the same as for the online tests: January 9 through July 16, 2018. Specific test administration schedules within that window were determined locally pursuant to the *California Code of Regulations*, Title 5, sections 855(a)(1), 855(a)(2), 855(b), and 855(c).

11.3. Test Assembly

Paper-pencil test versions are composed of PT items and items that are not based on performance tasks (non-PTs).

During the test development process, efforts were made to ensure that paper-pencil test items and online test items were comparable to the greatest extent possible. The paper-pencil test development involved evaluating the test blueprint and identifying which items could be successfully assessed in paper-pencil format. The paper-pencil item development process starts with looking at each technology-enhanced item that needs a replacement or modification.

A preliminary calibration report provided by the National Center for Research on Evaluation, Standards, & Student Testing (CRESST) found that no more than three items per grade level and content area from the online test item pool that appeared on paper-pencil tests without modifications were identified as functioning differently across the two modes (CRESST, 2015).

11.4. Test Administration

The *2017–18 California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced Paper-Pencil Test Administration Manuals* (CDE, 2018a) provide an overview of the summative assessment administration and supplement the *2017–18 CAASPP Smarter Balanced Online Test Administration Manual* (CDE, 2018b). The *Paper-*

Pencil Test Administration Manuals, available for each grade, are intended to familiarize test administrators with general rules for testing, how to prepare for the assessment, and what students experience in participating in the assessment.

Test preparation, administration, and security procedures must be followed so that all students will have an equal opportunity to demonstrate their academic achievement. Refer to *Chapter 5: Test Administration* for more information on procedures followed in 2017–18. Additionally, refer to subsection *5.4 Procedures to Maintain Standardization* for additional information about the staff involved with administering CAASPP assessments.

11.5. Universal Tools, Designated Supports, and Accommodations

Consistent with the online tests, designated supports, accommodations (subsection *2.5 Universal Tools, Designated Supports and Accommodations*) and unlisted resources (subsection *2.5.3 Unlisted Resources*) are assigned to individual students based on student need. Appendix 11.A presents counts and percentages of students using designated supports, accommodations, or unlisted resources. Note that “NA” indicates that the designated support, accommodation, or unlisted resource is not available for that test. The majority of students do not use any designated supports, accommodations, or unlisted resources.

11.6. Calibration and Scaling

Post-test calibration, equating, and scaling of the Smarter Balanced paper-pencil summative tests are conducted by CRESST by using data from paper-pencil tests administered by two member states of the Smarter Balanced Assessment Consortium. To produce scores for the paper-pencil tests that are on the same scale as the online tests, separate calibrations of the paper-pencil response data were conducted and then scaled to the online item bank. The “new” calibrations for the paper-pencil versions were established by calibrating samples of item response data from the paper-pencil administration; the “reference” calibrations were based on the Smarter Balanced Online Summative Assessment item bank that was established during the field test.

For the purpose of linking the paper-pencil forms to the official reporting scale derived from the online test mode, the paper-pencil test item parameter estimates are placed on the reference scale by using a set of anchor items that were not modified. Specifically, these unmodified items indicate these items may appear in either test delivery mode as-is without altering the construct; that is, the items parameter estimates should be invariant across the delivery mode.

The procedure used for equating the Smarter Balanced paper-pencil summative tests involves three parts: initial item calibration, anchor item evaluation, and final item calibration. Each of those procedures, as described in the next subsection, is applied to all tests. The calibrations were performed with the flexMIRT® item response modeling software (Cai, 2015).

11.6.1. Initial Calibration

The following steps are involved in the initial calibration to obtain item parameter estimates and model goodness-of-fit indices. The generalized partial credit (GPC) model was applied to both multiple-choice items and polytomously scored items. Refer to subsection

7.4.1 Total Test Scores in *Chapter 7: Scoring and Reporting* of this report for the mathematical formula of the GPC model.

1. The parameters of all unmodified items are fixed to the parameter values obtained from the online item pool.
2. The parameters of all modified items are freely estimated.
3. The latent variable density is estimated as an empirical histogram (refer, for example, to Woods, 2007; Houts & Cai, 2013) with estimated mean and variance from the "all" student population, including students taking online tests.

11.6.2. Anchor Item Evaluation

The purpose of anchor item evaluation is to select items that function similarly across both online and paper-pencil modes as anchors. By linking tests through these anchor items, paper-pencil test results are placed onto the online test scale and scores from the two modes should be comparable.

