California Department of Education Assessment Development & Administration Division

Computer-based Initial English Language Proficiency Assessments for California (ELPAC) Fall 2019 Field Test Technical Report

Contract #CN140284

Prepared for the California Department of Education by Educational Testing Service

Submitted December 30, 2020

Table of Contents

List of Appendices

List of Tables

Acronyms and Initialisms Used in the *Computer-based Initial English Language Proficiency Assessments for California Fall 2019 Field Test Technical Report*

Term	Definition
1PL	one-parameter logistic
AERA	American Educational Research Association
AIR	American Institutes for Research (now Cambium Assessment)
AIS	average item score
ALTD	Assessment & Learning Technology Development
AST	Administration and Scoring Training
CAASPP	California Assessment of Student Performance and Progress
CALPADS	California Longitudinal Pupil Achievement Data System
CalTAC	California Technical Assistance Center
CCR	*California Code of Regulations*
CDE	California Department of Education
CR	constructed response
DEI	Data Entry Interface
DFA	*Directions for Administration*
DIF	differential item functioning
EC	*Education Code*
EL	English learner
ELD Standards	English Language Development Standards
ELP	English language proficiency
ELPAC	English Language Proficiency Assessments for California
EO	English only
eSKM	Enterprise Score Key Management
ESSA	Every Student Succeeds Act
ETS	Educational Testing Service
IBIS	Item Banking Information System
ICC	item characteristic curve
IDEA	Individuals with Disabilities Act
IEP	individualized education program
IFEP	Initial fluent English proficient
IRT	item response theory
K	kindergarten
K–2	kindergarten through grade two
LEA	local educational agency
MC	multiple choice
MH	Mantel-Haenszel
MH-DIF	Mantel-Haenszel differential item functioning

Table of Acronyms and Initialisms *(continued)*

Term	Definition
ONE	Online Network for Evaluation
OTI	Office of Testing Integrity
PAR	Psychometric Analysis & Research
PCM	partial credit model
PPT	paper–pencil test
RFEP	reclassified fluent English proficient
SBE	State Board of Education
SCOE	Sacramento County Office of Education
SD	standard deviation
SMD	standardized mean difference
SSR	Student Score Report
STAIRS	Security and Test Administration Incident Reporting System
TBD	to be determined
TCC	test characteristic curve
TDS	test delivery system
TIPS	Technology and Information Processing Services
TOMS	Test Operations Management System
TRCS	Technology Readiness Checker for Students
UAT	user acceptance testing
USC	United States Code

Chapter 1 Introduction

This technical report focuses on the development, administration, psychometric analyses, and results of the computer-based Initial English Language Proficiency Assessments for California (ELPAC) field test. Chapter 1 provides an overview of both the computer-based Summative and Initial ELPAC field test administration, including background information, purposes of the field test, intended population, testing window, and an overview of the field test technical report. The remaining chapters of this report focus on aspects of the development, administration, and analysis of the computer-based Initial ELPAC field test.

1.1 ELPAC Overview

The ELPAC "is the required state test for English language proficiency (ELP) that must be given to students whose primary language is a language other than English. State and federal laws require that local educational agencies (LEAs) administer a state test of ELP to eligible students in kindergarten through grade twelve" (California Department of Education [CDE], 2019). California *Education Code (EC)* Section 313(a) requires that the assessment of ELP be done upon initial enrollment and annually thereafter until the LEA reclassifies the student as initial fluent English proficient (IFEP).

In November 2018, the State Board of Education (SBE) approved the plan to transition the paper–pencil ELPAC to a computer-based ELPAC. As part of the transition work to prepare for the operational computer-based ELPAC administration, Educational Testing Service (ETS) conducted a combined Initial and Summative ELPAC field test of the ELPAC items in an online environment. Participating schools were assigned to either a computer-based form of the Initial or Summative ELPAC or a mix of paper-based and computer-based versions of the oral or written language composites as part of a mode comparability study. The oral language composite was comprised of the Listening and Speaking domain assessments; the written language composite was comprised of the Reading and Writing assessments.

The computer-based ELPAC has replaced the paper–pencil Summative ELPAC as of February 2020 and replaced the paper–pencil Initial ELPAC on August 20, 2020. (Note that while the Writing domain for kindergarten through grade two is administered as a paper–pencil test, it is a component of the computer-based assessment.)

1.1.1 Initial ELPAC and Summative ELPAC

The ELPAC consists of two assessments: the Initial ELPAC and the Summative ELPAC. The Initial ELPAC identifies students who are potential English learners (ELs) who need extra help learning English and will need to be enrolled in an English language development program. Students identified as ELs after taking the Initial ELPAC go on to take the Summative ELPAC annually until reclassified. The Summative ELPAC is one piece of the evidence used to determine whether the student's English proficiency has improved to the point that the student can be redesignated as fluent English proficient (RFEP) or reclassified.

The Initial ELPAC is administered only once during a student's time in a California public school. The Summative ELPAC is administered annually to students in kindergarten through grade twelve who are identified as EL students.

Table 1.1 shows the differences between the Initial and Summative ELPAC.

Table 1.1 Differences Between the Initial and Summative ELPAC

Initial ELPAC	Summative ELPAC
This is an assessment used to identify a student as either an EL who needs support to learn English or as proficient in English.	This is an assessment used to measure the skills of EL students. The results will help the school or LEA determine if the student is ready to be reclassified as proficient in English.
This assessment is administered within 30 days of when the student enrolls in a California school for the first time.	This assessment is administered every spring from February 1 to May 31.
A student takes this test one time only.	A student takes this test annually until reclassified.
There is one test form.	The test form is revised annually.
There are six grade levels and grade spans: kindergarten, 1, 2, 3–5, 6–8, and 9–12.	There are seven grade levels and grade spans: kindergarten, 1, 2, 3–5, 6–8, 9–10, and 11–12.
The Speaking and Writing domains are locally scored by a trained ELPAC test examiner, whereas the Listening and Reading domains are machine scored. Raw scores are entered into the Data Entry Interface (DEI) and Teacher Hand Scoring System Local Scoring Tool. Student Score Reports (SSRs) are locally printed by designated staff.	The Speaking domain is locally scored, and raw scores are entered into the DEI. The Writing domain is scored by ETS. The Listening and Reading domains are machine scored. This is scored by ETS, and SSRs are provided by ETS to the LEAs.

1.1.2 Mode Comparability of the Computer-based Initial ELPAC

The goal of the mode comparability study was to establish links that preserve the substantive meaning of the reported score scale and allow valid comparisons and interpretations of the ELPAC paper-based and computer-based assessment scores. Comprehensive mode comparability analyses were supported by the Summative and Initial ELPAC field test designs. An overview of the designs and the plan to link computer-based scores to the paper-based scale is provided in table 2.1.

The methodology, analyses, and results of this study are outlined in the *Initial ELPAC Mode Comparability Memorandum* (ETS, 2020a), which relied on the Summative mode comparability data provided in the CDE report, *A Study of Mode Comparability for the Transition to Computer-based English Language Proficiency Assessments for California: Results from the Psychometric Analyses of Computer-based Assessment* (ETS, 2020b). Ultimately, the decision was made to use results from the common item linking design to link the computer-based scores to the paper-based scale.

1.1.3 ELPAC Computer-based Field Test Forms

Two test forms, comprised of all four domains, were created for the computer-based ELPAC field test. These forms supported a combined Initial and Summative ELPAC field test administration. Data from these forms was used for statistical analyses and scaling.

Form 1 comprised the Summative ELPAC field test form for all grade levels and grade spans. This form included computer-based items that aligned to the 2018–2019 Summative ELPAC blueprint and the current Initial ELPAC blueprint. Additional Speaking and Listening items were included in this form for all grade levels and grade spans to serve as vertical and horizontal linking items. The Writing domain was administered using paper Answer Books for students in kindergarten through grade two only.

Form 2 was the Initial ELPAC field test form. This form also included computer-based items that aligned to the 2018–2019 Summative ELPAC blueprint and the current Initial ELPAC blueprint. Additional items were included to allow some flexibility in choosing the most effective items for the 2019–2020 operational test forms. The Writing domain was administered using paper Answer Books for students in kindergarten through grade two only. Results in this report are based solely on form 2.

1.2 Purposes of the Field Test

There were three main purposes for the computer-based ELPAC field test. First, it provided an opportunity for the LEAs to become familiar with the computer-based format of the ELPAC. Second, it generated item-level statistics that could inform the test specifications for the operational computer-based versions of both the Initial and Summative ELPAC. Third, the field test provided data to link the computer-based scores to the paper-based scale. The field tests were not used to report individual student scores to LEAs.

1.3 Intended Population

Students in kindergarten through high school who had an English Language Acquisition Status of EL, IFEP, or “to be determined” were eligible to participate in the computer-based ELPAC field test. Student participation in the field test was voluntary, and it is important to be wary of direct comparisons of field test students and operational populations from past years.

1.4 Testing Window and Times

The computer-based ELPAC field test window occurred from October 1 through October 25, 2019, but was later extended through November 8, 2019, due to fire emergencies that affected testing throughout the state. LEAs were able to schedule their testing sessions according to local preference within this window.

Table 1.2 shows the number of items and the estimated time to complete each Initial ELPAC field test form. LEAs were advised to administer the Summative and Initial ELPAC field test forms over multiple test sessions or days.

Table 1.2 Number of Items and Estimated Testing Times for Field Test Forms

Variable	Summative Field Test Form	Initial Field Test Form
Number of Items for Kindergarten–Grade 2 (K–2)	K–2: 64–86 items	K–2: 63–88 items
Estimated Time for K–2	K: 75–85 minutes 1: 85–95 minutes 2: 110–120 minutes	K: 75–85 minutes 1: 85–95 minutes 2: 105–115 minutes

Table 1.2 *(continuation)*

Variable	Summative Field Test Form	Initial Field Test Form
Number of Items for Grade Levels 3–12	3–12: 90–95 items	3–12: 88–94 items
Estimated Time for Grade Levels 3–12	3–12: 175–210 minutes	3–12: 175–210 minutes

1.5 Preparation for Local Educational Agencies (LEAs)

LEA recruitment to participate in the field test began in March 2019 when invitation packets were sent to superintendents and LEA ELPAC coordinators. Incentives to participate included early registration for Administration and Scoring Trainings, additional seats for committing to field testing 40 or more students, and stipends based on the number of students tested.

To ensure the computer-based ELPAC field test was a successful experience for ELPAC students and test examiners, the Sacramento County Office of Education (SCOE) dedicated the first 10 Summative ELPAC Administration and Scoring Training dates, from late September through early October, to the LEAs that agreed to participate in the field test. SCOE also provided training presentations and videos, training sets and calibration quizzes for the Speaking domain, and Speaking rubrics on the Moodle website for LEA and school staff to access and use during local trainings. (Moodle is a free, learning-management, open-source software.)

ETS provided online resources, videos, and webcasts with detailed information on ELPAC test administration procedures. In addition, ETS provided test administration resources to schools and LEAs. These resources included detailed information on topics such as technology readiness, test administration, test security, accommodations, the test delivery system (TDS), and other general testing rules.

1.6 Groups and Organizations Involved with the ELPAC

1.6.1 State Board of Education (SBE)

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

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

1.6.2 California Department of Education (CDE)

The CDE oversees California's public school system, which is responsible for the education of more than 6,180,000 children and young adults in more than 10,500[1] schools. California aims to provide a world-class education for all students, from early childhood to adulthood. The CDE serves the state by innovating and collaborating as a team with educators, school staff, parents/guardians, and community partners to prepare students to live, work, and thrive in a highly connected world.

Within the CDE, the Instruction & Measurement Branch oversees programs promoting innovation and improving student achievement. Programs include oversight of statewide assessments and the collection and reporting of educational data (CDE, 2020b).

1.6.3 California Educators

A variety of California educators, including school administrators and teachers experienced in teaching EL students, were selected based on their qualifications, experiences, demographics, and geographic locations and were invited to participate in the ELPAC development process. In this process, California educators participated in tasks that included standard setting, score reporting, and scoring the constructed-response (CR) items.

1.6.4 Contractors

1.6.4.1 Primary Contractor—Educational Testing Service

The CDE and the SBE contract with ETS to develop and administer the ELPAC field test. As the prime contractor, ETS has the overall responsibility for working with the CDE to implement and maintain an effective assessment system and to coordinate the work of ETS with its subcontractors. Activities directly conducted by ETS include, but are not limited to, the following:

- Providing management of the program activities
- Providing tiered help desk support to LEAs
- Developing all ELPAC items
- Constructing, producing, and controlling the quality of ELPAC test forms and related test materials, including grade- and content-specific *Directions for Administration*
- Hosting and maintaining a website with resources for the ELPAC
- Developing, hosting, and providing support for the Test Operations Management System (TOMS)
- Processing student test assignments
- Completing all psychometric procedures

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

1.6.4.2 Subcontractor—American Institutes for Research (AIR)

ETS also monitors and manages the work of AIR (now Cambium Assessment), subcontractor to ETS for California online assessments. Activities conducted by AIR include the following:

- Providing the AIR proprietary TDS, including the Student Testing Interface, Test Administrator Interface, Teacher Hand Scoring System, DEI, secure browser, and practice and training tests
- Hosting and providing support for its TDS
- Scoring machine-scorable items
- Providing high-level technology help desk support to LEAs for technology issues directly related to the TDS

1.6.4.3 Subcontractor—Sacramento County Office of Education (SCOE)

ETS contracted with SCOE to manage all activities associated with recruitment, training, and outreach, including the following:

- Supporting and training county offices of education, LEAs, and charter schools
- Developing informational materials
- Recruiting and logistics for the field test
- Producing training videos

1.7 Systems Overview and Functionality

1.7.1 Test Operations Management System (TOMS)

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

- Assigning and managing ELPAC online user roles
- Managing student test assignments and accessibility resources
- Reviewing test materials orders and pre-identification services
- Viewing and downloading reports
- Providing a platform for authorized user access to secure materials such as *Directions for Administration,* ELPAC user information, and access to the *ELPAC Security and Test Administration Incident Reporting System* form and the Appeals module

TOMS receives student enrollment data and LEA and school hierarchy data from the California Longitudinal Pupil Achievement Data System (CALPADS) via a daily feed. CALPADS is "a longitudinal data system used to maintain individual-level data, including student demographics, course data, discipline, assessments, staff assignments, and other data for state and federal reporting."[2] LEA staff involved in the administration of the ELPAC field test—such as LEA ELPAC coordinators, site ELPAC coordinators, and ELPAC test

[2] From the CDE CALPADS web page at http://www.cde.ca.gov/ds/sp/cl/

examiners—were assigned varying levels of access to TOMS. A description of user roles is explained more extensively in the *Test Operations Management System User Guide* (CDE, 2020c).

1.7.2 Test Delivery System (TDS)

The TDS is the means by which the statewide online assessments are delivered to students. Components of the TDS include

- the Test Administrator Interface, the web browser–based application that allows test examiners to activate student tests and monitor student testing;
- the Student Testing Interface, on which students take the test using the secure browser;
- the secure browser, the online application through which the Student Testing Interface may be accessed and through which students are prevented from accessing other applications during testing; and
- the DEI, the web browser–based application that, for the computer-based fall field test, allowed test examiners to enter scores for the Speaking domain.

1.7.3 Training Tests

The training tests were provided to LEAs to prepare students and LEA staff for the computer-based ELPAC field test. These tests simulate the experience of the computer-based ELPAC. Unlike the computer-based ELPAC, the training tests do not assess standards, gauge student success on the operational test, or produce scores. Students may access them using a web browser, although accessing them through the secure browser permits students to take the tests using the text-to-speech embedded accommodation and to test assistive technology.

The purpose of the training tests is to allow students and administrators to quickly become familiar with the user interface and components of the TDS as well as with the process of starting and completing a testing session.

1.7.4 Constructed-Response (CR) Scoring Systems for Educational Testing Service (ETS)

CR items from the Writing domain in the TDS and from the paper–pencil test forms were routed to ETS' CR scoring system. CR items were scored by certified raters. Targeted efforts were made to hire California educators for human-scoring opportunities. Hired raters were provided in-depth training and certified before starting the human-scoring process. Human raters were supervised by a scoring leader and provided ELPAC scoring materials such as anchor sets, scoring rubrics, validity samples, qualifying sets, and condition codes for unscorable responses within the interface. The quality-control processes for CR scoring are explained further in *Chapter 8: Quality Control*.

1.8 Limitations of the Assessment

A limitation of the Initial ELPAC field test was the relatively small sample sizes for some grade levels and grade spans. This limitation will be discussed in more detail in chapter 6 of this report.

1.9 Overview of the Technical Report

This technical report addresses the characteristics of the computer-based ELPAC field test administered in fall of the 2019–2020 school year and contains nine additional chapters, as follows:

- Chapter 2 describes the procedures followed during item conversion to the computer-based administration, item review, and test assembly.
- Chapter 3 details the processes involved in the actual fall 2019 administration. It also describes the procedures followed to maintain test security throughout the test administration process.
- Chapter 4 provides information on the scoring processes. Also discussed is the development of materials such as scoring rubrics and range finding.
- Chapter 5 summarizes the statistical analysis plans for the fall 2019 field test.
- Chapter 6 summarizes the statistical analysis results for the fall 2019 field test, including
  - classical item analysis,
  - DIF analysis, and
  - item response theory calibration, linking, and scaling.
- Chapter 7 discusses the procedures designed to ensure the reliability and validity of score use and interpretations.
- Chapter 8 highlights the quality-control processes used at various stages of the computer-based ELPAC field test administration, including item development, test form development, test administration, scoring procedures, and psychometric analysis processes.
- Chapter 9 discusses the computer-based ELPAC field test post-test survey design, administration, and results.
- Chapter 10 details the ongoing means of program improvement.

References

California Department of Education. (2019). *English Language Proficiency Assessments for California (ELPAC)*. Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ep/

California Department of Education. (2020b, June). *Organization.* Sacramento, CA: California Department of Education. Retrieved from http://www.cde.ca.gov/re/di/or/

California Department of Education. (2020a, June). *State Board of Education responsibilities.* Sacramento, CA: California Department of Education. Retrieved from http://www.cde.ca.gov/be/ms/po/sberesponsibilities.asp

California Department of Education. (2020c) *Test Operations Management System User Guide, 2019–20.* Sacramento, CA: California Department of Education. Retrieved from https://www.elpac.org/s/pdf/CAASPP-ELPAC.toms-guide.2019-20.pdf

Educational Testing Service. (2020b). *A study of mode comparability for the transition to computer-based English Language Proficiency Assessments for California: Results from the psychometric analyses of computer-based assessment*. [Draft report]. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2020a). *A study of mode comparability for the transition to computer-based Initial English Language Proficiency Assessments for California: Results from the psychometric analyses of computer-based assessment*. [Draft memorandum]. Princeton, NJ: Educational Testing Service.

Chapter 2 Test Development

This chapter presents the detailed procedures of item conversion and field test assembly for the Initial English Language Proficiency Assessments for California (ELPAC) field test administration.

2.1 Overview

To prepare for the Initial ELPAC field test, several design tasks were needed prior to conducting item conversion and test development tasks. A high-level test design was developed (California Department of Education [CDE], 2019a), a usability pilot was conducted, task type conversion specifications were created (CDE, 2019b), and an item use plan was formed. The entire pool of 2,289 paper-based items was converted for computer-based administration based on these plans. All items used in the Initial ELPAC field test came from this pool.

All converted items were reviewed to ensure that they contained accurate content and formatting. The field test forms were reviewed to ensure that they conformed to the *Test Blueprints for the Initial ELPAC* (2019c).

2.2 Initial ELPAC Test Blueprints

Test blueprints were developed to describe the content of the Initial ELPAC. The test blueprints contain four tables with information about the task types in each of the four language domains of Listening, Speaking, Reading, and Writing. Task types are individual items or sets of items that required a student to perform an activity to elicit information about the student's English language proficiency (ELP).

