

# **Addendum to the High-Level Test Design for the California Spanish Assessment**

**Contract #CN150012**

**Prepared for the California Department of Education by ETS**

**Presented September 30, 2022**



Table of Contents

[1. Overview 3](#_Toc112857862)

[1.A. Background 3](#_Toc112857863)

[1.B. Key Assumptions for the Expansion 3](#_Toc112857864)

[2. Revised Achievement Level Descriptors 4](#_Toc112857865)

[3. Standards and Claims 4](#_Toc112857866)

[4. Item Development 5](#_Toc112857867)

[4.A. New Item Development 5](#_Toc112857868)

[4.B. Item Types 6](#_Toc112857869)

[5. Test Design 6](#_Toc112857870)

[5.A. Test Format 6](#_Toc112857871)

[5.B. Test Development Stages 6](#_Toc112857872)

[5.C. Field Test Design 6](#_Toc112857873)

[6. Psychometric Considerations 7](#_Toc112857874)

[6.A. Embedded Field Test 7](#_Toc112857875)

[6.B. Field Testing Sample Size and Threshold 8](#_Toc112857876)

[6.C. Psychometric Analyses 8](#_Toc112857877)

[6.C.1. Classical Item Analysis and Differential Item Functioning 8](#_Toc112857878)

[6.C.2. Interrater Reliability Analysis 9](#_Toc112857879)

[6.C.3. Dimensionality Analysis 9](#_Toc112857880)

[6.C.4. Item Response Model Analysis 9](#_Toc112857881)

[Appendix A: High-Level Test Development Timeline 10](#_Toc112857882)

[Appendix B: References 11](#_Toc112857883)

List of Tables

[Table 1. Embedded Field Test Design, Grades Three Through Eight 7](#_Toc112868440)

[Table 2. Embedded Field Test Design, High School 7](#_Toc112868441)

[Table 3. Number of Embedded Field Test Items 7](#_Toc112868442)

[Table 4. High-Level Test Development Timeline 10](#_Toc112868443)

## Overview

### Background

California is a state with great linguistic diversity. More than 40 percent of students in California speak a language other than English.[[1]](#footnote-2) Of these students, over 1.2 million speak Spanish.[[2]](#footnote-3) The student population in California includes students who are native speakers of Spanish and students who are learning Spanish as an additional language. California’s educational system includes instruction in Spanish in various forms. Thus, the California Spanish Assessment (CSA) was first designed and created in 2016 as a new computer-based assessment for students in grades three through eight and high school to measure students’ competency in Spanish in reading, writing mechanics, and listening. Currently, the CSA is part of the California Assessment of Student Performance and Progress System of assessments.

The purpose of this addendum is to articulate how the test design presented in the [*High-Level Test Design for the California Spanish Assessment*](https://www.caaspp.org/rsc/pdfs/CSA.high-level-test-design.pdf)(hereafter referred to as the “*CSA High-Level Test Design*”), which was approved by the State Board of Education (SBE) in September 2016, has been updated to include the expansion of the writing domain and addition of a speaking domain, as approved in the *California Assessment System Scope of Work* for the 2022–27 school years. The expansion of the writing domain and addition of a speaking domain will allow the CSA to be used, in part, to achieve the State Seal of Biliteracy, as originally intended (pursuant to California *Education Code* Section **60640[j]**).

### Key Assumptions for the Expansion

For planning and development purposes, ETS makes the following assumptions about expanding the CSA:

1. The assessment was originally developed with a focus on reading, writing mechanics, and listening. The updated assessment will include an additional writing component as well as speaking.
2. Field testing of the additional writing and speaking components will be handled by embedding field test items into the 2023–24 operational forms.
3. General achievement level descriptors (ALDs) will be updated as specified in [section 2](#_Revised_Achievement_Level).
4. The first operational administration of the expanded assessment will occur in the school year 2024–25.
5. A standard setting process will be designed, and an educator workshop will occur after the first operational administration of the expanded assessment.
6. Interest holders’ input will be a critical component of the development process; there will be a number of opportunities to provide input on test design by interest holders throughout the test development process.

