

California Department of Education Assessment Development and Administration Division



# California Assessment of Student Performance and Progress California Spanish Assessment Pilot Test Analysis Report

2017–18 Administration

Submitted May 14, 2018

Educational Testing Service



Contract No. CN150012

Table of Contents

[Chapter 1: Introduction 1](#_Toc179902650)

[1.1 Overview 1](#_Toc179902651)

[1.2 Purpose of the CSA Pilot Test 2](#_Toc179902652)

[1.3 Pilot Test Content 2](#_Toc179902653)

[1.4 Intended Population 2](#_Toc179902654)

[1.5 Testing Window and Times 2](#_Toc179902655)

[1.6 Overview of the Analysis Report 3](#_Toc179902656)

[References 4](#_Toc179902657)

[Chapter 2: Item Development and Assembly 5](#_Toc179902658)

[2.1 Overview 5](#_Toc179902659)

[2.2 Item Development Plan 5](#_Toc179902660)

[2.3 Item Types 5](#_Toc179902661)

[2.4 Test Assembly and Length 8](#_Toc179902662)

[References 9](#_Toc179902663)

[Chapter 3: Test Administration 10](#_Toc179902664)

[3.1 Pilot Test Administration 10](#_Toc179902665)

[3.2 Test Security and Confidentiality 10](#_Toc179902666)

[3.2.1. Office of Testing Integrity (OTI) 11](#_Toc179902667)

[3.2.2. Procedures to Maintain Standardization of Test Security 11](#_Toc179902668)

[3.2.3. Security of Electronic Files Using a Firewall 12](#_Toc179902669)

[3.2.4. Transfer of Scores via Secure Data Exchange 12](#_Toc179902670)

[3.2.5. Data Management 13](#_Toc179902671)

[3.2.6. Statistical Analysis 13](#_Toc179902672)

[3.2.7. Student Confidentiality 13](#_Toc179902673)

[3.2.8. Security and Test Administration Incident Reporting System (STAIRS) Process 13](#_Toc179902674)

[3.3 Accessibility Resources 14](#_Toc179902675)

[3.4 Participation 14](#_Toc179902676)

[3.5 Demographic Summaries 15](#_Toc179902677)

[References 16](#_Toc179902678)

[Appendix 3.A Demographic Summary 17](#_Toc179902679)

[Chapter 4: Summary Statistics for the 2017–18 Pilot Test Administration 20](#_Toc179902680)

[4.1 Overview 20](#_Toc179902681)

[4.2 Samples Used for Item Analyses 20](#_Toc179902682)

[4.3 Raw Scores Distributions 21](#_Toc179902683)

[4.4 Classical Item Analyses 25](#_Toc179902684)

[4.4.1. Item Difficulty Indices 25](#_Toc179902685)

[4.4.2. Item Discrimination 26](#_Toc179902686)

[4.4.3. Distribution of Item Scores 26](#_Toc179902687)

[4.4.4. Distractor Analyses 27](#_Toc179902688)

[4.4.5. Summary of Classical Item Analyses Flagging Criteria 27](#_Toc179902689)

[4.4.6. Classical Item Analyses Results Summary 28](#_Toc179902690)

[4.5 Response Time Analyses 32](#_Toc179902691)

[4.6 DIF Analyses 32](#_Toc179902692)

[4.6.1. DIF Procedure for Dichotomous Items 33](#_Toc179902693)

[4.6.2. DIF Procedure for Polytomous Items 34](#_Toc179902694)

[4.6.3. DIF Category Classification 34](#_Toc179902695)

[4.6.4. Items Exhibiting Significant DIF 35](#_Toc179902696)

[References 36](#_Toc179902697)

[Accessibility Information 37](#_Toc179902698)

[Alternative Text for Equation 4.1 37](#_Toc179902699)

[Alternative Text for Equation 4.2 37](#_Toc179902700)

[Alternative Text for Equation 4.3 37](#_Toc179902701)

[Alternative Text for Equation 4.4 37](#_Toc179902702)

[Alternative Text for Equation 4.5 37](#_Toc179902703)

[Alternative Text for Equation 4.6 37](#_Toc179902704)

[Appendix 4.A Classical Item Analyses 38](#_Toc179902705)

[Appendix 4.B Response Time Analysis 43](#_Toc179902706)

[Appendix 4.C Differential Item Functioning (DIF) 44](#_Toc179902707)

[Chapter 5: Constructed-Response (CR) Item Evaluation 45](#_Toc179902708)

[5.1 Evaluation of the Use of Tools in the Writing Prompt 45](#_Toc179902709)

[5.1.1. Use of the Cognitive Lab Form 46](#_Toc179902710)

[5.2 Analysis of the Use of Tools in the Writing Prompt 46](#_Toc179902711)

[Chapter 6: Discussions and Implications 47](#_Toc179902712)

[6.1 Summary of Findings 47](#_Toc179902713)

[6.2 Implications for the Field Test 47](#_Toc179902714)

Tables

[Acronyms and Initialisms Used in the CSA Pilot Test Analysis Report v](#_Toc179902715)

[Table 2.1 Number of Pilot Items Developed 5](#_Toc179902716)

[Table 2.2 Item Types for CSA Pilot Tests 6](#_Toc179902717)

[Table 2.3 ETS Pilot Test Form Review Process 8](#_Toc179902718)

[Table 2.4 Overview of Pilot Test Forms 8](#_Toc179902719)

[Table 3.1 Demographic Student Groups to Be Reported 15](#_Toc179902720)

[Table 3.A.1 Demographic Summary for Grade Band Three Through Five 17](#_Toc179902721)

[Table 3.A.2 Demographic Summary for Grade Band Six Through Eight 18](#_Toc179902722)

[Table 3.A.3 Demographic Summary for High School 19](#_Toc179902723)

[Table 4.1 Summary of Completion of the Pilot Test by Grade Band 20](#_Toc179902724)

[Table 4.2 Grades Three Through Five Raw Score Frequency Distribution 21](#_Toc179902725)

[Table 4.3 Grades Six Through Eight Raw Score Frequency Distribution 22](#_Toc179902726)

[Table 4.4 High School Raw Score Frequency Distribution 23](#_Toc179902727)

[Table 4.5 Summary Statistics of the Raw Scores by Grade Band 24](#_Toc179902728)

[Table 4.6 Summary Statistics of the Raw Scores per Grade Level for Grades Three Through Five 24](#_Toc179902729)

[Table 4.7 Summary Statistics of the Raw Scores per Grade Level for Grades Six Through Eight 24](#_Toc179902730)

[Table 4.8 Summary Statistics of the Raw Scores per Grade Level for High School 24](#_Toc179902731)

[Table 4.9 Item Difficulty and Item-Total Correlation by Grade Band 28](#_Toc179902732)

[Table 4.10 Summary of the Classical Item Statistics by Item Type for Grades Three Through Five 29](#_Toc179902733)

[Table 4.11 Summary of the Classical Item Statistics by Item Type for Grades Six Through Eight 29](#_Toc179902734)

[Table 4.12 Summary of the Classical Item Statistics by Item Type for High School 30](#_Toc179902735)

[Table 4.13 Summary of the Classical Item Statistics by Item Type and Response Type for Grades Three Through Five 30](#_Toc179902736)

[Table 4.14 Summary of the Classical Item Statistics by Item Type and Response Type for Grades Six Through Eight 31](#_Toc179902737)

[Table 4.15 Summary of the Classical Item Statistics by Item Type and Response Type for High School 31](#_Toc179902738)

[Table 4.16 DIF Categories for Dichotomous Items 34](#_Toc179902739)

[Table 4.17 DIF Categories for Polytomous Items 35](#_Toc179902740)

[Table 4.A.1. Dichotomous Items Statistics by Grade Band Three Through Five 38](#_Toc179902741)

[Table 4.A.2. Dichotomous Items Statistics by Grade Band Six Through Eight 39](#_Toc179902742)

[Table 4.A.3. Dichotomous Items Statistics by High School 40](#_Toc179902743)

[Table 4.A.4. Grades Three Through Five Polytomous Items Statistics 41](#_Toc179902744)

[Table 4.A.5. Grades Six Through Eight Polytomous Items Statistics 41](#_Toc179902745)

[Table 4.A.6. High School Polytomous Items Statistics 42](#_Toc179902746)

[Table 4.B.1 Total Testing Time (in Minutes) at Each Raw Score Interval 43](#_Toc179902747)

[Table 4.C.1 Gender DIF Classifications Summary by Grade Band 44](#_Toc179902748)

[Table 4.C.2 Items Exhibiting Significant DIF by Student Group 44](#_Toc179902749)

[Table 5.1 Summary of the Use of Spanish Characters in Pilot Constructed Responses 45](#_Toc179902750)

[Table 5.2 Summary of the Use of Spanish Special Characters in *¡Está fría la mañana!* in Cognitive Lab Constructed Responses 46](#_Toc179902751)

Acronyms and Initialisms Used in the CSA Pilot Test Analysis Report

| Abbreviation | Term |
| --- | --- |
| AIS | average item score |
| CAAs | California Alternate Assessments |
| CAASPP | California Assessment of Student Performance and Progress |
| CCR | California Code of Regulations |
| CDE | California Department of Education |
| CR | constructed response |
| CSA | California Spanish Assessment |
| DIF | differential item functioning |
| EL | English learner |
| ELA | English language arts/literacy |
| ETS | Educational Testing Service |
| IA | item analyses |
| IBIS | Item Banking Information System |
| ID | item identification |
| LEA | local educational agency |
| MC | multiple choice |
| MH | Mantel-Haenszel |
| OTI | Office of Testing Integrity |
| SD | standard deviation |
| SFTP | secure file transfer protocol |
| SMD | standardized mean difference |
| SS | Single select |
| STS | Standards-based Tests in Spanish |
| TE | technology enhanced |
| TOMS | Test Operations Management System |
| USC | United States Code |
| ZN | Zone |

**This page is intentionally left blank.**

## Introduction

### Overview

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.

During the 2017–18 administration, the overall CAASPP System had the following components:

* Smarter Balanced assessments and tools:
* Summative Assessments—Online assessments for English language arts/literacy (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 and 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
* CAA for Science second year pilot test in grades five, eight, and high school
* California Science Test field test in grades five, eight, and high school
* A primary language assessment, the Standards-based Tests in Spanish (STS) for Reading/‌Language Arts, in grades two through eleven (optional for all students )

As part of the CAASPP System of assessments, the California Spanish Assessment (CSA) is being developed as an optional assessment that will replace the STS. This new computer-based assessment for students in grades three through eight and high school is designed to measure a student’s Spanish skills in reading, writing mechanics, and listening for the purpose of:

* Providing student-level data in Spanish competency;
* Providing aggregate data that may be used for evaluating the implementation of Spanish language arts programs at the local level;
* Providing a high school measure suitable to be used, in part, for the California State Seal of Biliteracy.

Development of the CSA started in September 2016 with the State Board of Education’s approval of the high-level test design. Following item development and reviews with California educators, the CSA pilot test was administered to select LEAs during the fall of the 2017–18 administration.

### Purpose of the CSA Pilot Test

The purpose of the CSA pilot test was to evaluate the new computer-based CSA reading, writing mechanics, and listening items. The focus of the CSA pilot test was on the functionality of item types in the online test delivery system. Student responses to items were collected and reviewed to examine whether students could understand and respond as expected. In addition, cognitive labs were conducted with students in a one-on-one administration to gather additional feedback.

The CSA pilot test was intended to assess item performance and not student performance. Although data were collected at the item level, student scores were not reported for the CSA pilot test.

### Pilot Test Content

Educational Testing Service (ETS) assessment specialists developed items for the CSA pilot test that were aligned with the California *Common Core State Standards* en Español and consistent with the goals of California’s testing program. Items were written by ETS assessment developers, trained item writers familiar with assessment development in Spanish and specifically trained for the CSA, as well as California educators who received item writer training. All items were reviewed by ETS content and editorial staff, the California Department of Education, and a review panel comprised of California educators.