The test blueprints provide information about the number of items and points that are administered per task type within each grade level and domain. The test blueprints also provide two types of alignment between task types and the standards: "primary" and "secondary." Primary alignment indicates there is a close or strong match in terms of the language knowledge, skills, and abilities covered by both the task type and the standard. Secondary alignment indicates that there is a moderate or partial match between the standard and the item in terms of language knowledge, skills, and abilities.

In November 2015, the State Board of Education (SBE) approved the *Proposed Test Blueprints for the ELPAC* (CDE, 2015), which included some task types adapted from the California English Language Development Test items that were aligned with the 2012 *California English Language Development Standards, Kindergarten Through Grade 12* (2012 ELD Standards) (CDE, 2014a). After the SBE approved the *Proposed Test Blueprints for the ELPAC*, the first pilot of ELPAC items and the stand-alone sample field test of the Initial ELPAC was administered as a paper–pencil test. Analysis of the pilot and the stand-alone sample field test results led to modifications of the Initial ELPAC test blueprints. The names of some of the task types were changed and some of the task types were removed from the test blueprints.

The SBE approved the *Test Blueprints for the Initial ELPAC* (2019c) in March 2018 in advance of the first operational administration of the paper-based Initial ELPAC from July 1, 2018, through June 30, 2019. The same test blueprints were used to assemble tests for the computer-based field test administration.

2.3 High-Level Test Design

In 2016, the CDE authorized Educational Testing Service (ETS) to investigate theoretical and empirical literature about the advantages and potential challenges of computer-based assessments, as well as the suitability of the paper-based ELPAC task types for transition to computer-based assessment. The results were reported in *Considerations in the Transition of the ELPAC Paper-Pencil Tests to Computer-Based Assessments* (CDE, 2017), which provided recommendations for consideration when transitioning to a computer-based ELPAC and confirmed the suitability of the paper-based ELPAC task types for transition to a computer-based platform. The report found that the task types on the paper-based ELPAC were appropriate for measuring the 2012 ELD Standards and could be used on a computer-based platform with relatively modest adaptations to take advantage of that platform. This finding was supported by feedback from classroom educators that the existing ELPAC task types did an effective job of measuring student ELP consistent with how 2012 ELD Standards were being implemented in classrooms. Similarly, the model for administration for the computer-based ELPAC followed the model used for the paper-based ELPAC, including one-on-one assessment of students in kindergarten (K) and grade one for all domains and one-on-one administration of the Speaking domain in all grade levels.

In 2018, the CDE called for the transition of the paper-based ELPAC to the computer-based ELPAC. ETS provided plans for this transition in the *Proposed High-Level Test Design for the Transition to Computer-Based ELPAC* (CDE, 2019a). The document provided an overview of the assessment purposes, test-taking population, and test design for the computer-based ELPAC. The test design drew upon current best practices and the latest research findings, and it maintained consistency with California's *English Language Arts/English Language Development Framework* (CDE, 2014b). The test design described guiding principles for developing a computer-based assessment at K through grade twelve in the domains of Listening, Speaking, and Reading. In the domain of Writing, the design included development of computer-based assessments at grades three through twelve while retaining paper-based K through grade two (K–2) Writing assessments.

The *Proposed High-Level Test Design for the Transition to Computer-Based ELPAC* (2019ba) was presented to the SBE in May 2019. The SBE approved the high-level test design in May 2019 with the amendment that grade two students would be administered the Listening and Reading domains one-on-one with a test examiner instead of in small-group administrations.

2.4 Usability Pilot

As part of the transition work, ETS, in collaboration with the CDE and the Sacramento County Office of Education, conducted a small-scale usability pilot employing cognitive laboratory methodology (henceforth called "the usability pilot") on the ELPAC task types in an online environment. The study was conducted at the earliest stage of the development of the computer-based ELPAC prior to the large-scale conversion of paper-based ELPAC items to a computer-based format. The usability pilot methodology, findings, and recommendations were described in the *ELPAC Usability Pilot: A Final Report* (CDE, 2019a).

2.4.1 Participants

The study was limited to a small sample size due to its one-on-one, intensive data collection methodology. Thus, it is possible that other students with different characteristics not represented in the sample may experience different outcomes when interacting with the computer-based ELPAC.

Six schools across two local educational agencies (LEAs) participated in the study. The LEAs and schools were selected because they represented the key variables of interest. Specifically, recruitment efforts were made to ensure that students who had little experience in computer-based assessment (e.g., transitional K–2 and recently arrived English learner [EL] students) were included in the study. A small number of non-EL students who would be able to perform their grade-appropriate tasks also were included in the sampling criteria to allow researchers to identify any EL-specific difficulties in interacting with the computer-based assessment features.

Participating students represented diverse background characteristics in terms of their grade level, ELP level, home language, recently arrived EL status, computer familiarity, and disability status. A total of 19 test examiners and 107 students—89 EL and 18 non-EL—from transitional K to grade eight participated in the study. Of the 89 EL students, 13 were EL students with disabilities.

The rationale to exclude grades nine through twelve is that ELPAC task types were the same in grades six through eight and grades nine through twelve, and that linguistic and cognitive processes of students in grades six through eight and nine through twelve would be similar. That is, findings about the usability of computer-based assessment features based on the sample for grades six through eight would be applicable to the conversion of grades-nine-through-twelve materials for the computer-based assessment format. Furthermore, targeting the sample in this way greatly reduced strain on the LEA to provide participants.

Because the sample was not representative of geographic diversity across the state of California, widespread generalizations could not be made based on the results of the study. Still, the usability pilot provided valuable information on how to better improve the conversion of the ELPAC task types and computer-based administration.

2.4.2 Recommendations

Based on the findings, the following recommendations were made to guide test developers in appropriately converting the paper-based ELPAC to the computer-delivery format when preparing for the field test as well as for future operational administration of the computer-based ELPAC. The recommendations were also intended to enhance the usability of the platform, computer-based ELPAC items, and their administration materials for test users. The recommendations were as follows:

1. Improve test familiarity materials (tutorials, training tests, practice tests) to ensure students are prepared to take the computer-based ELPAC and test examiners are prepared to administer it
2. Create resource materials for educators and test examiners to help determine if students are ready to take the computer-based ELPAC under typical conditions
3. Allow students to listen only once to audio stimuli on the Listening domain

4. Maintain recorded audio files for Listening stimuli on the K and grade one Listening tests, similar to the grades two through eight Listening tests
5. Increase opportunity for familiarity and practice with accessibility resources for both test examiners and students
6. Provide appropriate supports to ensure students' level of familiarity with technology does not impede their ability to take the computer-based ELPAC
7. Simplify the Speaking administration to make the administration of the assessment and scoring easier for the test examiner
8. Improve the organization of the *Directions for Administration (DFAs)*
9. Enhance test examiner training on administration and scoring

Detailed results and proposed action items for each recommendation were provided in the *ELPAC Usability Pilot: A Final Report* (CDE, 2019d). In addition, an addendum was created to describe how the recommendations from the final report were implemented in preparation for the computer-based ELPAC field test. The addendum describes actions that were taken to implement the recommendations, along with the implementation dates. The actions are provided in *Chapter 10: Continuous Improvement*.

2.5 Task Type Conversion Process

In preparation for the Initial ELPAC field test, ETS carefully analyzed the best way to convert each task type for computer-based delivery and documented this analysis in the *Specifications for Conversion of ELPAC Task Types for Computer-Based Delivery* (CDE, 2019b). The specifications described the details of the process followed to prepare Initial ELPAC paper-based items for computer-based delivery, including the screen layout, the use of images, the use of audio, and the features of the *DFAs*. The *Specifications for Conversion of ELPAC Task Types for Computer-Based Delivery* was first used to guide the conversion of approximately 125 ELPAC items for the computer-based usability pilot and cognitive labs that were held in April 2019. The document was updated based on the recommendations of the usability pilot and then used to guide the conversion of the entire pool of over 2,200 paper-based ELPAC items for the computer-based field test.

The pool of over 2,200 paper–pencil items underwent a rigorous conversion and review process. The items were converted according to the *Specifications for Conversion of ELPAC Task Types for Computer-Based Delivery*. Item-level directions were updated to make them appropriate for a computer-based administration.

When necessary, new audio files were recorded. All audio files were recompressed into two file types: audio files for Windows products and audio files for iPads and other iOS products. In addition, the black-and-white graphics that had been used in paper-based administrations were converted to color graphics that were compliant with the Web Content Accessibility Guidelines 2.1 (World Wide Web Consortium, 2018).

All updated text, audio files, and graphics files were entered in appropriate layouts within the ETS Item Banking Information System (IBIS). Assessment specialists familiar with the layout of the computer-based items reviewed each converted item to ensure that the text, audio, and graphics all functioned correctly in the IBIS item previewer. The converted items were then provided to the CDE for review within IBIS.

CDE staff provided ETS with comments regarding any needed revisions. The items were revised and members of the CDE ensured that any revisions were implemented accurately before the converted items were approved for use.

The high-level test design and the usability pilot guided the development of the *Specifications for Conversion of ELPAC Task Types for Computer-Based Delivery*. Based on the specifications, the Listening, Speaking, Reading, and Writing domains were administered in the Initial ELPAC field test as described in subsection *2.5.1 Listening Domain* through subsection *2.5.4 Writing Domain*.

2.5.1 Listening Domain

During the computer-based Initial ELPAC field test, K–2 students sat one-on-one with a test examiner. This allowed the test examiners to provide one-on-one support to operate the computer tools. At grades three through twelve, students progressed through the test independently. Students were able to play the Listening stimuli once unless they had an individualized education program or a Section 504 plan that allowed them to listen to the audio stimuli more than once. All students were able to play the directions, questions, and answer options multiple times.

2.5.2 Speaking Domain

In the Speaking domain, test examiners continued to administer items one-on-one to students, maintaining the interview style that was used in the paper-based ELPAC. On the computer-based ELPAC, however, students viewed images that accompanied items on a computer screen rather than in a printed Test Book. Test examiners continued to assign scores to student responses in the moment. On the computer-based ELPAC, however, there were two interfaces: in addition to the computer screen that students used to view stimuli and record their spoken responses, test examiners had a Data Entry Interface (DEI) into which they entered scores.

The computer-based ELPAC also used voice-capture technology to capture student responses to support the review of examiner-assigned scores.

2.5.3 Reading Domain

For the Reading domain, passages and items were presented on the computer-based ELPAC much as they appeared on the paper-based ELPAC. Directions on the computer-based ELPAC were presented as follows: The directions for K and grade one were read aloud by the test examiner from printed *DFAs*. For grades two through twelve, directions for the Reading domain were presented only as on-screen text without audio recordings. Item-level directions appeared on the same screen as the Reading stimulus.

2.5.4 Writing Domain

For the Writing domain, K–2 students wrote their responses in pencil in scannable Answer Books. The student experience remained paper-based to allow for the administration of items that aligned with the 2012 ELD Standards and conformed to best practices for literacy instruction in K–2. Scannable Answer Books were returned to ETS for scoring.

For students in grades three through twelve, the Writing test was taken solely on the computer. Students progressed through the Writing test independently and entered their responses using a keyboard. The directions were presented via audio recordings and as text on the screen. Students were able to replay the directions and item audio.

2.6 Item Use Plan

All items that were administered during the computer-based Initial ELPAC field test came from the existing paper–pencil item pool of over 2,200 items that was converted for computer-based administration. To the extent possible, items from the 2018–2019 paper-based Initial ELPAC were selected for field testing. All but two of the items that were field-tested came from the 2018–2019 paper-based Initial ELPAC. One grade span three through five Writing item was replaced in response to feedback that the item was not accessible to students who were deaf or hard of hearing. In addition, one grade span nine through twelve Reading item was replaced with another item from the same set of Reading items.

2.7 Task Types and Features

2.7.1 Task Types

The Initial ELPAC field test contained 23 task types. Each task type required a student to perform an activity to elicit information about the student's ELP. Each task type had one or more items that aligned with the 2012 ELD Standards. While the 2012 ELD Standards are organized according to three modes of communication (collaborative, interpretive, and productive communication), federal Title I requirements of the Every Student Succeeds Act (ESSA) of 2015 call for scores to be reported according to the four language domains of Listening, Speaking, Reading, and Writing (ESSA, 200.6[h][1][ii]).

The Listening domain of the Initial ELPAC had five task types, the Speaking domain had five task types, the Reading domain had eight task types, and the Writing domain had five task types. When a task type required the use of integrated language skills, such as Listening and Speaking, the task type was classified according to the language skill used to provide the response. For instance, the task type *Summarize an Academic Presentation* required a student to listen to a presentation and then summarize the presentation by speaking to the test examiner. Because the student provided the summary as a spoken response, the task type was classified as a Speaking task type.

The next subsections describe the task types used to assess ELP within each domain of the Initial ELPAC.

2.7.1.1 Listening Task Types

Listening task types for the Initial ELPAC assessed the ability of a student to comprehend spoken English (conversations, discussions, and oral presentations) in a range of social and academic contexts. Students listened to a stimulus and then demonstrated their ability to actively listen by answering multiple-choice (MC) questions. Students heard audio recordings of the Listening stimuli. The following are descriptions of the stimuli provided for the five Listening task types:

- **Listen to a Short Exchange, K through grade twelve:** Students heard a two-turn exchange between two speakers and then answered a question about the exchange.
- **Listen to a Classroom Conversation, grades three through twelve:** Students heard a multiple-turn conversation between two speakers and then answered three questions about the conversation.
- **Listen to a Story, K through grade five:** Students heard a multiple-turn conversation between two speakers and then answered three questions about the conversation.

- **Listen to an Oral Presentation, K through grade twelve:** Students heard an oral presentation on an academic topic and then answered three to four questions about the presentation.
- **Listen to a Speaker Support an Opinion, grades six through twelve:** Students heard an extended conversation between two classmates. In the conversation, one classmate made an argument in support of an opinion or academic topic. After listening to the conversation, students answered four questions.

2.7.1.2 Speaking Task Types

Speaking task types for the Initial ELPAC assessed the ability of a student to express information and ideas and to participate in grade-level conversations and class discussions. All task types included one or more constructed-response (CR) items. Test examiners scored student responses in the moment using scoring rubrics. The following are descriptions of the five Speaking task types:

- **Talk About a Scene, K through grade twelve:** The student was presented with an illustration of a familiar scene. The test examiner first asked three who-, what-, and when-type questions about the scene. The test examiner then administered three items intended to generate longer responses.
- **Speech Functions, grades three through twelve:** Students stated what they would say in a situation described by the test examiner.
- **Support an Opinion, K:** The student listened to a presentation about two activities, events, materials, or objects, and was asked to give an opinion about why one was better than the other. Students viewed a picture of the choices for context and support.
- **Retell a Narrative, K–2:** The student listened to a story that followed a series of pictures, and then the student used the pictures to retell the story.
- **Summarize an Academic Presentation, grade one through grade twelve:** The student listened to an academic presentation while looking at a related picture(s). The student was prompted to summarize the main points of the presentation using the illustration(s) and key terms of the presentation, if provided.

2.7.1.3 Reading Task Types

Reading task types for the Initial ELPAC assessed the ability of a student to read, analyze, and interpret a variety of grade-appropriate literary and informational texts. The following are descriptions of the eight Reading task types:

- **Read-Along Word with Scaffolding, K:** With scaffolding from the test examiner, the student provided the individual letter names and the initial letter sound for a decodable word. The student then answered a comprehension question about the word.
- **Read-Along Story with Scaffolding, K and grade one:** The student listened and followed along as the test examiner read aloud a literary text accompanied by three pictures for context and support. The student then answered a series of comprehension questions about the story.
- **Read-Along Information, grade one:** The student listened and followed along as the test examiner read aloud an informational text accompanied by three pictures for

context and support. The student then answered a series of comprehension questions about the information.

- **Read and Choose a Word, grades one and two:** The student read three words and chose the word that matched a picture.
- **Read and Choose a Sentence, grades two through twelve:** The student read three or four sentences and chose the sentence that best described a picture.
- **Read a Short Informational Passage, grades two through twelve:** The student read a short informational text and answered MC questions related to the text.
- **Read a Literary Passage, grade two:** The student read a literary text and answered MC questions related to the text.
- **Read an Informational Passage, grades three through twelve:** The student read an informational text and answered MC questions related to the text.

2.7.1.4 Writing Task Types

Writing task types for the Initial ELPAC assessed the ability of a student to write literary and informational texts to present, describe, and explain information. The following are descriptions of the five Writing task types:

- **Label a Picture—Word, with Scaffolding, K and grade one:** With scaffolding from the test examiner, the student wrote labels for objects displayed in a picture.
- **Write a Story Together with Scaffolding, K–2:** With scaffolding from the test examiner, the student collaborated with the test examiner to jointly compose a short literary text by adding letters, words, and a sentence to a story.
- **Describe a Picture, grades two through five:** At grade two, the student looked at a picture and wrote a brief description about what was happening. At grades three through five, the student looked at a picture and was prompted to examine a paragraph written by a classmate about what was happening in the picture. The student was asked to expand, correct, and combine different sentences written by a classmate before completing the final task of writing a sentence explaining what the students will do next.
- **Write About an Experience, grades six through twelve:** The student was provided with a common topic, such as a memorable classroom activity or event, and was prompted to write about the topic.
- **Justify an Opinion, grades three through twelve:** The student was asked to write an essay providing a position and appropriate supporting reasons about a school-related topic.

2.8 Item Review Process

Before Initial ELPAC items were designated as field-test ready, the draft versions underwent a thorough ETS internal review process, including two content reviews, a fairness review, and an editorial review; external reviews by item review panels; and a CDE review and final approval. This section describes the review process.

2.8.1 Educational Testing Service (ETS) Content Review

On all items ETS developed, content-area assessment specialists conducted two content reviews of items and stimuli. Assessment specialists verified that the items and stimuli were in compliance with ETS's written guidelines for clarity, style, accuracy, and appropriateness for California students as well as in compliance with the approved item specifications. Assessment specialists reviewed each item in terms of the following characteristics:

- Relevance of each item to the purpose of the test
- Match of each item to the Item Writing Guidelines for the ELPAC
- Match of each item to the principles of quality item writing
- Match of each item to the identified standard or standards
- Accuracy of the content of the item
- Readability of the item or passage
- Grade-level appropriateness of the item
- Appropriateness of any illustrations, graphs, or figures

Assessment specialists checked each item against its classification codes, both to evaluate the correctness of the classification and to confirm that the task posed by the item was relevant to the outcome it was intended to measure. The reviewers were able to accept the item and classification as written, suggest revisions, or recommend that the item be discarded. These steps occurred prior to the CDE's review.

2.8.2 ETS Editorial Review

After content-area assessment specialists reviewed each item, a group of specially trained editors also reviewed each item in preparation for consideration by the CDE and participants at the item review meeting. The editors checked items for clarity, correctness of language, appropriateness of language for the grade level assessed, adherence to the style guidelines, and conformity with accepted item writing practices.

2.8.3 ETS Sensitivity and Fairness Review

ETS assessment specialists who were specially trained to identify and eliminate questions that contain content or wording that could be construed to be offensive to, or biased against, members of specific ethnic, racial, or gender groups conducted the next level of review (ETS, 2014). These trained staff members reviewed every item before the CDE reviews and item review meetings.

The review process promoted a general awareness of, and responsiveness to, the following:

- Cultural diversity
- Diversity of background, cultural tradition, and viewpoints to be found in the test-taking populations
- Changing roles and attitudes toward various groups
- Role of language in setting and changing attitudes toward various groups

- Contributions of diverse groups (including ethnic and minority groups, individuals with disabilities, and women) to the history and culture of the United States and the achievements of individuals within these groups
- Item accessibility for ELs

All items drafted by California educators and ETS contractors went through internal ETS reviews, including two content reviews, an editorial review, and a fairness and sensitivity review. The items were then submitted to the CDE for review and approval.