## Revised General Achievement Level Descriptors

The general ALDs are generic descriptors of student performance expectations that provide the range expected in each performance level. The descriptors provide a snapshot of student achievement in the current school year.

With the expansion of the CSA, speaking will be added to the descriptors. Additionally, the word “mechanics,” a qualifier in the writing domain’s current ALD text, will be removed because the writing domain will be fully assessed after the expansion is implemented operationally.

* **Level 3:** Students at Level 3 demonstrate a **high degree** of grade-appropriate Spanish literacy in reading, writing, listening, and speaking, and *are on track* for being literate in Spanish by high school graduation.
* **Level 2:** Students at Level 2 demonstrate a **moderate degree** of grade-appropriate Spanish literacy in reading, writing, listening, and speaking, but *require further development to be on track* for being literate in Spanish by high school graduation.
* **Level 1:** Students at Level 1 demonstrate a **limited degree** of grade-appropriate Spanish literacy in reading, writing, listening, and speaking, and *require substantial development* *before being considered on track* for being literate in Spanish by high school graduation.

## Standards and Claims

The CSA is aligned to the *California Common Core State Standards (CCSS) en Español*, which are a translated and linguistically augmented version of the *California* *CCSS for English Language Arts & Literacy*.

The *California CCSS en Español* are organized into the following domains:

* Reading standards
* Writing standards
* Speaking and listening standards
* Language standards

Using the domains as a guide for the test design, the SBE adopted the following claims for the CSA in September 2016:

* Claim for grades three through eight: Students can demonstrate progress toward a high level of competency in attaining reading/language arts skills and practices through Spanish.
* Claim for high school: Students can demonstrate a high level of competency in attaining reading/language arts skills and practices through Spanish.

The SBE also adopted the following Spanish language arts competency claims for all grade levels and the high school grade band:

* Reading: Students can read, analyze, and interpret a variety of texts and genres through Spanish.
* Writing: Students can write texts for a range of purposes and audiences to accurately and convincingly present, describe, and explain ideas through Spanish.
* Listening: Students can comprehend spoken Spanish in a range of contexts.

With the expansion of the CSA into the speaking domain, ETS recommends the addition of the following claim for all grade levels and the high school grade band:

* Speaking: Students can speak Spanish to accurately and convincingly present, describe, and explain ideas for a range of purposes and audiences.

## Item Development

### New Item Development

ETS assessment specialists will develop samples of new item types for speaking and writing for the California Department of Education’s (CDE’s) review and approval. These samples will be used to update the item specifications and training materials in collaboration with the CDE. The new item specifications will be used when ETS begins development work on items for all domains in the assessment.

All items will be aligned with the *California CCSS* *en Español* and consistent with the goals of California’s testing program. Items will be written by ETS assessment developers and trained item writers familiar with assessment development in Spanish and specifically trained for the CSA as well as California educators who have received item writer training, including updated training on new item types for speaking and writing. All items will be reviewed by ETS content and editorial staff, the CDE, and a review panel composed of California educators. ETS will leverage the current best practices developed on other California assessments that feature constructed responses (CRs) when designing CSA speaking and writing items.

### Item Types

ETS will develop machine-scorable and human-scorable item types for computer-based administration by leveraging the most current assessment innovations conducive to assessing reading/language arts skills.

The assessment includes both stand-alone items and passage-based items; all items may contain a stimulus (e.g., a passage, video, or image). Some of the items have technology-enhanced interactions. These interactions include having a student respond by typing an answer, completing a table, selecting from a drop-down list, etc.

The expansion of the CSA will include the administration of human-scorable items for writing and speaking, which will necessitate the development of scoring rubrics. The use of artificial intelligence to score written responses will be explored, as advances in this field are well underway.

## Test Design

### Test Format

The expanded CSA will continue to be a linear test delivered online under untimed testing conditions. It will be untimed to allow students sufficient time to complete the test. Testing time estimates will be reviewed and adjusted if necessary and provided to local educational agencies (LEAs) for scheduling purposes. In view of California’s desire for a Spanish reading/language arts assessment that measures a high level of competency demonstrated by students who are on track for exiting public instruction as biliterate graduates, the high school CSA will continue to feature complex passages and tasks that can aid local decisions about eligibility for the State Seal of Biliteracy.