There were three general fixed forms administered in three grade spans (i.e., grades three through five, six through eight, and high school). Each grade span assessed had one general fixed form, where students were given the same questions regardless of their responses or ability.

Each test form consisted of three stand-alone items and three passages containing six to nine items for each passage. Overall, there were 26 items per form including constructed-response (CR) prompts.

### Intended Population

The CSA pilot test was administered to 2,268 students within a selected number of local educational agencies (LEAs). The assessed population for the CSA pilot test was all students in grades four through eight, and high school who receive instruction in Spanish in California, and who seek a measure that recognizes their Spanish-specific reading, writing mechanics, and listening skills.

### Testing Window and Times

The CSA pilot test was administered during a testing window from September 18, 2017, to October 6, 2017, and the cognitive labs were conducted from October 3 through October 5, 2017, in a one-on-one administration to gather additional feedback.

Similar to other CAASPP assessments, the CSA pilot was untimed for test takers. A student could take the CSA pilot test within the testing window over as many days as required to meet a student’s needs (*California Code of Regulations,* Title 5*,* Education, Division 1, Chapter 2, Subchapter 3.75, Article 2*,* Section 855[a][3]).

### Overview of the Analysis Report

This analysis report addresses the characteristics of the CSA pilot test administered in fall 2017. It contains five additional chapters as follows:

* Chapter 2 discusses the detailed procedures of item development, item review, and pilot test assembly for the 2017–18 administration. In particular, new item types and features that differ from traditional item types are described.
* Chapter 3 describes the details of administering the CSA pilot forms, as well as the LEA participation and demographic summaries.
* Chapter 4 summarizes the results of the analyses for the CSA pilot administration, including classical item analyses, test completion analyses, response time analyses, and differential item functioning analyses.
* Chapter 5 summarizes the evaluation of responses that were gathered from the CR items in the CSA pilot test forms.
* Chapter 6 summarizes the findings from the 2017–18 CSA pilot test administration and discusses the implications for the 2018–19 CSA field test administration.

### Reference

California Department of Education. *California* *Code of Regulations, Title 5,* Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, § 855. Sacramento, CA: California Department of Education.

## Item Development and Assembly

This chapter discusses the detailed procedures of item development and pilot test assembly for the California Spanish Assessment (CSA) pilot administration. In particular, new item types and features that differ from traditional item types are described.

### Overview

Educational Testing Service (ETS) developed 115 pilot test items across the three grade bands (i.e., grades three through five, six through eight, and high school) and delivered them to the California Department of Education (CDE) via the ETS Item Banking Information System (IBIS). The items developed were designed to be engaging to the student population and represented a wide variety of item types. All items for the CSA pilot tests were developed in accordance with the *ETS Standards for Quality and Fairness* (2014) across all phases of item and test development. While under initial development, the assessment materials, including items, passages, constructed-response (CR) prompts, and listening stimuli, were kept on password-protected ETS computers and secure internal network drives. Audio recordings were produced as electronic audio files and delivered to the CDE for review. All secure documents needed for CDE review that were not available in IBIS were delivered to the CDE via the Tumbleweed secure file transfer protocol server.

### Item Development Plan

The initial item development plan for the CSA pilot was to develop sufficient items to observe student responses to individual item types based on stimuli students listen to and read. This constituted a sample of the new item types to appear on the CSA. It was not necessary for these items to reflect the distribution of domains in the blueprint. Table 2.1 shows the number of items developed in each of the domains of reading, writing mechanics, and listening. The total number of machine-scorable items developed (115) is greater than the number to be administered (75) because overage was built in. ETS develops overage to account for the potential rejection of items during item review meetings. If item review meeting participants determine that certain items are not appropriate for the pilot test, the overage ensures that the minimum item counts for the pilot test forms will be satisfied.

Table 2. Number of Pilot Items Developed

|  |  |
| --- | --- |
| **Domain** | **Number** |
| Listening | 18 |
| Reading | 36 |
| Writing Mechanics | 61 |
| **Total Number of Items** | **115** |

### Item Types

ETS developed a variety of technology-enhanced (TE) item types that required the student to respond to a question in different ways from typical selected-response items. In addition to TE items, CR items were also used in the assessment. Items may contain a stimulus (e.g., a passage, audio, or image).

For TE items, they require the students to respond by typing an answer, completing a graph, dragging a response to a designated area, using a drop-down box selection, or selecting multiple areas (also known as “hot spots”) in a graphic. All TE item types were designed to be machine scorable. For CR items, detailed descriptions of the response evaluations are included in Chapter 5.

The following item types were included in the 2017–18 pilot tests:

* Multiple choice (MC) (single select and multiple select)
* Zone (single select and multiple select)
* Inline choice list (Single select and multiple select)
* Numeric
* Grid (single select and multiple select)
* Match (single select and multiple select)
* Composite
* Extended text

Table 2.2 provides a detailed description of the items used in the pilot. TE items are identified with an asterisk (\*) in the table.

Table 2. Item Types for CSA Pilot Tests

| Item Type | Response Type | Description |
| --- | --- | --- |
| MC | Multiple choice single select  | Item that generally consists of a stem and list of choices; test taker can select only one choice to respond. May also include a stimulus.  |
| MC | Multiple choice multiple select  | Item that generally consists of a stem and list of choices; test taker can select one or more choices to respond. May also include a stimulus.  |
| Hot Spot | Zone single select\* | Item where the answer choices are predefined “hotspots” on an image. When the test taker selects (clicks) on the spot, the selection is highlighted, shaded, or outlined in red. The test taker selects one zone to respond. |
| Hot Spot | Zone multiple select\*  | Item where the answer choices are predefined “hotspots” on an image. When the test taker selects (clicks) on the spot, the selection is highlighted, shaded, or outlined in red. The test taker selects two or more zones to respond. |
| MC | Inline choice list single select\* | The stem contains a single blank, and the test taker must fill the blank by selecting a choice from its corresponding choice list.  |
| MC | Inline choice list multiple select\* | The stem contains two or more blanks, and test taker must fill each blank by selecting a choice from the corresponding choice lists.  |
| Short CR | Numeric\*  | The test taker responds by filling in a single entry box with a numeric value. The entry box may be stand-alone, in line with text, or displayed on top of an image. |
| MC | Grid single select\* | Test taker responds by marking a single cell in a table grid. |
| Drag & Drop | Match single select\* | The test taker responds by dragging and dropping a single choice (“source”) into the appropriate location (“target”). For the CSA items, students do not drag items, they simply select (click) the source and then the target area, and the source snaps to the target area.There are four main varieties of this item type:* Target Table—text-based sources with targets arranged in table structure
* Target Passage—text-based sources with targets arranged in paragraphs of text
* Target Positions—text-based sources with targets arranged on top of an image
* Image Map—image-based sources, and both sources and targets are arranged on top of an image
 |
| Drag & Drop | Match multiple select\* | The test taker responds by dragging and dropping two or more choices (“sources”) into the appropriate locations (“targets”). For the CSA items, students do not drag items, they simply select (click) the source and then the target area, and the source snaps to the target area.There are four main varieties:1. Target Table—text-based sources with targets arranged in table structure
2. Target Passage—text-based sources with targets arranged in paragraphs of text
3. Target Positions—text-based sources with targets arranged on top of an image
4. Image Map—image-based sources, and both sources and targets are arranged on top of an image

These varieties allow for the following scenarios: * Exact matching (i.e., ordering)
* Sources correctly placed in multiple different targets
* Reuse sources
* Reuse targets
* Partial scoring
 |
| All item types except CR | Composite\* | The test taker completes multiple tasks based on a combination of machine-scored items. |
| Short CR and Extended CR | Extended Text\* | Extended Text consists of a stem in which the test taker must provide a written response, usually in the form of an essay. |

### Test Assembly and Length

Following the item review meeting, the CDE conducted three rounds of item review. ETS assessment specialists worked closely with the CDE to select items and assemble pilot test forms once the item review was complete.

The pilot test forms were assembled so that they covered a variety of item types, item difficulties, cognitive levels, and key distributions. The forms were evaluated via the ETS review process shown in table 2.3 and reviewed and approved by the CDE.

Table 2. ETS Pilot Test Form Review Process

| Steps | Task |
| --- | --- |
| 1. Test Assembly
 | Assessment specialists select test items that meet the specifications, are fair, and reflect appropriate content coverage. These items are collected in the item bank so they can be tracked as a unit. |
| 1. Senior Review
 | An assessment specialist with content-area expertise, who did not assemble the test, reviews all of the items and checks for content-related issues (e.g., incorrect keys, overlapping content, cueing of one item by another) and other concerns (e.g., confirming that the items match the test framework). The assessment specialist also verifies that the test meets content and statistical specifications. |
| 1. Senior Fresh-Eyes Review
 | Every new test form goes through a “fresh-eyes” review. During this review, a senior-level content expert, who has never encountered the form, reviews it carefully for any content errors that may have been missed during earlier stages of review. |
| 1. Certification
 | Once these reviews are completed and the test form is judged to be free from errors, ETS certifies the test form and sends it to be packaged for device delivery. |

Table 2.4 shows the total number of items needed per grade band to accommodate the pilot tests. These tests were not aligned with the CSA test blueprint. The estimated duration for the pilot test was 35 to 75 minutes per grade band.

Table 2. Overview of Pilot Test Forms

| Form Details | Grades 3–5 | Grades 6–8 | High School |
| --- | --- | --- | --- |
| Number of Forms | 1 | 1 | 1 |
| Number of Listening Passages per Form | 1 | 1 | 1 |
| Number of Reading Passages per Form | 1 | 1 | 1 |
| Number of Writing Passages per Form | 1 | 1 | 1 |
| Listening Passage-based Items | 6 | 6 | 6 |
| Reading Passage-based Items | 9 | 9 | 9 |
| Writing Passage-based Items | 7 | 7 | 7 |
| Number of Stand-alone Items per Form | 3 | 3 | 3 |
| Number of Writing CR Items per Form | 1 | 1 | 1 |
| Estimated Testing Time | 35–45 minutes | 35–45 minutes | 55–75 minutes |

### Reference

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

## Test Administration

### Pilot Test Administration

As the focus of the pilot test was to test the item types rather than the content, test forms were developed for grade bands. Additionally, because the California Spanish Assessment (CSA) is a summative assessment, participating students took the grade-band test that reflected the grade they were enrolled in for the 2016–17 school year. For example, students enrolled in grade six in fall 2017 took the grades three through five assessment and students enrolled in grade nine in fall 2017 took the grades six through eight assessment.

In accordance with the procedures for all online California Assessment of Student Performance and Progress (CAASPP) assessments, local educational agencies (LEAs) identified test administrators to administer the CSA pilot test and entered them into the Test Operations Management System (TOMS). Educational Testing Service (ETS) provided LEA staff with the appropriate training materials, such as test administration manuals and/or Webcasts, to ensure that the LEA staff and test administrators understood how to administer the computer-based CSA pilot test.

Once the pilot test administration window opened, each participating LEA/school locally determined which dates they wished to administer the pilot test. Students reported to the testing classroom or center and were provided a computer or testing device to take the assessment. The pilot test utilized the same secure browser and online testing platform as all of the CAASPP assessments. The students received initial direction from the test administrator as well as item-level directions as needed.

In addition to the computer administration, ETS conducted one-on-one cognitive labs in which a Spanish-speaking ETS researcher worked with a Spanish-speaking student at selected schools from San Juan Unified School District. ETS researchers employed a think-aloud approach followed by a retrospective cognitive interview to collect information about how students interpreted and responded to the items, and how they used the online tools or supports. A total of five bilingual researchers participated in this cognitive lab study.