A series of calibrations identical to the "initial" calibration are performed but with the parameters of one unmodified item at a time freely estimated. The parameters of all other unmodified items are fixed to their prior estimates from the online item pool. As in the initial calibration, the parameters of all modified items are freely estimated, along with the population distribution's mean, variance, and shape.

To decide whether each unmodified item should be retained or rejected as an anchor in the final calibration for the paper-pencil forms, the parameter estimates from the online item pool administration and the parameter estimates from the initial calibration are used to compute the expected score functions for the two modes of test administration. The two expected score functions—for the computer-based and paper-pencil administrations—are plotted, and differences in item functioning across the two modes are quantified by computing a weighted Area Between the Curves (wABC; refer to Hansen, Cai, Stucky, Tucker, Shadel, & Edelen, 2014). Any items with a wABC value greater than 0.150 were rejected as anchors.

11.6.3. Final Calibration

For tests in which any unmodified item is rejected as an anchor, a final calibration is conducted using the approach described in subsection *11.6.1 Initial Calibration*, except that the parameters of all rejected anchor items are freely estimated. Parameters of the modified items also are freely estimated. The parameter estimates from this final calibration are used in scoring the paper-pencil forms. In this way, paper-pencil version scores are placed on the online test scale.

11.7. Scoring

As in the CAASPP Smarter Balanced online assessments, student item responses in the paper-pencil forms are scored and individual student scores are calculated (i.e., overall scale scores and claims and subscores) based on the scored item responses. The same scoring specifications and procedures as in the online assessments are followed except that all the CR items in the paper-pencil versions are human-scored; no artificial intelligence machine scoring is used. However, because of the small student sample sizes, particularly in the upper grades and certain student groups, caution should be taken when interpreting some of the summary statistics.

11.7.1. Total Score Distributions and Achievement Levels

Summary statistics that describe student performance on each test are presented in Table 11.1. Included in the table are the number of students administered each test and the means and standard deviations of student scores expressed in terms of both scale scores and theta scores. Only students with valid scores are included in this table. “Valid score” means the student records were not flagged as “not scored,” and the students were enrolled in the same grade as they were tested.

Table 11.1 Mean and Standard Deviation of Total Theta and Scale Scores for CAASPP Smarter Balanced Paper-Pencil Summative Assessments

Content Area/Grade	No. of Students	Scale Score Mean	Scale Score Standard Dev.	Theta Score Mean	Theta Score SD
ELA 3	421	2431	95	-0.90	1.11
ELA 4	397	2482	87	-0.30	1.02
ELA 5	281	2534	87	0.30	1.02
ELA 6	84	2570	78	0.72	0.91
ELA 7	47	2564	91	0.65	1.06
ELA 8	51	2615	88	1.24	1.03
ELA 11	14	2587	75	0.91	0.87
Mathematics 3	420	2415	76	-1.26	0.96
Mathematics 4	397	2471	75	-0.55	0.95
Mathematics 5	286	2500	76	-0.19	0.96
Mathematics 6	82	2547	99	0.41	1.25
Mathematics 7	48	2538	93	0.29	1.17
Mathematics 8	44	2621	89	1.33	1.13
Mathematics 11	15	2510	135	-0.06	1.71

The number and the percentage of students in each achievement level and the numbers and the percentages which meet or exceed the standard are shown in Table 11.2.

Table 11.2 Percentages and Counts of Students in Achievement Levels for CAASPP Smarter Balanced Paper-Pencil Summative Assessments

Content Area/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	120	29%	68	16%	96	23%	137	33%	233	55%
ELA 4	82	21%	82	21%	117	29%	116	29%	233	59%
ELA 5	46	16%	50	18%	96	34%	89	32%	185	66%
ELA 6	7	8%	16	19%	38	45%	23	27%	61	73%
ELA 7	10	21%	11	23%	17	36%	9	19%	26	55%
ELA 8	4	8%	8	16%	27	53%	12	24%	39	76%
ELA 11	2	14%	7	50%	4	29%	1	7%	5	36%
Mathematics 3	138	33%	118	28%	109	26%	55	13%	164	39%
Mathematics 4	84	21%	154	39%	94	24%	65	16%	159	40%
Mathematics 5	91	32%	91	32%	56	20%	48	17%	104	36%
Mathematics 6	18	22%	20	24%	21	26%	23	28%	44	54%
Mathematics 7	14	29%	11	23%	16	33%	7	15%	23	48%
Mathematics 8	4	9%	11	25%	14	32%	15	34%	29	66%
Mathematics 11	10	67%	2	13%	2	13%	1	7%	3	20%

* May not exactly match the sum of percent *Standard Met* and percent *Standard Exceeded* due to rounding

Detailed score distribution information is available in appendix 11.B. Table 11.B.1 and Table 11.B.2 show the estimated distributions of theta scores for each test. Table 11.B.3 and Table 11.B.4 present selected percentiles of the ELA and mathematics scale score distributions. Table 11.B.5 through Table 11.B.18 present frequency distributions of scale scores for each test.