### Test Development Stages

To support the 2024–25 operational launch of the expanded CSA, several activities will take place over a three-year time span:

* 2022–23: Test design, blueprint development, and item and task development
* 2023–24: Field testing of full-write and speaking items and prompts
* 2024–25: First operational administration of full-write and speaking prompts; standard setting

### Field Test Design

Forms with embedded field test items, including the new speaking and writing items, will be administered for grades three, four, five, six, seven, and eight, and for high school, during the 2023–24 school year. Table 1 shows a tentative embedded field test design for grades three through eight.

Table 1. Embedded Field Test Design, Grades Three Through Eight

|  |  |  |
| --- | --- | --- |
| Item Type | # of Items Taken by Each Student | Estimated Testing Time for Each Student |
| Speaking CR items (human-scored) | 8 items | 60–90 minutes |
| Full-write CR items (human-scored) | 1 item | 45–60 minutes |
| Writing non-CR items (machine-scored) | 1 item | 1–2 minutes |
| **Estimated Total:** | **10 items** | **105–150 minutes** |

Table 2 shows a tentative embedded field test design for high school.

Table 2. Embedded Field Test Design, High School

|  |  |  |
| --- | --- | --- |
| Item Type | # of Items Taken by Each Student | Estimated Testing Time for Each Student |
| Speaking CR items (human-scored) | 12 items | 60–90 minutes |
| Full-write CR items (human-scored) | 2 items | 60–80 minutes |
| Writing non-CR items (machine-scored) | 2 items | 1–2 minutes |
| **Estimated Total:** | **16 items** | **120–170 minutes** |

The field testing of CR items entails an average increase in overall testing time of two hours for grades three through eight and two and one half hours for high school, based on the estimates in table 1 and table 2. Therefore, if the CSA before the expansion took one and one half to two hours, with the expansion, the overall field testing time would potentially range from four to four and one half hours.

ETS is committed to field-testing the number of items shown in table 3.

Table 3. Number of Embedded Field Test Items

|  |  |  |
| --- | --- | --- |
| Item Type | # of Items Field-Tested per Grade, Grades 3–8 | # of Items Field-Tested, High School |
| Speaking CR items (human-scored) | 24 items | 36 items |
| Full-write CR items (human-scored) | 3 items | 6 items |
| Writing non-CR items (machine-scored) | 3 items | 6 items |
| **Estimated Total:** | **30 items** | **48 items** |

## Psychometric Considerations

### Embedded Field Test

As noted in [section 5C](#_Field_Test_Design), the 2023–24 operational CSA operational forms will include an embedded field test design for the field testing of the new writing and speaking CR items. Under this design, the newly developed writing and speaking CR items are embedded in an operational form and are analyzed with other operational items in the form. Therefore, the new CR items will be linked to the operational base scale through psychometric analysis of calibration and linking. Logistically, the embedded field test is a cost-efficient way to collect data for the field test items, which does not require an additional off-season test for students and test administrators.

The spring 2024 embedded field test will feel like the operational 2025 year (window is the same, form breakdown will be similar to operational, effort will be there, good testing experience for students with new item type). ETS will have the same test-taking population. In addition, the embedded field testing of items can occur on an ongoing basis for each operational test administration to support a mutually agreed upon refresh rate. All newly developed writing and speaking CR items can be placed and spiraled at the student level to achieve random samples for analysis.

Psychometric analysis and linking procedures can be accomplished with an embedded field test design. Item response theory (IRT) calibration and linking procedures can be used to calibrate, link, and scale the new CR items to the operational base scale.

### Field Testing Sample Size and Threshold

As an optional assessment, the key challenge for a CSA field test administration at the various grade levels and the high school grade band is the small sample size. In general, larger test samples occur in the lower grades three through five. As enrollment in Spanish instruction attenuates at the middle and upper grades, the test-taking sample size decreases as the grade level increases. In the embedded field test model, the psychometric threshold of the testing sample for item analysis is the same for each grade level and the high school grade band, (e.g., a minimum of 200 students per item and desirable 300 students per item per form, excluding the accommodated form). To receive reliable analysis results, ETS would adhere to this threshold of the sample size for item analysis for the field tests of new writing and speaking CR items.