### Test Security and Confidentiality

All tests within the CAASPP System are secure. For the CSA pilot test, every person who had access to test materials maintained the security and confidentiality of the tests. ETS’s internal Code of Ethics required that all test information, including tangible materials associated with the CSA pilot test (e.g., test questions and test results), confidential files, processes, and activities be kept secure. To ensure security for all tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI).

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 as follows:

* Test delivery
* Security of electronic files using a firewall
* Transfer of scores via secure data exchange
* Data management
* Statistical analysis
* Student confidentiality

All tests within the CAASPP System, as well as the confidentiality of student information, are 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 standard describes the measures intended to prevent potential test security incidents prior to testing and the actions that were taken to handle security incidents occurring during or after the testing window using the Security and Test Administration Incident Reporting System (STAIRS) process.

For the CSA pilot, every person who worked with the assessments and/or received testing information was responsible for maintaining the security and confidentiality of the tests, including California Department of Education (CDE) staff, ETS staff, ETS subcontractors, LEA assessment coordinators, school assessment coordinators, students, parents, teachers, and cooperative educational service agency staff. ETS’s Code of Ethics required that all test information, including tangible materials (e.g., test items), confidential files (e.g., those containing personally identifiable student information), and processes related to test administration (e.g., the configurations of secure servers) be kept secure. ETS has systems in place that maintained tight security for test items and test results, as well as for student data.

There were zero test security incidents that were reported for the CSA pilot test.

#### 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 the *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.

#### Procedures to Maintain Standardization of Test Security

Test security requires the accounting of all secure materials—including online summative test items and student data—before, during, and after each test administration. The LEA CAASPP coordinator is responsible for keeping all electronic 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.

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/or passages generated by the print-on-demand feature of the test delivery system (CDE, 2017a).

The following measures ensure the security of CAASPP System assessments:

* LEA CAASPP coordinators and test site coordinators must sign and submit a “CAASPP Test Security Agreement for LEA CAASPP coordinators and CAASPP test site coordinators” form to the California Technical Assistance Center before ETS can grant the coordinators access to TOMS. (California Code of Regulations, Title 5 [5 CCR], Education, Division 1, Chapter 2, Subchapter 3.75, Article 1, Section 859[a])
* Anyone having access to the testing materials must sign and submit 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 is the responsibility of every participant in the CAASPP System to immediately report any violation or suspected violation of test security or confidentiality. The CAASPP test site coordinator must report to the LEA CAASPP coordinator. The LEA CAASPP coordinator must report to the CDE within 24 hours of the incident. (5 *CCR*, Section 859[e])

#### Security of Electronic Files Using a Firewall

A firewall software is used at ETS to prevent unauthorized entry to files, e-mail, and other organization-specific information. All ETS data exchanges and internal e‑mail 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.

#### Transfer of Scores via Secure Data Exchange

Due to the confidential nature of test results, ETS uses secure file transfer protocol (SFTP) and encryption for all data file transfers, including student data files. 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.

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

Data are always transmitted to the SFTP server in an encrypted format; test data are never sent via e-mail. The SFTP server is used as a conduit for the transfer of files; secure test data are only temporarily stored on the shared SFTP server.

#### Data Management

ETS 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 these data are collected during operational testing. Only individuals with the appropriate credentials can access these 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.

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

All stored test content and student data are encrypted. Industry-standard secure protocols are used to transfer test content and student data from the ETS internal data center to any external systems. 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).

#### Statistical Analysis

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 extracts the data from the database and performs quality control procedures (e.g., the values of all variables are as expected) before passing files to the ETS statistical analysis group. 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.

#### Student Confidentiality

To meet 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. ETS takes every precaution to prevent any of this information from becoming public or being used for anything other than for testing and score reporting purposes. These procedures are applied to all documents in which student demographic data appear, such as technical reports.

#### 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, compromise the reliability and validity of test results (CDE, 2016c). 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 must verify that all test security and summative administration incidents are documented by filling out the secure STAIRS form for reporting, which contains selectable options to guide coordinators in their submittal. Incidents are then resolved when the LEA CAASPP coordinator or CAASPP test site coordinator either files 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 e-mail in response to the STAIRS form submittal.

The following types of STAIRS reports are 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 Statewide Student Identifier used (i.e., intentionally switched)
* Restoring a test that had been reset
* Student unable to review previous answers (i.e., 20-minute pause rule)

Prior to the pilot test administration, ETS and the CDE agreed that the following test security incidents would apply to the CSA pilot test:

* Administration Error
* Cheating/Accessing Unauthorized Devices
* Disruption or Technical Issue
* Exposing Secure Materials
* Incorrect Statewide Student Identifier Used
* Student Disruption

However, zero test security incidents were reported during the CSA pilot test.

### Accessibility Resources

The CSA pilot test offered commonly used accessibility resources universally available through the CAASPP online testing platform, where applicable for the tested construct. Some of these features could include a highlighter, the ability to mark an item for future review, and the ability to visually zoom the computer display in (making the display larger) or out (making the display smaller).

The following embedded universal tools were available to students during the CSA pilot test:

* Breaks;
* Digital notepad;
* Expandable passages;
* Expandable items;
* Highlighter;
* Keyboard navigation;
* Line reader;
* Mark for review;
* Strikethrough;
* Writing tools (e.g., bold, italic, bullets, undo/redo) (full write items);
* Zoom (in/out)

### Participation

The goal of the pilot recruitment was to select three or more LEAs that cover the North, Central, and South regions, and to select a minimum of 300 students per grade band to participate in the pilot test.

ETS used information available in the California Longitudinal Pupil Achievement Data System to put together an initial list of LEAs that serve high concentrations of Spanish speaking students, including students that are enrolled in California schools for fewer than 12 months. The CDE and its stakeholders reviewed the initial list and identified LEAs to recruit for pilot test participation.

As a result, a total of six LEAs agreed to participate in the CSA pilot test: Coachella Valley Unified, Palmdale Elementary, Porterville Unified, San Bernardino City Unified School District, Woodland Joint Unified, and San Juan Unified School District (cognitive labs).

### Demographic Summaries

The number and the percent of students for selected groups with completed test scores are provided starting on page 17 in table 3.A.1, table 3.A.2, and table 3.A.3 for each test per grade-band (i.e., grades three through five, six through eight, and high school). In the tables, students are grouped by demographic characteristics, including gender, ethnicity, English-language fluency, economic status (disadvantaged or not), special education services status, and length of enrollment in U.S. schools, as shown in table 3.1.

Table 3. Demographic Student Groups to Be Reported

| Student Group | Definition |
| --- | --- |
| Gender | * Male
* Female
 |
| Ethnicity | * American Indian or Alaska Native
* Asian
* Native Hawaiian or Other Pacific Islander
* Filipino
* Hispanic or Latino
* Black or African American
* White
* Two or more races
 |
| English-language Fluency | * English only
* Initially fluent English proficient
* English learner (EL)
* 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
 |
| Enrollment in U.S. Schools | * Less than 12 months
* 12 months or more
 |

### 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. (2017d). *Security incidents and appeals procedure guide, 2016–17 administration.* Sacramento, CA: California Department of Education.

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

### Appendix 3.A Demographic Summary

Note: The percentages of student groups may not sum to 100 due to rounding.

Table 3.A.1 Demographic Summary for Grade Band Three Through Five

| Demographic | Number Tested | Percent |
| --- | --- | --- |
| Students completed the test | 1,193 | 100.00 |
| Male | 574 | 48.11 |
| Female | 619 | 51.89 |
| American Indian or Alaska Native | 4 | 0.34 |
| Asian  | 0 | 0.00 |
| Native Hawaiian or Other Pacific Islander | 2 | 0.17 |
| Filipino | 1 | 0.08 |
| Hispanic or Latino | 1,096 | 91.87 |
| Black or African American | 36 | 3.02 |
| White | 46 | 3.86 |
| Two or more races | 3 | 0.25 |
| Unknown | 5 | 0.42 |
| English only | 425 | 35.62 |
| Initially fluent English proficient | 28 | 2.35 |
| English learner (EL) | 359 | 30.09 |
| Reclassified fluent English proficient | 373 | 31.27 |
| To be determined | 1 | 0.08 |
| English proficiency unknown | 7 | 0.59 |
| No special education services | 1,128 | 94.55 |
| Special education services | 65 | 5.45 |
| Not economically disadvantaged | 728 | 61.02 |
| Economically disadvantaged | 465 | 38.98 |
| In U.S. schools < 12 months | 15 | 1.26 |
| In U.S. schools ≥ 12 months | 1,178 | 98.74 |

Table 3.A.2 Demographic Summary for Grade Band Six Through Eight

| Demographic | Number Tested | Percent |
| --- | --- | --- |
| Students completed the test | 585 | 100.00 |
| Male | 294 | 50.26 |
| Female | 291 | 49.74 |
| American Indian or Alaska Native | 2 | 0.34 |
| Asian  | 0 | 0.00 |
| Native Hawaiian or Other Pacific Islander | 1 | 0.17 |
| Filipino | 2 | 0.34 |
| Hispanic or Latino | 543 | 92.82 |
| Black or African American | 5 | 0.85 |
| White | 24 | 4.10 |
| Two or more races | 1 | 0.17 |
| Unknown | 7 | 1.20 |
| English only | 122 | 20.85 |
| Initially fluent English proficient | 10 | 1.71 |
| EL | 209 | 35.73 |
| Reclassified fluent English proficient | 228 | 38.97 |
| To be determined | 0 | 0.00 |
| English proficiency unknown | 16 | 2.74 |
| No special education services | 548 | 93.68 |
| Special education services | 37 | 6.32 |
| Not economically disadvantaged | 257 | 43.93 |
| Economically disadvantaged | 328 | 56.07 |
| In U.S. schools < 12 months | 19 | 3.25 |
| In U.S. schools ≥ 12 months | 566 | 96.75 |

Table 3.A.3 Demographic Summary for High School

| Demographic | Number Tested | Percent |
| --- | --- | --- |
| Students completed the test | 476 | 100.00 |
| Male | 201 | 42.23 |
| Female | 275 | 57.77 |
| American Indian or Alaska Native | 0 | 0.00 |
| Asian  | 1 | 0.21 |
| Native Hawaiian or Other Pacific Islander | 0 | 0.00 |
| Filipino | 1 | 0.21 |
| Hispanic or Latino | 466 | 97.90 |
| Black or African American | 0 | 0.00 |
| White | 4 | 0.84 |
| Two or more races | 1 | 0.21 |
| Unknown | 3 | 0.63 |
| English only | 33 | 6.93 |
| Initially fluent English proficient | 13 | 2.73 |
| EL | 54 | 11.34 |
| Reclassified fluent English proficient | 369 | 77.52 |
| To be determined | 0 | 0.00 |
| English proficiency unknown | 7 | 1.47 |
| No special education services | 466 | 97.90 |
| Special education services | 10 | 2.10 |
| Not economically disadvantaged | 221 | 46.43 |
| Economically disadvantaged | 255 | 53.57 |
| In U.S. schools < 12 months | 5 | 1.05 |
| In U.S. schools ≥ 12 months | 471 | 98.95 |

## Summary Statistics for the 2017–18 Pilot Test Administration

### Overview

This chapter summarizes the results of the item- and test-level analyses from the 2017–18 California Spanish Assessment (CSA) pilot test administration. Analyses include the following:

* Raw score distributions on the test level
* Classical item analyses
* Response time analyses
* Differential item functioning (DIF) analyses

### Samples Used for Item Analyses

The pilot test was delivered to all eligible students in grades four through twelve. Because the pilot test was administered in fall of 2017, the fourth graders in fall are generally equivalent to the third graders in spring. Therefore, students in grades four through twelve are eligible for the CSA pilot test instead of students starting in grade three. Educational Testing Service (ETS) administered all items to the participating students per grade band (i.e., grades three through five, grades six through eight, and high school).