11.7.2. Claim Score Distributions and Performance Levels

Table 11.C.1 through Table 11.C.4 in appendix 11.C show the number of items presented within each claim, number of students with valid scores in each claim, and the means and standard deviations of student scores expressed in terms of both scale scores and theta scores. The number of students in each claim performance level as well as the percentage of students in that claim performance level are reported in Table 11.C.5 through Table 11.C.8. Note that the percentage is shown as “NA” when there are no students in a performance level for a claim.

11.7.3. Group Scores

Statistics summarizing student performance by content area and grade for selected demographic groups of students are provided in appendix 11.D, in Table 11.D.1 through Table 11.D.14 for each test, and for each test claim in Table 11.D.15 through Table 11.D.28. Note that statistics are reported only for samples that are comprised of 11 or more students; statistics are presented in the tables as "NA" for samples fewer than 11. The percentage is shown as "NA" when there are no students in a performance level for a claim.

11.8. Analyses

This section summarizes the item-parameter values, reliability and conditional standard error of measurement (CSEM), and correlations between content areas calculated for the Smarter Balanced Paper-Pencil Summative Assessments. Note that statistics should not be assumed to generalize, due to the small numbers of students in the analyses. Additionally, because of the small sample size in paper-pencil tests, some analyses that were reported for the online summative tests are not reported for the for paper-pencil tests. These analyses include but are not limited to reliability of performance classifications and interrater reliability and agreement.

11.8.1. IRT Parameter Values

Parameter estimates for the paper-pencil versions of the 2017–18 CAASPP Smarter Balanced operational items were obtained using the procedure described in subsection *11.6 Calibration and Scaling*. 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 presented.

Table 11.E.1 through Table 11.E.14 in appendix 11.E present univariate statistics (mean, standard deviation, minimum, and maximum) of the scaled item response theory (IRT) *a*-values. For each test, the results are presented for all items in the test and for the items in each claim. Table 11.E.15 through Table 11.E.28 present the univariate statistics of the IRT *b*-values for all items in the test and for the items in each claim.

Table 11.E.29 and Table 11.E.30 show the distributions of *a*-values of non-PT items in each test across 10 intervals. Table 11.E.31 and Table 11.E.32 present the distributions of *b*-values of non-PT items across 16 intervals. The mode of each distribution is highlighted and indicated using an asterisk. Table 11.E.33 and Table 11.E.34 show the distribution of *a*-values for the PT items. Table 11.E.35 and Table 11.E.36 show the distribution of *b* values for the PT items. Parameter values of all PT items are presented in Table 11.E.37 through Table 11.E.50.

11.8.2. Reliability Analyses

This subsection presents results of the reliability analyses of test scores and claim scores for the population as a whole and for selected student groups. Refer to subsection *8.5.2 Marginal Reliability* for the description and calculation of reliability. Similar to the reliability analyses conducted for the CAASPP online test, students assigned to the lowest or highest obtainable scale score were excluded.

Table 11.3 gives the total score reliability for theta, the mean, standard deviation, and standard error of measurement (SEM) for the theta and scale scores for each of the 14 tests. Only students with complete records were included in this table. A student's record for the test is not considered complete unless the student completed at least 10 non-PT items and at least one PT item.

Table 11.3 Summary Statistics for Scale Scores and Theta Scores, Reliabilities, and SEMs for CAASPP Smarter Balanced Paper-Pencil Summative Assessments

Content Area and Grade	No. of Students	Reliability	Scale Score Mean	Scale Score SD	Scale Score SEM	Theta Score Mean	Theta Score SD	Theta Score SEM
ELA 3	420	0.93	2430	95	25.04	-0.91	1.10	0.29
ELA 4	394	0.91	2481	86	25.69	-0.32	1.00	0.30
ELA 5	275	0.91	2530	85	24.91	0.25	0.99	0.29
ELA 6	83	0.89	2568	76	25.71	0.70	0.89	0.30
ELA 7	46	0.90	2560	87	27.12	0.60	1.02	0.32
ELA 8	47	0.90	2602	79	24.66	1.09	0.92	0.29
ELA 11	14	0.87	2587	75	26.66	0.91	0.87	0.31
Mathematics 3	416	0.90	2415	74	23.00	-1.26	0.93	0.29
Mathematics 4	391	0.91	2468	72	21.56	-0.59	0.91	0.27
Mathematics 5	284	0.90	2499	75	23.39	-0.20	0.94	0.29
Mathematics 6	81	0.91	2545	97	28.40	0.38	1.23	0.36
Mathematics 7	48	0.90	2538	93	29.51	0.29	1.17	0.37
Mathematics 8	41	0.88	2607	77	26.96	1.17	0.97	0.34
Mathematics 11	14	0.86	2526	124	46.97	0.14	1.57	0.59