As this threshold can be met easily in the lower grades and might be challenging in higher grades, especially in the high school grade band, it is possible to increase the number of field test items in each version of the regular forms and reduce the number of versions in high school to support the goal that items that are field-tested have sufficient student responses to be analyzed as the minimum and desirable student counts mentioned previously. For example, three versions of a regular form may be appropriate in grades three through eight, each containing one CR item for writing and speaking. In the high school grade band, two versions of a regular form could be administered, each containing two CR items in each domain.

### Psychometric Analyses

#### **Classical Item Analysis and Differential Item Functioning**

Classical item analysis will be conducted to evaluate the performance of all newly developed writing and speaking CR items with respect to item difficulty and item discrimination. The following flagging rules for these statistics will be used to identify items that are not performing as expected:

* A *p*-value (or mean of item score) less than 0.2 or greater than 0.95
* An item with a polyserial correlation less than 0.20
* An item with nonresponse rates greater than 5 percent
* An item that has less than 3 percent of the students at any score level (e.g., 0, 1, 2)

Differential item functioning (DIF) analyses will be performed on the newly developed writing CR items and speaking CR items. Because of the CSA sample size, only gender DIF will be analyzed and C-DIF will be flagged. If fewer than 200 students respond to a CR field test item, item analysis and DIF analysis may not be conducted.

#### **Interrater Reliability Analysis**

The new writing and speaking CR items will be scored by a single rater with 10 percent back-reading, which means 10 percent of the responses to CR items will be scored independently by a second rater. Data from items with two raters will be used to evaluate interrater reliability of scoring. The statistics for interrater reliability include the percentage of perfect agreement and adjacent agreement between the two raters, and the quadratic weighted kappa statistic (QWK). QWK is a statistic used to measure the degree of association between two ratings with values ranging from 0.0 (indicating no agreement) to 1.0 (indicating perfect agreement).

CR items will be flagged if any of the following conditions occur:

* Adjacent plus exact agreement < 0.80
* QWK < 0.70

#### **Dimensionality Analysis**

When writing and speaking CR items are included in the test form, it is unknown whether the unidimensional assumption can be held. Thus, ETS will conduct a dimensionality analysis to examine whether the CSA is unidimensional or multidimensional. Results of the study will inform how the items should be calibrated and scores be reported.

#### **Item Response Model Analysis**

IRT models will be continuously used to analyze writing and speaking CR items and the CSA forms. The selection of the specific IRT models will be based on the results of the aforementioned dimensionality analysis. Furthermore, even if the dimensionality analysis confirms unidimensional features of the CSA scale, the continuity of the current scale established in the 2019 will be investigated carefully. Scale scores and achievement levels will be produced and reported.

## Appendix A: High-Level Test Development Timeline

Table 4. High-Level Test Development Timeline

|  |  |
| --- | --- |
| Activity | Date |
| SBE action on the *Addendum to the* *High-Level Test Design for the California Spanish Assessment*, revised test blueprints, and general ALDs | September 2022 |
| Administration of the embedded field test | 2023–24 administration |
| Dimensionality study | Spring and summer 2024 |
| Administration of the operational test of the newly expanded CSA | 2024–25 administration |
| Standard setting | Summer 2025 |
| SBE action on the threshold scores and reporting ALDs | September 2025 |

## Appendix B: References

Finegan, E., & Biber, D. (2001). Register variation and social dialect variation: The register axiom. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 235–‍68). Cambridge: Cambridge University Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–‍74.

Valdés, G., Fishman, J., Chávez, R., & Pérez, W. (2006). *Developing minority language resources: The case of Spanish in California*. Clevedon, UK: Multilingual Matters.

1. California Department of Education, DataQuest, Language Census Data for 2012–13 [↑](#footnote-ref-2)
2. EDFacts/California Consolidated State Performance Report, 2012–13 and 2013–14 [↑](#footnote-ref-3)