Upon the completion of the pilot test window, it was found that the participating high schools did not test any ninth grade students. Thus, there were no ninth grade students that took the grades six through eight test. During the recruiting process, some of the high school local educational agencies (LEAs) mentioned that it would be difficult to test ninth grade students because most of them would have just begun their first year of a Spanish course in August, and they would be unsure of what level of Spanish instruction they had received in the previous year. Other LEAs mentioned that they would not test their ninth grade students because of the coinciding English Language Proficiency Assessments for California (ELPAC) initial assessment that would have to be administered to their ELs.

All analyses in this report were performed using only test takers with complete responses for all items. Table 4.1 includes the number of students taking each test. For each grade band, registered students exhibited very high pilot test completion rates.

Table 4. Summary of Completion of the Pilot Test by Grade Band

| Grade Band | Total Number of Students Completed the Test | Percent Completion | Total Number of Registered Test Takers |
| --- | --- | --- | --- |
| 3–5 | 1,193 | 99.67 | 1,197 |
| 6–8 | 585 | 98.82 | 592 |
| High School | 476 | 99.37 | 479 |

### Raw Scores Distributions

For all of the CSA pilot tests, the total test raw score is defined as total points obtained for machine-scorable items. Table 4.2, table 4.3, and table 4.4 show the raw score distributions of each test form.

Table 4. Grades Three Through Five Raw Score Frequency Distribution

| Raw Score | Freq. | Percent | Cumulative Percent |
| --- | --- | --- | --- |
| 0 | 0 | 0.00 | 0.00 |
| 1 | 0 | 0.00 | 0.00 |
| 2 | 0 | 0.00 | 0.00 |
| 3 | 1 | 0.08 | 0.08 |
| 4 | 2 | 0.17 | 0.25 |
| 5 | 6 | 0.50 | 0.75 |
| 6 | 7 | 0.59 | 1.34 |
| 7 | 9 | 0.75 | 2.10 |
| 8 | 15 | 1.26 | 3.35 |
| 9 | 14 | 1.17 | 4.53 |
| 10 | 30 | 2.51 | 7.04 |
| 11 | 39 | 3.27 | 10.31 |
| 12 | 46 | 3.86 | 14.17 |
| 13 | 36 | 3.02 | 17.18 |
| 14 | 51 | 4.27 | 21.46 |
| 15 | 52 | 4.36 | 25.82 |
| 16 | 64 | 5.36 | 31.18 |
| 17 | 80 | 6.71 | 37.89 |
| 18 | 73 | 6.12 | 44.01 |
| 19 | 63 | 5.28 | 49.29 |
| 20 | 66 | 5.53 | 54.82 |
| 21 | 67 | 5.62 | 60.44 |
| 22 | 74 | 6.20 | 66.64 |
| 23 | 55 | 4.61 | 71.25 |
| 24 | 68 | 5.70 | 76.95 |
| 25 | 71 | 5.95 | 82.90 |
| 26 | 68 | 5.70 | 88.60 |
| 27 | 52 | 4.36 | 92.96 |
| 28 | 26 | 2.18 | 95.14 |
| 29 | 29 | 2.43 | 97.57 |
| 30 | 15 | 1.26 | 98.83 |
| 31 | 4 | 0.34 | 99.16 |
| 32 | 10 | 0.84 | 100.00 |
| 33 | 0 | 0.00 | 100.00 |

Table 4. Grades Six Through Eight Raw Score Frequency Distribution

| Raw Score | Freq. | Percent | Cumulative Percent |
| --- | --- | --- | --- |
| 0 | 0 | 0.00 | 0.00 |
| 1 | 0 | 0.00 | 0.00 |
| 2 | 0 | 0.00 | 0.00 |
| 3 | 2 | 0.34 | 0.34 |
| 4 | 2 | 0.34 | 0.68 |
| 5 | 7 | 1.20 | 1.88 |
| 6 | 11 | 1.88 | 3.76 |
| 7 | 14 | 2.39 | 6.15 |
| 8 | 18 | 3.08 | 9.23 |
| 9 | 36 | 6.15 | 15.38 |
| 10 | 40 | 6.84 | 22.22 |
| 11 | 33 | 5.64 | 27.86 |
| 12 | 34 | 5.81 | 33.68 |
| 13 | 29 | 4.96 | 38.63 |
| 14 | 45 | 7.69 | 46.32 |
| 15 | 40 | 6.84 | 53.16 |
| 16 | 34 | 5.81 | 58.97 |
| 17 | 36 | 6.15 | 65.13 |
| 18 | 34 | 5.81 | 70.94 |
| 19 | 26 | 4.44 | 75.38 |
| 20 | 33 | 5.64 | 81.03 |
| 21 | 34 | 5.81 | 86.84 |
| 22 | 22 | 3.76 | 90.60 |
| 23 | 21 | 3.59 | 94.19 |
| 24 | 11 | 1.88 | 96.07 |
| 25 | 9 | 1.54 | 97.61 |
| 26 | 8 | 1.37 | 98.97 |
| 27 | 2 | 0.34 | 99.32 |
| 28 | 2 | 0.34 | 99.66 |
| 29 | 2 | 0.34 | 100.00 |
| 30 | 0 | 0.00 | 100.00 |
| 31 | 0 | 0.00 | 100.00 |
| 32 | 0 | 0.00 | 100.00 |

Table 4. High School Raw Score Frequency Distribution

| Raw Score | Freq. | Percent | Cumulative Percent |
| --- | --- | --- | --- |
| 0 | 0 | 0.00 | 0.00 |
| 1 | 0 | 0.00 | 0.00 |
| 2 | 0 | 0.00 | 0.00 |
| 3 | 1 | 0.21 | 0.21 |
| 4 | 0 | 0.00 | 0.21 |
| 5 | 4 | 0.84 | 1.05 |
| 6 | 4 | 0.84 | 1.89 |
| 7 | 6 | 1.26 | 3.15 |
| 8 | 13 | 2.73 | 5.88 |
| 9 | 21 | 4.41 | 10.29 |
| 10 | 17 | 3.57 | 13.87 |
| 11 | 27 | 5.67 | 19.54 |
| 12 | 27 | 5.67 | 25.21 |
| 13 | 32 | 6.72 | 31.93 |
| 14 | 37 | 7.77 | 39.71 |
| 15 | 43 | 9.03 | 48.74 |
| 16 | 30 | 6.30 | 55.04 |
| 17 | 37 | 7.77 | 62.82 |
| 18 | 35 | 7.35 | 70.17 |
| 19 | 38 | 7.98 | 78.15 |
| 20 | 34 | 7.14 | 85.29 |
| 21 | 22 | 4.62 | 89.92 |
| 22 | 24 | 5.04 | 94.96 |
| 23 | 11 | 2.31 | 97.27 |
| 24 | 9 | 1.89 | 99.16 |
| 25 | 1 | 0.21 | 99.37 |
| 26 | 2 | 0.42 | 99.79 |
| 27 | 0 | 0.00 | 99.79 |
| 28 | 1 | 0.21 | 100.00 |
| 29 | 0 | 0.00 | 100.00 |
| 30 | 0 | 0.00 | 100.00 |
| 31 | 0 | 0.00 | 100.00 |

In addition, table 4.5 presents the number of machine-scorable items, the number of students tested, and the summary statistics of each grade-band form. Table 4.6, table 4.7, and table 4.8 present the summary statistics of the machine-scorable items per grade level for each grade-band form. The summary statistics presented include the mean and standard deviation (SD). Note that in table 4.6, table 4.7, and table 4.8, the enrolled grades in 2016–‍17 year are used. For grades three through five, students in higher grade levels outperformed students in lower grade levels. For grades six through eight, students in grade seven outperformed students in grade six. There were no eighth grade students—ninth-graders in the fall of 2017—that participated in the pilot test. For high school, very few students in grade eleven participated. Overall, grade ten students outperformed grade nine students.

Table 4. Summary Statistics of the Raw Scores by Grade Band

| Grade Band | N Items[[1]](#footnote-2) | N Points | N Students | Mean | Mean as % of Total | SD |
| --- | --- | --- | --- | --- | --- | --- |
| 3–5 | 25 | 33 | 1,193 | 19.44 | 58.91 | 5.85 |
| 6–8 | 25 | 32 | 585 | 15.26 | 47.69 | 5.28 |
| High School | 25 | 31 | 476 | 15.67 | 50.54 | 4.48 |

Table 4. Summary Statistics of the Raw Scores per Grade Level for Grades Three Through Five

| Grade\* | N Students | Mean | Mean as % of Total | SD |
| --- | --- | --- | --- | --- |
| 3 | 418 | 16.71 | 50.64 | 5.63 |
| 4 | 375 | 19.79 | 59.97 | 5.40 |
| 5 | 400 | 21.97 | 66.58 | 5.23 |
| Overall | 1,193 | 19.44 | 58.91 | 5.85 |

Table 4. Summary Statistics of the Raw Scores per Grade Level for Grades Six Through Eight

| Grade\* | N Students | Mean | Mean as % of Total | SD |
| --- | --- | --- | --- | --- |
| 6 | 264 | 14.18 | 44.31 | 4.90 |
| 7 | 321 | 16.15 | 50.47 | 5.42 |
| Overall | 585 | 15.26 | 47.69 | 5.28 |

Table 4. Summary Statistics of the Raw Scores per Grade Level for High School

| Grade\* | N Students | Mean | Mean as % of Total | SD |
| --- | --- | --- | --- | --- |
| 9 | 218 | 14.72 | 47.48 | 4.48 |
| 10 | 230 | 16.56 | 53.42 | 4.39 |
| 11 | 28 | 15.68 | 50.58 | 3.88 |
| Overall | 476 | 15.67 | 50.54 | 4.48 |

### Classical Item Analyses

For all pilot-tested items that are machine scorable, classical item analyses are used to evaluate the item performance with respect to item difficulty, item discrimination, and distractor analysis. In addition, the distributions of score categories on key-based selected-response items and rule-based machine-scored items are also included in the classical item analyses results. Lastly, the associated flagging rules of these statistics are used to identify items that are not performing as expected. Items scored as one (correct) or zero (incorrect) are referred to as dichotomous items. Items with maximum score greater than one are called polytomous items.

#### Item Difficulty Indices

For dichotomous items, item difficulty is indicated by its *p*-value, which is the proportion of students who answer the item correctly. The range of *p*-values is from 0.00 to 1.00. Items with high *p*-values are easier items; those with low *p*-values are more difficult items. Dichotomous items are flagged for review if they have *p*-values above 0.95 (i.e., too easy) or below 0.10 (i.e., too difficult).

The formula for the *p*-value for a dichotomous item is presented in equation 4.1. *Refer to the* [*Alternative Text for Equation 4.1*](#_Alternative_Text_for) *for a description of this equation.*

 (4.1)

where,

 is the number of students that answered item *i* correctly, and

 is the total number of students who were presented with item *i*.

For polytomous items, the difficulty is indicated by the average item score (AIS). The AIS can range from 0.00 to the maximum total possible points for an item. Desired AIS values for polytomous items generally fall within the range of 30 percent to 70 percent of the maximum obtainable item score; items with values outside this range are flagged for review. To facilitate the interpretation, the AIS values for polytomous items are often expressed as the proportion of the maximum possible score, which are equivalent to the *p-*values of dichotomous items.

The formula for the *p-*value for a polytomous item is presented in equation 4.2. *Refer to the* [*Alternative Text for Equation 4.2*](#_Alternative_Text_for_1) *for a description of this equation.*

 (4.2)

where,

 is the score assigned for a given polytomous item *i* and student *j*,

*Max (Xi)* is the maximum possible score for item *i*, and

 is the total number of students who were presented with item *i*.