Intercorrelations, reliability estimates and theta-based SEMs for the claims are presented in Table 11.F.1 through Table 11.F.14 in appendix 11.F. The reliability estimates across claims vary significantly according to the number of items as well as the types of content standards that are included in each claim.

Reliabilities and theta-based SEMs for the total test scores and the claim scores are reported for each student group analysis. Table 11.F.15 through Table 11.F.20 present the overall test reliabilities for student group defined by student gender, economic status, special education services status, English language fluency, primary ethnicity, and migrant status. Table 11.F.21 and Table 11.F.22 present the reliabilities for the student groups based on primary ethnicity within economic status.

The next set of tables, Table 11.F.23 through Table 11.F.92, present the claim-level reliabilities for the student groups. Table 11.F.23 through Table 11.F.36 present the claim-level reliabilities for the student groups based on gender, economic status, and migrant status. Table 11.F.37 through Table 11.F.50 show the same analyses for the student groups based on special education services status and English language fluency. Table 11.F.51 through Table 11.F.64 present results for the student groups based on primary ethnicity of the students. The last set of tables, Table 11.F.65 through Table 11.F.92, present the claim-level reliabilities for the student groups based on primary ethnicity within economic status.

Note that the reliabilities are reported only for samples that include 11 or more students. In cases where the sample size is smaller than 11, reliabilities are presented in the tables as “NA.” The reliability estimates for some of the student groups are negative due to small variation in scale scores and large CSEMs for extreme score values. These negative reliabilities and their associated SEMs are presented as “NA.”

11.8.3. CSEM Distributions

This subsection presents CSEM distributions for the total test scores and the mean CSEM for each achievement level. Table 11.4 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. The CSEM is presented as “NA” if there are no students at the cut point for a certain achievement level.

Table 11.4 Scale Score CSEM at Achievement-level Cut Points for CAASPP Smarter Balanced Paper-Pencil Summative Assessments

Content Area and Grade	Standard Nearly Met Min SS	Standard Nearly Met CSEM	Standard Met Min SS	Standard Met CSEM	Standard Exceeded Min SS	Standard Exceeded CSEM
ELA 3	NA	NA	NA	NA	2490	23
ELA 4	2416	25	2473	24	2533	25
ELA 5	NA	NA	NA	NA	2582	25
ELA 6	NA	NA	NA	NA	NA	NA
ELA 7	2479	27	NA	NA	2649	26
ELA 8	NA	NA	NA	NA	NA	NA
ELA 11	NA	NA	NA	NA	NA	NA
Mathematics 3	NA	NA	2436	17	NA	NA
Mathematics 4	2411	20	2485	17	2549	17
Mathematics 5	NA	NA	NA	NA	2579	18
Mathematics 6	NA	NA	NA	NA	NA	NA
Mathematics 7	NA	NA	NA	NA	NA	NA
Mathematics 8	NA	NA	NA	NA	NA	NA
Mathematics 11	NA	NA	NA	NA	NA	NA

Table 11.5 presents the average CSEMs in each achievement level by content area and grade level.

Table 11.5 Average CSEM of Scale Scores in Each Achievement Level for CAASPP Smarter Balanced Paper-Pencil Summative Assessments

Content Area and Grade	Standard Not Met	Standard Nearly Met	Standard Met	Standard Exceeded
ELA 3	28.24	22.72	22.00	23.61
ELA 4	28.83	25.00	24.53	25.77
ELA 5	26.67	24.00	24.63	26.42
ELA 6	28.86	26.13	25.55	27.27
ELA 7	29.20	26.09	25.65	27.63
ELA 8	30.75	26.25	26.00	28.00
ELA 11	33.50	29.14	28.00	31.00
Mathematics 3	21.85	17.89	17.00	17.49
Mathematics 4	22.23	18.13	17.00	17.63
Mathematics 5	26.54	20.87	18.20	17.91
Mathematics 6	32.56	23.00	20.48	20.77
Mathematics 7	38.00	26.64	21.56	20.00
Mathematics 8	35.50	28.18	24.29	21.58
Mathematics 11	49.44	31.00	24.00	22.00

Scale score CSEM distributions are shown in Table 11.G.1 through Table 11.G.14 of appendix 11.G. The plots of the CSEMs conditional for scale scores are also presented in this appendix, in Figure 11.G.1 through Figure 11.G.14. In the figures, the vertical axis is defined as the CSEMs and the horizontal axis is designated as scale scores, which is a common metric for tests within the same content area. Each data point represents an individual student.