2.8.4 California Educator Review

Each newly developed item was reviewed during two educator meetings in 2016: a Content Review Panel meeting and a Bias and Sensitivity Review Panel meeting. Items that eventually appeared on the Initial ELPAC and the Summative ELPAC were reviewed at the meetings, which were held at the Sacramento County Office of Education.

A total of 42 educators participated in the Content Review Panel meeting from August 1, 2016, through August 5, 2016. Six educators performed content reviews at each of the seven grade levels and grade spans. A total of 18 educators attended the Bias and Sensitivity Review Panel meeting from August 3, 2016 through August 5, 2016. Six educators performed bias and sensitivity reviews in each of three groups: one group for K–2 items, a second group for items for grade levels three through eight, and a third group for items for grade levels nine through twelve.

During the Content Review Panel meeting, California educators considered whether each item would appropriately measure the aligned standard(s), whether each item was appropriate for the designated grade level or grade span, and whether each item was presented clearly and effectively. MC items were also reviewed to ensure that each one had a single best key and distractors that were all plausible yet wrong. In addition, CR items were reviewed to make sure that each prompt would elicit a response that allowed students to demonstrate their language abilities, as described by the 2012 ELD Standards (CDE, 2014a).

During the Bias and Sensitivity Review Panel meeting, educators considered whether each item was free of content that was potentially biased against, or offensive to, any identified group, such as students from other countries or students who are deaf or hard of hearing. If an item contained potentially biased or offensive content, the educators considered whether the item could be revised to remove the potentially biased or offensive content.

Educators at both the Content Review Panel meeting and the Bias and Sensitivity Review Panel meeting had the option of making one of three decisions regarding each item: approve the item as is, approve the item with revisions, or reject the item. Whenever an item was approved with revisions, educators specified the revisions needed to text or images and the reasons for the proposed revisions.

After the educator meetings, CDE staff reviewed the proposed revisions and made final decisions as to whether each educator's proposed revisions should be implemented. ETS assessment specialists then applied the CDE-approved revisions. After the items were revised, CDE staff confirmed that revisions were entered correctly and approved the items for use as field test items.

2.9 Test Assembly

ETS assessment specialists assembled the Initial ELPAC field tests, which were reviewed and approved by the CDE. This process began with the creation of test development specifications, which described the content characteristics, psychometric characteristics, and quantity of items to be used in the Initial ELPAC field test. ETS created the test development specifications that the CDE reviewed and approved.

After the test development specifications were approved, ETS assessment specialists assembled the tests in IBIS according to the specifications. IBIS then generated form planners, which are spreadsheets containing essential item information such as the number of items, the alignment of items according to the 2012 ELD Standards, and the keys to MC items. ETS assessment specialists and psychometricians reviewed the form planners before they were delivered to the CDE for review. The CDE reviewed and approved the form planners after ETS revised the form planners as needed.

2.9.1 Field Test Forms

The Initial ELPAC field test form was administered as a computer-based field test in preparation for the 2020–2021 Initial ELPAC operational administration.

This subsection describes the composition of the entire combined—Summative ELPAC and Initial ELPAC—field test. However, the analyses in this report focus on the performance of items from the Initial ELPAC field test. Separate reports described the performance of items in the Summative ELPAC field test: its technical report (ETS, 2020a) and *A Study of Mode Comparability for the Transition to Computer-based ELPAC: Results from the Psychometric Analyses of Computer-based Assessment* (ETS, 2020b). The results of the mode comparability analyses for the Initial ELPAC data are provided in the in the CDE report, *A Study of Mode Comparability for the Transition to Computer-based Initial English Language Proficiency Assessments for California: Results from the Psychometric Analyses of Computer-based Assessment* (2020c).

The combined field test included a total of five field test forms per grade level or grade span: K, grade one, grade two, grade span three through five, grade span six through eight, grade span nine and ten, and grade span eleven and twelve. The following list provides descriptions of the five field test forms:

1. **Mode Comparability Study 1 (C1):** This was a paper reprint of the 2018–2019 Summative ELPAC paper–pencil test (PPT) form. Note that the Writing domain of Form C1 was not administered at K–2 because the responses for the computer-based ELPAC were already paper-based only for these grade levels.
2. **Mode Comparability Study 2 (C2):** This was a computer-based form consisting of oral language composite items—that is, items from the Listening and Speaking domains—that were computer renderings of all oral language items in the 2017–2018 Summative ELPAC, with additional items to allow its alignment to the adjusted Summative ELPAC test blueprints.
3. **Mode Comparability Study 3 (C3):** This was a computer-based form consisting of written language composite items—that is, items from the Reading and Writing domains—that were computer renderings of all written language items in the 2017–2018 Summative ELPAC, with additional items to allow its alignment to the adjusted Summative ELPAC test blueprints. Note that Form C3 at K–2 did not contain any

Writing items because the responses for the computer-based ELPAC remained paper-based only for these grade levels.

4. **Summative Field Test Form (F1):** This was a preassembled 2019–2020 computer-based Summative form aligned with the 2019-adjusted Summative ELPAC blueprints. It was also aligned to the *Test Blueprints for the Initial ELPAC* (CDE, 2019c). Additional oral language items were included to serve as vertical and horizontal linking items.
5. **Initial Field Test Form (F2):** This was a computer-based form aligned with both the Initial ELPAC test blueprints (the current operational Initial ELPAC form) and adjusted Summative ELPAC test blueprints. Additional written language items were included to serve as vertical and horizontal linking items.

Table 2.1 shows the test form configurations.

Table 2.1 Field Test Forms Descriptions

Variable	Mode Comparability Study 1 (C1)	Mode Comparability Study 2 (C2)	Mode Comparability Study 3 (C3)	Summative Field Test Form (F1)	Initial Field Test Form (F2)
Form Purpose	Compare student performance on the PPT and computer-based ELPAC	Compare student performance on the PPT and computer-based ELPAC	Compare student performance on the PPT and computer-based ELPAC	Field test items to be used on the 2020–2021 Summative ELPAC	Field test items to be used on the 2020–2021 Initial ELPAC
Domains	Writing (grades 3–12 only), Listening, Speaking, and Reading	Listening and Speaking	Writing (grades 3–12 only) and Reading	Listening, Speaking, Reading, and Writing	Listening, Speaking, Reading, and Writing
Test Format	PPT	Computer-based ELPAC	Computer-based ELPAC	Computer-based ELPAC with a PPT for K–2 Writing	Computer-based ELPAC with a PPT for K–2 Writing
Administration Plan	A sample takes C1 Listening and Speaking, as well as C2 Listening and Speaking; a separate sample takes C1 Writing (grades 3–12 only) and Reading, as well as C3 Writing (grades 3–12 only) and Reading.	A sample takes C1 Listening and Speaking, as well as C2 Listening and Speaking.	A sample takes C1 Writing (grades 3–12 only) and Reading, as well as C3 Writing (grades 3–12 only) and Reading.	A sample takes F1 only.	A sample takes F2 only.
Linking Plan	Anchored back to ELPAC reporting scale	Horizontal linking with F1	Horizontal linking with F2	In Listening and Speaking, horizontal linking with C2 and F2 plus vertical linking across all grades	In Reading and Writing, horizontal linking with C3 and F1 plus vertical linking across all grades

Assessment specialists at ETS developed form planners showing the number of items to be administered at each grade and domain. The form planners underwent standard ETS reviews, including a psychometric review, a content review, a fresh-perspective review, and an editorial review. The form planners were sent to the CDE in April 2019 for review and approval before items were exported to the American Institutes for Research (AIR) (now Cambium Assessment), the test delivery system vendor, in June 2019. After AIR developed the field test forms in the delivery platform, ETS and the CDE conducted user acceptance testing (UAT) in July and August 2019. The CDE approved the UAT for the forms before they were administered.

One Initial ELPAC field test form (F2) was developed for each grade level and grade span: K, grades one and two, and grade spans three through five, six through eight, and nine through twelve. Table 2.2 shows the numbers of items in each domain.

Table 2.2 Numbers of Items in Initial Field Test Form

Grade Level or Grade Span	Listening Items	Speaking Items	Reading Items	Writing Items
K	20	11	20	12
Grade 1	25	11	29	13
Grade 2	25	14	31	10
Grade span 3–5	33	15	37	8
Grade span 6–8	29	14	28	9
Grade span 9–12	26	14	38	9

2.9.2 Considerations for Fall Testing Window

The Initial ELPAC testing window runs the entire year (July 1–June 30), with the bulk of its testing volume taking place at the beginning of the school year, typically August through October. The computer-based ELPAC field test window occurred from October 1 through October 25, 2019, but was later extended through November 8, 2019, due to fire emergencies that affected testing throughout the state. The field test administration took place during the early portion of a normal operational testing window.

2.9.3 Psychometric Review

The ETS Psychometric Analysis & Research (PAR) group reviewed the field test forms to ensure that they aligned with the field test design. The PAR review also ensured that the field test forms conformed to the Initial ELPAC test blueprints.

Six field test forms were reviewed, one for each grade level and grade span. These were identified as F2 forms as described in subsection 2.9.1. The following criteria were evaluated for each form:

- All items from the 2018–2019 Initial ELPAC paper-based forms were included.
- The forms aligned with the Summative ELPAC and Initial ELPAC blueprints.
- The forms contained the same number of items as described in table 1.1.

The number of items and total score points, for each task type, were aggregated within each domain. These summary counts were then compared with the associated values in the blueprint. The psychometricians determined that each of the six forms contained enough items and score points, across task types, to meet the form requirements specified by the

Summative ELPAC and Initial ELPAC blueprints. Additionally, it was confirmed that all items from the 2018–2019 paper-based Initial ELPAC were included in the field test forms.

The psychometric review established that the field test forms met the expected criteria.

2.9.4 CDE Review

The CDE used a three-stage gatekeeper process to review all test materials. Test materials for review and approval by the CDE included form planners, *DFAs*, K–2 Writing Answer Books, student-facing items in the test delivery system, and DEI items for the entry of Speaking scores. All test materials were approved before they were posted for use.

For the reviews of form planners, *DFAs*, and K–2 Writing Answer Books, ETS initiated the review by submitting materials to the CDE via the gatekeeper system, along with the criteria for the review. CDE consultants performed the initial review and returned comments and requests for revisions to ETS. ETS staff then revised the materials as requested and returned them to the CDE consultants, who then reviewed the updated materials. If the test materials needed additional revisions, they were returned to ETS for further modifications.

Once CDE consultants found the test materials met the review criteria, the CDE consultants submitted the test materials to the CDE administrator for approval. Test materials that were approved with revisions were revised by ETS and resubmitted for approval. Test materials that were not approved needed significant revisions and had to be submitted to the consultants again before they could be resubmitted to the CDE administrator for approval.

For the reviews of student-facing items for the test delivery system and the DEI items for the entry of Speaking scores, CDE staff conducted a two-stage UAT. During the first stage, CDE staff reviewed the computer-based content and entered any needed revisions in a log. AIR staff updated the items based on the comments and provided them to CDE staff for a second review. All issues with the computer-based items were resolved before they were approved for the field test administration.

References

California Department of Education. (2014a). *2012 California English language development standards: Kindergarten through grade 12*. Sacramento, CA: California Department of Education. Retrieved from http://www.cde.ca.gov/sp/el/er/documents/eldstndspublication14.pdf

California Department of Education. (2014b). *English language arts/English language development framework*. Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ci/rl/cf/

California Department of Education. (2015). *Proposed test blueprints for the English Language Proficiency Assessments for California.* Approved by the California State Board of Education in November 2015. Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/be/ag/ag/yr15/documents/nov15item12.doc

California Department of Education. (2017). *Considerations in the transition of the English Language Proficiency Assessments for California paper-pencil tests to computer-based assessments.* Sacramento, CA: California Department of Education. Retrieved from: https://www.cde.ca.gov/ta/tg/ep/documents/elpaccbareporttagged.pdf

California Department of Education. (2019d). *ELPAC usability pilot: A final report (with addendum).* [Unpublished report]. Sacramento, CA: California Department of Education.

California Department of Education. (2019a). *Proposed high-level test design for the transition to computer-based English Language Proficiency Assessments for California*. Approved by the California State Board of Education in May 2019. Sacramento, CA: California Department of Education.

California Department of Education. (2019b). *Specifications for conversion of ELPAC task types for computer-based delivery.* [Unpublished report]. Sacramento, CA: California Department of Education.

California Department of Education. (2019c). *Test blueprints for the Initial English Language Proficiency Assessments for California*. Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ep/documents/elpacinitialbluprt.pdf

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

Educational Testing Service. (2020a). *Computer-based Summative English Language Proficiency Assessments for California (ELPAC) fall 2019 field test technical report.* [Unpublished report]. Sacramento, CA: California Department of Education.

Educational Testing Service. (2020b). *A Study of mode comparability for the transition to computer-based ELPAC: Results from the psychometric analyses of computer-based assessment, 2019–2020 Administration*. [Draft report]. Sacramento, CA: California Department of Education.

Educational Testing Service. (2020c). *A study of mode comparability for the transition to computer-based Initial English Language Proficiency Assessments for California: Results from the psychometric analyses of computer-based assessment*. [Draft report]. Princeton, NJ: Educational Testing Service.

S. 1177—114th Congress: Every Student Succeeds Act. 2015. Title 1, Part A, Subpart A, Subject Group 127 Standards and Assessments, Section 200.6 Inclusion of all students.

World Wide Web Consortium. (2018). Web Content Accessibility Guidelines 2.1. WC3. Retrieved from: https://www.w3.org/TR/WCAG21/

Chapter 3 Test Administration

This chapter provides the details of administering the computer-based Initial English Language Proficiency Assessments for California (ELPAC) field test, as well as test security, accessibility resources, participation, and demographic summaries.

3.1 Field Test Administration

All local educational agencies (LEAs) were required to attend the statewide Summative 2019–20 ELPAC Administration and Scoring training in fall 2019. Of the 20 statewide trainings planned, the first 10 were dedicated to the LEAs participating in the field test and were spread across the state, covering northern, central, and southern California, as well as the San Francisco Bay Area.

In accordance with the procedures for all California assessments, LEAs identified test examiners to administer the ELPAC field 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, videos, and webcasts, to ensure that the LEA staff and test examiners understood how to administer the computer-based ELPAC field test.

The field test was designed for one-on-one administration between a single student and a test examiner for kindergarten through grade two in three domains and group administration for grades three through twelve. The exceptions were the Speaking domain, which was administered one-on-one for all grade levels, and the Writing domain, which had optional small-group administration for grade two.

Students were provided with a computer or testing device on which to take the assessment. Test examiners used a separate computer or testing device on which to access the Test Administrator Interface and manage the testing session. The ELPAC field test used the same secure browser and online testing platform as all the California Assessment of Student Performance and Progress (CAASPP) assessments.

Test examiners were required to use the *Directions for Administration (DFAs)*, housed in TOMS, to administer tests to students. For the field test, there was a combined *DFA* for the Listening, Reading, and Writing domains and a separate *DFA* for the Speaking domain. The last page of the Speaking domain *DFA* contained a student score sheet that was provided for optional use by the test examiner to record a student's Speaking scores in the moment. This student score sheet could then be used to enter the student's Speaking scores into the Data Entry Interface. The other option for test examiners was to enter the student's Speaking scores directly into the DEI during the administration of the Speaking domain.

3.2 Test Security and Confidentiality

All testing materials for the fall 2019 Initial ELPAC field test—Test Books, Answer Books, *Examiner's Manuals,* and *DFAs*—were considered secure documents. Every person having access to test materials was required to maintain the security and confidentiality of the test materials. ETS' Code of Ethics requires that all test information, including tangible materials (e.g., test booklets, test questions, test results), confidential files, processes, and activities are kept secure.

To ensure security for all tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI). In the pursuit of enforcing secure practices, ETS and the OTI strive

to safeguard the various processes involved in a test development and administration cycle. For the fall 2019 Initial ELPAC field test, those processes included the following:

- Test development
- Item and data review
- Item banking
- Transfer of forms and items to the California Department of Education (CDE) and American Institutes for Research (now Cambium Assessment)
- Security of electronic files using a firewall
- Printing and publishing
- Test administration
- Test delivery
- Processing and scoring
- Data management
- Statistical analysis
- Student confidentiality

3.2.1 Educational Testing Service's Office of Testing Integrity (OTI)

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

The OTI's mission is to

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

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

3.2.2 Procedures to Maintain Standardization of Test Security

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

keeping student information confidential, and making sure the site ELPAC coordinators and ELPAC test examiners are properly trained regarding security policies and procedures.

The site ELPAC coordinator is responsible for mitigating test security incidents at the test site, keeping test materials secure, and reporting incidents to the LEA ELPAC coordinator.

The ELPAC test examiner is responsible for reporting testing incidents to the site ELPAC coordinator, keeping test materials secure, and securely destroying printed and digital media for *DFAs* (CDE, 2019a).

The following measures ensured the security of the ELPAC:

- LEA ELPAC coordinators and site ELPAC coordinators must have electronically signed and submitted an ELPAC *Test Security Agreement* in 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 have electronically signed and submitted an ELPAC *Test Security Affidavit* in TOMS before receiving access to any testing materials (5 *CCR*, Section 859[c]).

In addition, it was the responsibility of every participant in the Initial ELPAC field test administration to immediately report any violation or suspected violation of test security or confidentiality. The ELPAC test examiner reported to the site ELPAC coordinator or LEA ELPAC coordinator, who then submitted the incident using the Security and Test Administration Incident Reporting System (STAIRS)/Appeals process. Breach incidents were to be reported by the LEA ELPAC coordinator to the California Technical Assistance Center (CalTAC) and entered into STAIRS within 24 hours of the incident (5 *CCR*, Section 859[e]).

3.2.3 Security of Electronic Files Using a Firewall

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

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

3.2.4 Data Management

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

All stored test content and student data is encrypted. ETS complies with the Family Educational Rights and Privacy Act (20 *United States Code [USC]* § 1232g; 34 *Code of*

Federal Regulations Part 99) and the Children's Online Privacy Protection Act (15 *USC* §§ 6501–6506, P.L. No. 105–277, 112 Stat. 2681–1728).

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

3.2.5 Statistical Analysis on Secure Servers

Immediately following submission of the fall 2019 Initial ELPAC field test results into the test delivery system, either computer-based or scanned paper-based,[3] results were transmitted to scoring systems for human and machine scoring. For paper-based results, several quality control checks were implemented. These included verifying there was no damage to the Answer Books prior to scanning as well as capturing issues such as double marks and inconsistencies between pre-identification labels and marked information. All responses were securely stored using the latest industry standards. Human scoring occurred through the ETS trained network of human raters.

After constructed-response (CR) items were scored, the Information Technology team at ETS extracted data files from the secure file transfer protocol site and loaded them into a database that contained results from both the multiple-choice and CR items. Final scoring of results from all item types was conducted by the Enterprise Score Key Management scoring system.

The ETS Data Quality Services staff extracted the data from the database and performed quality-control procedures before passing files to the ETS Psychometric Analysis & Research (PAR) group. The PAR group kept all data files on secure servers. This data was then used to conduct all statistical analyses. All staff members involved with the data adhered to the ETS Code of Ethics and the ETS Information Protection Policies to prevent any unauthorized access to data.

3.2.6 Student Confidentiality

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

3.2.7 Security and Test Administration Incident Reporting System (STAIRS) Process

The LEA ELPAC coordinator or site ELPAC coordinator was responsible for immediately reporting all testing incidents and security breaches. The online ELPAC STAIRS form, available in TOMS, was the starting point for LEA ELPAC coordinators and site ELPAC coordinators to report a test security incident or other testing issue. For the Initial ELPAC field test, all computer-based test irregularities were required to be reported in STAIRS. For kindergarten through grade two paper-based Writing irregularities, the test examiner only needed to fill in the *Test Irregularity* circle on the back cover of the Writing Answer Book.

If an irregularity or security breach occurred at the school, the test examiner was required to report the incident to the site ELPAC coordinator, who would then report the incident to the

[3] Only the Writing domain for kindergarten through grade two was administered on paper.