#### Item Discrimination

Item discrimination evaluates how well an item distinguishes between low and high ability students, and is generally referred to as “item-total correlation.” The expectation is that high ability students will outperform low ability students on a good discriminating item. For the CSA pilot test, we should interpret with caution because the total score is not aligned to the blueprint therefore it is not well defined. In general, the item-total correlation statistic describes the relationship between students’ performance on a specific item and their performance on the total test and is calculated as the correlation coefficient between the item score and total score. A relatively high item-total correlation coefficient value is desired, as it indicates that students with higher scores on the overall test tended to perform better on the item. In general, item-total correlation ranges from -1.0 (for a perfect negative relationship) to 1.0 (for a perfect positive relationship). A negative item-total correlation typically signifies a problem with the item, as the students with higher scores on the overall test generally are getting the item wrong or a low score and the students with lower scores on the overall test are getting the item correct or a high score.

Specifically, the polyserial correlation is used as the index of item-total correlation for both polytomous and dichotomous items. Statistically, polyserial correlations are based on a polyserial regression model (Olsson, 1979; Drasgow, 1988), which assumes that performance on an item is determined by the examinee’s position on an underlying latent variable that is normally distributed at a given criterion score level. Based on this approach, the polyserial correlation can be estimated as presented in equation 4.3. *Refer to the* [*Alternative Text for Equation 4.3*](#_Alternative_Text_for_2) *for a description of this equation.*

 (4.3)

where,

 is the standard deviation of the students’ total test scores as a criterion score, and

 is the item parameter to be estimated from the data, with the estimate denoted as , using maximum likelihood estimation. It is a regression coefficient (slope) for predicting the continuous version of an item score onto the continuous version of the total score.

There are as many regressions as there are boundaries between scores with all sharing a common slope, *β*. For a polytomous item, there are *m-*1 regressions, where *m* is the number of score points on the item. Beta (*β*) is the slope for all *m*-1 regressions. Desired polyserial correlation values of items are positive and larger than 0.20.

#### Distribution of Item Scores

For polytomous items, examination of the distribution of scores helps evaluate how well the item functions. If no students achieved the highest possible score, the item may not be functioning as expected. The item may be confusing, poorly worded, or just unexpectedly difficult, or students may not have had an opportunity to learn the content.

#### Distractor Analyses

The quality of distractors is an important component of an item’s overall quality. Distractors should be clearly incorrect, but at the same time plausible and attractive to lower-ability students. For the CSA pilot test, the following distractor analyses were conducted to evaluate the quality of distractors.

* The percentage of students at each response option is calculated for the highest-performing 20 percent of students. If the percentage of students who selected a distractor is greater than the percentage of students who selected a key for the high-performing group, the item should be examined to determine if it has multiple correct answers or the wrong key (i.e., the item is miskeyed).
* The polyserial correlation is calculated for each response option. While the key should have a positive polyserial correlation with the criterion score, the distractors should exhibit negative polyserial correlations (i.e., lower ability students would likely choose the distractors, while the higher ability students would not). Items with positive distractor-total correlations were flagged for review, as these items may have multiple correct answers, be miskeyed, or have other content issues.

#### Summary of Classical Item Analyses Flagging Criteria

In summary, items are flagged for review if the item analysis yields any of the following results:

* **Difficulty flag** indicates extreme values of the proportion-correct (for dichotomous items) or the proportion of the possible maximum points earned (for polytomous items).
* A value less than 0.10 for dichotomous items or 0.30 for polytomous items suggests that the item might be too difficult.
* A value greater than 0.95 for dichotomous items or 0.70 for polytomous items suggests that the item might be too easy.
* **Discrimination flag** indicates that the item does not discriminate effectively between high- and low-ability students. Items with a polyserial correlation less than 0.20 are flagged.
* **Omit flag** is set for dichotomous items with nonresponse rates greater than five percent and polytomous items with nonresponse rates greater than 20 percent.
* **Distractor flag** is used for any distractors having positive correlation with the criterion score.
* **Miskey flag** is used for multiple-choice (MC) items when more of the high-ability examinee group—the top 20 percent of examinees on the total test—choose any distractor rather than choosing the response keyed as correct.
* **Underrepresented score point flag** is used for any item that has less than three percent of the students at any score level.

ETS’s psychometric staff and assessment development staff carefully reviewed each of the flagged items during and at the end of the item analyses. All confirmed flagged items will be summarized and then sent to the California Department of Education (CDE).

#### Classical Item Analyses Results Summary

The summary statistics of the classical item analyses, which include the means and ranges of overall item difficulty and item-total correlation for all machine-scorable items, is presented in table 4.9 for each grade band. Table 4.10, table 4.11, and table 4.12 present the summary statistics of the classical item analyses by item type for each grade band. Table 4.13, table 4.14, and table 4.15 present the summary statistics of the classical item analyses by item type and response type for each grade band. The results of classical item analyses for each item in each grade band is presented in table 4.A.1 for dichotomous items and table 4.A.2, table 4.A.3, and table 4.A.4 for polytomous items in Appendix 4.A. The item statistics, including the *p*-value for dichotomous items or AIS for polytomous items, polyserial correlation, omit rates, and the distribution of score points on each polytomous item are listed in those tables.

For all pilot tests within each grade band, the average item difficulties and discriminations were reasonable. The following is true for grades three through five:

* The most difficult items were of the Match item type.
* The easiest items were of the Numeric item type.
* The most discriminating items were of the Zones item type.
* The least discriminating items were of the Grid item type.

The following is true for grades six through eight:

* The most difficult items were of the Match item type.
* The easiest items were of the Numeric, Zone, and Grid item types.
* The most discriminating items were of the Match item type.
* The least discriminating items were of the InlineChoice item type.

Finally, the following is true for high school:

* The most difficult items were of the Numeric item type.
* The easiest items were of the Composite item type.
* The most discriminating items were of the InlineChoice item type.
* The least discriminating items were of the Numeric item type.

Table 4. Item Difficulty and Item-Total Correlation by Grade Band

| Grade Band | No. of Items | No. of Examinees | Mean Difficulty | Mean Item-Total Corr. | Min. Difficulty | Min. Item-Total Corr. | Max. Difficulty | Max. Item-Total Corr. |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3–5 | 25 | 1,193 | 0.57 | 0.53 | 0.22 | 0.22 | 0.86 | 0.77 |
| 6–8 | 25 | 585 | 0.47 | 0.49 | 0.11 | 0.14 | 0.77 | 0.75 |
| High School | 25 | 476 | 0.47 | 0.42 | 0.07 | 0.11 | 0.80 | 0.71 |

Table 4. Summary of the Classical Item Statistics by Item Type for Grades Three Through Five

| Item Type | No. of Items | No. of Examinees | Mean *p*-value | Mean Polyserial | Min. *p*-value | Min. Polyserial | Max. *p*-value | Max. Polyserial |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grid | 7 | 1,193 | 0.50 | 0.41 | 0.22 | 0.22 | 0.75 | 0.56 |
| InlineChoice | 4 | 1,193 | 0.64 | 0.63 | 0.42 | 0.51 | 0.83 | 0.69 |
| Match | 4 | 1,193 | 0.47 | 0.48 | 0.38 | 0.37 | 0.60 | 0.64 |
| MC | 6 | 1,193 | 0.57 | 0.58 | 0.30 | 0.40 | 0.77 | 0.71 |
| Numeric | 2 | 1,193 | 0.75 | 0.56 | 0.65 | 0.51 | 0.86 | 0.61 |
| Zone | 2 | 1,193 | 0.73 | 0.69 | 0.71 | 0.61 | 0.75 | 0.77 |

Table 4. Summary of the Classical Item Statistics by Item Type for Grades Six Through Eight

| Item Type | No. of Items | No. of Examinees | Mean *p*-value | Mean Polyserial | Min. *p*-value | Min. Polyserial | Max. *p*-value | Max. Polyserial |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Composite | 2 | 585 | 0.42 | 0.46 | 0.31 | 0.38 | 0.54 | 0.53 |
| Grid | 4 | 585 | 0.52 | 0.45 | 0.33 | 0.20 | 0.62 | 0.58 |
| InlineChoice | 4 | 585 | 0.49 | 0.36 | 0.39 | 0.18 | 0.68 | 0.57 |
| Match | 4 | 585 | 0.40 | 0.66 | 0.11 | 0.59 | 0.58 | 0.75 |
| MC | 4 | 585 | 0.42 | 0.57 | 0.17 | 0.49 | 0.65 | 0.72 |
| Numeric | 2 | 585 | 0.52 | 0.43 | 0.27 | 0.16 | 0.77 | 0.71 |
| Zone | 5 | 585 | 0.52 | 0.43 | 0.41 | 0.14 | 0.62 | 0.62 |

Table 4. Summary of the Classical Item Statistics by Item Type for High School

| Item Type | No. of Items | No. of Examinees | Mean *p*-value | Mean Polyserial | Min. *p*-value | Min. Polyserial | Max. *p*-value | Max. Polyserial |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Composite | 2 | 476 | 0.63 | 0.40 | 0.49 | 0.32 | 0.76 | 0.48 |
| Grid | 4 | 476 | 0.39 | 0.45 | 0.07 | 0.18 | 0.80 | 0.64 |
| InlineChoice | 4 | 476 | 0.57 | 0.61 | 0.34 | 0.59 | 0.68 | 0.63 |
| Match | 4 | 476 | 0.54 | 0.46 | 0.33 | 0.28 | 0.76 | 0.71 |
| MC | 7 | 476 | 0.42 | 0.34 | 0.15 | 0.19 | 0.67 | 0.49 |
| Numeric | 2 | 476 | 0.28 | 0.28 | 0.20 | 0.11 | 0.37 | 0.45 |
| Zone | 2 | 476 | 0.54 | 0.36 | 0.50 | 0.28 | 0.58 | 0.44 |

Table 4. Summary of the Classical Item Statistics by Item Type and Response Type for Grades Three Through Five

| Item Type | Response Type | No. of Items | No. of Examinees | Mean *p*-value | Mean Polyserial | Min. *p*-value | Min. Polyserial | Max. *p*-value | Max. Polyserial |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grid | GridMS | 4 | 1,193 | 0.44 | 0.38 | 0.22 | 0.22 | 0.57 | 0.51 |
| Grid | GridSS | 3 | 1,193 | 0.56 | 0.45 | 0.41 | 0.23 | 0.75 | 0.56 |
| InlineChoice | InlineChoiceMS | 3 | 1,193 | 0.71 | 0.67 | 0.58 | 0.65 | 0.83 | 0.69 |
| InlineChoice | InlineChoiceSS | 1 | 1,193 | 0.42 | 0.51 | 0.42 | 0.51 | 0.42 | 0.51 |
| Match | MatchMS | 2 | 1,193 | 0.49 | 0.52 | 0.38 | 0.40 | 0.60 | 0.64 |
| Match | MatchSS | 2 | 1,193 | 0.45 | 0.43 | 0.40 | 0.37 | 0.49 | 0.48 |
| MC | InlineTextChoice | 1 | 1,193 | 0.30 | 0.40 | 0.30 | 0.40 | 0.30 | 0.40 |
| MC | InlineTextChoiceMS | 1 | 1,193 | 0.69 | 0.57 | 0.69 | 0.57 | 0.69 | 0.57 |
| MC | MCMS | 2 | 1,193 | 0.69 | 0.68 | 0.60 | 0.65 | 0.77 | 0.71 |
| MC | MCSS | 2 | 1,193 | 0.54 | 0.58 | 0.43 | 0.48 | 0.64 | 0.69 |
| Numeric | Numeric | 2 | 1,193 | 0.75 | 0.56 | 0.65 | 0.51 | 0.86 | 0.61 |
| Zone | ZoneSS | 2 | 1,193 | 0.73 | 0.69 | 0.71 | 0.61 | 0.75 | 0.77 |

Table 4. Summary of the Classical Item Statistics by Item Type and Response Type for Grades Six Through Eight

| Item Type | Response Type | No. of Items | No. of Examinees | Mean *p*-value | Mean Polyserial | Min. *p*-value | Min. Polyserial | Max. *p*-value | Max. Polyserial |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Composite | Composite | 2 | 585 | 0.42 | 0.46 | 0.31 | 0.38 | 0.54 | 0.53 |
| Grid | GridMS | 1 | 585 | 0.52 | 0.56 | 0.52 | 0.56 | 0.52 | 0.56 |
| Grid | GridSS | 3 | 585 | 0.51 | 0.42 | 0.33 | 0.20 | 0.62 | 0.58 |
| InlineChoice | InlineChoiceMS | 3 | 585 | 0.52 | 0.36 | 0.39 | 0.18 | 0.68 | 0.57 |
| InlineChoice | InlineChoiceSS | 1 | 585 | 0.40 | 0.39 | 0.40 | 0.39 | 0.40 | 0.39 |
| Match | MatchMS | 3 | 585 | 0.40 | 0.69 | 0.11 | 0.65 | 0.58 | 0.75 |
| Match | MatchSS | 1 | 585 | 0.38 | 0.59 | 0.38 | 0.59 | 0.38 | 0.59 |
| MC | InlineTextChoice | 1 | 585 | 0.49 | 0.72 | 0.49 | 0.72 | 0.49 | 0.72 |
| MC | InlineTextChoiceMS | 1 | 585 | 0.17 | 0.53 | 0.17 | 0.53 | 0.17 | 0.53 |
| MC | MCSS | 2 | 585 | 0.51 | 0.52 | 0.37 | 0.49 | 0.65 | 0.54 |
| Numeric | Numeric | 2 | 585 | 0.52 | 0.43 | 0.27 | 0.16 | 0.77 | 0.71 |
| Zone | ZoneMS | 1 | 585 | 0.57 | 0.52 | 0.57 | 0.52 | 0.57 | 0.52 |
| Zone | ZoneSS | 4 | 585 | 0.50 | 0.41 | 0.41 | 0.14 | 0.62 | 0.62 |

Table 4. Summary of the Classical Item Statistics by Item Type and Response Type for High School

| Item Type | Response Type | No. of Items | No. of Examinees | Mean *p*-value | Mean Polyserial | Min. *p*-value | Min. Polyserial | Max. *p*-value | Max. Polyserial |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Composite | Composite | 2 | 476 | 0.63 | 0.40 | 0.49 | 0.32 | 0.76 | 0.48 |
| Grid | GridMS | 2 | 476 | 0.08 | 0.32 | 0.07 | 0.18 | 0.09 | 0.46 |
| Grid | GridSS | 2 | 476 | 0.69 | 0.59 | 0.58 | 0.53 | 0.80 | 0.64 |
| InlineChoice | InlineChoiceMS | 2 | 476 | 0.64 | 0.61 | 0.60 | 0.59 | 0.68 | 0.62 |
| InlineChoice | InlineChoiceSS | 2 | 476 | 0.50 | 0.61 | 0.34 | 0.60 | 0.66 | 0.63 |
| Match | MatchMS | 2 | 476 | 0.55 | 0.54 | 0.33 | 0.37 | 0.76 | 0.71 |
| Match | MatchSS | 2 | 476 | 0.52 | 0.38 | 0.36 | 0.28 | 0.69 | 0.48 |
| MC | InlineTextChoice | 3 | 476 | 0.30 | 0.27 | 0.15 | 0.19 | 0.44 | 0.32 |
| MC | InlineTextChoiceMS | 1 | 476 | 0.36 | 0.41 | 0.36 | 0.41 | 0.36 | 0.41 |
| MC | MCMS | 1 | 476 | 0.54 | 0.38 | 0.54 | 0.38 | 0.54 | 0.38 |
| MC | MCSS | 2 | 476 | 0.58 | 0.40 | 0.49 | 0.31 | 0.67 | 0.49 |
| Numeric | Numeric | 2 | 476 | 0.28 | 0.28 | 0.20 | 0.11 | 0.37 | 0.45 |
| Zone | ZoneSS | 2 | 476 | 0.54 | 0.36 | 0.50 | 0.28 | 0.58 | 0.44 |

### Response Time Analyses

The length of time it takes students to complete a test is recorded and analyzed to build a profile describing what a typical testing event looks like for each test. In addition, variability in testing time is investigated to determine whether a student’s testing time should be viewed as unusual or irregular for further investigation. It should be noted that the CSA assessments are untimed.

In these analyses, all students who completed the test are considered. The testing population is partitioned into quartiles based on machine-scorable items. The descriptive statistics—for example, the number of students, mean, standard deviation, minimum and maximum, percentiles—of the time required to complete the total test are computed for each of the four quartile groups for each tested grade band (i.e., grades three through five, six through eight, and high school).

Appendix 4.B summarizes results of testing time analysis. Table 4.B.1 provides descriptive statistics of total testing time for the full student population at each ability level. The unit of testing time is the minute; for example, in table 4.B.1, the median (i.e., 50th percentile) of the testing time for grades three through five Q1 group is 35.89 minutes. Overall, there is no consistent pattern in testing times across the grade bands, given that the CSA pilot tests were untimed. Based on the medians of total testing time from the test for grades three through five, students in the highest ability level (4th quartile, Q4) spend less time completing the test than students in the other groups. For grades six through eight and high school, students in the lowest ability level (1st quartile, Q1) spent less time completing the tests than students in the other groups.

### DIF Analyses

Analyses of DIF can provide evidence of the degree to which an item score interpretation or use is valid for individuals who differ in their demographic characteristics. An item may be biased if it contains content or language that is differentially familiar to student groups. It is important, however, to recognize that item performance differences flagged for DIF might be related to actual difference in relevant knowledge or skills (group impact) or statistical Type I error, which might falsely assert DIF exists for an item. As a result, DIF statistics are used to identify potential item bias. Subsequent reviews by content experts and bias/sensitivity experts are required to determine the source and meaning of item performance differences.

DIF analyses were performed on all machine-scorable items. In examining the DIF between groups, the reference group is often designated as the group that is assumed to have an advantage, while the focal group refers to the group anticipated to be disadvantaged by the test. The sample size requirements for the DIF analyses were 100 in the smaller of either group and 400 in the combined focal and reference groups. These sample size requirements are based on standard operating procedures with respect to DIF analyses at ETS to ensure reliable DIF results can be obtained.

#### DIF Procedure for Dichotomous Items

The DIF analyses for dichotomous items utilized the Mantel-Haenszel (MH) DIF statistic (Mantel & Haenszel, 1959; Holland & Thayer, 1988). For this method, students are classified to relevant student groups of interest (e.g., gender or ethnicity). Students at each total score level in the focal group (e.g., females) are compared with examinees at each total score level in the reference group (e.g., males). The common odds ratio—that is, the proportion of correct response over the proportion of incorrect response—is estimated across all levels of matched student ability using the formula in equation 4.4 (Dorans & Holland, 1993). The resulting estimate is interpreted as the relative probability of success on a particular item for members of two groups when matched on ability, as presented in equation 4.4. *Refer to the* [*Alternative Text for Equation 4.4*](#_Alternative_Text_for_3) *for a description of this equation.*

 (4.4)

where,

*m* = the number of score categories of the total test,

*Rrm* = the number of students in the reference group who answer the item correctly,

*Wrm* = the number of students in the reference group who answer the item incorrectly,

*Rfm* = the number of students in the focal group who answer the item correctly,

*Wfm* = the number of students in the focal group who answer the item incorrectly, and

*Ntm* = the total number of students.

To facilitate the interpretation of MH results, the common odds ratio is frequently transformed to the delta scale using the formula presented in equation 4.5 (Holland & Thayer, 1988). *Refer to the* [*Alternative Text for Equation 4.5*](#_Alternative_Text_for_4) *for a description of this equation.*

 (4.5)

Positive values indicate DIF in favor of the focal group (i.e., positive DIF items are differentially easier for the focal group), whereas negative values indicate DIF in favor of the reference group (i.e., negative DIF items are differentially easier for the reference group).

#### DIF Procedure for Polytomous Items

The standardization DIF (Dorans & Schmitt, 1993; Zwick, Thayer, & Mazzeo, 1997; Dorans, 2013) in conjunction with the Mantel chi-square statistic (Mantel, 1963; Mantel & Haenszel, 1959) is calculated for polytomous items. The standardized mean difference (SMD) compares the item means of the two groups after adjusting for differences in the distribution of students across all items and is calculated using the formula presented in equation 4.6. *Refer to the* [*Alternative Text for Equation 4.6*](#_Alternative_Text_for_5) *for a description of this equation.*

 (4.6)

where,

*X =* the criterion score (total raw score),

*Y =* the item score,

*M* = the number of score levels on X,

*D* = the difference in the distribution of students at score level m,

*Nrm* = the number of students in the reference group at score level m,

*Nfm* = the number of students in the focal group at score level m,

*Er* = the expected item score for the reference group, and

*Ef* = the expected item score for the focal group.

A positive SMDvalue means that, conditional on the criterion score, the focal group has a higher mean item score than the reference group (i.e., the item is differentially easier for the focal group). In contrast, a negative SMD value means that, conditional upon the criterion score, the focal group has a lower mean item score than the reference group (i.e., the item is differentially harder for the focal group).

#### DIF Category Classification

Based on the DIF statistics and significance tests, items are classified into three categories and assigned values of A, B, or C. Category A items contain negligible DIF, Category B items exhibit slight to moderate DIF, and Category C items possess moderate to large DIF values.

The flagging criteria for dichotomous items are presented in table 4.16; the flagging criteria for polytomous items are provided in table 4.17. In these tables, SMDis standardized mean DIF, and SD is total group standard deviation of item score.

Table 4. DIF Categories for Dichotomous Items

| DIF Category | Criteria |
| --- | --- |
| A (negligible) | * Absolute value of MH D-DIF is not significantly different from zero, or is less than one.
* Positive values are classified as “A+” and negative values as “A-.”
 |
| B (moderate) | * Absolute value of MH D-DIF is significantly different from zero but not from one, and is at least one; OR
* Absolute value of MH D-DIF is significantly different from one, but is less than 1.5.
* Positive values are classified as “B+” and negative values as “B-.”
 |
| C (large) | * Absolute value of MH D-DIF is significantly different from one, and is at least 1.5.
* Positive values are classified as “C+” and negative values as “C-.”
 |

Table 4. DIF Categories for Polytomous Items

| DIF Category | Criteria |
| --- | --- |
| A (negligible) | * Mantel Chi-square statistic is not significantly different from zero (at the 0.05 level) or |SMD/SD| ≤ 0.17
 |
| B (moderate) | * Mantel Chi-square statistic is significantly different from zero (at the 0.05 level) and 0.17< |SMD/SD| ≤ 0.25
 |
| C (large) | * Mantel Chi-square statistic is significantly different from zero (at the 0.05 level) and |SMD*/SD*| > 0.25
 |

#### Items Exhibiting Significant DIF

DIF analyses for the gender group were conducted for the CSA pilot test on each grade band. [Appendix 4.C](#_Appendix_4.C_Differential) provides detailed DIF results. Table 4.C.1 shows the distributions of items across the DIF category classifications for each grade band. In addition, “Small N” indicates that the DIF analysis was not performed due to insufficient sample size in table 4.C.1. Table 4.C.2 lists the C-DIF item flagged across all tests. Among all machine-scorable items, only one item was flagged as negative C-DIF item in the high school test. A formal DIF panel review revealed that the item did not show any content flaw. Therefore the C-DIF item was not removed from the item bank.

### References

Dorans, N. J. (2013). ETS contributions to the quantitative assessment of item, test, and score fairness. *ETS Research Report Series*, i–38.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–65). Hillsdale, NH: Lawrence Erlbaum Associates, Inc.

Drasgow F. (1988). Polychoric and polyserial correlations. In L. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 7, pp. 69–74). New York: Wiley.

Holland, P. W., & Thayer, D. T. (1988). *An alternative definition of the ETS delta scale of item difficulty* (Research Report 85–43). Princeton, NJ: Educational Testing Service.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58*, 690–700.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analyses of data from retrospective studies of disease. *Journal of the National Cancer Institute,* *22*, 719–48.

Olsson, U. (1979) Maximum likelihood estimation of the polychoric correlation coefficient *Psychometrika, 44,* 443–460*.*

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*(4), 321–344.

### Accessibility Information

#### Alternative Text for Equation 4.1

P-value sub dich equals the fraction with the numerator the sum of X sub ic and the denominator N sub I end fraction. *(Return to* [*equation 4.1*](#EQ4_1)*.)*

#### Alternative Text for Equation 4.2

P-value sub poly equals the fraction with the numerator X sub ij and the denominator N sub i times Max of X sub I end fraction. *(Return to* [*equation 4.2*](#EQ4_2)*.)*

#### Alternative Text for Equation 4.3

r sub polyreg equals the fraction Beta sub hat times S tot divided by the square root of Beta sub hat squared times s sub tot squared plus 1. *(Return to* [*equation 4.3*](#EQ4_3)*.)*

#### Alternative Text for Equation 4.4

Alpha sub MH equals the numerator open parenthesis the sum sub m of R sub rm times W sub fm divided by N sub tm close parenthesis divided by the denominator open parenthesis the sum sub m of R sub fm times W sub rm divided by N sub tm closed parenthesis. *(Return to* [*equation 4.4*](#EQ4_4)*.)*

#### Alternative Text for Equation 4.5

MH D - DIF equals negative 2.35 times the natural logarithm open bracket alpha sub MH close bracket. *(Return to* [*equation 4.5*](#EQ4_5)*.)*

#### Alternative Text for Equation 4.6

SMD equals the fraction with numerator the sum from m equals 1 to M of N sub fm times E sub f of Y from X equals m and denominator the sum from m equals 1 to M of N sub fm end fraction minus the fraction with numerator the sum from m equals 1 to M of N sub fm times E sub r of Y from X equals m and denominator the sum from m equals 1 to M of N sub fm end fraction equals the fraction with the numerator the sum from m equals 1 to M of D sub fm and the denominator m equals1 to M of N suf fm end fraction. *(Return to* [*equation 4.6*](#EQ4_6)*.)*

### Appendix 4.A Classical Item Analyses

Notes:

1. An “r” indicates polyserial correlation, which is a statistical index of the item-total correlation.
2. A hyphen (“-”) in table 4.A.1, table 4.A.2, and table 4.A.3 indicates polytomous items, the statistics for which are presented in table 4.A.4, table 4.A.5, and table 4.A.6.

Table 4.A.1. Dichotomous Items Statistics by Grade Band Three Through Five

| Item Sequence | *p*-value | r | Omit Rate |
| --- | --- | --- | --- |
| 1 | - | - | - |
| 2 | - | - | - |
| 3 | - | - | - |
| 4 | 0.60 | 0.64 | 0.00 |
| 5 | 0.75 | 0.61 | 0.00 |
| 6 | 0.53 | 0.56 | 0.00 |
| 7 | - | - | - |
| 8 | 0.64 | 0.69 | 0.00 |
| 9 | 0.86 | 0.61 | 0.00 |
| 10 | - | - | - |
| 11 | 0.58 | 0.66 | 0.00 |
| 12 | - | - | - |
| 13 | 0.71 | 0.77 | 0.00 |
| 14 | 0.40 | 0.37 | 0.00 |
| 15 | 0.75 | 0.56 | 0.00 |
| 16 | 0.38 | 0.40 | 0.00 |
| 17 | 0.49 | 0.48 | 0.00 |
| 18 | 0.42 | 0.51 | 0.00 |
| 19 | 0.41 | 0.23 | 0.00 |
| 20 | - | - | - |
| 21 | 0.22 | 0.22 | 0.00 |
| 22 | 0.65 | 0.51 | 0.00 |
| 23 | 0.30 | 0.40 | 0.00 |
| 24 | - | - | - |
| 25 | 0.43 | 0.48 | 0.00 |

Table 4.A.2. Dichotomous Items Statistics by Grade Band Six Through Eight

| Item Sequence | *p*-value | r | Omit Rate |
| --- | --- | --- | --- |
| 1 | - | 0.47 | 0.00 |
| 2 | - | 0.59 | 0.00 |
| 3 | - | 0.71 | 0.00 |
| 4 | 0.60 | 0.49 | 0.00 |
| 5 | 0.75 | - | - |
| 6 | 0.53 | 0.16 | 0.00 |
| 7 | - | - | - |
| 8 | 0.64 | 0.49 | 0.00 |
| 9 | 0.86 | 0.62 | 0.00 |
| 10 | - | 0.75 | 0.00 |
| 11 | 0.58 | 0.65 | 0.00 |
| 12 | - | 0.58 | 0.00 |
| 13 | 0.71 | - | - |
| 14 | 0.40 | 0.54 | 0.00 |
| 15 | 0.75 | 0.14 | 0.00 |
| 16 | 0.38 | 0.66 | 0.00 |
| 17 | 0.49 | 0.20 | 0.00 |
| 18 | 0.42 | 0.39 | 0.00 |
| 19 | 0.41 | 0.38 | 0.00 |
| 20 | - | - | - |
| 21 | 0.22 | 0.53 | 0.00 |
| 22 | 0.65 | - | - |
| 23 | 0.30 | - | - |
| 24 | - | - | - |
| 25 | 0.43 | 0.72 | 0.00 |

Table 4.A.3. Dichotomous Items Statistics by High School

| Item Sequence | *p*-value | r | Omit Rate |
| --- | --- | --- | --- |
| 1 | - | - | - |
| 2 | 0.36 | 0.41 | 0.00 |
| 3 | - | - | - |
| 4 | 0.37 | 0.45 | 0.00 |
| 5 | 0.50 | 0.44 | 0.00 |
| 6 | 0.66 | 0.63 | 0.00 |
| 7 | 0.32 | 0.19 | 0.00 |
| 8 | - | - | - |
| 9 | 0.09 | 0.46 | 0.00 |
| 10 | 0.58 | 0.28 | 0.00 |
| 11 | 0.67 | 0.49 | 0.00 |
| 12 | - | - | - |
| 13 | 0.33 | 0.37 | 0.00 |
| 14 | 0.80 | 0.64 | 0.00 |
| 15 | 0.34 | 0.60 | 0.00 |
| 16 | - | - | - |
| 17 | 0.58 | 0.53 | 0.00 |
| 18 | 0.07 | 0.18 | 0.00 |
| 19 | 0.20 | 0.11 | 0.00 |
| 20 | - | - | - |
| 21 | 0.44 | 0.29 | 0.00 |
| 22 | 0.36 | 0.28 | 0.00 |
| 23 | 0.69 | 0.48 | 0.00 |
| 24 | 0.15 | 0.32 | 0.00 |
| 25 | 0.49 | 0.31 | 0.00 |

Table 4.A.4. Grades Three Through Five Polytomous Items Statistics

| Item ID | Item Type | Response Type | Item Seq | AIS | r | Flag | Omit Rate | Prop. of 0 Points | Prop. of 1 Point | Prop. of 2 Points |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| VH625639 | Grid | GridMS | 1 | 0.51 | 0.51 |  | 0.00 | 0.20 | 0.58 | 0.22 |
| VH627156 | MC | MCMS | 2 | 0.60 | 0.65 |  | 0.00 | 0.19 | 0.41 | 0.40 |
| VH625626 | InlineChoice | InlineChoiceMS | 3 | 0.71 | 0.65 | H | 0.00 | 0.14 | 0.29 | 0.57 |
| VH626984 | MC | InlineTextChoiceMS | 7 | 0.69 | 0.57 |  | 0.00 | 0.07 | 0.48 | 0.45 |
| VH625004 | MC | MCMS | 10 | 0.77 | 0.71 | H | 0.00 | 0.06 | 0.34 | 0.60 |
| VH624176 | Grid | GridMS | 12 | 0.57 | 0.44 |  | 0.00 | 0.11 | 0.65 | 0.24 |
| VH681095 | Grid | GridMS | 20 | 0.48 | 0.35 |  | 0.00 | 0.22 | 0.59 | 0.19 |
| VH627245 | InlineChoice | InlineChoiceMS | 24 | 0.83 | 0.69 | H | 0.00 | 0.09 | 0.17 | 0.74 |

Table 4.A.5. Grades Six Through Eight Polytomous Items Statistics

| Item ID | Item Type | Response Type | Item Seq | AIS | r | Flag | Omit Rate | Prop. of 0 Points | Prop. of 1 Point | Prop. of 2 Points |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| VH633257 | InlineChoice | InlineChoiceMS | 5 | 0.39 | 0.18 | R | 0.00 | 0.34 | 0.54 | 0.12 |
| VH633269 | Grid | GridMS | 7 | 0.52 | 0.56 |  | 0.00 | 0.25 | 0.45 | 0.30 |
| VH633652 | InlineChoice | InlineChoiceMS | 13 | 0.68 | 0.57 |  | 0.00 | 0.13 | 0.37 | 0.49 |
| VH681132 | InlineChoice | InlineChoiceMS | 20 | 0.48 | 0.31 |  | 0.00 | 0.28 | 0.49 | 0.23 |
| VH681128 | Composite | Composite | 22 | 0.54 | 0.53 |  | 0.00 | 0.20 | 0.52 | 0.28 |
| VH633368 | Composite | Composite | 23 | 0.31 | 0.38 |  | 0.00 | 0.60 | 0.19 | 0.21 |
| VH633370 | Zone | ZoneMS | 24 | 0.57 | 0.52 |  | 0.00 | 0.13 | 0.60 | 0.27 |

Table 4.A.6. High School Polytomous Items Statistics

| Item ID | Item Type | Response Type | Item Seq | AIS | r | Flag | Omit Rate | Prop. of 0 Points | Prop. of 1 Point | Prop. of 2 Points |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| VH623955 | InlineChoice | InlineChoiceMS | 1 | 0.68 | 0.62 |  | 0.00 | 0.14 | 0.36 | 0.49 |
| VH623956 | Match | MatchMS | 3 | 0.76 | 0.71 | H | 0.00 | 0.12 | 0.24 | 0.64 |
| VH623361 | Composite | Composite | 8 | 0.76 | 0.48 | H | 0.00 | 0.09 | 0.30 | 0.61 |
| VH624167 | InlineChoice | InlineChoiceMS | 12 | 0.60 | 0.59 |  | 0.00 | 0.23 | 0.33 | 0.43 |
| VH681183 | Composite | Composite | 16 | 0.49 | 0.32 |  | 0.00 | 0.30 | 0.42 | 0.28 |
| VH681176 | MC | MCMS | 20 | 0.54 | 0.38 |  | 0.00 | 0.10 | 0.71 | 0.18 |

### Appendix 4.B Response Time Analysis

Note the following about table 4.B.1:

* Raw scores for machine-scorable items were used to partition students into quartiles.
* All students who completed the test and have unrounded test time greater than 0 are included.

Table 4.B.1 Total Testing Time (in Minutes) at Each Raw Score Interval

| Grade Band | Raw Score Interval | N | Mean | Standard Deviation | Min. | Max. | % Pt. 1 | % Pt. 10 | % Pt. 25 | % Pt. 50 | % Pt. 75 | % Pt. 90 | % Pt. 99 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3–5 | 0–14 | 256 | 38.95 | 14.97 | 3.61 | 99.01 | 12.06 | 21.68 | 28.87 | 35.89 | 47.69 | 58.39 | 86.49 |
| 3–5 | 15–19 | 332 | 41.17 | 15.07 | 13.31 | 111.12 | 18.41 | 25.90 | 31.59 | 38.15 | 46.89 | 59.18 | 102.94 |
| 3–5 | 20–23 | 262 | 38.87 | 14.33 | 14.06 | 91.60 | 18.70 | 24.45 | 29.10 | 35.73 | 45.65 | 55.54 | 90.59 |
| 3–5 | 24–33 | 343 | 33.79 | 12.00 | 12.08 | 84.20 | 14.68 | 20.27 | 25.25 | 31.75 | 39.47 | 48.84 | 73.45 |
| 6–8 | 0–10 | 130 | 46.24 | 19.26 | 5.92 | 139.55 | 6.16 | 26.81 | 34.17 | 44.76 | 55.25 | 67.43 | 122.59 |
| 6–8 | 11–14 | 141 | 51.44 | 15.74 | 9.20 | 121.39 | 17.05 | 32.78 | 42.94 | 50.02 | 59.50 | 70.61 | 105.94 |
| 6–8 | 15–18 | 144 | 51.62 | 16.49 | 22.12 | 102.32 | 22.77 | 32.65 | 41.32 | 48.79 | 59.54 | 75.39 | 100.04 |
| 6–8 | 19–32 | 170 | 46.77 | 12.39 | 15.63 | 79.33 | 23.82 | 32.18 | 36.94 | 44.99 | 55.40 | 64.27 | 77.96 |
| High School | 0–11 | 93 | 35.93 | 11.50 | 8.29 | 63.22 | 8.29 | 22.62 | 29.86 | 34.33 | 43.84 | 51.20 | 63.22 |
| High School | 12–15 | 139 | 41.65 | 13.71 | 15.49 | 85.47 | 18.05 | 26.11 | 33.40 | 39.08 | 47.18 | 59.73 | 82.52 |
| High School | 16–18 | 102 | 38.35 | 10.40 | 20.42 | 74.30 | 20.79 | 25.82 | 31.82 | 36.74 | 44.69 | 52.33 | 66.15 |
| High School | 19–31 | 142 | 40.00 | 10.10 | 21.11 | 79.07 | 21.50 | 29.32 | 32.81 | 38.46 | 46.23 | 52.37 | 71.03 |

### Appendix 4.C Differential Item Functioning (DIF)

Table 4.C.1 Gender DIF Classifications Summary by Grade Band

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **DIF Category** | **3–5: N** | **3–5: Pct** | **6–8: N** | **6–8: Pct** | **High School: N** | **High School: Pct** |
| C- | 0 | 0.00 | 0 | 0.00 | 1 | 4.00 |
| B- | 1 | 4.00 | 1 | 4.00 | 0 | 0.00 |
| A- | 10 | 40.00 | 9 | 36.00 | 13 | 52.00 |
| A+ | 14 | 56.00 | 15 | 60.00 | 9 | 36.00 |
| B+ | 0 | 0.00 | 0 | 0.00 | 2 | 8.00 |
| C+ | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Small N | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| **Total** | **25** | **100.0** | **25** | **100.0** | **25** | **100.0** |

Table 4.C.2 Items Exhibiting Significant DIF by Student Group

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Grade Band** | **Item ID** | **N Focal** | **N Reference** | **SMD** | **Comparison** | **In Favor Of** |
| High School | VH624167 | 275 | 201 | -0.21 | Male – Female | Male |

## Constructed-Response (CR) Item Evaluation

This chapter summarizes the evaluation of the responses that were gathered from the CR items in the California Spanish Assessment (CSA) pilot test forms. As with the rest of the pilot test, the focus was to evaluate students’ use of the online testing environment to demonstrate their knowledge, skills, and abilities at the moment they keyboarded their response in Spanish. The summary of the CR items evaluation provides data about the extent to which students used the tools available to them to provide the characters shown in table 5.1.

### Evaluation of the Use of Tools in the Writing Prompt

Each pilot form contained a writing prompt, to which each student was to write a response. Of the 2,311 responses received, 450 were reviewed, 150 from each grade span. The subset of responses that were neither blank, copies of the prompt, nor written in English were analyzed to gauge whether a student used Spanish characters—i.e., composed letters using a tilde (i.e., ñ), acute accent mark (e.g., á, é, etc.), or dieresis (i.e., ü), and inverted punctuation (i.e., ¿ or ¡), as appropriate—to produce grammatical written Spanish. In addition to evaluating the presence/absence of these elements, Educational Testing Service (ETS) documented inconclusive and alternative use, defined as follows:

1. Inconclusive (Inc.) use means that the student’s Spanish writing sample did not employ wording that needed any of the characters.
2. Alternative (Alt.) use indicates students typed in such a way that it was clear they knew one or more Spanish characters were needed, but they employed a different means to convey that information such as typing “an~os” instead of *años*, “vergu:enza” instead of *vergüenza*, or “!si'!” instead of *¡sí!*

Those data are presented in table 5.1.

Table 5. Summary of the Use of Spanish Characters in Pilot Constructed Responses

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade Band** | **Total *N*** | **Not Used *N*** | **Not Used %** | **Used *N*** | **Used %** | **Inc. Use *N*** | **Inc. Use%** | **Alt. Use *N*** | **Alt. Use %** |
| 3–5 | 136 | 84 | 62% | 34 | 25% | 13 | 10% | 5 | 4% |
| 6–8 | 143 | 105 | 73% | 28 | 20% | 8 | 6% | 2 | 1% |
| High School | 144 | 96 | 67% | 41 | 28% | 6 | 4% | 1 | 1% |
| **Total/Mean** | **423** | **285** | **67%** | **103** | **24%** | **27** | **6%** | **8** | **2%** |

#### Use of the Cognitive Lab Form

A subset of the students in table 5.1 were administered a cognitive lab form instead of a pilot form. While the writing prompt was the same as the pilot form’s prompt, a subset of the cognitive lab students were asked by their proctor to type a sample sentence that includes Spanish characters: *¡Está fría la mañana!* When present, this sentence was analyzed separately but similarly to the overall response previously described; due to the presence of Spanish characters in the cognitive lab sentence to be typed, “inconclusive use” was not a category used in the data presented in table 5.2.

Table 5. Summary of the Use of Spanish Special Characters in *¡Está fría la mañana!* in Cognitive Lab Constructed Responses

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade Band** | **Sample Sentence Present Total *N*** | **Not Used *N*** | **Not Used %** | **Used *N*** | **Used%** | **Alt. Use *N*** | **Alt. Use %** |
| 3–5 | 21 | 3 | 14% | 18 | 86% | 0 | 0% |
| 6–8 | 6 | 0 | 0% | 6 | 100% | 0 | 0% |
| High School | 8 | 0 | 0% | 7 | 88% | 1 | 13% |
| **Total/Mean** | **35** | **3** | **5%** | **31** | **89%** | **1** | **4%** |

### Analysis of the Use of Tools in the Writing Prompt

The choice to use correct characters when composing written Spanish is ultimately governed by widely varying input, from the Spanish surrounding students in their daily life to the implementation of rigorous standards and the reinforcement received in the classroom.

The analysis of the pilot responses clearly indicates that a large majority of test takers chose not to include the Spanish characters in their own typewritten responses. Data from the cognitive lab, though, suggests many test takers probably would have been able to include the characters if specifically prompted to do so.

In order to elicit the best performance from students completing CSA writing prompts on the field test, which includes the appropriate use of Spanish characters, ETS advocates keeping the availability and use of Spanish characters fresh in the mind of students. Therefore, ETS will insert a brief reminder page that goes over the use of special characters prior to the CR item on the field test forms. This reminder, located on the screen immediately preceding the writing prompt, can orient or reorient students to the special character menu available on all writing prompts via the Ω ([**Omega**]) button.

## Discussions and Implications

### Summary of Findings

This chapter summarizes the findings from the results of the item- and test-level analyses from the 2017–18 California Spanish Assessment (CSA) pilot test administration. Based on the results of classical item analyses presented in Chapter 4, the average item difficulties and discriminations for all pilot tests within each grade band were reasonable.

GridMS items presented a challenge for test takers in grades three through five, but were less of a challenge for students in other grades. This may be related to the fact that the test for grades three through five was administered to those students who were just becoming literate users of Spanish; those students had difficulty with GridMS items, likely because half of them were Writing items and three-quarters of them were 2-point items. Conversely, test takers in grades three through five seemed to have a consistently easy time with Zone and Numeric item types.

While curriculum was not the main goal of the pilot test, some conclusions can be made:

* **Writing Mechanics and Conventions was collectively the most difficult portion of the curriculum across all pilot test forms.**
* The average *p*-values for dichotomous items in Writing Mechanics and Conventions were lower than in other domains.
* The average item scores for polytomous items in Writing Mechanics were lower than other domains in both upper grade-band forms.
* **Listening appears to be less difficult at the lower grade levels.**
* The average *p*‑values for dichotomous items in the Listening domain were higher for grades three through five than for other two grade bands.
* Average item scores for polytomous items in Listening were higher for grades three through five and grades six through eight than for high school.

### Implications for the Field Test

Results obtained from the 2017–18 administration provide Educational Testing Service with actionable items for supporting the fall 2018–19 field test administration, described next.

The California Spanish Assessment (CSA) pilot test was administered to a select number of local educational agencies (LEAs) in the 2017–18 administration. As ETS and the CDE continues the implementation of the CSA into the CAASPP System of assessments, training and communication will be a focal point moving forward. Because the CSA is a brand new assessment and completely voluntary, ETS will continue to provide statewide training specific to the CSA to LEA staff and test administrators so they are prepared to administer the test. Training will include CSA-specific testing interface video tutorials in Spanish and English, a field test administration manual, and a field test administration Webcast.

Furthermore, like the other CAASPP assessments, in order to continue familiarizing students with the CSA items, ETS will develop and provide training tests with a variety of item types. Training tests include a small set of sample test questions that allow students and test administrators to learn how to interact with the different item types, available accessibility features, and test administration instructions. Training tests are typically not grade-specific nor are they available as full-length tests. Each grade-band CSA training test will include approximately 10 items. LEAs will be encouraged to use the training tests to prepare students to become more familiar with utilizing the technology and technology-enhanced items prior to taking the field test.

In the interest of increasing the number of items that are “born accessible”—i.e., items that are as universally accessible by all populations as possible—ETS is investigating the construct-irrelevant use of item types that have no discernable difference from traditional test questions. Therefore, ETS is reducing the development of both Grid single-select items as well as text-based Zone items, since they function much as multiple choice items.

Collectively, the Writing Mechanics and Conventions items proved to be the most challenging domain of the CSA pilot.  When more items are field tested for eventual operational use, there exists the possibility that Writing Mechanics items may acquire statistics that bar their use. As a stopgap measure for that eventuality, ETS will develop and field-test stand-alone items that may later be inserted into a test form in need of additional Writing Mechanics items with statistics in order to meet the blueprint.

The cognitive labs revealed that some students who were generally feeling quite confident in their speaking and listening skills struggled with the literacy tasks in front of them; some had only the most rudimentary grasp of Spanish reading and writing. In order to ease these students into the CSA field test forms, the first passage-based items a student will encounter will be based on listening passages.

The analysis of typewritten responses to the pilot CR prompts revealed that, on average, 67 percent of students did not type Spanish special characters (e.g., ñ, ¿, ü) in words which call for their use. The cognitive labs, on the other hand, revealed that 89 percent of students asked specifically to do so were able to reproduce the characters in a sample sentence. Because the CSA field test includes CR prompts, the responses of which will be centrally scored, ETS will insert a reminder screen immediately before the CR set/item in order to encourage the inclusion of the Spanish-specific characters via the American Institutes for Research pop-up menu.

1. CR items are excluded from the analysis. [↑](#footnote-ref-2)