11.8.4. Correlations between Content Area Test Scores

Table 11.6 provides the correlations between scores on the 2017–18 ELA and mathematics paper-pencil tests and the numbers of students on which these correlations are based. Sample sizes for individual tests are shown in bold and indicated with an asterisk; the numbers of students on which the correlations are based are shown on the lower left without bolding. The correlations are provided in the upper right. Results are based on all students with valid scale scores and are provided by grade.

In general, students' ELA scores correlated moderately with their mathematics scores. Due to very small test volumes in many demographic groups, the correlations are not presented between content areas for student groups.

Table 11.6 Correlations between Content Areas for All Students with Paper-Pencil Tests

Content Area and Grade	Sample Size	R and Sample Size
ELA 3	***421**	0.78
Mathematics 3	416	***420**
ELA 4	***397**	0.74
Mathematics 4	396	***397**
ELA 5	***281**	0.68
Mathematics 5	279	***286**
ELA 6	***84**	0.68
Mathematics 6	80	***82**
ELA 7	***47**	0.78
Mathematics 7	47	***48**
ELA 8	***51**	0.77
Mathematics 8	44	***44**
ELA 11	***14**	0.59
Mathematics 11	14	***15**

Notes:

- Numbers not in **bold** font on the left side are the sample sizes to calculate the correlations.
- Sample sizes of the individual assessments are in bold font.
- R denotes the correlation coefficient; these are decimals that begin with “0” (zero).

References

Cai, L. (2015). *FlexMIRT: Flexible multilevel item factor analysis and test scoring* [Computer software]. Seattle, WA: Vector Psychometric Group.

California Department of Education. (2018a). *2017–18 CAASPP Smarter Balanced paper-pencil test administration manual*. Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.gr5-sb-ppt-tam.2017-18.pdf

California Department of Education. (2018b). *2017–18 CAASPP Smarter Balanced online test administration manual.* Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.online_tam.2017-18.pdf

CRESST. (August, 2015). *Initial report on the calibration of paper and pencil forms*. Los Angeles, CA.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. S., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS smoking item banks. *Nicotine and Tobacco Research, 16, Supplement 3,* S175-S189.

Houts, C. R., & Cai, L. (2013). *FlexMIRT user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring.* Chapel Hill, NC: Vector Psychometric Group.

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and psychological measurement, 67*, 73–87.

Chapter 12: Continuous Improvement

The fourth 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 2017–18. Throughout the past four 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.

The California Department of Education (CDE) and Educational Testing Service (ETS) reviewed the findings of the Human Resources Research Organization (HumRRO), which is the independent evaluator for the CAASPP System. HumRRO has conducted several evaluations since the Smarter Balanced field test administration and has reported its findings to the CDE and the State Board of Education. The CDE and ETS addressed HumRRO's findings and feedback as part of the overall effort to improve the CAASPP program. HumRRO reports and ETS' responses are generally posted on the CDE CAASPP Technical Reports and Studies web page at https://www.cde.ca.gov/ta/tg/ca/caasspprptstudies.asp.

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.

12.1. Test Delivery and Administration with the Smarter Balanced Assessment Consortium

12.1.1. Performance Task (PT) Administration

Since the 2014–15 administration, the Smarter Balanced online assessments have been administered using an adaptive design that incorporates nonadaptive performance tasks. For the first two administrations, in 2014–15 and 2015–16, ETS assigned PTs randomly at the school level because the PTs were accompanied by a classroom activity that was done prior to administration. The random-selection lists were sent to local educational agencies (LEAs) for test preparation and to the American Institutes for Research (AIR), which prepared the PTs for delivery through its test delivery engine.

For the 2016–17 administration and beyond, Smarter Balanced made classroom activities optional for both ELA and mathematics. Consequently, because it was not necessary to assign a grade level's classroom activity (and therefore, a PT) to a school, PTs were, instead, randomly assigned at the student level, lessening the testing burden on both students and teachers. Furthermore, having the randomization occur at student level increased the diversity of the student groups responding to each of the different PTs.