LEA ELPAC coordinator. Testing irregularities relate to incidents that occurred during the administration of the ELPAC that were likely to impact the reliability and validity of test interpretations.

Testing irregularities included, *but were not limited to,* the following:

- Cheating by students
- Failing to follow test administration directions
- Rushing students through the test or parts of the test
- Coaching students, including, *but not limited to,* the following:
  - Discussing questions with students before, during, or after testing
  - Giving or providing any clues to the answers
- Administering the wrong grade level or grade span test to a student or using mismatched test materials
- Writing on the scannable Answer Book by a test examiner that would cause the Answer Book to be unscorable and therefore need transcription to a new Answer Book
- Leaving instructional materials on walls in the testing room that may assist students in answering test questions
- Allowing students to have additional materials or tools (e.g., books, tables) that are **not** specified in an individualized education program (IEP), Section 504 plan, or approved by the CDE as an allowed testing accommodation

Security breaches included, *but were not limited to,* the following*:*

- Site ELPAC coordinators, test examiners, proctors, or students using electronic devices such as cell phones during testing
- Posting pictures of test materials on social media sites
- Missing test materials
- Copying or taking a photo of any part of the test materials
- Permitting eligible students access to test materials outside of the testing periods
- Failing to maintain security of all test materials
- Sharing test items or other secure materials with anyone who has not signed the *ELPAC Test Security Affidavit*
- Discussing test content or using test materials outside of training and administration
- Allowing students to take the test out of the designated testing area
- Allowing test examiners to take test materials home
- Allowing untrained personnel to administer the test

If an incident occurred, the LEA ELPAC coordinator was instructed to enter the incident in STAIRS within 24 hours of the incident. Depending on the type of incident submitted, either the CDE or CalTAC would review the form to determine whether the testing issue required additional action by the LEA.

3.3 Universal Tools, Designated Supports, and Accommodations for Students with Disabilities

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

The CDE's Matrix Four (CDE, 2019b) is intended for school-level personnel and IEP and Section 504 plan teams to select and administer the appropriate universal tools, designated supports, and accommodations as deemed necessary for individual students.[4]

The computer-based Initial ELPAC field test offered commonly used accessibility resources available for paper–pencil test (PPT) administration as non-embedded resources and through the CAASPP online testing platform as embedded and non-embedded resources, where applicable for the tested construct.

3.3.1 Universal Tools

Universal tools are available to all students by default, although they can be disabled if a student finds them distracting. Each universal tool falls into one of two categories: embedded and non-embedded. Embedded universal tools are provided through the student testing interface (through the secure browser), although they can be turned off by a test examiner. Students who were assigned to take the paper–pencil field test form did not have access to embedded universal tools.

The following embedded universal tools were available to students testing in the secure browser:

Breaks
Digital notepad
Expandable passages
Expandable items
Highlighter
Keyboard navigation
Line reader (grades three through twelve)
Mark for review (grades two through twelve)
Strikethrough (grades three through twelve)
Writing tools (grades three through twelve)
Zoom (in or out)

[4] This technical report is based on the version of Matrix Four that was available during the computer-based Initial ELPAC 2019 fall field test.

The following non-embedded universal tools were available to students testing in the secure browser:

Breaks
Oral clarification of test directions by the test examiner in English
Scratch paper
Test navigation assistant

The following non-embedded universal tools were available to students taking the paper–pencil field test forms:

Breaks
Highlighter
Line reader (grades three through twelve)
Mark for review (grades two through twelve)
Oral clarification of test directions by the test examiner in English
Scratch paper
Strikethrough (grades three through twelve)

3.3.2 Designated Supports

Designated supports are available to all students and must be set by an LEA ELPAC coordinator or site ELPAC coordinator in the test settings in TOMS. The designated supports each fall into one of two categories: embedded and non-embedded. Embedded designated supports are provided through the student testing interface (through the secure browser).

The following embedded designated supports were available to students testing in the secure browser:

Color contrast
Masking
Mouse pointer (size and color)
Pause or replay audio—Listening domain
Pause or replay audio—Speaking domain
Permissive mode
Print size
Streamline
Turn off any universal tool(s)

The following non-embedded designated supports were available to students testing in the secure browser:

Amplification
Color contrast
Color overlay
Designated interface assistant
Magnification
Medical supports
Noise buffers
Print on demand
Read aloud for items (Writing domain)
Separate setting

The following non-embedded designated supports were available to students taking the paper–pencil Writing domain field test for kindergarten through grade two:

Amplification
American Sign Language or Manually Coded English
Color overlay
Magnification
Masking
Medical supports
Noise buffers
Read aloud for items (Writing domain)
Separate setting

3.3.3 Accommodations

Accommodations are changes in procedures or materials that increase equitable access during the ELPAC assessments and are available to students who have a documented need for the accommodation(s) via an IEP or Section 504 plan. Assessment accommodations generate valid assessment results for students who need them; they allow these students to show what they know and can do. Accommodations do not compromise the learning expectations, construct, grade-level standard, or intended outcome of the assessments. For the computer-based field test, embedded accommodations were not available.

The following non-embedded accommodations were available to students testing in the secure browser:

Alternate response options
Scribe
Speech-to-text

The following non-embedded accommodations were available to students taking the paper–pencil Writing domain field test for kindergarten through grade two:

Alternate response options
American Sign Language or Manually Coded English
Breaks
Scribe

3.3.4 Resources for Selection of Accessibility Resources

The full list of the universal tools, designated supports, and accommodations that are available for the ELPAC are documented in Matrix Four (CDE, 2019b). Part 1 of Matrix Four lists the embedded and non-embedded universal tools available for ELPAC testing. Part 2 of Matrix Four includes the embedded and non-embedded designated supports that are available for ELPAC testing. Part 3 of Matrix Four includes the embedded and non-embedded accommodations available for ELPAC testing. School-level personnel, IEP teams, and Section 504 teams used Matrix Four when deciding how best to support the student's test-taking experience.

3.3.5 Delivery of Accessibility Resources

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

Non-embedded resources are not part of the technology platform for online ELPAC testing. Examples of non-embedded resources include magnification, noise buffers, and the use of a scribe.

3.3.6 Unlisted Resources

Unlisted resources are non-embedded supports that may be provided if specified in an eligible student's IEP or Section 504 plan. Unlisted resources were not available for the computer-based Initial ELPAC field test.

3.4 Participation

Because student participation in the combined—Summative ELPAC and Initial ELPAC—field test was voluntary, the goal of the field test recruitment was to have as many eligible students and LEAs as possible participate. In spring 2019, a recruitment email was sent to the LEAs that had schools meeting the minimum threshold requirement of having a school with at least 20 English learner (EL) students in a grade level or grade span. The overall goal was to recruit approximately 56,000 students statewide for participation and have LEAs and schools that are geographically representative and diverse.

3.4.1 Rules for Including Student Responses in Analyses

Two sets of criteria were used to prepare student response data for statistical analyses. The first criterion was student EL status. The second criterion was the attemptedness indicators. Only EL students were included for the item and differential item functioning (DIF) analyses and item response theory (IRT) calibrations for the Initial ELPAC field test.

Attemptedness rules were applied to data where students responded to relatively few items. For initial data, students had to obtain at least one item score in each of the four domains to be kept in the final samples for item and DIF analyses. These rules were also applied to generate item response matrices to conduct IRT calibrations.

3.5 Demographic Summaries

The number and the percentage of students for selected groups with completed Initial ELPAC field test scores are provided, for all grade levels and grade spans, in table 3.A.1 through table 3.A.6 of appendix 3.A. Grade spans reflect students' enrolled grade spans during the 2019–2020 school year. For purposes of comparison, also provided in these demographic tables are the number and the percentage of students for selected groups with completed 2018 Initial ELPAC paper-based test scores.

In the tables, students are grouped by demographic characteristics, including gender, ethnicity, five identified countries of origin, 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. The tables in appendix 3.A show consistent patterns. For all grade levels and grade spans, female students accounted for about a half of the field test samples. It was also found that 80 percent or more of the students were Hispanic or Latino,

except for grade span nine through twelve, which reported 69 percent Hispanic or Latino. In terms of English proficiency, 86 to 94 percent of the test takers were ELs.

Students whose country of origin was identified as likely having limited access to technology were of particular concern in the transition from PPT to computer-based assessments. It was important that these students be able to participate in the new computer-based Initial ELPAC. However, all groups involved in supporting this transition recognized that appropriate resources were critical to help ensure that lack of prior technology access did not serve as a barrier to students' ability to do their best on these tests. In anticipation of the students coming from the five identified countries of origin where access to computers might be limited, as well as students who are technology novices in general, ETS and the CDE developed the Technology Readiness Checker for Students. This online resource was designed to help educators determine a student's familiarity with navigating an online interface. The purpose of the tool is for educators to better understand what kind of supports a student may need to increase technology familiarity, to understand what kind of support the student may need during the assessment, or both. The percentage of students coming from the five identified countries of origin where access to computers might be limited varied from less than 1 percent to about 11 percent, across grade levels and grade spans (refer to table 3.A.1 through table 3.A.41).

The demographic information for students taking each Initial ELPAC field test form looked similar to the distributions of the population of Initial ELPAC test takers in 2018–2019. These are reported in appendix 11 of the *2018–2019 Initial ELPAC Technical Report* (CDE, 2020). Across grade levels and grade spans, male students accounted for 50 to 60 percent of ELPAC test takers in both the 2018–2019 Initial ELPAC PPT and the field test data. Across the grades, both sets of data contained more than 75 percent of Hispanic or Latino students. The percentage of students not receiving special education services for the field test sample was very similar to that observed for the 2018–2019 population, 96 to 100 percent.

Table 3.1 lists the demographic student groups reported.

Table 3.1 Demographic Student Groups to Be Reported

Category	Student Groups
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 • Unknown

Table 3.1 *(continuation)*

Category	Student Groups
Five Identified Countries of Origin	• Guatemala • Honduras • Colombia • El Salvador • Afghanistan
English-Language Fluency	• Initial fluent English proficient (IFEP) • English learner (EL) • To be determined (TBD)
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 • Duration unknown
Migrant Status	• Migrant • Nonmigrant

3.6 Training Test

The training tests were provided to LEAs to prepare students and LEA staff for the computer-based Initial ELPAC field test. These tests simulated the experience of the computer-based ELPAC. Unlike the computer-based ELPAC, the training tests did not assess standards, gauge student success on the operational test, or produce scores. Students could access the training tests using a secure browser; this permitted them to take the tests using the text-to-speech embedded accommodation and to test assistive technology.

The purpose of the training tests was to allow students and administrators to quickly become familiar with the user interface and components of the test delivery system as well as with the process of starting and completing a testing session.

References

California Department of Education. (2019a). *Computer-based ELPAC field test administration manual.* Sacramento, CA: California Department of Education. Retrieved from https://elpac.org/s/pdf/ELPAC.Field-Test-Administration-Manual.2019-20.pdf

California Department of Education. (2019b). *Matrix Four: Universal tools, designated supports, and accommodations for the English Language Proficiency Assessments for California*. Retrieved from https://www.cde.ca.gov/ta/tg/ca/accessibilityresources.asp

California Department of Education. (2020). *Initial English Language Proficiency Assessments for California technical report 2018–2019 administration*. [Unpublished report]. Sacramento, CA: California Department of Education.

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

Chapter 4 Scoring

This chapter summarizes constructed-response scoring at the item level for the computer-based Initial English Language Proficiency Assessments for California (ELPAC) field test.

4.1 Overview of Human Scoring for Constructed-Response (CR) Items

Speaking and Writing domains contain CR items; Listening and Reading domains do not include CR items.

Speaking CR items were scored locally by test examiners during the field test. Writing CR items from the test delivery system were routed to Educational Testing Service's (ETS') CR scoring systems. Writing items were scored by certified raters. Targeted efforts were made to hire California educators for human-scoring opportunities. Hired raters were provided in-depth training and certified before starting the human-scoring process. Human raters were supervised by a scoring leader and provided scoring materials such as scoring rubrics, anchor sets, and training samples within the interface. Writing responses for kindergarten through grade two students were entered in Writing Answer Books and then shipped to ETS for scoring. The quality-control processes for CR scoring are explained further in *Chapter 8: Quality Control*.

4.2 Sampling Process

Sampling procedures were not applied to the scoring of computer-based ELPAC CR items in the field test phase; all items were scored.

4.3 Scoring Rubric Development

For the previous paper–pencil Initial ELPAC, ETS' Assessment & Learning Technology Development (ALTD) group developed rubrics for the scoring of each Speaking and Writing CR task type. The same rubrics were used to score the computer-based Initial ELPAC field test.

Rubrics were edited as needed based on feedback from the California Department of Education (CDE) and California educators during the range finding process for the computer-based field test. Changes from the paper–pencil test (PPT) rubrics were made for clarification and to address keyboarding errors in grades three through twelve. Proposed rubric revisions underwent internal ETS ALTD review and CDE review, resulting in revisions to one Speaking rubric and several Writing rubrics. Only minor revisions were made, which were not expected to yield differences in the scoring process for the paper–pencil responses and computer-based responses.

4.4 Range Finding

Soon after receiving Writing responses from California schools, ETS and the Sacramento County Office of Education facilitated a range finding event at the Hilton Sacramento Arden West from October 28, 2019 through November 1, 2019. The range finding event enlisted 30 California educators to select responses for each Writing prompt that exemplified each score point on each rubric. These responses were then made into sample sets for training, calibrating (qualifying), and monitoring raters (scorers). In the process, some samples were

also annotated by California educators to explain how the rubrics applied to each response sample, resulting in a particular score.

The following steps describe how the range finding process was implemented for Writing. Note that range finding was not needed for Speaking; existing samples from PPTs had previously been selected by California educators and approved by the CDE. These samples were used to train and qualify local test examiners to score Speaking responses on the computer-based field test.

1. ETS staff used the rubrics (scoring guides) to initially prescore responses representing each score point on each item's rubric. The number of responses selected varied by prompt and were based on the number of points and the prompts that were preselected for certifying and training raters. The prescored responses formed a pool of potential samples from which California educators selected samples for the various purposes summarized in table 4.1.

Table 4.1 Computer-based ELPAC Field Test Sample Selection for Human Scoring Procedures

Sample Type	Purpose	Number of Sets and Samples in Sets	Configuration of Sets
Training	Training samples with annotations for rater training and scoring practice	• Two sets per task type per grade span	Varied depending on the task type and grade span, but generally two to three samples for each score point per set
Benchmarks or Anchors	Benchmark samples with annotations that represent exemplar responses at each score point on the rubric	• One set of samples per unique prompt per grade span	One to three samples for each score point
Calibration	Calibration samples for periodically qualifying raters to score a particular task type at a particular grade	• Two sets per task type per grade span • Mixed score points	One to three samples for each score point per set

2. Responses were reviewed and selected by several panels of California educators (with support from ETS ALTD staff) using the ETS Online Network for Evaluation (ONE) system at the range finding event. Educators also wrote annotations, or short notes, with each score point to explain why a response earned a particular rating. Annotations help raters make explicit connections between the scoring guide and responses, thus informing their careful and accurate scoring of responses. ETS provided the CDE with the scored samples, annotations, and recommendations for which responses would be used as benchmarks or anchor samples.

3. CDE and ETS content experts reviewed the samples, scores, and rationale for all benchmark or anchor samples to agree upon the scores and samples to use for specific sets. The annotations for the samples also were reviewed and refined as needed. The CDE made final decisions about samples to be used as anchors or benchmarks and about proposed changes to rubrics.
4. ETS created all final sample sets in the ONE system and used these samples as part of a system of training and controls for verifying the quality and consistency of pilot scoring.

4.5 Rater Recruitment and Certification Process

The rater pool was recruited from the same pool of raters that scored the Initial ELPAC Writing operational test. ETS recruited a pool of eligible raters experienced in scoring English language assessments. These raters underwent an extensive training for ELPAC content before participating in scoring.

4.6 Rater and Scoring Leader Training

ETS selected scoring leaders to oversee a group of raters during the scoring process. Scoring leaders were experienced raters who had demonstrated high scoring accuracy from previous scoring projects at ETS and were invited to act as a scoring leader on a project. For the 2019 ELPAC field test administration, the scoring leader backread (read behind), guided, and retrained raters as needed. Scoring leaders monitored the small group of raters on a shift, usually up to 10 to 12 raters, to assist Scoring and Reporting Operations with scoring quality.

4.6.1 Training for Scoring Leaders

ETS assessment specialists previously conducted virtual training sessions for scoring leaders by means of conference calls using online conferencing tools. The purpose of the training was to discuss the duties of scoring leaders and to provide specific grade-level guidance on particular prompts. The training included guidance on using condition codes that are applied to nonscorable responses (such as blank [B]), communication with raters, how to monitor raters, and other information necessary for their role during scoring.

4.6.2 Training for Raters

Training for raters occurred within the ONE system. Raters were provided ONE system training documents as well as program-specific information to which they could refer at any time. Prior to scoring, raters were given a window of time to review all training materials in the system and practice scoring using the prescored training sets. After raters completed a training set, they were provided with annotations for each response as a rationale for the rating assigned.

The scoring training provided for each potential rater was designed using materials developed by ETS and followed the three-step progression noted in the following subsections.

4.6.2.1 Step One: Review the Scoring Guide and Benchmarks

Training for scoring began with an overview of the CDE-approved scoring guide, or rubric, and benchmarks. The raters accessed the scoring guide and benchmarks in ONE in the same manner that the resources would be accessed during operational scoring. The

benchmarks had annotations associated with them to call the raters' attention to specific content in the sample responses.

4.6.2.2 Step Two: Score Training Sets

After orientation to the scoring guide and the benchmark function, raters progressed through an online content training in the ONE system, in which they reviewed several sets of sample responses, assigned scores, and received feedback on their scores based on ratings for each response and applicable supporting annotation. Training sets, also called feedback sets, were samples of responses that provided the rater annotations after each sample was completed. The feedback sets for the 2019 ELPAC field test administration contained a mixed set of sample responses for each score point on the rubric as well as feedback in the form of annotations after a rater submitted a score.

4.6.2.3 Step Three: Set Calibration

Calibration is a system-supported control to ensure raters meet a specified standard of accuracy when scoring a series of prescored responses. Raters calibrated before they were allowed to score, meaning they scored a certain percentage of responses accurately from a set of responses called a calibration set. The passing percentage was determined by the program and number of responses in a set.

In general, calibration occurred whenever a rater began to score a particular task type for a particular grade span. Raters were allowed two chances to calibrate successfully. If raters met the standard on the first attempt, they proceeded directly to scoring responses. If raters were unsuccessful, they might have reviewed training sets and attempted to calibrate again with a new calibration set. If they were unsuccessful after both attempts, they were not allowed to score that task type.

Calibration can also be used as a means to control rater and group drift, which are changes in behavior that affect scoring accuracy between test administrations. Ongoing calibration can be used throughout a scoring season to check scoring accuracy on prescored sets of responses. In the case of the 2019 ELPAC field test, calibration occurred once every three days per task type scored per grade span.

4.7 Scoring Monitoring and Quality Management

Approximately 10 percent of responses were double-scored as a check for rater consistency. Raters were not aware when a second scoring was occurring and so did not have access to the first score.

In addition to the calibration function described previously, raters were monitored closely for the quality of their scoring throughout the scoring window. During a scoring shift, scoring leaders read behind raters at a rate of up to 10 percent of the responses scored by each individual rater to determine if raters were applying the scoring guide and benchmarks accurately and consistently. When necessary, the scoring leader redirected the rater by referencing the rubric, benchmarks, or both the rubric and benchmarks to explain why a response should have received a different score.

4.8 Rater Productivity and Reliability

The ONE system offers a comprehensive set of tools that the scoring leaders and scoring management staff used to monitor the progress and accuracy of individual raters and raters in aggregate. Reports produced to show rater productivity and performance presented how many responses a rater scored during a shift.