12.1.2. Increased Field Test PT Sample Size

Smarter Balanced has been field testing PTs to a group of randomly selected students during the operational test administration since 2016–17. Approximately 5,000–7,000 students per test were assigned to participate in the Smarter Balanced PT field tests during the 2016–17 administration. For the 2017–18 administration, approximately 9,000 students per test were assigned to participate in the PT field tests.

Because the sample size increased in the 2017–18 administration, the target student populations were represented better. Consequently, results for the field tested PTs have generated more robust item statistics.

12.1.3. Shorter Paper-Pencil Forms

Paper-pencil versions of the Smarter Balanced Summative Assessments have been shortened. These shorter forms continue to meet content blueprints and reliability requirements.

For the 2014–15 through 2016–17 administrations, Form 1, which consisted of about 52 ELA items and 41 mathematics items, was administered to students. For the 2017–18 administration, Form 3 was used, which consisted of about 44 ELA items and 39 mathematics items.

Reliability comparisons for the 2016–17 and 2017–18 paper-pencil test administrations show that both forms 1 and 3 achieved similar reliability, with a minimum of 0.86 and a maximum of 0.93. Similar reliability and shorter forms reduce testing time for students, thus improving the efficiency of the paper-pencil tests. Refer to the Smarter Balanced technical reports for more information on the improvement of paper-pencil forms (Smarter Balanced, 2016a, 2016b, 2017, 2018a).

12.2. ETS Administration and Delivery

12.2.1. Post-Test Survey

The CAASPP program annually solicits feedback from CAASPP stakeholders through the CAASPP Post-Test Survey. LEA and test site staff, as well as test administrators and test examiners, were invited to participate in the 2017–18 CAASPP Post-Test Survey. More than 10,000 California educators provided specific, actionable insights about their testing experience. Some of these suggestions (e.g., feedback about the training, improvements to TOMS and the test delivery system, and additional video tutorials) were acted upon. A majority of survey respondents and focus group participants overall reported experiencing adequate preparation and training, resulting in generally smooth and successful CAASPP assessment administrations.

12.2.2. Additional Training Resources

To address respondents' confusion regarding the assignment and use of embedded universal tools, designated supports, and accommodations, ETS updated existing training materials, including how-to videos about each available resource. Additional videos were added and some existing videos were updated. ETS promoted these expanded training materials via multiple channels, including by providing links in manuals and mentions during workshops and webcasts.

12.3. Hand Scoring

12.3.1. Document Summative Assessment Scoring Activities

Constructed-response scoring information is captured within the ETS and Measurement Incorporated (MI) scoring systems. Reports for quality monitoring during the scoring process are produced and reviewed by MI scoring and ETS Assessment Development (AD) staff. In addition, validity detail documents provide statistics for each validity sample for each item.

During the 2017–18 administration, the same metrics were used to evaluate the validity samples scored after enough data had accumulated. ETS and MI used at least 30 validity papers covering the full range of scores, although supplemental samples were added as needed. Validity sets were monitored throughout the administration and post-administration periods for performance.

Improvements that occurred in 2017–18 included that rater agreement and validity statistics for rater agreement were monitored each week. Scoring leaders provided feedback to ETS AD to determine what adjustments to training or samples were to be made.

Planned improvements for 2018–19 will include the following:

- The quality monitoring plan will be formally documented.
- Scoring leader performance indicator panels will be implemented to allow easier access to quantitative feedback regarding individual raters.

ETS summarizes and documents the rater training process in subsection *7.3 Rater Training* of the annual *CAASPP Smarter Balanced Technical Report*. Improved training and scoring documentation within the scoring systems features the following:

- System training documents and videos for raters and score leaders covering navigating the system for training and scoring
- Training directions for raters, organized by item type (The directions outline the training sets to be reviewed for each item type.)
- Procedures for scoring use of condition codes and processing crisis alert responses
- Functions such as the following:
  - Escalation from scoring leaders through the scoring hierarchy, from group scoring leaders to chief scoring leaders, and to ETS AD content experts, as needed, for review and appropriate action
  - End-of-shift reports submitted by scoring leaders capturing item feedback each day (This data is used by ETS AD for continuous feedback and improvement. Scoring leaders can also capture item feedback on their end-of-shift reports each day.)
  - Training sets completed by raters before beginning scoring a particular item
  - Annual documentation review to capture changes to the systems, policies and procedures

12.3.2. Compliance with Web Accessibility Standards

In ETS's distributed constructed-response (CR) scoring system, the Online Network for Evaluation, responses are rendered via image or text viewers and audio and video players with standard features that allow raters to enhance the visibility and volume of constructed responses. ETS engages its Accessibility, Standards, and Assistive Technology Research Group to conduct accessibility reviews, ensuring systems comply with Web Content Accessibility Guidelines (WCAG) 2.0 and Section 508 accessibility standards.