Chapter 5 Analysis Plans

This chapter presents the data analysis plans that were conducted using the computer-based Initial English Language Proficiency Assessments for California (ELPAC) field test data.

5.1 Data Collection Plan

One test form was administered in the fall 2019 computer-based Initial ELPAC field test phase. This was a preassembled 2019–2020 computer-based form aligned with the 2019–2020 adjusted Summative and Initial ELPAC blueprints. Returning English learner (EL) students and newcomer students, regardless of their EL status at the time of the field test administration, were eligible to take the Initial ELPAC. Analysis results from the field test administration were used to create the 2020–2021 computer-based operational forms.

5.1.1 Form Assignment

The Initial ELPAC field test comprised only one form for each grade level and grade span. Consequently, a form assignment process was not needed.

5.1.2 Challenges in Sample Recruitment

Local educational agencies (LEAs) were encouraged to enroll multiple schools to participate in the computer-based Initial ELPAC field test. Educational Testing Service (ETS) and the Sacramento County Office of Education (SCOE) identified LEAs that were eligible to participate based on their having 20 or more students in a grade level or grade span. Eligible LEAs were asked to provide SCOE with voluntary numbers of students by school and grade level or grade span. Some challenges for sample recruitment were as follows:

- The planned field test window was relatively brief, only three weeks from October 1 to October 25, 2019. The testing window was later extended through November 8, 2019, because of the impacts of fire emergencies that affected testing throughout the state.
- The field test window overlapped with the operational paper–pencil Initial ELPAC testing window.
- The training window started at the same time as the opening of the field test window. Schools participating in the field test needed to complete training and immediately start testing students within the testing window.

Table 5.1 shows the case counts originally targeted for the fall 2019 computer-based Initial and Summative ELPAC field tests.

Table 5.1 Target Case Counts for the Fall 2019 Computer-based ELPAC Field Tests

Grade	Initial Grade Level or Grade Span	Target N
TK	K	800
K	K	1,200
1	1	1,200
2	2	2,400
3	3–5	2,400
4	3–5	800
5	3–5	800
6	6–8	800
7	6–8	800
8	6–8	800
9	9–12	800
10	9–12	1,500
11	9–12	1,500
12	9–12	1,500
Total	N/A	17,300

Note: Transitional kindergarten = TK, kindergarten = K.

5.1.3 Form Assignment Principles

Unlike the computer-based Summative ELPAC field test, only one form of the Initial ELPAC field test was administered to each grade level and grade span. Consequently, form assignment principles were not needed.

To deal with challenges related to sample recruitment, the following decisions were made:

- Each school within an LEA was targeted for testing and could be assigned more than one grade level or grade span. For example, a school might have participated in the kindergarten test and the grade span three through five test.
- Because recruitment counts were lower than expected, participation targets were adjusted to reflect this reality for each grade or grade span.

5.1.4 Student Roster Selection

Student rosters were developed for the computer-based Initial ELPAC field test to provide structure for participating schools.

For each school, up to 50 percent more students per grade level or grade span were included in the roster than were pledged by individual schools at each specific grade level or grade span. This was done to help ensure sufficient N counts were maintained should some targeted students not be able to take the field test. Individual student records obtained from the California Longitudinal Pupil Achievement Data System (CALPADS) in August 2019 were used for roster selection. Previous Initial ELPAC performance was used to evaluate

whether the roster of students selected for participation was representative of EL students in the state.

The sample was stratified in terms of students' disability status (Individuals with Disabilities Education Act [IDEA] indicator = yes). This stratification was used to ensure that students with disabilities were included in the field test. However, students with three primary disability types—Intellectual Disability, Visual Impairment, and Deaf-Blindness—were excluded from the roster because the computer-based ELPAC field test did not include appropriate accommodations for these students. Previous ELPAC performance, as well as gender, home language, and other demographic information available in CALPADS, were used to evaluate whether the roster of students was representative of the state EL population.

At the request of the California Department of Education, a student's country of origin was considered as a proxy for technology exposure while developing the student roster. Students who indicated Afghanistan, El Salvador, Honduras, Guatemala, and Colombia as their country of origin were given priority to be included in the roster. It was anticipated that students from these five countries would have limited exposure to technology compared to students from the United States and other countries. Early in the recruiting process, the goal was to identify LEAs that might have large numbers of students from these five countries. Ultimately, to ensure sufficient volumes of student responses, SCOE encouraged all LEAs to participate in this study with as many students as possible. Demographic summaries, including whether students were from these five countries of interest, are provided in appendix 3.A. These demographic tables indicate the total number of students who took the test and the number of students included in analyses after the data cleaning rule for initial item analysis was applied.

5.2 Data Analysis Plan for the Initial ELPAC

5.2.1 Data Cleaning

Data was collected during the fall 2019 field test administration to support the preequating of Initial test forms. Data was screened to evaluate what constituted a valid case for analysis purposes. The rule applied to the data was this: Remove all test takers who did not obtain at least one item score in each of the four domains.

This rule was applied to both the classical test theory and differential item functioning (DIF) analyses. Table 3.A.1 through table 3.A.6 present the number of students participating in the fall 2019 Initial ELPAC field test and the number of students who were included in the analyses. Additional rules were applied to item response theory (IRT) analyses of the oral composite data. Those rules are described in subsection *5.2.5*.

Omitted or not-reached responses were handled in the same way in all statistical analyses (item analysis, DIF, IRT). In these analyses, omits, no responses, and multiple-grid responses from administered forms were treated as incorrect responses.

5.2.2 Classical Test Theory Analyses

Many of the statistics that are commonly used for evaluating assessments, such as *p*-values, point-biserial correlations, DIF classifications, and reliability coefficients, arise from classical test theory. These classical item analyses were conducted for each item across all domains. Detailed results of these item analyses are presented in appendix 6.A and are summarized in the tables in chapter 6.

5.2.2.1 Description of Classical Item Analysis Statistics

The classical item analyses include item difficulty indices (i.e., *p*-values) and item-total correlation indices (i.e., point-biserial correlations). Flagging rules associated with these statistics identify items that are not performing as expected. The omit rate for each item, the proportion of test takers choosing each distractor, the correlation of each distractor with the total score, and the distribution of students at each score point for the polytomous or constructed-response (CR) items are also included in the classical item analyses.

5.2.2.1.1 Item Difficulty

For dichotomous or multiple-choice (MC) items, item difficulty is indicated by the *p*-value, which is the proportion of students who answer an item correctly. The range of *p*-values is from 0.00 to 1.00. Items with higher *p*-values are easier items; those with lower *p*-values are more difficult items.

The formula for *p*-value for an MC item is

$$p - value_{MC} = \frac{\sum X_{ij}}{N_i} \tag{5.1}$$

Refer to the Alternative Text for Equation 5.1 for a description of this equation.

where,

X_{ij} is the score received for a given MC item i for student j, and

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

For CR items, 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. To facilitate interpretation, the AIS values for CR items are often expressed as the proportion of the maximum possible score, which is analogous to the *p*-values of dichotomous items.

For CR items, the *p-value* is defined as

$$p\text{–}value_{CR} = \frac{\sum X_{ij}}{N_i \times Max(X_i)} \tag{5.2}$$

Refer to the Alternative Text for Equation 5.2 for a description of this equation.

where,

X_{ij} is the score received for a given CR item i for student j,

$Max\ (X_i)$ is the maximum score for item i, and

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

Additional analyses for CR items include examination of score distribution. If no students achieved the highest possible score, the item may not be functioning as expected. The item may be confusing, not well-worded, unexpectedly difficult, or students may not have had an opportunity to learn the content. Items with a low percentage (e.g., less than 3%) of students who obtained any possible item score would be flagged for further review. Items with few students achieving a particular score may pose problems during the IRT calibrations. Consequently, these items need to be carefully reviewed and possibly excluded from item calibration analyses.

5.2.2.1.2 Item-Total Correlation

An important indicator of item discrimination is the point-biserial correlation (i.e., item-total correlation), defined as the correlation between student scores on an individual item and student "total" scores on the test (after excluding the scores from the item being analyzed).

To calculate point-biserial correlations by domain, the total scores are the domain scores, rather than the total test scores. The item-total correlation ranges from -1.0 (a perfect negative relationship) to 1.0 (a perfect positive relationship). A relatively high positive item-total correlation is desired, as it indicates that students with higher scores on the test tended to perform better on the item than students with lower test scores. A negative item-total correlation signifies a potential problem with the item, because it indicates that more students with low scores on the test are answering the item correctly than students with high scores on the test; this may indicate a scoring key issue.

To avoid artificially inflating the correlation coefficients, the contribution of the item being analyzed is removed from the calculation of the total score when calculating each of the point-biserial correlations. Thus, performance on each Listening item was correlated with the total Listening score minus the score on the item in question. Likewise, performance on each Reading item was correlated with the total Reading score minus the score on the item in question and so on for the Speaking and Writing items. Desired values for this correlation are positive and larger than 0.20.

5.2.2.2 Summary of Classical Item Analysis Flagging Criteria

Items were flagged based on the classical item statistics using the criteria described in table 5.2.

Table 5.2 Item Flagging Criteria Based on Classical Item Analyses

Flag Type	Criteria
A	Low *p*-values (less than .25)
D	MC items with proportionally higher ability students selecting a distractor over the key
H	High *p*-values (greater than .95)
O	High percent of omits (greater than 5%)
R	Low item-total correlation (less than .20)

5.2.2.3 Omit Rates

Data from tests that measure constructs other than language proficiency are typically analyzed to evaluate whether items have high omit rates. This sometimes indicates an issue with the presentation or wording of the item, which results in many students omitting that item. Relatively high omit rates for tests such as the Initial ELPAC may be expected; students with minimal familiarity with English are likely to omit a substantial number of items. Nevertheless, ELPAC items with omit rates of 5 percent or more were flagged for further investigation to ensure no issues were found with these items.

5.2.3 Differential Item Functioning (DIF) Analyses

DIF analyses for gender and ethnicity were performed for all items with the scored item files used for classical item analysis. If an item performs differentially across identifiable student groups—for example, by gender or ethnicity—when students are matched on ability, the item may be measuring something other than the intended construct (i.e., possible evidence of bias).

It is important, however, to recognize that item performance differences flagged for DIF might be related to actual differences in relevant knowledge or skills between student groups (i.e., impact) or statistical Type I error, which might falsely find DIF in an item. As a result, DIF analysis is used mainly as a statistical tool to identify *potential* item bias. Subsequent reviews by content experts and bias and sensitivity experts are required to determine the source and meaning of performance differences.

The Initial ELPAC DIF procedures used were the Mantel-Haenszel (MH) procedure (1959) for MC items and the standardized mean difference (SMD) procedure (Dorans, 1989) for CR items.

The Mantel-Haenszel differential item functioning (MH-DIF) statistic was calculated for MC items (Mantel & Haenszel, 1959; Holland & Thayer, 1985). For this procedure, the examinees were assigned to a focal group (female; non-Hispanic or non-Latino), which is typically of prime interest, and a reference group (male; Hispanic or Latino).

Each group is then further divided into *k* matched ability groups, often based on the total test raw score. For example, all examinees obtaining a raw score of 10 represent one matched ability group. Then for an item, *j*, the data from the *k*th level of reference and focal group members can be arranged as a 2 × 2 table, as shown in table 5.3.

Table 5.3 Mantel-Haenszel Data Structure

Group	Item *j* Correct	Item *j* Incorrect	Total
Reference Group	A_k	B_k	n_{Rk}
Focal Group	C_k	D_k	n_{Fk}
Total Group	R_k	W_k	n_{Tk}

The MH odds ratio estimate, α_{MH}, for item *j* compares the two groups in terms of their odds of answering the item correctly and is given as follows:

$$\alpha_{MH} = \frac{\sum_k \frac{A_k D_k}{N_{Tk}}}{\sum_k \frac{B_k C_k}{N_{Tk}}} \tag{5.3}$$

Refer to the Alternative Text for Equation 5.3 for a description of this equation.

The odds ratio estimate is rescaled to the ETS delta scale (Holland & Thayer, 1985) using the following transformation:

$$\Delta_{MH} = -2.35 \log_e(\alpha_{MH}) \tag{5.4}$$

Refer to the Alternative Text for Equation 5.4 for a description of this equation.

The index MH D-DIF, Δ_{MH} , is negative when the item is more difficult for members of the focal group than it is for comparable members of the reference group. DIF items will be flagged when MH D-DIF, Δ_{MH} , is significantly greater than 1.0 and has an absolute value of 1.5 or greater; all efforts will be made to exclude these items from use in future forms construction.

5.2.3.1 DIF Procedure for Polytomous Items

For CR items, the SMD was used (Dorans & Schmitt, 1991). For items with s score levels and k matched ability groups, the SMD is calculated using the following formula:

$$SMD = \sum_{k=1}^{K} \frac{n_{k,\,focal}}{n_{focal}} \left(\frac{\sum_{s=0}^{S} s \cdot n_{ks,\,focal}}{n_{k,\,focal}} - \frac{\sum_{s=0}^{S} s \cdot n_{ks,reference}}{n_{k,reference}} \right) \tag{5.5}$$

Refer to the Alternative Text for Equation 5.5 for a description of this equation.

Mantel and Haenszel's (1959) chi-squared statistic for polytomous items will also be calculated, as with the *p*-value associated with it, $p_{\chi^2_{MH}}$. CR item *j* will be flagged when the absolute value of $\frac{SMD_j}{sd_j}$ is greater than .25 and $p_{\chi^2_{MH}}$ is less than .05 (based on a rule described in Zwick, Thayer, & Mazzeo, 1997). sd_j is the standard deviation of the item score calculated for the combined focal or reference sample. All efforts will be made to exclude items flagged by this rule from use in future forms construction.

5.2.3.2 DIF Categories and Definitions

DIF category descriptions are the same for dichotomous and polytomous items, but the underlying calculations vary somewhat. Table 5.4 and table 5.5 provide the specific rules used to evaluate DIF for dichotomous and polytomous items.

Table 5.4 DIF Categories for MC Items

DIF Category	Definition
A (negligible)	• MH D-DIF is not significantly different from 0 at the 0.05 level (i.e., the *p*-value of MH_Chi_Sq > 0.05), or |MH DDIF ≤ 1.
B (slight to moderate)	• MH D-DIF is significantly different from 0 and |MH D-DIF| is greater than 1, and • Either MH D-DIF is not significantly different from 1 or |MH D-DIF| is greater than 1.50.
C (moderate to large)	• MH D-DIF is significantly different from 1 at the 0.05 level and is at least 1.50.

Table 5.5 DIF Categories for CR Items

DIF Category	Definition
A (negligible)	• Mantel chi-square *p*-value is ≥ 0.05; or • The absolute value of |SMD/SD| is ≤ 0.17.
B (slight to moderate)	• Mantel chi-square *p*-value is < 0.05; and • The absolute value of |SMD/SD| is greater than 0.17 and less than or equal to 0.25.
C (moderate to large)	• Mantel chi-square *p*-value is < 0.05; and • The absolute value of |SMD/SD| is > 0.25.

Note: Value for |SMD/SD| are rounded to two decimal places before being evaluated.

5.2.4 Response Time Analyses

ELPAC assessments are untimed, but test examiners need guidance on test duration they might anticipate as they schedule administrations.

5.2.4.1 Item Level Analyses

Timing information is collected by the delivery platform for each "page" or computer screen that is presented to test takers. Information about the time required to answer a single question is available for items that appear on a page alone. Time required to answer all questions on a page is available when multiple items appear on a page.

5.2.4.2 Total Test Analyses

Total test administration durations or response times were calculated by summing the page durations for all items in the Initial ELPAC field test. Summary information regarding total test response times is presented in subsection *6.1.6 Response Time Analyses*. Table 6.B.1 in appendix 6.B provides summary statistics of response times at the first, tenth, twenty-fifth, fiftieth, seventy-fifth, ninetieth, and ninety-ninth percentiles. Total test response times calculated for the fiftieth and ninetieth percentiles provide administrators with an indicator of how much time students require on average, as well as how much time might be needed for students who require more time.

5.2.5 Item Response Theory (IRT) Calibration

For dichotomous items, the one-parameter logistic (1PL) IRT model was used for the Initial ELPAC item calibration:

$$P_i(\theta_j) = \frac{\exp(D(\theta_j - b_i)}{1+\exp(D(\theta_j - b_i)} \quad (5.6)$$

Refer to the Alternative Text for Equation 5.6 for a description of this equation.

$P_i(\theta_j)$ is the probability of student with proficiency θ_j giving the correct answer on item i. This formula produces an item characteristic curve (ICC), which graphically describes how an item performs. The value, b_i, reflects an item's difficulty, with larger values indicating a more difficult item. D is a scaling constant that leads the ICC to reflect a normal distribution.

The partial credit model (PCM) (Masters, 1982) was used for polytomous items. The mathematical formula of the PCM, presented in the formulation used previously for generalized partial credit, is the following:

$$P_{ih}(\theta_j) = \begin{cases} \dfrac{\exp\left(\sum_{v=1}^{h} D(\theta_j - b_i + d_{iv})\right)}{1 + \sum_{c=1}^{n_i} \exp\left(\sum_{v=1}^{c} D(\theta_j - b_i + d_{iv})\right)}, & \text{if score } h = 1, 2, ..., n_i \\ \dfrac{1}{1 + \sum_{c=1}^{n_i} \exp\left(\sum_{v=1}^{c} D(\theta_j - b_i + d_{iv})\right)}, & \text{if score } h = 0 \end{cases} \quad (5.7)$$

Refer to the Alternative Text for Equation 5.7 for a description of this equation.

where,

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

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

b_i is the location parameter for item i;

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

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

5.2.6 Linking Procedures for the Initial ELPAC

As part of the Initial ELPAC transition from paper-based to computer-based assessments, it was of critical importance to accurately place the computer-based scores onto the paper-based scale. A common item equating design was used.

The sets of linking items were used to place the computer-based scores onto the paper–pencil test (PPT) scale using the Stocking-Lord equating method (Stocking & Lord, 1983). Each linking item had computer-based item parameter estimates from the IRT calibrations of the field test forms. They also had PPT item parameter estimates from the Initial ELPAC item bank. The software STUIRT (Kim & Kolen, 2004) was used to find the Stocking-Lord constants necessary to perform the required linear transformations.

In performing linking to place the fall 2019 field test items onto the operational scale, it became clear that the Initial ELPAC assessment scale transformations were unusual for the oral skills composite. The form of the transformations appeared to differ for items associated with the Listening and Speaking domains. In reviewing the distributions of raw scores for these domains, it was discovered that students who participated in past operational administrations of these tests earned zero or near-zero scores at much higher rates than students participating in the fall 2019 field test administration. These differences were too large to be ignored and an effort was made to screen the data to better align the data supporting the fall 2019 field test and the historical item calibrations. This was accomplished in two steps:

1. Remove all test takers who earned a raw score of zero on either the Listening or the Speaking domains
2. If needed, remove all test takers whose differences between Listening and Speaking raw scores were in the top or bottom 5 percent of the score difference distribution

Updated item calibrations and subsequent analyses were performed for the Listening and Speaking domain tests after step 1 was performed. Updated item calibrations and subsequent analyses were performed for the Listening and Speaking domains for grade two and grade spans three through five, six through eight, and nine through twelve after step 2 was performed. These updated calibrations were performed for data from both the fall 2019 field test and the 2017–2018 administration that defined the operational reporting scale for each test. Specifically, the analyses included the following:

1. Performed updated item calibrations for 2017–2018 operational test forms
2. Created updated raw-to-theta value tables for 2017–2018 test forms based on the updated item parameter estimates
3. Added the updated theta scores to the historical scoring tables, so that new theta values and historical scale scores were linked at each raw score point on the 2017–2018 test forms
4. Used linear interpolation to define theta values for each scale score point between the values that were present on the 2017–2018 forms
5. Performed updated item calibrations for the fall 2019 field test forms
6. Repeated the Stocking and Lord procedure to put the fall 2019 items onto the newly updated 2017–2018 theta scale
7. Created updated raw-to-theta value tables for the preequated 2020–2021 forms based on the updated item parameter estimates
8. Entered the theta values from step 7 into the table created in step 4 to obtain updated scale score values at each raw score point

The implementation and results of the psychometric analysis plans described in this chapter are provided in chapter 6.

References

Dorans, N. J. (1989). Two new approaches to assessing differential items functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, 3, 217–33.

Dorans, N. J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach. (ETS Research Report 91-47.) Princeton, NJ: Educational Testing Service. Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd edition) (pp. 105–46). New York: Macmillan.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (ETS RR-85-43). Princeton, NJ: ETS.

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.

Kim, S., & Kolen M. J. (2004). STUIRT: *A computer program for scale transformation under unidimensional item response theory models.* University of Iowa. Version 1.0.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 207–10.

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)*,* 321–44.

Accessibility Information

Alternative Text for Equation 5.1

The p-value for item i is equal to the sum of the ith item scores across all j students divided by the total number of students who were presented with item i.

Alternative Text for Equation 5.2

The p-value for item i is equal to the sum of the ith item scores across all j students divided by product of the total number of students who were presented with item i and the maximum score available for item i.

Alternative Text for Equation 5.3

Alpha sub MH is equal to a fraction where the numerator is the sum over all k of a fraction where the numerator is A sub k multiplied by D sub k and the denominator is n sub Tk. The denominator is equal to a fraction where the numerator is the sum over all k of a fraction where the numerator is B sub k times C sub k and the denominator is N sub Tk.

Alternative Text for Equation 5.4

Delta sub MH is equal to the product of negative two point three five and natural logarithm of alpha sub MH.

Alternative Text for Equation 5.5

SMD is equal to the summation over k from 1 to capital K of the product of two factors. The first factor is a fraction where the numerator is n sub k, focal. The denominator of the first factor is n sub focal. The second factor is the difference between two fractions. The numerator of the first fraction is the summation over all s from 0 to S of s minus n sub ks, focal. The denominator of the first fraction is the n sub k, focal. The numerator of the second fraction is the summation over all s from 0 to S of s minus n sub ks, reference. The denominator of the first fraction is the n sub k, reference.

Alternative Text for Equation 5.6

P sub j of theta sub j is equal to a ratio where the nominator is the natural exponent of a constant D times the difference between theta sub j and b sub j and the denominator is one plus the natural exponent of a constant D times the difference between theta sub j and b sub j.

Alternative Text for Equation 5.7

If score h equals 1, 2, up to n sub i, then P sub ih open parenthesis theta sub j closed parenthesis is equal to fraction where the numerator has the exponential of the summation of v from 1 to h of D times a sub i times open parenthesis theta sub j minus b sub i plus d sub iv closed parenthesis. The denominator is 1 plus the summation of c from 1 to n sub I of the exponential of sum of v from 1 to c of D times a sub i times open parenthesis theta sub j minus b sub i plus d sub iv closed parenthesis.

If score h equals 0, then P sub ih open parenthesis theta sub j closed parenthesis is equal to fraction where the numerator is 1. The denominator is 1 plus the summation of c from 1 to n sub I of the exponential of sum of v from 1 to c of D times a sub i times open parenthesis theta sub j minus b sub i plus d sub iv closed parenthesis.

Chapter 6 Analysis Results

6.1 Initial ELPAC Results

This chapter summarizes the item- and test-level results of the psychometric analyses for the fall 2019 computer-based Initial English Language Proficiency Assessments for California (ELPAC) field test. These analyses include classical item analyses, response time analyses, differential item functioning (DIF), item response theory (IRT), and linking analyses.

6.1.1 Overview

The descriptions of these analyses are provided in chapter 5. They include classical item analyses, response time analyses, DIF, IRT, and linking analyses. Most of the items included in the field test had item statistics within the ranges described in chapter 5. Items with classical statistics outside of the flagging criteria were identified and reviewed collectively by Educational Testing Service's (ETS') psychometric and content teams.

All tables of analytic results are presented in appendix 6. The sections in this chapter describe the field test data and results of each of the analyses.

6.1.2 Samples Used for the Analyses

In general, analyses included in the technical report are based on all valid students' scores in the field test samples. An exception occurred for the samples used for all reliability analyses (i.e., classification accuracy and consistency and coefficient alpha). Students included in these analyses were screened to ensure

- they attempted at least half of the items in each relevant domain for the corresponding composite and overall reliability calculations, and
- no student had a raw score of zero.

As shown in table 6.D.1 in appendix 6.D, across the grade levels and composites, there were at least 900 students taking the assessment who could be included in the IRT analyses. The one-parameter logistic (1PL) IRT model was used to calibrate the Initial ELPAC field test data; therefore, these sample sizes were sufficient.

6.1.3 Raw Score Distributions

For all ELPAC field tests, the total test raw score is defined as the total points obtained for all machine-scorable items and hand-scored, constructed-response (CR) items combined. Appendix 6.A contains the raw score frequency distributions and summary statistics tables by form and by grade level or grade span (table 6.A.1 through table 6.A.18).

The average of oral and written composite raw scores by grade levels or grade spans ranged from 45 percent to 65 percent of the maximum possible raw score (table 6.A.19 and table 6.A.20). The written composite scores tended to have lower means as a percent of the maximum possible raw scores, compared to oral composite scores. Similarly, the average of overall raw scores by grade levels or grade spans was about 48 percent to 64 percent of the maximum possible raw score (table 6.A.21).

6.1.4 Results of Classical Item Analyses

ETS psychometric and content assessment staff carefully reviewed each of the items flagged after the 2019 Initial ELPAC field test administration. These results were summarized and submitted to the California Department of Education (CDE) for approval and then were entered into the item bank and used by the content assessment team for future operational test assembly.

This subsection presents tables of the classical item analysis results for the 2019 test items. Table 6.A.22 in appendix 6.A presents the overall *p*-value and item-total correlation information by grade level or grade span. Across the grade levels, grade spans, and domains, there was a range of item difficulty and item-total correlations. The total test item difficulties ranged from 0.20 to 0.98 and the total test item-total correlations ranged from -0.08 to 0.93. Items with difficulty values less than 0.25 or greater than 0.95 and items with item-total correlations less than 0.20 (i.e., outside the preferred range of classical item statistics) were flagged for additional review.

Across all grade levels and grade spans, eight items were flagged for *p*-values less than 0.25 and two items were flagged for *p*-values greater than 0.95. Five items were flagged for item-total correlation less than 0.20 and two of those items were identified as having negative item-total correlations. All flagged items were reviewed by ETS content specialists to ensure there were no issues with these items or the corresponding answer keys. All flagged items, with the exception of the two items having negative item-total correlations, were determined to be appropriate for operational use. The two items with the negative item-total correlations—one item in Reading for grade span three through five and one item in Reading for grade span nine through twelve—were labeled as “Do Not Use” and will not be used in future Initial ELPAC operational forms.

Summary statistics of the item analyses for the Initial ELPAC for each task type, by grade level and grade span, are presented in table 6.A.23 through table 6.A.28 in appendix 6.A. The classical statistics for each dichotomous item are presented in table 6.A.29 through table 6.A.34. Results for polytomous items are shown in table 6.A.35 through table 6.A.40. The tables indicated most item statistics were within reasonable ranges.

6.1.5 Differential Item Functioning (DIF) Results

DIF analyses were conducted for student response data from the fall 2019 Initial ELPAC field test items with sufficient sample sizes. The sample size requirements for the DIF analyses were 100 in the smaller of either the focal group or the reference 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.

6.1.5.1 Classification

DIF analyses were conducted on each test for designated comparison groups if there were sufficient numbers of students in each group. Groups were defined based on the demographic variables of gender and race or ethnicity. These comparison groups are specified in table 6.1.

Table 6.1 Student Groups for DIF Comparison

DIF Type	Focal Group	Reference Group
Gender	Female	Male
Ethnicity	Non-Hispanic or non-Latino	Hispanic or Latino

6.1.5.2 Items Exhibiting Significant DIF

Based on the DIF statistics and significance tests, items were classified into three categories and assigned values of "A," "B," or "C." Category A items contained negligible DIF, category B items exhibited slight-to-moderate DIF, and category C items possessed moderate-to-large DIF. Items with a category of C were carefully reviewed by content bias and fairness panels to determine whether these items could be considered unfair to some student groups. Items evaluated as being free of bias were then entered into the item pool for the creation of future test forms. The classification included an indication of which group had higher performance: "-" indicated that the reference group had higher item performance, and "+" indicated that the focal groups' item performance was higher.

The results of the DIF analyses are provided in appendix 6.C. Three of the 481 items analyzed were flagged for potential DIF (i.e., category C DIF items) when comparing females to males (table 6.C.1 and table 6.C.3). Two of the items were flagged for the Listening domain in kindergarten; one item favored females and the other item favored males. The third item was flagged for the grade two Reading domain and favored females.

Nine items were flagged when comparing non-Hispanic, non-Latino students to Hispanic or Latino students (table 6.C.2 and table 6.C.4), where three of these nine items favored the Hispanic or Latino students. One item was flagged for the grade two Listening domain and favored the non-Hispanic or non-Latino students. The other eight items were flagged in the Reading domain: two items for kindergarten and six items for grade span nine through twelve. Both kindergarten Reading items favored the non-Hispanic or non-Latino students. For the grade span nine through twelve items, three of the items favored the non-Hispanic or non-Latino students while the other three items favored the Hispanic or Latino students.

All items flagged for significant DIF values were reviewed by ETS content specialists. No issues were identified that would be expected to introduce bias or impact fairness to students. Consequently, these items were deemed appropriate for future use on operational test forms.

6.1.6 Response Time Analyses

Item response time for the computer-based Initial ELPAC field test was collected and is summarized in table 6.B.1. The table shows descriptive statistics of response time by grade level or grade span and raw score interval based on quartiles. Also reported is the time taken by students in each of seven percentiles of response time (first, tenth, twenty-fifth, fiftieth, seventy-fifth, ninetieth, and ninety-ninth percentiles). The minimum testing time for the whole test was 4.5 minutes for low-performing students in kindergarten. The maximum testing time for the whole test was about 6.5 hours for grade spans three through five. The median testing time varied from about one-half hour to two hours. Students from lower grades tended to complete the tests in less time than those from higher grades. Note that the Initial ELPAC is an untimed test.

6.1.7 IRT Results for the Initial ELPAC

Two unidimensional IRT scales were developed for each grade level or grade span during the calibration stage:

1. The composite oral language scale was comprised of the Listening and Speaking assessments.

2. The composite written language scale was comprised of the Reading and Writing assessments.

The 1PL model was used to calibrate dichotomous items. The generalized partial credit model was used to calibrate polytomous items.

Table 6.D.1 in appendix 6.D shows the number of items and sample sizes for each IRT calibration. The sample sizes appeared to be adequate for fitting a 1PL model to the data. The minimum number of students was 686, for the oral composite of grade span nine through twelve. The written composite for grade span three through five had the maximum sample size of 1,632 students.

Summaries of the IRT *b*-value parameter estimates for the oral and written composites are shown in table 6.2 and table 6.3, respectively. The mean, standard deviation (SD), minimum, and maximum values are presented. The parameter estimates for each item, by grade level or grade span, are reported in table 6.D.2 through table 6.D.7 for the oral composite and in table 6.D.8 through table 6.D.13 for the written composite. This appendix also contains frequency distributions of IRT *b*-value parameter estimates for the oral and written composites in table 6.D.14 and table 6.D.15, respectively.

Across the grade levels and grade spans, items from the oral composite had mean difficulty estimates varying from -6.10 to 1.51. The range for written composite items was from -4.14 to 1.98. IRT *b*-parameter estimates in the range of -4.0 to +4.0 are typically viewed as psychometrically acceptable. As presented in table 6.D.14 and table 6.D.15, six items from the oral composite had *b*-values less than -4.0 and one item from the written composite had a *b*-value less than -4.0. Consequently, the majority of items were within the acceptable range for *b*-parameter estimates.

Most of the means of the estimates for each grade level and grade span were below zero, indicating that, overall, the assessments were relatively easy. Note that the Initial ELPAC assessments are not vertically scaled. Thus, it is not expected that the *b*-parameter estimates will increase across grade levels and grade spans.

Table 6.2 IRT *b*-values for Oral Language Composite by Grade Level or Grade Span

Grade Level or Grade Span	Domain	N Items	Mean	SD	Minimum	Maximum
K	Listening	20	-0.85	1.03	-3.08	1.21
K	Speaking	11	-0.62	0.82	-1.64	0.66
1	Listening	25	-1.09	0.83	-2.68	0.31
1	Speaking	11	-1.73	1.01	-3.41	-0.06
2	Listening	25	-1.85	1.03	-3.68	0.76
2	Speaking	14	-2.40	1.25	-4.37	-0.75
3–5	Listening	33	-0.82	1.05	-2.70	1.51
3–5	Speaking	15	-2.69	1.43	-6.10	-0.70
6–8	Listening	29	-1.12	0.98	-3.49	0.30
6–8	Speaking	14	-2.18	1.34	-4.93	-0.26
9–12	Listening	26	-1.02	0.73	-2.64	0.43
9–12	Speaking	14	-2.11	1.03	-4.09	-0.97

Table 6.3 IRT *b*-values for Written Language Composite by Grade Level or Grade Span

Grade Level or Grade Span	Domain	N Items	Mean	SD	Minimum	Maximum
K	Reading	20	-0.15	1.16	-2.36	1.98
K	Writing	12	1.40	0.50	0.33	1.97
1	Reading	29	-0.24	1.01	-2.07	1.44
1	Writing	13	-0.53	1.55	-3.38	1.15
2	Reading	31	-0.73	0.99	-4.14	0.92
2	Writing	10	-0.70	0.96	-2.64	0.32
3–5	Reading	37	0.39	0.74	-1.67	1.86
3–5	Writing	8	-0.13	0.37	-0.64	0.38
6–8	Reading	28	0.39	0.88	-2.49	1.41
6–8	Writing	9	-0.57	0.38	-0.99	0.07
9–12	Reading	37	-0.13	0.68	-2.60	0.79
9–12	Writing	9	-0.56	0.41	-1.08	0.24

6.1.7.1 Horizontal Linking to the Initial ELPAC Paper-based Scale

As described in the CDE report, *A Study of Mode Comparability for the Transition to Computer-based English Language Proficiency Assessments for California: Results from the Psychometric Analyses of Computer-based Assessment* (ETS, 2020a), alternative methods for linking were investigated before the decision was made to apply the common item linking design. This method was used to transform computer-based ELPAC scores to the paper–pencil (PPT) scale.

The results of this linking process for the Initial ELPAC are described in the in the CDE report, *A Study of Mode Comparability for the Transition to Computer-based Initial English Language Proficiency Assessments for California: Results from the Psychometric Analyses of Computer-based Assessment* (2020b).

To evaluate the quality of the linking or common items, plots were created to compare the computer-based ELPAC field test and PPT item parameter estimates for each common item across the grade levels and grade spans. Common items with extreme *b*-parameter estimates or large root mean square deviations between the new (computer-based) and reference (PPT) parameter estimates were removed from the linking item sets. Across the domains and grade levels and grade spans, 22 common items were removed. The oral composite for grade two had the largest number of linking items excluded, where 4 out of 16 items were removed.

Following exclusion of problematic linking items, the final linking item sets accounted for approximately 26 to 61 percent of the total items on the Initial ELPAC forms. These percentages were considered sufficient to support the linking analyses.

The final sets of linking items were used to place the computer-based scores onto the PPT scale using the Stocking-Lord equating method (Stocking & Lord, 1983). Each linking item had computer-based item parameter estimates from the IRT calibrations of the Initial ELPAC field test forms. They also had PPT item parameter estimates from the Initial ELPAC

item bank. The software STUIRT (Kim & Kolen, 2004) was used to find the Stocking-Lord constants necessary to perform the required linear transformations.

6.1.7.2 Characteristic Curves by Grade Levels or Grade Spans

Unlike the Summative ELPAC, the Initial ELPAC is not vertically scaled. Therefore, it is not appropriate to show the Initial ELPAC test characteristic curves (TCCs) across grade levels and grade spans within a single figure. In appendix 6.D, figure 6.D.1 through figure 6.D.6 present the TCCs for the oral composite for each grade level and grade span. Figure 6.D.7 through figure 6.D.12 provide the corresponding TCCs for the written composite. These curves look reasonable as they are positioned in the middle of the score scales.

6.2 Constructed-Response (CR) Item Analysis

6.2.1 Interrater Agreement

To monitor the consistency of human-scored ratings assigned to student responses, approximately 10 percent of the CR items received a second rating. The two sets of ratings were then used to compute statistics describing the consistency (i.e., reliability) of the ratings. This interrater consistency or reliability is described by the percentage of agreement between two raters.

6.2.1.1 Percentage Agreement

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

Table 6.4 shows, for all writing items, the average percent exact, adjacent, and discrepant score agreement for each grade level and grade span, by the number of maximum score points. With only a few exceptions, the percent exact across all grade levels and grade spans, given the maximum score points, met the qualification standard used to monitor ELPAC CR scoring (refer to table 7.2 of the *Summative English Language Proficiency Assessments for California Technical Report* [CDE, 2020]). When the standard was not met, ETS staff reviewed the prompt's training materials, made revisions when necessary, and retrained raters using the revised materials.

Table 6.4 Interrater Reliability

Grade Level or Grade Span	Number of Score Points	Average of Percent Exact	Average of Percent Adjacent	Average of Percent Discrepant
K	All Writing Items	97.79	2.14	0.07
K	1-pt score Items	98.46	1.54	0.00
K	2-pt score Items	97.13	2.73	0.15
1	All Writing Items	74.29	25.35	0.36
1	2-pt score Items	81.25	18.75	0.00
1	3-pt score Items	73.41	26.19	0.40

Table 6.4 *(continuation)*

Grade Level or Grade Span	Number of Score Points	Average of Percent Exact	Average of Percent Adjacent	Average of Percent Discrepant
2	All Writing Items	79.78	19.79	0.43
2	2-pt score Items	90.79	9.13	0.08
2	3-pt score Items	75.94	23.45	0.61
2	4-pt score Items	61.02	38.18	0.80
3–5	All Writing Items	71.01	28.57	0.41
3–5	2-pt score Items	73.91	25.90	0.19
3–5	3-pt score Items	73.76	25.89	0.34
3–5	4-pt score Items	62.35	36.69	0.96
6–8	All Writing Items	71.38	28.28	0.33
6–8	2-pt score Items	78.98	20.90	0.12
6–8	3-pt score Items	70.48	29.30	0.21
6–8	4-pt score Items	61.28	38.02	0.70
9–12	All Writing Items	68.76	30.83	0.42
9–12	2-pt score Items	75.03	24.81	0.15
9–12	3-pt score Items	66.73	32.92	0.34
9–12	4-pt score Items	60.81	38.35	0.84

6.3 Limitations and Caveats for Data Interpretation

As discussed in chapter 3 and section *6.1 Initial ELPAC Results*, the data collected and analyzed from the field test phase presented some limitations that should be taken into account when interpreting the results reported in this chapter.

It should be noted that, although the demographic information of student samples participating in the field test looked similar to the population taking the assessment in 2018–2019, the timing of the field test window differed compared to the typical Initial ELPAC testing window. The typical Initial ELPAC testing window is year-round, starting July 1 and ending June 30 of the following year. More than 80 percent of the students usually take the Initial ELPAC forms from July to September. Meanwhile, the fall 2019 field test window occurred from October 1 through November 8, 2019.

The difference in testing windows may have resulted in some lack of representativeness between the English proficiency of students who took the 2018–2019 Initial ELPAC forms and the Initial field test forms. Another potential limitation for interpretation of results was that no distinction was made between English learner (EL) students and newcomer students who recently entered the United States. EL students having even a few months of school education may be expected to perform better than newcomer students.

An additional factor was that the Initial ELPAC field test contained more items than a typical Initial ELPAC form. For most of the grade levels and grade spans, and for the Listening, Speaking, and Reading domains, the field test forms had more than twice the number of items specified in the Initial ELPAC test blueprints. The gap in the testing window and the difference in the test length between the field test and the Initial ELPAC operational forms should be taken into consideration when interpreting the field test results reported in this chapter.

References

California Department of Education. (2020). *Summative English Language Proficiency Assessments for California technical report 2018–2019 administration*. [Draft report]. Sacramento, CA: California Department of Education.

Educational Testing Service. (2020a). *A study of mode comparability for the transition to computer-based English Language Proficiency Assessments for California: Results from the psychometric analyses of computer-based assessment 2019–2020 administration*. [Draft report]. Sacramento, CA: California Department of Education.

Educational Testing Service. (2020b). *A study of mode comparability for the transition to computer-based Initial English Language Proficiency Assessments for California: Results from the psychometric analyses of computer-based assessment*. [Draft report]. Princeton, NJ: Educational Testing Service.

Kim, S., & Kolen M. J. (2004). STUIRT*: A computer program for scale transformation under unidimensional item response theory models*. University of Iowa. Version 1.0.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 207–10.

Chapter 7 Reliability and Validity

This chapter provides reliability and validity evidence to support the interpretation of English Language Proficiency Assessments for California (ELPAC) scores and results of the field test analyses.

7.1 Evidence Based on Test Content

Evidence based on test content refers to traditional forms of content validity evidence, such as the rating of test specifications and test items (Crocker et al., 1989; Sireci, 1998), as well as alignment methods for educational tests that evaluate the interactions between curriculum frameworks, testing, and instruction (Rothman et al., 2002; Bhola, Impara, & Buckendahl, 2003; Martone & Sireci, 2009).

Chapter 2 of this report describes the planning, design, and development considerations undertaken to facilitate the transition from paper-based to computer-based administration of the Initial ELPAC, while continuing to ensure the assessment remained fair, reliable, and valid for its intended purposes. The corresponding test blueprints were revised and later reviewed to identify where minor adjustments could be made to appropriately use computer-based delivery. Another consideration made was to increase the amount of information collected at the upper range of English language proficiency, again while continuing to ensure the assessment remained fair, reliable, and valid for its intended purposes.

As described in section *3.5 Demographic Summaries*, in anticipation of some students having very little, if any, access to computers, Educational Testing Service (ETS) and the California Department of Education (CDE) developed the Technology Readiness Checker for Students. This is an online resource designed to help educators determine a student's familiarity with navigating an online interface. The purpose of the tool is for educators to better understand what kind of supports a student may need to increase technology familiarity, to understand what kind of support the student may need during the assessment, or both. This type of resource helps to ensure that students are being evaluated on their English proficiency rather than their experience with technology.

7.2 Evidence Based on Internal Structure

Validity evidence based on *internal structure* refers to the statistical analysis of item and score subdomains to investigate the primary and secondary (if any) dimensions measured by an assessment. Procedures for gathering such evidence include correlational analyses.

Evidence collected from the fall 2019 field test data supported the oral and written composites that are currently used to report Initial ELPAC scores. Correlations were calculated using data from the fall 2019 computer-based field test to examine the relationship between the four content domains and the two composites of the assessment. Additionally, various types of reliability analyses were conducted. The purposes of these analyses were to obtain validity evidence to support the continuation of the reporting scales for the computer-based ELPAC and to support reliable and valid interpretation of Initial ELPAC test scores.

7.2.1 Correlations Between Domain and Composite Scores

Using student raw scores from Initial ELPAC forms, correlation coefficients between the four domain scores and two composite scores were calculated for the Initial ELPAC forms. Table 7.A.1 through table 7.A.6 in appendix 7.A present the correlation coefficients for each grade level and grade span. The results showed moderate-to-strong relationships between domains. Correlations ranged from 0.396 between the Speaking and Writing domains for kindergarten to 0.846 between the Speaking and Writing domains for grade span nine through twelve.

7.2.2 Reliability Estimates, Overall and by Student Groups

The results of the reliability analyses for the overall ELPAC scores for all students within each grade level are presented in the last column of table 7.B.1 in appendix 7.B. The overall results indicate that the reliability estimates for Initial ELPAC total test scores across grade levels were within acceptable ranges, from 0.87 to 0.94. Reliability estimates for 8 out of 13 grade levels were 0.90 or higher.

When the analysis was conducted by student groups within each grade level, as shown in table 7.B.2 through table 7.B.14, the lowest overall reliability estimate observed was 0.84 for economically disadvantaged students in grade six (table 7.B.8). The highest overall estimate was 0.93 for a number of student groups: Asian students in 6 of the 13 grade levels, males in grade eleven, and grade eleven students not receiving special education. Reliability estimates of domains and composites, as well as decision accuracy and consistency reliability estimates, are discussed in the next subsections.

7.2.3 Domain and Composite Reliability Estimates

The results of reliability analyses for the four domain scores and two composite scores are presented in table 7.B.1. The reliability estimates for each domain of the test were somewhat low to high, ranging from 0.47 for grade three Reading to 0.89 for grade ten Speaking.

Speaking and Writing domains had higher reliability estimates than the Listening and Reading domains. This finding is consistent with reliability results from the 2018–2019 paper–pencil operational administration. For the oral and written composite scores, the reliability estimates were moderate to high, ranging from 0.67 to 0.92 across grade levels.

7.2.4 Decision Classification Analyses

While the reliabilities of performance-level classifications, which are criterion referenced, are related to the reliabilities of the test scores on which they are based, they are not exactly the same. Glaser (1963) was among the first to draw attention to this distinction, and Feldt and Brennan (1989) extensively reviewed the topic. While test reliability evaluates the consistency of test scores, decision classification reliability evaluates the consistency of classification.

Consistency in classification represents how well two versions of an assessment with equal difficulty agree in the classification of students (Livingston & Lewis, 1995). This is estimated by using actual response data and total test reliability from an administered form of the assessment from which two parallel versions of the assessment are statistically modeled, and classifications are compared. Decision consistency, then, is the extent to which the test classification of examinees into mastery levels agrees with classifications based on a hypothetical parallel test. The examinees' scores on the second form are statistically modeled.

Note that the values of all indices depend on several factors, such as the reliability of the actual test form, distribution of scores, number of threshold scores, and location of each threshold score. The probability of a correct classification is the probability that the classification the examinee received is consistent with the classification that the examinee would have received on a parallel form. This is akin to the exact agreement rate in interrater reliability. The expectation is that this probability would be high.

Decision accuracy is the extent to which the test's classification of examinees into levels agrees with the examinees' true classification. The examinees' true scores—and, therefore, true classification—are not known, but can be modeled. Consistency and accuracy are important to consider together. The probability of accuracy represents the agreement between the observed classification based on the actual test form and true classification, given the modeled form. These methods were applied to the Initial ELPAC fall 2019 field test data.

Commonly used indices for decision consistency and accuracy include (a) decision consistency and accuracy at each threshold score, (b) overall decision consistency and accuracy across all threshold scores, and (c) coefficient kappa.

Cohen's kappa (Fleiss & Cohen, 1973) represents the agreement of the classifications between two parallel versions of the same test, taking into account the probability of a correct classification by chance. It measures how the test contributes to the classification of examinees over and above chance classifications. In general, the value of kappa is lower than the value of the probability of correct classification because the probability of a correct classification by chance is larger than zero.

The methodology used for estimating the reliability of classification decisions is described in Livingston and Lewis (1995). These calculations are implemented using the ETS-proprietary computer program RELCLASS-COMP (Version 4.14).

7.2.4.1 Reliability of Classification Accuracy and Consistency

The results of decision accuracy and consistency at each threshold proficiency level for each language composite, as well as for overall scores, are presented in table 7.B.15 through table 7.B.20 in appendix 7.B for all grade levels and grade spans. Tables 7.B.15 through table 7.B.17 provide classification accuracy, while table 7.B.18 through table 7.B.20 show classification consistency.

At each threshold, the classification at adjacent performance levels appeared to be acceptably reliable and consistent. Classification accuracy ranged from 0.82 to 0.96, while classification consistency ranged from 0.75 to 0.95, with most values at or above 0.85. These values are similar to the classification accuracy and consistency estimates reported in the *2018–2019 Initial ELPAC Technical Report* (CDE, 2020).

Table 7.B.21 presents the comprehensive classification accuracy and consistency results, for both the composite and overall scores. For both classification accuracy and consistency, the grade span six through eight written composite had the lowest accuracy and consistency classification reliabilities, 0.76 and 0.67, respectively. The written composite for kindergarten had the highest classification reliabilities with 0.88 for accuracy and 0.84 for consistency. The overall accuracy reliability estimates ranged from 0.82 to 0.85, while overall consistency estimates ranged from 0.74 to 0.79.

7.3 Evidence Based on Consequences of Testing

Evidence based on *consequences of testing* refers to the evaluation of the intended and unintended consequences associated with a testing program. Examples of evidence based on testing consequences include investigations of adverse impact, evaluation of the effects of testing on instruction, and evaluation of the effects of testing on issues such as high school dropout rates. With respect to educational tests, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014) stress the importance of evaluating test consequences:

> When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described by those who mandate the tests. It is also the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences as feasible. Consequences resulting from the use of the test, both intended and unintended, should also be examined by the test developer and/or user. (AERA et al., 2014, p. 195)

Investigations of testing consequences relevant to the Initial ELPAC may include correction of classification from English learner to initial fluent English proficient or vice versa. Results from the Initial ELPAC may be used for instructional planning.

Unintended consequences, such as changes in instruction, diminished morale among teachers and students, increased pressure on students leading to increased dropout rates, or the pursuit of college majors and careers that are less challenging can be evaluated. These sorts of investigations require information beyond what is currently available to the Initial ELPAC program.

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.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*, 21–29.

California Department of Education. (2020). *Initial English Language Proficiency Assessments for California technical report 2018–2019 administration*. [Unpublished report]. Sacramento, CA: California Department of Education.

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*, 179–94.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd edition) (pp. 105–46). New York: Macmillan.

Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613–19.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist,* 18. 519–32

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement, 32*, 179–97.

Martone, A., & Sireci, S. G. (2009). *Evaluating alignment between curriculum, assessments, and instruction. Review of Educational Research, 4,* 1332–61.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing.* [Technical Report 566]. Washington, DC: Center for the Study of Evaluation.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5,* 299–321.

Chapter 8 Quality Control

The California Department of Education (CDE) and Educational Testing Service (ETS) implemented rigorous quality-control procedures throughout the item development, test development, administration, scoring, analyses, and reporting processes for the computer-based Initial English Language Proficiency Assessments for California (ELPAC) field test. As part of this effort, ETS staff worked with the ETS Office of Professional Standards Compliance, which publishes and maintains the *ETS Standards for Quality and Fairness* (ETS, 2014). These *Standards* support the goals of delivering technically sound, fair, and useful products and services; and assisting the public and auditors in evaluating those products and services. Quality-control procedures are outlined in this chapter.

8.1 Quality Control of Item Development

The pool of over 2,200 paper–pencil items underwent rigorous item development processes. The items were created according to the *Specifications for Conversion of ELPAC Task Types for Computer-Based Assessment* (CDE, 2019) and entered in appropriate layouts within the ETS Item Banking Information System (IBIS). Assessment specialists who were familiar with the layout of the computer-based items reviewed each item to ensure that the text, audio, and graphics all functioned correctly in the IBIS item previewer. The items were then provided to the CDE for review within IBIS. CDE staff provided ETS with comments regarding any necessary revisions. The items were revised and CDE staff ensured that any revisions were implemented accurately before the CDE approved the items for use.

After the CDE approved the items, ETS assessment specialists performed a final review of the items in IBIS, called final content review. During this review, an assessment specialist who was familiar with the Initial ELPAC task types performed an independent review of each item to ensure that the item content, metadata, graphics, and audio files were all accurate. The assessment specialist also reviewed comments that were made during previous reviews to ensure that they were implemented. Items were reviewed and approved at final content review before they were exported to the test delivery system vendor.

Once the items were with the test delivery system vendor, item-level quality checks were performed. Items were reviewed within the test delivery system vendor's item banking system to ensure that all item content and graphics were accurately displayed and audio files played correctly. ETS assessment specialists performed a side-by-side check of each item in IBIS next to each item in the test delivery system vendor's item bank to ensure that items contained accurate content and functioned correctly. Any issues were resolved prior to quality-control checks of the test forms in the test delivery system.

8.2 Quality Control of Test Form Development

ETS conducted multiple levels of quality-assurance checks on each constructed field test form to ensure it met the form-building specifications. Both ETS Assessment & Learning Technology Development and Psychometric Analyses & Research (PAR) staff reviewed and confirmed the accuracy of forms before the test forms were put into production for administration in the field test. Detailed information related to test assembly can be found in section *2.9 Test Assembly*.

In particular, the assembly of all test forms went through a certification process that involved various checks, including verifying that

- all keys were correct,
- answers were scored correctly in the item bank and incorrect answers were scored as incorrect,
- all items aligned with a standard,
- all content in the item was correct,
- distractors were plausible,
- multiple-choice (MC) item options were parallel in structure,
- language was grade-level appropriate,
- no more than three MC items in a row had the same key,
- all art was correct,
- there were no errors in spelling or grammar, and
- items adhered to the approved style guide.

Reviews were also conducted for functionality and sequencing of items in the test delivery system during the user acceptance testing (UAT) process. Three sets of UAT were performed: the first was performed by the test delivery system vendor, the second was performed by ETS, and the third was performed by the CDE. CDE staff made a final quality check to ensure that all issues that were identified during UAT were resolved prior to the release of the field test forms.

8.3 Quality Control of Test Administration

During the computer-based ELPAC field test administration, every person who either worked with the assessments, communicated test results, or received testing information was responsible for maintaining the security and confidentiality of the tests, including CDE staff, ETS staff, ETS subcontractors, local educational agency ELPAC coordinators, site ELPAC coordinators, ELPAC test examiners, and teachers.

ETS' Code of Ethics requires that all test information, including tangible materials (e.g., test items and test books), confidential files (e.g., those containing personally identifiable student information), and processes related to test administration (e.g., the packing and delivery of test materials) are kept secure. For the fall 2019 computer-based ELPAC field test, ETS had systems in place that maintained tight security for test items, test books, and test results, as well as for student data. Refer to chapter 5 of the *2018–2019 Initial ELPAC Technical Report* for the processes used to maintain security and confidentiality of test items and results (CDE, 2020).

To ensure security for all the tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI). As described in subsection *3.2.1 Educational Testing Service's Office of Testing Integrity (OTI)*, the mission of the OTI is to oversee quality assurance of all ETS testing programs and to safeguard the various processes throughout the test development and administration cycles.

8.4 Quality Control of Scoring

8.4.1 Human Scoring

8.4.1.1 Quality Control in the Scoring Process

In general, the ELPAC scoring design is based on a team of 10 to 12 raters scoring one item type at a time and one item at a time under the supervision of a scoring leader. Scoring leaders were supervised by group scoring leaders. Each group scoring leader was responsible for multiple teams in a grade level or grade span.

Each rater calibrated for an item type prior to scoring any response by passing the corresponding calibration test. The team scored multiple items of a similar type per shift. Once all responses of the same type were scored, each rater calibrated for a new item type. Each rater worked independently on the rater's own device to read each student response and entered a score for each response.

8.4.1.2 Quality Control Related to Raters

ETS developed a variety of procedures to control the quality of ratings and monitor the consistency of scores provided by raters. These procedures specified rater qualifications and procedures for rater certification and rater calibration. Raters were required to demonstrate their accuracy by passing a certification test before ETS assigned them to score a specific assessment and by passing a shorter, more focused calibration test before scoring for a specific grade and item type. The calibration results were valid for three days or until a rater switched to a different grade or item type. Rater certification and calibration are key components in maintaining quality and consistency.

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

8.4.1.3 Rater Qualification

Raters met the following requirements prior to being hired:

- A bachelor's degree was required.
- Teachers currently teaching English were preferred.
- Scoring experience was preferred.
- Graduate students and substitute teachers were encouraged to apply.
- Retired California educators with a California teaching credential who were not current classroom teachers were eligible; these educators must live in California.
- Candidates completed rater training and achieved qualifications through the certification process.

Table 8.1 provides a summary of the human scorers who participated in the computer-based Initial ELPAC field test.

Table 8.1 Summary of Characteristics of ETS Human Raters Scoring ELPAC Assessments

Characteristic	N	%
Experience teaching in a kindergarten through grade twelve (K–12) school	341	25
Currently works in a K–12 school in California	98	7
Others—Not meeting any of the previous criteria	917	68
Total raters scoring in 2019–2020	**1,356**	**100**

California educators should have met the following qualifications:

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

All scoring leaders and raters were required to qualify before scoring and were informed of what they were expected to achieve to qualify (refer to *4.1.5 Rater and Scoring Leader Training* for a more complete description of this training).

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

Upon completion of the required training sets, raters moved on to a consequential calibration set that determined rater eligibility for operational scoring of a particular item type. Calibration (qualification) sets had minimum agreement levels that were enforced, and nonadjacent scores were not allowed.

Responses in calibration and qualification sets had been scored previously by scoring experts, who came to a consensus on the score for each response. The standards for a rater to achieve qualification for scoring, provided in table 8.2, were applied in terms of the percent of exact agreement with consensus scores. The standards applied differ by the score point range.

Table 8.2 Rater Qualification Standards for Agreement with Consensus Scores

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

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

8.4.1.3.1 Monitoring Raters

ETS staff created performance scoring reports so that scoring leaders could monitor the daily human-scoring process and plan any retraining activities, if needed.

For monitoring rater accuracy, scoring leaders scored a subset of responses already scored by each individual rater to determine if raters were applying the scoring guide and benchmarks accurately and consistently. Scoring leaders did this at a rate of approximately 10 percent, and targeted raters who exhibited weaker scoring performance. Scoring leaders discussed score discrepancies on these responses using the rubric, benchmarks, or both the rubric and benchmarks. This process is referred to as back-reading.

For monitoring interrater reliability, 10 percent of the student responses that had already been scored by the raters were randomly selected for a second scoring and assigned to raters by the scoring system.

The second rater was unaware of the first rater's score. The evaluation of the response from the second rater was compared to that of the first rater. Scoring leaders and chief scoring leaders provided second reads during their shifts for additional quality review.

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

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

8.4.2 Interrater Reliability Results

At least 10 percent of the test responses to constructed-response (CR) Writing items were scored independently by a second reader. The statistics for interrater reliability for all items at all grades are presented in table 6.4. These statistics include the percentage of exact agreement and adjacent agreement between the two raters.

ETS used the following criteria to monitor the consistency or reliability of scores assigned to CR Writing items that were scored by a second reader. This information served to provide additional rater training if needed. Polytomous items were flagged if any of the following conditions occurred:

The sum of exact agreement and adjacent agreement < 0.80
Exact agreement < 0.60

Dichotomous items were flagged if the following condition occurred:

Exact agreement < 0.80

Table 8.3 shows the number of items flagged by content area, grade level or grade span, and scoring method. These are the items flagged from the interrater reliability results using the criteria described in the previous paragraph; items flagged by item analyses or differential item functioning analyses are not included in this table. Out of 39 Writing items, only two polytomous items were flagged across all grade levels and grade spans. No dichotomous items were flagged.

Table 8.3 Number of CR Items Flagged, by Grade Level or Grade Span, in the Fall 2019 Computer-based Initial ELPAC Field Test

Scoring Method	Content Area	Grade Level or Grade Span	Flagged Polytomous Items	Flagged Dichotomous Items	Total Flagged Items	Total Number of Scored Items	Percentage Flagged
Human to Human	Writing	K	0	0	0	4	0.0%
Human to Human	Writing	1	0	N/A	0	5	0.0%
Human to Human	Writing	2	0	N/A	0	4	0.0%
Human to Human	Writing	3–5	0	N/A	0	8	0.0%
Human to Human	Writing	6–8	1	N/A	1	9	11.1%
Human to Human	Writing	9–12	1	N/A	1	9	11.1%

8.5 Quality Control of Psychometric Processes

8.5.1 Development of Scoring Specifications

A number of measures were taken to ascertain that the scoring keys were applied to the student responses as intended and that student scores were computed accurately. ETS built and reviewed the scoring system models based on scoring specifications developed by ETS and approved by the CDE. Machine-scored item responses and demographic information were collected by ETS from the Answer Books. Human-scored item responses were sent electronically to the ETS Online Network for Evaluation system for scoring by trained, qualified raters. Record counts were verified against the counts obtained during security check-in from the document processing staff to ensure all students were accounted for in the file.

Once the record counts were reviewed, the machine-scored item responses were scored using the appropriate answer key. In addition, the student's original response string was stored for data verification and auditing purposes.

The scoring specifications contained detailed scoring procedures, along with the procedures for determining whether a student attempted a test and whether that student response data should be included in the statistical analyses and calculations for computing summary data. Standard quality inspections were performed on all data files, including the evaluation of each student data record for correctness and completeness. Student results were kept confidential and secure at all times.

8.5.2 Development of Scoring Procedures

The ETS Enterprise Score Key Management (eSKM) scoring system uses scoring procedures specified by psychometricians and provides scoring services. The eSKM system produces the official student scores of record. Following scoring, a series of quality-control checks were carried out by ETS psychometricians to ensure the accuracy of each score.

8.5.2.1 Enterprise Score Key Management System (eSKM) Processing

ETS developed two independent and parallel scoring structures to produce students' scores: the eSKM scoring system, which collected, scored, and delivered individual students' scores to the ETS reporting system; and the parallel scoring system developed by ETS Technology and Information Processing Services (TIPS), which scored individual students' responses. The two scoring systems independently applied the same scoring algorithms and specifications.

ETS psychometricians verified the eSKM scoring by comparing all individual student scores from TIPS and resolving any discrepancies. This parallel processing is an internal quality-control step and is in place to verify the accuracy of scoring. Students' scores were reported only when the two parallel systems produced identical results.

If scores did not match, the mismatch was investigated by ETS' PAR and eSKM teams and resolved. The mismatch could be a result of a CDE decision not to score an item because a problem was identified with the item or rubric. In these cases, ETS applied a problem item notification status to the item so that it would not be scored in the eSKM system. This parallel system of monitoring student scores in real time was designed to continually detect mismatches and track remediation.

Finally, data extracts were sent to ETS' Data Quality Services for data validation. Following validation, the student response statistical extracts were made available to the psychometricians for analyses. These processes were followed to help ensure the quality and accuracy of scoring and to support the transfer of scores into the database of the student records scoring system before data was used for analyses.

References

California Department of Education. (2019). *Specifications for conversion of ELPAC task types for computer-based delivery.* [Unpublished report]. Sacramento, CA: California Department of Education.

California Department of Education. (2020). *Initial English Language Proficiency Assessments for California technical report 2018–2019 administration*. [Unpublished report]. Sacramento, CA: California Department of Education.

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

Chapter 9 Post-test Survey

This chapter describes the development and administration of the post-test survey sent to local educational agency (LEA) English Language Proficiency Assessments for California (ELPAC) coordinators, site ELPAC coordinators, and ELPAC test examiners; and the results of analyses of their responses. The post-test survey applied to both the Summative and Initial ELPAC field test administrations. This section contains the same information included in the *Summative ELPAC Field Test Technical Report* (CDE, 2020).

9.1 Overview

During the fall 2019 computer-based ELPAC field test, Educational Testing Service (ETS) administered a post-test survey to LEAs. The purpose of the survey was to gather information on the clarity of the *Directions for Administration (DFAs)*, knowledge and use of training tests and the Technology Readiness Checker for Students (TRCS), student interaction with the online test delivery system, knowledge and use of accessibility resources, and overall administration experience.

9.2 Test Examiner Survey

The responses to the test examiner survey provided additional insight into the student test-taking experience and administration of the computer-based ELPAC field test. The feedback from the survey will help in the development and administration of the ELPAC operational tests. The test examiners completed their survey via SurveyGizmo, an online survey software tool.

The survey questions used during the administration and the results are included in appendix 9.

9.2.1 Survey Design and Questionnaire Development

The post-test survey was developed by program management staff at ETS in consultation with the California Department of Education (CDE). The CDE provided guidance in terms of the length of the survey and the number and focus of the questions.

The goal of the survey was to gain insights from the field for potential future improvement of the computer-based test administration and assessment processes overall. This survey was hosted on SurveyGizmo.com, a website with survey-creation and hosting services.

9.2.2 Survey Administration

LEAs were invited via email to participate in the post-test survey during the fourth week of testing. A link to the survey on the SurveyGizmo website was included in the communication. Feedback was collected from 675 educators who participated in the field test. The breakdown of respondents who participated in the survey by role was 179 LEA ELPAC coordinators, 281 site ELPAC coordinators, and 430 ELPAC test examiners. The total number of respondent roles exceeded 675 because respondents were able to select all applicable roles.

9.2.3 Summary of Test Examiner Survey Results

Previous ELPAC surveys had focused on paper–pencil test administration and been combined with California Assessment of Student Performance and Progress (CAASPP) questions as part of the *CAASPP and ELPAC Post-Test Survey*. This post-test survey, instead, focused on the computer-based administration.

Overall, educators indicated the training and resource materials that were provided for the field test were adequate. ELPAC educators also felt that the information and directions provided in the *Field Test Administration Manual* and *DFAs* were clear, although there were multiple comments about the directions being too wordy or too long. There were some notable challenges for educators, especially with the new process for administering the Speaking domain, but educators noted that sufficient opportunity to practice would help with preparing both test examiners and students for the first computer-based Initial ELPAC operational administration in August 2020.

Survey respondents reported experiencing adequate training for the ELPAC field test administration. However, some educators shared that it would have been better if there had been more time to digest training information before the beginning of the field test administration window.

Prior to the start of the computer-based ELPAC field test, the training tests and TRCS were made available as new resources. The feedback from educators on the usability of the training tests was that they were helpful or very helpful. The majority of educators did not use the TRCS with students prior to the field test.

Respondents provided feedback on their experience with the systems regarding the administration, how students interacted with the test delivery system, and the quality of the audio that was being played through the test delivery system. Over half of the educators participating in the survey reported never having issues with logging on to the Test Administrator Interface or with the student logging on to the test delivery system.

Eight-four percent of survey participants responded that the test directions were clear or very clear, whereas 16 percent of respondents indicated the test directions were somewhat clear or not clear. From the latter group of respondents, ETS was able to collect additional information that offered the opinions that directions were too wordy and lengthy, directions did not make it clear for students to know when to select the [**Next**] button, and some vocabulary in the directions were high-level words for some grades.

When asked about students in grades four through twelve navigating the platform independently, the majority of the responses indicated students independently navigated the system in the Listening, Reading, and Speaking domains. About 41 percent of the respondents indicated students experienced no difficulty with typing responses in the Writing domain. The grade levels that did have trouble typing their responses were grades four and five.

Educators were asked to provide feedback on the audio quality of the recorded files for the Listening, Speaking, and Writing domains. Over 71 percent of respondents indicated having no issues with the audio files for these domains.

The field test featured embedded universal tools and embedded designated supports, which are new for the ELPAC. The majority of the respondents did not help the students access the universal tools, whether during one-on-one or group administration. Additionally, 49 percent of educators were not familiar with Matrix Four or that enhanced accessibility resources were allowed and available for the computer-based ELPAC. Remedial action is recommended in *Chapter 10 Continuous Improvement*.

The CDE and ETS will continue their outreach efforts to LEAs to provide test administration support for ELPAC administrations. ETS also will use focus groups, surveys, and evaluations to continually identify areas for improvement for the overall ELPAC-related processes, systems, and resources.

A summary of the survey results is included in the *2019–20 ELPAC Post-Field Test Administration Survey and Focus Group Report* (CDE, 2019).

References

California Department of Education. (2019). *2019–20 English Language Proficiency Assessments for California (ELPAC) post-field test administration survey and focus group report.* [Unpublished report]. Sacramento, CA: California Department of Education.

Educational Testing Service. (2020). *Computer-based Summative English Language Proficiency Assessments for California (ELPAC) fall 2019 field test technical report.* [Unpublished report]. Sacramento, CA: California Department of Education.

Chapter 10 Continuous Improvement

The field test administration of the computer-based English Language Proficiency Assessments for California (ELPAC) took place in fall 2019. Since its inception, continuous efforts have been made to improve the computer-based ELPAC. This chapter presents the procedures used to gather information to improve the computer-based ELPAC as well as strategies to implement possible improvements.

10.1 Item and Test Development

As part of the transition from the paper–pencil tests (PPTs) to the computer-based ELPAC, Educational Testing Service (ETS), in collaboration with the California Department of Education (CDE) and the Sacramento County Office of Education (SCOE), conducted a small-scale usability pilot study. Cognitive laboratory methodology was used to investigate the ELPAC task types in an online environment.

The study was conducted in the early stage of development of the computer-based ELPAC prior to the large-scale transition of PPT items to a computer-based format. Detailed results and proposed action items for each recommendation were provided in the *ELPAC Usability Pilot: A Final Report* (CDE, 2019a). In addition, an addendum was created to describe how the recommendations from the final report were implemented in preparation for the computer-based ELPAC field test.

The following list describes the nine recommendations and the actions that were taken to implement the usability pilot recommendations:

1. **Improve Test Familiarity Materials—**Improve test familiarity materials (tutorials, training tests, practice tests) to ensure students are prepared to take, and test examiners are prepared to administer, the computer-based ELPAC:
    - Training tests and tutorials were released in September 2019, before the October 2019 field test administration.
    - The Technology Readiness Checker for Students (TRCS) was created for students to engage in common activities using a technological platform. Guidelines also were created to provide teachers and test examiners with suggestions for additional resources that a student might need based on the results of the TRCS report.
    - Resources such as a technical specifications manual and test administration manual were released ahead of the field test.
    - Translated test directions were provided in the 18 most popular languages spoken in California as an available support to orient students to each domain.
    - The new Speaking *Directions for Administration (DFAs)* included student and test examiner practice questions as part of the voice-capture check in the test delivery system. There were also instructions related to voice capture.
    - Local educational agency (LEA) trainers and test examiners—who attended the Administration and Scoring Training (AST) for the field test and Initial ELPAC administrations—were instructed to bring a mobile device to the training so they could practice test administration using the training tests.

- Use of the test delivery platform was incorporated into educator training during the in-person AST.
- Administration videos were shown during the AST. The videos were made available for LEAs to use in their local training. The videos showed the administration and scoring of the Speaking domain, including the Data Entry Interface (DEI), one-on-one kindergarten through grade two administration, and group administration for grades three through twelve.
- LEA trainers and test examiners who attended the AST received printed materials and videos that communicated the changes and new features of the computer-based ELPAC.
- Communications around preparing technology for the computer-based ELPAC, new embedded accessibility resources, and use of the TRCS were developed and disseminated based on the timing of specific releases.
- Full-length practice tests were released in November 2019 before the February 1, 2020, opening of the Summative ELPAC operational administration window.

2. **Create Educator Resource Materials—**Create resource materials for educators and test examiners to help determine if students are ready to take the computer-based ELPAC:
   - An online resource, the TRCS, was created to help educators determine a student's familiarity with using a technological platform.
3. **Allow Single-Listen for Listening Stimuli—**Allow students to listen only once to audio stimuli on the Listening test:
   - The Listening settings were updated to limit the playback of the Listening stimuli to one time. Students with a designated support for audio replay for Listening could replay a stimuli multiple times during the practice test and all operational assessments.
4. **Deliver Recorded Audio Files for the Listening Test Through the Testing Interface—**Maintain recorded audio files for Listening stimuli on the kindergarten and grade one Listening tests, like the grades two through eight Listening tests:
   - The training tests, the practice tests, and all operational tests included audio files for kindergarten and grade one students.
   - The audio files for kindergarten and grade one students were updated to direct the student to point to the answer when the options are pictures. For text options, students were directed to say their answer.
5. **Increase Accessibility Resource Familiarity—**Increase opportunity for familiarity and practice of accessibility resources for both test examiners and students:
   - Two products with accessibility resources were released. Training tests and tutorials were released in September 2019, before the October 2019 field test. Practice tests were released in November 2019 before the February 1, 2020, start of the Summative ELPAC operational administration window.

- Listening, Reading, and Writing *DFAs* contained language in the "Before Testing" and "During Testing" portions of the front matter that addressed these subjects as appropriate for each grade. Examples of bullets from the front matter included the following:
  - If desired, set up any additional resources (e.g., large mouse cursor) to facilitate administration of the computer-based ELPAC.
  - Where appropriate, use the universal tools (zoom, line reader, etc.) introduced during test examiner training and described in Matrix Four.
  - To minimize risk of unforeseen usability challenges, use the resources built into the platform, not affordances of the specific device, to adjust settings (e.g., zoom using the test delivery system, not the track pad or touch screen).

6. **Increase Technology Familiarity—**Provide appropriate supports to ensure students' level of familiarity with technology does not impede their ability to take the computer-based ELPAC:
   - Two new resources were added to Matrix Four to assist students who did not have enough experience with technology to navigate through the test delivery system alone and to assist students who could not enter their responses without support. In June 2019, the test navigation assistant was added as a non-embedded universal tool and the designated interface assistant was added as a non-embedded designated support. Additionally, print-on-demand was added as an embedded designated support so students who may not have been comfortable reading on the computer screen had the opportunity to print the items, if the test examiner felt this was necessary.
   - A document entitled *ELPAC Accessibility Resources for Operational Testing* (CDE, 2019b) was created that covered guidelines for the use of accessibility resources. This was communicated to the field when the ELPAC regulations were approved in September 2019.
7. **Simplify the Administration of the Speaking Test—**Simplify the Speaking administration to make test administration and scoring easier for the test examiner:
   - Speaking *DFAs* were developed specific to each grade level or grade span, allowing the test examiner to read test directions and questions and have access to rubrics, anchor samples, and prompting guidelines for test administration. The *DFAs* included a score sheet that test examiners used to score in the moment and then entered the Speaking scores into the DEI upon completion of the administration. The Speaking *DFAs* were available as PDFs and could be downloaded for optional printing.
   - The Speaking *DFA* had two diagramed options for seating arrangements for the test examiner and student.
   - The Speaking *DFA* incorporated directions for the test examiner to begin the audio recording of Speaking responses. For each test question, a microphone icon was placed before the "say" statement to provide an indicator and reminder to the test examiner to begin the recording.

8. **Improve the *DFAs*—**Improve the organization of the *DFAs:*
    - The Speaking *DFAs* were set up by task type and the administration directions were embedded within the test examiner script. Notes to the test examiner and prompting guidelines were placed within each task type and, if appropriate, each test question.
    - Checks were performed to ensure consistency between the test delivery system and the *DFAs*. The *DFAs* were organized to place scripts, prompting, and pointing all on the same page. For each test question, a microphone icon was placed before the "say" statement to provide an indicator and reminder to the test examiner to begin the recording.
9. **Enhance Training for Test Examiners—**Enhance administration and scoring training for test examiners:
    - Twenty-two day-long statewide trainings were held for LEAs from September through November 2019. The training incorporated test administration for kindergarten through grade twelve and included videos of students and test examiners on the computer-based platform. Most of the training focused on the administration and scoring of the Speaking domain.
    - LEA ELPAC trainers and test examiners who attended the AST were instructed to bring an electronic device to the training to practice the administration using the training tests.
    - The training had participants watch a video of the one-on-one kindergarten through grade two administration and participants logged on to the kindergarten training tests for practice.
    - Training videos were created to demonstrate exemplary administration models and then were shown during the trainings.

10.2 Test Delivery and Administration

10.2.1 Postadministration Survey

During the fall 2019 computer-based ELPAC field test administration, ETS administered a post-test survey to LEAs. The survey focused on gathering information on the clarity of the *DFAs*, knowledge and use of training tests and the TRCS, student interaction with the online test delivery system, knowledge and use of accessibility resources, and overall administration experience.

In response to the LEA feedback, ETS implemented the following improvements for the 2019–2020 operational administration:

- Updated *DFAs* and the *Test Administration Manual* with more concise wording and less repetition
- Promoted the availability of the TRCS, training tests, and practice tests to help LEAs and students prepare for the operational assessment

- Promoted the availability of the DEI demonstration video in Moodle
- Clarified and provided additional communication on the use of universal tools, designated supports, and accommodations by promoting the Student Accessibility Resources web page on the ELPAC website, at https://www.elpac.org/test-administration/accessibility-resources/
- Translated test directions are a non-embedded designated support where a biliterate adult trained in the *Directions for Administration* can read the test directions to a student with limited English skills. Note that this designated support does not include reading any part of a test question to the student. ETS added 14 more languages to the current three translated test directions for the ELPAC on the basis of feedback from the focus group with California educators who participated in the usability pilot and the field test; the additional languages included the following:
  - Arabic
  - Armenian
  - Farsi
  - Hindi
  - Hmong
  - Japanese
  - Khmer
  - Korean
  - Mandarin
  - Punjabi
  - Russian
  - Tagalog
  - Telugu
  - Urdu

10.2.2 Training and Communication

Training and communication will be focal points moving forward as ETS continues work on the computer-based Initial ELPAC. ETS will continue to provide timely communication for each critical component of the ELPAC administration, including material order dates and deadlines and training schedules. ETS will continue to work with SCOE to emphasize the importance and necessity of training, along with providing statewide training to LEA staff so they are prepared to administer the test. Training will continue to focus on local scoring of the Speaking domain and Writing domain for the Initial ELPAC.

ETS will continue to support familiarizing students with the ELPAC items using practice and training tests and informational videos. Parent/Guardian engagement continues to be an important factor for student participation and familiarization. To that end, ETS will work with the CDE to increase communication and information targeted at parents. Communications will also encourage LEAs to use the practice and training tests to prepare students to become more familiar with the computer-based Initial ELPAC.

10.3 Human Scoring

During field tests, double scoring percentages will be set to ensure a minimum of 2,500 double-scored responses per item to ensure adequate sample sizes for future rater reliability analyses. Additionally, sets of validity samples will be created and deployed during field test scoring to evaluate raters. Validity sample sets include responses that have been

prescored by scoring experts who came to a consensus on the score. Evaluating raters' agreement with consensus scores on validity responses is a measure of scoring accuracy that will help ensure scoring quality of the ELPAC field test items and more closely mirror Initial ELPAC operational scoring conditions.

On the basis of score quality monitoring by prompt, training materials will be reexamined and updated, if necessary. Raters will be retrained using the updated materials to improve rater accuracy and agreement This will be performed when a prompt's interrater reliability or validity exact agreement rate falls below the standard threshold.

10.4 Psychometric Analysis

As the computer-based Initial ELPAC transitions from a field test to operational administrations beginning in late summer 2020, the Psychometric Analysis & Research team will continue to maintain best practices to ensure quality of psychometric results and look for ways to streamline and improve psychometric processes.

10.5 Accessibility

With the launch of the computer-based ELPAC, students have access to a much larger range of accessibility resources during testing than those allowed as part of the PPT ELPAC administrations. The field test phase provided an opportunity to evaluate the embedded and non-embedded universal tools and designated supports, as well as to consider the embedded and non-embedded accommodations that will be available as part of the online test delivery system. Unlike the paper–pencil administrations, for computer-based testing, the LEA staff will assign and verify designated supports and accommodations in TOMS prior to the student testing. Universal tools will be available to all students in the online interface.

References

California Department of Education. (2019b). *ELPAC accessibility resources for operational testing.* Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ep/documents/accessibilityresources.docx

California Department of Education. (2019a). *ELPAC Usability Pilot: A final report (with addendum).* Sacramento, CA: California Department of Education.