To prioritize the development of future accessibility enhancements, it is essential to identify the types of accommodations that are most urgently required by end users. ETS is planning discussions with the CDE about the accessibility capabilities that would be most beneficial to expanding CAASPP scoring opportunities to a wider community of California educators. ETS is committed to compliance with WCAG 2.0.

In MI's system, Virtual Scoring System (VSC) users are able to increase the font size of student online responses, as well as zoom, pan, and adjust contrast for paper scanned documents. VSC scoring applications can be navigated using the keyboard, in addition to a mouse, as the primary user input device. Once specific capabilities have been agreed upon with the CDE, ETS will work with MI to implement them in the VSC where possible.

12.3.3. Documenting and Revisiting Summative Assessment Item Flagging Criteria

At least ten 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 statistic was calculated to reflect the level of improvement beyond the chance level in the consistency of scoring in chapter 8. Refer to appendix 8.G for detailed information.

In 2017–18, ETS identified items that did not meet the requirements for interrater agreement using the flagging criteria developed by Smarter Balanced and documented the flagged items in subsection *7.2.4 Interrater Reliability Results*.

12.3.4. 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 CDE on the counts of California educators and California residents participating in both the existing rater pool and as potential raters in the recruitment pipeline. ETS' Strategic Workforce Solutions has the capability, through its applicant tracking system, to collect additional background information on all applicants. As a part of the current recruiting process for CAASPP raters, ETS and MI gather rater responses to the following questions:

- Are you fluent in Spanish and interested in scoring assessments in Spanish?
- Do you have experience teaching in a kindergarten through grade twelve (K–12) school?
- Do you currently work in a K–12 school in California?
- Do you have experience teaching English as a second or foreign language?

Systemically, ETS Strategic Workforce Solutions analyzes the data received in its application process and uses the answers to these questions to support the development of a strong, qualified workforce.

Documentation of the qualifications of the rater pool will be produced annually.

12.4. Psychometric Analyses

12.4.1. Smarter Balanced Item Pool Verification

ETS has verified the Smarter Balanced packages (item pool) each year since the first CAASPP Smarter Balanced administration in 2014–15 to verify the appropriateness of all scoring information such as item points, item parameters, and claim and target standard classifications.

During the 2018–19 Smarter Balanced package verification, ETS psychometricians added item flagging criteria to include standard errors for each item's item response theory (IRT) *b*-parameter estimates. Specifically, if the standard error corresponding to the *b*-parameter estimate is 2.0 or larger, ETS psychometricians will notify the Smarter Balanced Assessment Consortium of this issue and whether the item should remain in the operational item pool.[9].

12.4.2. Scoring Verification Process

Since the Smarter Balanced field test was administered in 2014–15, ETS has established two independent and parallel scoring systems to produce and verify student scores: the Enterprise Score Key Management scoring system and a parallel scoring system used by the ETS Psychometrics, Analysis, and Research (PAR) team. The parallel systems score each individual student independently, applying the same scoring algorithms and specifications with different programming software. PAR evaluates parallel scoring results from these two systems to ensure all scores for a student from the two systems are identical with acceptable tolerances.

Next, before the Smarter Balanced scores are reported to schools and LEAs, the ETS PAR team conducts a comprehensive statistical review using all available data that passed through the parallel scoring verification.

When the results of the comprehensive analyses show all scores are accurate and score distributions are reasonable and consistent with expectations, those results are sent to PAR leadership for review and approval. After that, Smarter Balanced scores undergo CDE pilot review. The reporting gate for Smarter Balanced scores opens as soon as the CDE approves the scores.

These verification activities effectively prevented scoring errors in previous administrations. In particular, the parallel scoring procedures resulted in a scoring accuracy of 100 percent in the past. However, because of a small sample size, a problematic item in grade eight mathematics was overlooked during the May 2018 verification but was subsequently discovered in August 2018, after the test was completed. The sample available for score verification in May was simply not sufficient enough to raise error flags for the item.

To identify possible items with concerns more effectively for the 2018–19 and subsequent CAASPP administrations, the PAR team has implemented the following new verification procedures to increase the rigor of scoring verification:

- ETS added item flagging criterion to the comprehensive statistical verification process. Based on the test results of the verification sample, if the percentage of correct scores for an item is too high or too low, this item will be placed on a watch list for continuous monitoring. When the sample size increases to 1,000 students and if the percent-correct score is still outside a reasonable range (.05 through .95), this item will be

[9] Generally, standard errors associated with IRT parameter estimates tend to be small (below 0.5). Note that during the package verification, nine items were identified for having standard errors associated with the difficulty parameter exceeding 2.0. ETS psychometricians consulted with Smarter Balanced as to whether those items should remain in the pool. The guidance from Smarter Balanced was to retain those items in the pool.

flagged formally and ETS will communicate with the CDE and Smarter Balanced regarding this item.

- In addition to verifying student scores on a monthly basis after the scoring gate opens, the comprehensive statistical review will continue as well. In particular, biweekly monitoring of average item scores for all items in the bank will occur.
- The comprehensive statistical review will be added to the verification of the final complete data files after the test window closes, along with the parallel scoring quality control. The results from the final comprehensive statistical review will be sent to the CDE for evaluation.

12.4.3. Average Conditional Standard Error of Measurement (CSEM) Based on 2017–18 Administration Results

Because the Smarter Balanced summative assessments use item pattern scoring to estimate student abilities and the conditional standard error of measurement (CSEM), there are unique estimates for each response pattern. For some response patterns, more uncertainty or random error will exist. This effect is evident at the upper and lower ends of the reporting scale, where items administered to students might not match a student's true ability level (e.g., a low performing student is administered hard items that are too difficult or a high performing student is administered easy items that are too easy). In these instances, the scale score CSEM may not be well-estimated.

After the 2014–15 test administration, this issue was discussed with the CAASPP Technical Advisory Group (TAG). With the TAG's approval, the average CSEMs at each scale score point were produced based on the 2014–15 administration results and reported in the 2014–15, 2015–16, and 2016–17 CAASPP Smarter Balanced technical reports (CDE, 2016, 2017, and 2018). The average CSEM reduced the level of uncertainty associated with individual CSEMs for each student. The CSEMs at the extreme ends of the ability continuum can be more accurately estimated on a cluster of students with similar abilities rather than one or two students with identical response patterns.

Based on the 2017–18 administration results, ETS recalculated the CSEM for each scale score point across grades and content areas. Those refreshed CSEMs are reported in chapter 8 of the 2017–18 technical report. The average CSEMs based on the 2017–18 data do not show big differences from the average CSEMs based on the 2014–15 data.

12.5. Accessibility

Like all CAASPP assessments, the Smarter Balanced Summative Assessments are administered using the test delivery system created by AIR for the Smarter Balanced assessments. As such, implementation of new online universal tools, designated supports, and accommodations are provided by Smarter Balanced (Smarter Balanced, 2018b) and aligned with the test delivery system.

The following changes will be implemented during the 2018–19 Smarter Balanced administration:

- Streamline will be reassigned as an embedded designated support.
- "Medical device" will be a new non-embedded designated support for all assessments.
- The Highlighter universal tool will be available in four colors.

- Scratch paper includes the use of non-embedded digital graph paper.
- Burmese is now among the embedded translation glossaries available as a designated support for the mathematics assessment.

References

California Department of Education. (2016). *CAASPP Smarter Balanced technical report, 2014–15 administration.* Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ca/documents/caaspp14techrpt.pdf

California Department of Education. (2017). *CAASPP Smarter Balanced technical report, 2015–16 administration.* Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ca/documents/sb16sbtechrpt.pdf

California Department of Education. (2018). *CAASPP Smarter Balanced technical report, 2016–17 administration.* Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ca/documents/sbac17techrpt.pdf

Smarter Balanced Assessment Consortium. (2016a). *Smarter Balanced Assessment Consortium: 2013–14 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from 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. Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf

Smarter Balanced Assessment Consortium. (2017). *Smarter Balanced Assessment Consortium: 2015–16 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/2015-16-summative-technical-report.pdf

Smarter Balanced Assessment Consortium. (2018a). *Smarter Balanced Assessment Consortium: 2016–17 technical report.* Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from http://portal.smarterbalanced.org/library/en/2016-17-summative-assessment-technical-report.pdf

Smarter Balanced Assessment Consortium. (2018b). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines.* Los Angeles, CA: Smarter Balanced Assessment Consortium and National Center on Educational Outcomes. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf