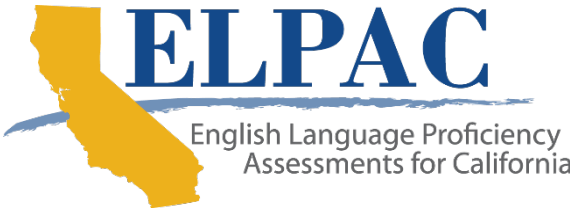

Computer-based Summative English Language Proficiency Assessments for California (ELPAC) Fall 2019 Field Test Technical Report

Submitted May 21, 2021

Educational Testing Service

Contract #CN150012

Table of Contents

List of Appendices

List of Tables

List of Figures

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

Term	Definition
2PL	two-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
CBA	computer-based assessment
CCR	*California Code of Regulations*
CDE	California Department of Education
CELDT	California English Language Development Test
CR	constructed response
DEI	Data Entry Interface
DFA	*Directions for Administration*
DHOH	deaf or hard of hearing
DIF	differential item functioning
EC	*Education Code*
EL	English learner
ELA	English language arts/literacy
ELD	English language development
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
ETS	Educational Testing Service
GPC	generalized partial credit
IBIS	Item Banking Information System
IEP	individualized education program
IFEP	initial fluent English proficient
IRT	item response theory
K	kindergarten
LEA	local educational agency
MC	multiple choice
MH	Mantel-Haenszel
ONE	Online Network for Evaluation

Table of Acronyms and Initialisms *(continuation)*

Term	Definition
OTI	Office of Testing Integrity
PAR	Psychometric Analysis & Research
PPT	paper–pencil test
RFEP	redesignated fluent English proficient
RMSD	root mean square difference
SBE	State Board of Education
SCOE	Sacramento County Office of Education
SD	standard deviation
SMD	standardized mean difference
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
TK	transitional kindergarten
TOMS	Test Operations Management System
TRCS	Technology Readiness Checker for Students
UAT	user acceptance testing
USC	United States Code
VI	visually impaired

Chapter 1: Introduction

This technical report focuses on the development, administration, psychometric analyses, and results of the computer-based Summative 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 the computer-based Summative 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 English proficient.

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 computer-based ELPAC has replaced the paper–pencil Summative ELPAC as of February 2020 and will replace the paper–pencil Initial ELPAC on August 20, 2020.

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 English learners (ELs) to be enrolled in an English language development program. Students identified as ELs on the Initial ELPAC go on to take the Summative ELPAC. 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 the California public school system. The Summative ELPAC is administered annually to students in kindergarten through grade twelve who have been 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 grades and grade spans: kindergarten, 1, 2, 3–5, 6–8, and 9–12.	There are seven grades 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. Student Score Reports are now available electronically and can be 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. Student Score Reports are provided by ETS electronically to the LEAs.

1.1.2. ELPAC Paper-based Mode Comparability Forms

The mode comparability aspect of the computer-based ELPAC field test was supported by random assignment of schools to several conditions (refer to table 2.4). Form C1 was a reprint of the 2018–2019 Summative ELPAC paper test form. Some schools were randomly assigned to Form C1 and the computer-based equivalent of the Listening and Speaking domains, referred to as Form C2. Schools assigned to these forms administered the Listening and Speaking domains in both the paper- and computer-based formats. Other schools were randomly assigned to Form C1 and the computer-based equivalent of the Reading and Writing domains, called Form C3 (refer to table 2.4). Schools assigned to these forms administered the Reading and Writing domains in both the paper and computer-based formats. The Writing domain for kindergarten through grade two was administered only on paper. Consequently, students in these grades were not administered the computer-based C3 version of the assessment for the mode comparability study.

The goal of the mode comparability study was to establish reliable interpretations of linking between the ELPAC paper–pencil test (PPT) scores and the computer-based assessment scores. The study adopted various measurement invariance and linking approaches, which included mode differential item functioning (DIF) analysis, single and equivalent group designs, and common item equating. The purpose of the study was twofold. First, it aimed to investigate whether the delivery modes (i.e., PPT and computer-based testing) would

impact student performance. Second, it provided an opportunity for ETS to explore options to link the computer-based test scores back to the ELPAC PPT scales.

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. Data from these forms was used for statistical analyses and scaling.

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

Form 2 was the Initial ELPAC field test form. This form also included computer-based items that aligned to the 2018–2019 Summative ELPAC blueprints and the current Initial ELPAC blueprints. Additional Reading and Writing items were included in this form for grades spans three through five, six through eight, and nine through twelve to serve as vertical and horizontal linking items. The Writing domain was administered only on paper Answer Books for students in kindergarten through grade two.

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, initial fluent English proficient, or “to be determined” were eligible to participate in the computer-based ELPAC field test. Student participation in the field test was voluntary.

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. LEAs with schools administering the mode comparability PPT version of the field test were asked to return completed Answer Books weekly.

Table 1.2 shows the number of items and the estimated time to complete each field test form for computer-based assessments, shown here as "CBA"; and paper–pencil tests, shown here as "PPT." LEAs were advised to administer the Summative and Initial field test forms over multiple test sessions or days.

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

Variable	Oral Language Mode Comparability Form (Forms C1 and C2)	Written Language Mode Comparability Form (Forms C1 and C3)	Summative Field Test Form (Form F1)	Initial Field Test Form (Form F2)
Number of Items for K–Grade 2	32–39 CBA items 30–35 PPT items	17–32 CBA items* 22–33 PPT items*	64–86 items	63–88 items
Estimated Time for K–Grade 2	K: 25 minutes (CBA) K: 30 minutes (PPT) 1: 35 minutes (CBA) 1: 35 minutes (PPT) 2: 50 minutes (CBA) 2: 50 minutes (PPT)	K: 30–40 minutes (CBA) K: 35–45 minutes (PPT) 1: 35–45 minutes (CBA) 1: 30–40 minutes (PPT) 2: 45–55 minutes (CBA) 2: 40–50 minutes (PPT)	K: 75–85 minutes 1: 85–95 minutes 2: 110–120 minutes	K: 75–85 minutes 1: 85–95 minutes 2: 105–115 minutes
Number of Items for Grade Levels 3–12	45–46 CBA items 32–35 PPT items	43–48 CBA items 32 PPT items	90–95 items	88–94 items
Estimated Time for Grade Levels 3–12	3–12 CBA: 65–70 minutes 3–12 PPT: 50–55 minutes	3–12 CBA: 110–140 minutes 3–12 PPT: 90–120 minutes	3–12: 175–210 minutes	3–12: 175–210 minutes

*Reading only

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 of the federal Every Student Succeeds Act and the state's Public School Accountability Act, which measures 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 defining the purpose and scope of the assessment, assessment design, item development, standard setting, score reporting, and scoring the constructed-response (CR) items. Details about the California educators who participated in tasks involving the ELPAC can be found in table 2.1.

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

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
- Processing student test assignments
- Completing all psychometric procedures

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, 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 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

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

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 PPT 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

Students who are identified as EL students must be tested annually during the annual Summative ELPAC assessment window—February 1 through May 31—until they are reclassified as RFEP. Because the Summative ELPAC is the ELP assessment developed pursuant to *EC* Section 60810, scores from the Summative ELPAC are one set of criteria used to determine whether individual students qualify for RFEP status. Results from the Summative ELPAC may also be used to plan for instruction.

One limitation of this field test was that the testing window for summative ELPAC was shifted from spring to fall. If English proficiency of students changed over the summer, the sample taking the assessment in the fall was less likely to be representative of the typical summative student population.

Another limitation of this field test was the small sample sizes for some test forms. 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 development, test construction, 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 and describes the process of test assembly, including the content being measured, as well as the content and psychometric criteria. 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 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. (2020a, June). *Organization.* Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/re/di/or/

California Department of Education. (2020b, June). *State Board of Education responsibilities.* Sacramento, CA: California Department of Education. Retrieved from https://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://mytoms.ets.org/

Chapter 2: Item Development and Test Assembly

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

2.1. Overview

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

In addition, approximately 360 new items were developed for computer-based administration. Most items used in the field test came from the pool of 2,289 converted items. A relatively small number of the 360 newly developed items were selected for use in the field test. 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 Summative ELPAC and to verify that they had sufficient vertical and horizontal linking items to support psychometric analyses.

2.2. Summative ELPAC Test Blueprints

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 (CELDT) 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 Summative Assessment, was administered. Analysis of the pilot and the stand-alone sample field test results led to modifications of the Summative 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 revised Summative ELPAC test blueprints in September 2017.

Test blueprints were developed to describe the content of the Summative 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 2018, the SBE approved plans to transition the ELPAC from a paper-based assessment to a computer-based assessment. As part of the transition work, the Summative ELPAC test blueprints were reviewed to determine where minor adjustments could be made to appropriately use computer-based delivery and increase the amount of information collected at the upper range of ELP, while continuing to ensure the assessment remains fair and valid for its intended purposes.

The most substantial revisions to the Summative ELPAC test blueprints were the addition of two existing task types to grade one and grade two. The task type of *Listen to a Classroom Conversation* was added at grade one and grade two because the introduction of Listening audio files at those grades made it possible for students to listen to conversations between two speakers. *Write About an Experience* was added at grade one and grade two to collect more information at the upper range of ELP because it was similar to *Short Compositions*, which had been administered at those grades in the ELPAC's predecessor ELP assessment, the CELDT. In addition, a second *Speaking—Retell a Narrative* item was added at kindergarten and a second *Speaking—Summarize an Academic Presentation* item was added at grades one through twelve to collect more information at the upper range of English language proficiency.

The SBE approved the revisions to the computer-based Summative ELPAC test blueprints in May 2019.

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 grades.

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, 2019b). 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 Writing assessments.

The *Proposed High-Level Test Design for the Transition to Computer-Based ELPAC* 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 (SCOE), 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, 2019d).

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 through grade two students and recently arrived English learners [ELs]) 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 across six schools from two LEAs participated in the study. Of the 89 EL students, 13 were EL students with disabilities.

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 test
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 of 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 Summative 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* (ETS, 2019d). The specifications described the details of the process followed to prepare Summative 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 and 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 Summative 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 Summative ELPAC field test, K through grade two students sat one-on-one with a test examiner. This allowed the test examiners to provide one-on-one support to operate the computer. 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 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 tests were read aloud by the test examiner from printed *DFAs*. The directions for the grade two Reading test were presented as on-screen text that was read aloud by the test examiner.

For the grades three through twelve tests, directions for the Reading domain were presented as on-screen text along with audio recordings. Students were directed to listen to the audio recordings while reading the directions themselves. Item-level directions appeared on the same screen as the Reading stimulus.

2.5.4. Writing Domain

For the Writing domain, K through grade two 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 through grade two. 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

The items that were administered during the computer-based Summative ELPAC field test came from two sources: the existing paper–pencil item pool of over 2,200 items that was converted for computer-based administration; and the set of approximately 360 items that were newly developed from 2018 to 2019.

The majority of the computer-based ELPAC items came from the existing pool of paper-based ELPAC items, which were converted for use on the computer-based ELPAC. During the creation of the computer-based field test forms, preference was given to use of converted items that had performed well statistically in prior paper-based administrations over newly developed items because it seemed more likely that the converted items would perform well statistically.

2.7. Item Development Plan

A small number of new items created during 2018 and 2019 were administered in the Summative ELPAC field test. The new items were grade one and grade two *Listen to a Classroom Conversation* and *Write about an Experience* items. These newly developed items needed to be used because these task types were not administered at grade one and grade two in the paper-based administrations.

In partnership with SCOE, ETS convened ELPAC item writer trainings and item review panels to develop test items for the Summative ELPAC. Select California educators were trained to write new items for the Summative ELPAC. In addition, ETS trained a small group of experienced contractors to draft Summative ELPAC items. After the items went through ETS internal and CDE reviews, California educators reviewed the items during Content Review Panel and Bias and Sensitivity Review Panel meetings.

This section describes how California educators were selected and the process used to develop new items for the Summative ELPAC.

2.7.1. Selection of Item Writers

California educators were recruited through email communications and by letter. To ensure broad representation, an email message and letter announcing the opportunities to write items and to review items were sent by the CDE to the following groups:

- The CDE's ELPAC listserv (includes CELDT District Coordinators and Title III county leads)
- The Bilingual Coordinators Network

- The CDE's California Assessment of Student Performance and Progress Coordinator listserv
- The CDE's All Assessment listserv
- The ELPAC Technical Advisory Group

The email and letter directed applicants to fill in an online application in SurveyMonkey, a third-party, online survey provider. The application allowed California educators to apply for any or all of the events. The information from the application was loaded into a database that was used for the review and selection process.

During the selection process, applications were selected from current and retired California educators who had the following minimum qualifications:

- Bachelor's degree
- Expertise in language acquisition or experience teaching EL students in K through grade twelve
- Knowledge of, and experience working with, the 2012 ELD Standards

Additional desirable qualifications included the following:

- A teaching credential authorization for English language development, specially designed academic instruction in English, or content instruction delivered in the primary language (e.g., Cross-cultural, Language, and Academic Development Certificate; or Bilingual, Cross-cultural, Language, and Academic Development Certificate)
- Specialized teaching certification in reading (e.g., Reading Certificate or Reading and Language Arts Specialist Certificate)
- Experience writing or reviewing test items for standardized tests, especially tests for EL students in K through grade twelve
- Recent experience administering the CELDT

Selections were made to ensure representation from different cultural and linguistic groups, various-sized LEAs and county offices of education, and different geographical regions of the state, and with regard to the travel budget allowable in the contract. ETS and SCOE made preliminary selections, which were reviewed by the CDE, adjusted as needed, and then approved. Twenty-one educators were selected for item writer training, along with 14 alternates. Twenty-one educators were selected for Content Review Panels, along with 22 alternates. Eighteen educators were selected for Bias and Sensitivity Review Panels, along with 18 alternates.

Table 2.1 shows the educational qualifications, present occupation, and credentials of the individuals who participated in an ELPAC item writer training or item review panel.

Table 2.1 ELPAC Item Writer Training (IWT) and Item Review Panel (IRP) Qualifications, by Meeting Type and Total

Qualification Type	Qualification	IWT	IRP	Total
Occupation	Classroom teacher	5	14	19
Occupation	English learner or literacy specialist or coach	9	17	26
Occupation	School administrator	4	1	5
Occupation	LEA or county office employee	0	7	7
Highest degree earned	Bachelor's degree	1	7	8
Highest degree earned	Master's degree	16	26	42
Highest degree earned	Doctorate	1	6	7
K–12 teaching credential	Elementary Teaching (multiple subjects)	13	23	36
K–12 teaching credential	Secondary Teaching (single subject)	4	15	19
K–12 teaching credential	Special Education Teaching	0	1	1
K–12 teaching credential	Language Development Specialist	1	3	4
K–12 teaching credential	English Learner (CLAD, BCLAD)	9	13	22
K–12 teaching credential	Other	7	3	10

Note: Numbers may not match the totals because participants may have multiple occupations or teaching credentials, or are currently working toward earning their highest degree. The information is self-reported and may not reflect all the experience and earned credentials.

SCOE contacted and invited the participants and contacted the alternates as necessary. Once all participants confirmed, SCOE notified those who were not selected.

2.7.2. Item Writing by Educators

Item writer training for California educators was divided into two meetings, each of which lasted two days.

A total of 21 educators were trained to develop items during the item writer training meetings in 2018. Nine educators from K through grade two were trained on Monday and Tuesday, October 8 and 9, 2018. Twelve educators from grades three through twelve were trained on Wednesday and Thursday, October 10 and 11, 2018.

The educators represented a mix of rural, suburban, and urban LEAs.

2.7.2.1 Introduction to Item Writing

During each of the two-day meetings, educators received training and then drafted ELPAC items. At the start of day one, a PowerPoint presentation was used to provide information to the educators about topics regarding the ELPAC and item development. Topics covered during the presentation included an overview of the ELPAC, general principles of item development, a review of the 2012 ELD Standards, the overall item development process, and the process for drafting and submitting items. After the PowerPoint presentation, ETS trainers provided educators with examples of task types that are shared across grade levels and grade spans.

ETS trainers facilitated brainstorming sessions, during which educators listed topics that served as a basis for item development. Educators were asked to propose topics for item content that are covered during prior grades to ensure that topics were appropriate. After brainstorming, educators worked as a whole group to assign topics to appropriate grade levels or grade spans. Educators then split up into grade-level groups to draft items corresponding to the topics from their brainstorming session. This pattern was followed for all domains (Listening, Speaking, Reading, and Writing).

2.7.2.2 Process

After educators divided into their grade-level groups, ETS trainers provided them with *Item Writing Guidelines for the ELPAC* (CDE, 2018), sample items, and item templates. The *Item Writing Guidelines for the ELPAC* provided details about the type of information that is required when drafting items, such as the length of any Listening stimuli or Reading passages, the number of items within the set, and the types of English language knowledge, skills, and abilities to be assessed by the items.

The sample items were developed by ETS assessment specialists to serve as examples of the task types to be developed. The item templates were Word files that contained areas for entering information. The item templates assured that items were drafted in a standardized manner and that all needed item information was entered. ETS trainers used the *Item Writing Guidelines for the ELPAC*, sample items, and item templates as training materials to provide clear expectations regarding the information needed when drafting each task type, as well as the level of quality that was expected.

All items developed by educators were drafted according to assignments that were given during the item writer training meetings. Educators were not given assignments to be completed after the meetings.

2.7.3. Item Writing by Contractors

In 2018, ETS assessment specialists worked with five contractors (i.e., outside item writers) who were fully trained, experienced item writers with a record of developing quality items for other ETS English language assessments. Because there was a limited amount of time to train California educators to develop Listening and Reading sets, ETS contractors developed the Listening task types with relatively long stimuli and the Reading task types with relatively long passages. The focus of the contractors was to develop the following task types:

- *Listening—Listen to a Story*
- *Listening—Listen to an Oral Presentation*
- *Reading—Read a Literary Passage*
- *Reading—Read an Informational Passage*

The contractors delivered all items to a secure ETS server. After ETS confirmed receipt of the files, contractors were prompted to delete the files from their personal devices.

2.8. Task Types and Features

2.8.1. Task Types

The Summative ELPAC field test contained 27 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 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 Summative ELPAC had five task types, the Speaking domain had six task types, the Reading domain had nine task types, and the Writing domain had seven 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 Summative ELPAC.

2.8.1.1 Listening Task Types

Listening task types, for the Summative ELPAC, assessed the ability of an EL 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 one 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.

2.8.1.2 Speaking Task Types

Speaking task types, for the Summative ELPAC, assessed the ability of an EL 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 six 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 two through twelve:** Students stated what they would say in a situation described by the test examiner.
- **Support an Opinion, K through grade twelve:** 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. At K, grade one, grade two, and grade span three through five, students viewed a picture of the choices for context and support.
- **Retell a Narrative, K through grade five:** The student listened to a story that followed a series of pictures, and then the student used the pictures to retell the story.
- **Present and Discuss Information, grades six through twelve:** The student viewed a graph, chart, or image that provided information. The student was prompted to read the information and then asked to respond to two prompts. The first prompt asked for a summary of the information in the graph, chart, or image. The second prompt asked for the student to state whether a claim was supported or unsupported based on the information in the graph or chart.
- **Summarize an Academic Presentation, K 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.

2.8.1.3 Reading Task Types

Reading task types, for the Summative ELPAC, assessed the ability of an EL to read, analyze, and interpret a variety of grade-appropriate literary and informational texts. The following are descriptions of the nine 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:** 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, K:** 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, grade one:** The student read three words and chose the word that matched a picture.
- **Read and Choose a Sentence, grades one through five:** The student read three or four sentences and chose the sentence that best described a picture.
- **Read a Short Informational Passage, grades one through twelve:** The student read a short informational text and answered MC questions related to the text.
- **Read a Student Essay, grades three through twelve:** The student read an informational essay presented as if written by a peer and answered a set of MC questions related to the essay.
- **Read a Literary Passage, grades one through twelve:** The student read a literary text and answered MC questions related to the text.
- **Read an Informational Passage, grades one through twelve:** The student read an informational text and answered MC questions related to the text.

2.8.1.4 Writing Task Types

Writing task types, for the Summative ELPAC, assessed the ability of an EL to write literary and informational texts to present, describe, and explain information. The following are descriptions of the seven Writing task types:

- **Label a Picture—Word, with Scaffolding, K:** With scaffolding from the test examiner, the student wrote labels for objects displayed in a picture.
- **Write a Story Together with Scaffolding, K through grade two:** 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.
- **Write an Informational Text Together, grades one and two:** With scaffolding from the test examiner, the student listened to a short informational passage and then collaborated with the test examiner to jointly compose a text about the passage by writing a dictated sentence and an original sentence about the topic.
- **Describe a Picture**
  - **Grades one and two:** The student looked at a picture and wrote a brief description about what was happening.
  - **Grades three through twelve:** 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 one 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.

- **Write About Academic Information, grades three through twelve:** The student interpreted academic information from a graphic organizer created for a group project and answered two questions about it.
- **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.

2.9. Item Review Process

Before Summative 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.

2.9.1. 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.

2.9.2. 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.

2.9.3. 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 EL students

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. Table 2.2 shows the number of items that were developed in 2018–2019.

Table 2.2 Number of Items Developed in 2018–2019

Domain	K	1	2	3–5	6–8	9–10	11–12	Total
Listening	15	23	22	12	12	17	0	101
Speaking	6	4	6	9	7	19	6	57
Reading	11	24	26	21	23	22	23	150
Writing	11	10	11	4	6	8	4	54
Total	**43**	**61**	**65**	**46**	**48**	**66**	**33**	**362**

A total of 362 items were presented for educator reviews as described in the following subsection.

2.9.4. California Educator Review

Each newly-developed item was reviewed during two educator meetings: a Content Review Panel meeting and a Bias and Sensitivity Review Panel meeting.

Two trainings for the panel participants were conducted during the meetings and prior to the item reviews: educators serving on the Content Review Panel were trained from February 12–13, 2019. Educators serving on the Bias and Sensitivity Review Panel were trained from February 14–15, 2019.

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, 2014).

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.

Table 2.3 provides the status of the items after the 2019 item review panel meetings.

Table 2.3 Status of Items After the 2019 Item Review Panel Meetings

Grade Level or Grade Span	Approved As Is	Approved with Revisions	Rejected
Kindergarten	11	32	0
Grade 1	16	45	0
Grade 2	0	61	4
Grade span 3–5	3	43	0
Grade span 6–8	2	46	0
Grade span 9–10	9	53	4
Grade span 11–12	1	32	0
Totals:	**42**	**312**	**8**

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. In 2019, 98 percent of the 362 items were approved. Educators enhanced the quality of the item pool by providing suggestions for revising items during Content Review Panel meetings and Bias and Sensitivity Review Panel meetings.

2.10. Test Assembly

ETS assessment specialists assembled the Summative 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 Summative 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.

2.10.1. Field Test Forms

Summative ELPAC field test forms were administered within a combined field test administration that included three activities:

1. Mode comparability study of student data resulting from a paper–pencil form and a computer-based form
2. Summative ELPAC computer-based field test in preparation for the 2019–2020 Summative ELPAC operational administration
3. Initial ELPAC computer-based field test in preparation for the 2020–2021 Initial ELPAC operational administration

This subsection describes the composition of the entire combined field test. However, the analyses in this report focus on the performance of items in the comparability study and the Summative ELPAC field test. A separate report will describe the performance of items in the Initial ELPAC field test.

The combined field test included a total of five field test forms per grade 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 through grade two (K–2) because the responses for the computer-based ELPAC were already paper-based only for these grades.
2. **Mode Comparability Study 2 (C2):** This was a computer-based form consisting of oral language items 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 items 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 grades.
4. **Summative Field Test Form (F1):** This was a preassembled 2019–2020 computer-based Summative form aligned with the 2019-adjusted Summative blueprints. It was also aligned to the *Initial Assessment Test Blueprints for the ELPAC*. 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 2.4 shows the test form configurations.

Table 2.4 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

The length of the summative form F1 was approximately 40 percent longer than the number of items in the adjusted Summative ELPAC test blueprint because of the need for horizontal and vertical linking items. Table 2.5 shows the approximate number of items in the Listening and Speaking computer-based forms (C2 and F1) compared to the number of items in the adjusted Summative ELPAC test blueprints.

Table 2.5 Approximate Number of Listening and Speaking Items in the Field Test Forms

Grade Level or Grade Span	Form	Field Test Listening Items	Listening Blueprint	Field Test Speaking Items	Speaking Blueprint
K	F1	28	20	13	9
K	C2	28	20	13	9
Grade 1	F1	31	22	13	9
Grade 1	C2	31	22	13	9
Grade 2	F1	31	22	17	12
Grade 2	C2	31	22	17	12
Grades 3–5	F1	31	22	17	12
Grades 3–5	C2	31	22	17	12
Grades 6–8	F1	31	22	17	12
Grades 6–8	C2	31	22	17	12
Grades 9–10	F1	31	22	17	12
Grades 9–10	C2	31	22	17	12
Grades 11–12	F1	31	22	17	12
Grades 11–12	C2	31	22	17	12

Table 2.6 shows the approximate number of items in the Reading and Writing computer-based forms (C3 and F1) compared to the number of items in the adjusted Summative ELPAC test blueprints.

Table 2.6 Approximate Number of Reading and Writing Items in the Field Test Forms

Grade Level or Grade Span	Form	Field Test Reading Items	Reading Blueprint	Field Test Writing Items	Writing Blueprint
K	F1	20	14	11	8
K	C3	20	14	0	8
Grade 1	F1	29	21	10	7
Grade 1	C3	29	21	0	7
Grade 2	F1	36	26	9	6
Grade 2	C3	36	26	0	6
Grades 3–5	F1	36	26	9	6
Grades 3–5	C3	36	26	9	6
Grades 6–8	F1	36	26	9	6
Grades 6–8	C3	36	26	9	6
Grades 9–10	F1	36	26	9	6
Grades 9–10	C3	36	26	9	6
Grades 11–12	F1	36	26	9	6
Grades 11–12	C3	36	26	9	6

One or two task types were selected from each domain for horizontal linking. To the greatest extent possible, horizontal linking occurred across each of the computer-based forms—C2, C3, and F1. The vertical linking items included items in selected task types from adjacent grade levels (one grade above and one grade below). The items or tasks used from adjacent grades were selected by evaluating the item statistics obtained during previous paper-based administrations to ensure that items would provide appropriate information. For example, items that were considered difficult for the grade level above would not be suitable as vertical linking items because they were likely to be even more challenging for the target grade. In general, easier item types from the grade level above the target grade, and more difficult items from the grade level below the target grades, were considered good candidates for vertical linking.

Assessment specialists at ETS developed form planners showing the number of items to be administered at each grade and domain, along with the horizontal linking and vertical linking plans. 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 comparability and 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.

2.11. Field Test Design

For each grade or grade span, forms C1, C2, and C3 were used for the mode comparability study, where individual students took an oral language (i.e., Listening and Speaking) or written language (i.e., Reading and Writing) test in both paper-based and computer-based formats in a counterbalanced design. That is, these three forms were administered in the configuration of C1/C2, C2/C1, C1/C3, and C3/C1 to reduce the practice effect caused by taking the same content twice within the testing window. Further rationale on the use of these forms may be found in section *5.1 Data Collection Plan*.

2.11.1. Considerations for Fall Testing Window

The operational Summative ELPAC testing window occurs annually from February 1 through May 31.

The ELPAC mode comparability study was conducted in October 2019; thus, decisions regarding when to administer the field test needed to consider the timing shift between the operational administrations and the field test. PPT items designated for the Summative ELPAC at grade *X* were administered to students in grade *X* + 1 for the fall computer-based field test.

Table 2.7 presents the content included in the field test forms. It should be noted that the ELPAC at the K level includes transitional K (TK) and K students. TK students in fall 2019 who were newcomers were not included in the administration of Forms C1, C2, C3, and F1.

Table 2.7 Grade Level of Test Content in the Summative Field Test Form

Students (Enrollment as of Fall 2019)	Forms C1–C3 and F1
TK	N/A
K	Summative K
1	Summative K
2	Summative Grade 1
3	Summative Grade 2
4	Summative Grades 3–5
5	Summative Grades 3–5
6	Summative Grades 3–5
7	Summative Grades 6–8
8	Summative Grades 6–8
9	Summative Grades 6–8
10	Summative Grades 9–12
11	Summative Grades 9–12
12	Summative Grades 9–12

2.11.2. Psychometric Review

The ETS Psychometric Analysis & Research (PAR) group reviewed the test forms to ensure that they aligned with the field test design. PAR staff checked Form C2 and Form C3 to ensure that they contained computer-based items that corresponded with the paper-based items in C1 for the purposes of the comparability study analyses. PAR ensured that Form F1 conformed to the Summative ELPAC test blueprints. They checked the number of horizontal linking items across Forms C2, C3, and F1, as well as the number of vertical linking items, to ensure that the linking would allow for appropriate analyses after the field test administration.

The following criteria were used to review forms C2 and C3:

- Do the forms include all the items from the 2017–2018 Summative ELPAC?
- Do the forms contain about the same number of items as described in table 2.5 (for form C2) and table 2.6 (for form C3)?
- Do the forms include appropriate numbers of vertical linking and horizontal linking items?

The following criteria were used to review form F1:

- Do the forms align with the Summative ELPAC and the Initial ELPAC test blueprints?
- Do the forms contain about the same number of items as described in table 2.5 and table 2.6?
- Do the forms include appropriate numbers of vertical linking and horizontal linking items?

The following steps were taken when reviewing different versions of the forms in reference to the criteria:

1. Items from C2 and C3 were matched with the final form planners of the 2017–2018 Summative ELPAC to make sure C2 and C3 contained all the oral and written language items of 2017–2018 forms, respectively.
2. Item counts and score points for each task type were summarized for form F1.
3. The statistics for form F1 were compared both with the counts in table 2.5 and table 2.6, and with the adjusted Summative ELPAC and the Initial ELPAC test blueprints.
4. The numbers of vertical linking and horizontal linking items were aggregated from the form planners. Once PAR staff realized there were too few linking items, they proposed that assessment specialists at ETS add more items into the linking list so these items accounted for at least 20 percent of the total numbers of items.
5. PAR staff approved the forms only when all the aforementioned criteria were met.

In short, all the forms C2, C3, and F1 used for the field test were aligned with the field test design and can support reliable linking analyses.

2.11.3. California Department of Education (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 Data Entry Interface (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 https://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*. 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. (2018). *Item writing guidelines for the English Language Proficiency Assessments for California.* [Unpublished report]. Sacramento, CA: California Department of Education.

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

California Department of Education. (2019a). *Test blueprints for the Summative English Language Proficiency Assessments for California*. Approved by the California State Board of Education in May 2019. Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ep/documents/elpacsummativebluprt.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

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 Summative English Language Proficiency Assessments for California (ELPAC) field test forms, as well as test security, accessibility resources, participation, and demographic summaries.

3.1. Field Test Administration

The computer-based Summative ELPAC field test window for fall 2019 was October 1 through November 8, 2019. The field test was open to all local educational agencies (LEAs) that had more than 20 English learner (EL) students within a grade level within a school. LEAs that agreed to participate were provided with a specific roster of students to test in each grade or grade span. Due to the timing of the field test, paper–pencil test items designated for the Summative ELPAC at grade *X* were administered to students in grade *X* + 1. LEAs were assigned to one of the three testing conditions (i.e., C1 + C2, C1 + C3, or F1).

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 grades, and the Writing domain, which had an optional group administration for grade two. The Writing domain was excluded for students in kindergarten through grade two (K–2) in schools assigned to the mode comparability paper–pencil field test form and the Writing computer-based field test form because this domain is assessed on paper only for these grades.

Students assigned to the computer-based field test forms 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 computer-based 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 log the student's Speaking scores that were later entered into the Data Entry Interface.

LEAs or schools assigned to administer the mode comparability paper–pencil ELPAC forms were provided with printed *Examiner's Manuals*, Test Books, Answer Books, pre-identification labels, and group identification sheets and were asked to return paper-based materials at least once a week.

Individual student scores, school-level scores, and score reports were not available to the LEAs for the computer-based ELPAC field test.

3.2. Test Security and Confidentiality

All testing materials for the fall 2019 computer-based ELPAC field test—Test Books, Answer Books, *Examiner's Manuals,* and *DFAs*—were considered secure documents. Every person with 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 be kept secure.

To ensure security for all tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI).

In an effort to enforce secure practices, ETS and the OTI strive to safeguard the various processes involved in a test development and administration cycle. For the fall 2019 computer-based 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 online and paper-based 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 *Directions for Administration* (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]).
- Anyone having access to the testing materials, but not having access to TOMS, must have signed the ELPAC *Test Security Affidavit for Non-TOMS Users*, which was available as a PDF on the ELPAC website, before receiving access to any testing materials.

In addition, it was the responsibility of every participant in the 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]).

There were no real or suspected violations of test security or confidentiality reported for the computer-based Summative ELPAC fall field test.

3.2.3. Test Security Monitoring

The LEA and school testing staff are responsible for maintaining the security and confidentiality of testing materials and devices during the testing window and reporting any irregularities or breaches that occur.

Normally, ETS performs site visits and testing procedure audits during testing windows. However, site visits and audits were not made for the LEAs participating in the field test due to the short testing window.

3.2.4. 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.

3.2.5. 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.

3.2.6. Statistical Analysis on Secure Servers

Immediately following submission of the fall 2019 computer-based ELPAC field test results into the test delivery system, either computer-based or scanned paper-based, 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 (eSKM) 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.

3.2.7. Student Confidentiality

To meet the requirements of the Every Student Succeeds Act as well as state requirements, LEAs must collect demographic data about students' ethnicity, disabilities, parent/guardian education, and so forth. ETS took every precaution to prevent any of this information from becoming public or being used for anything other than evaluation of the field test items. 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 (PPT).

3.2.8. Security and Test Administration Incident Reporting System 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 field test, all computer-based test irregularities and PPT irregularities, except for the K–2 Writing domain, were required to be reported in STAIRS. For K–2 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 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.

Potential testing irregularity types 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 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

Potential security breach types 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.

There were no STAIRS cases filed during the computer-based Summative ELPAC fall field test.

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. 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.[3]

The computer-based ELPAC field test offered commonly used accessibility resources available for the paper–pencil field test 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)

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)

[3] This technical report is based on the version of Matrix Four that was available during the computer-based ELPAC 2019 fall field test. Note that Matrix Four has since been combined with Matrix One to form a single accessibility resources matrix for California assessments.

- 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. Each designated support falls into one of two categories: embedded and non-embedded. Embedded designated supports are provided through the student testing interface (through the secure browser). Students that were assigned to take the paper–pencil ELPAC field test did not have access to embedded designated supports.

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 field test forms:

- Amplification
- American Sign Language or Manually Coded English
- Color overlay
- Magnification
- Masking
- Medical supports
- Noise buffers
- Pause or replay audio—Listening domain

- Pause or replay audio—Speaking domain
- 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. Instead, two activities were planned for the October through November 2019 timeframe to capture feedback on students using embedded accommodations: an ELPAC Accessibility Cognitive Laboratory Study for visually impaired (VI) and deaf or hard of hearing (DHOH) students; and an ELPAC Accessibility Pilot open to all interested LEAs. Participants were asked to administer the ELPAC training test to their VI students, DHOH students, or any students who required embedded accommodations, and to share their experiences using the embedded resources.

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 field test forms:

- Alternate response options
- American Sign Language or Manually Coded English
- Breaks
- Scribe
- Word processor (Writing domain) (grades three through twelve)

3.3.4. Resources for Selection of Accessibility Resources

The full list of the universal tools, designated supports, and accommodations that are available in ELPAC online and paper assessments 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.

In April 2019, ETS conducted the ELPAC usability pilot, which gave LEAs and students the opportunity to practice assigning and using embedded accommodations in the online training tests (ETS, 2019). The pilot used cognitive lab methodology to gather information on accommodated ELPAC testing materials for students who are visually impaired or who are deaf or hard of hearing. Student and test examiner feedback from the usability pilot, as well as recommendations for assessment, are available as a published report (CDE, 2020a).

Table 3.1 through table 3.3 list the non-embedded and embedded designated supports, non-embedded accommodations, and number of students assigned to each resource in TOMS by LEAs participating in the field test. (Because the test delivered was a field test and not an operational assessment, data on the use of accessibility resources within the student test delivery system was not collected.)

Table 3.1 Non-embedded Designated Supports Assigned in TOMS for the Field Test

Non-embedded Designated Support	N Students per Resource
Amplification	7
Color contrast	14
Color overlay	10
Designated interface assistant	14
Magnification	23
Medical supports	0
Noise buffers	30
Print on demand	0
Read aloud for items (Writing domain)	39
Separate setting	136
Total Resources Assigned:	**273**

Table 3.2 Embedded Designated Supports Assigned in TOMS for the Field Test

Embedded Designated Support	N Students per Resource
Color contrast	20
Masking	24
Mouse pointer (size and color)	17
Pause or replay audio (Listening and Speaking domains)	502
Permissive mode	1
Print size	30
Streamline	5
Turn off any universal tool(s)	0
Total Resources Assigned:	**599**

Table 3.3 Non-embedded Accommodations Assigned in TOMS for the Field Test

Non-embedded Accommodation	N Students per Resource
Alternate response options	0
Scribe	16
Speech-to-text	7
Total Resources Assigned:	**23**

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 Summative ELPAC field test.

3.4. Participation

Because student participation in the computer-based 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 met the minimum threshold requirement of having a school with at least 20 EL students in a grade 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.

The target population for the Summative ELPAC was existing EL students.

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 Summative assessments.

Attemptedness rules were applied to data where students responded to relatively few items. For summative data, students had to respond to at least four Listening items, three Speaking items, five Reading items, and two Writing items 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 test scores for forms C1, C2, C3, and F1 are provided, for all grades and grade spans, in table 3.A.1 through table 3.A.41 of appendix 3.A. Grade spans reflect students' enrolled grade spans during the 2019–2020 school year.

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 3.4. The tables in appendix 3.A show consistent patterns. Across the forms and for all grades and grade spans, female students accounted for about a half of the field test samples. It was also found that approximately 80 percent of the students were Hispanic or Latino. In terms of English proficiency, more than 90 percent of the test takers were EL students. The remaining students were either students who were classified as initial fluent English proficient (IFEP) or students whose classification as EL or IFEP was to be determined (TBD) at the time of the administration.

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 Summative 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 demonstrate their language proficiency 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. 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 (refer to table 3.A.1 to 3.A.41).

The demographic information for students taking each Summative ELPAC field test form looked similar to the distributions of the population of Summative ELPAC test takers in 2019. These are reported in appendix 11 of the *2018–2019 Summative ELPAC Technical Report* (CDE, 2020b). Across grades and grade spans, male students accounted for 50 to 60 percent of ELPAC test takers in both the 2018–2019 Summative ELPAC PPT and the field test data. 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 appeared to be higher than the percentage for the 2019 population, which was around 75 percent. Meanwhile, nearly 100 percent of the students taking part in the field test did not receive special education services. This difference occurred because there were no embedded accommodations to serve students receiving special education services during the field test phase.

Table 3.4 lists the demographic student groups reported.

Table 3.4 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
Five Identified Countries of Origin	1. Guatemala 2. Honduras 3. Colombia 4. El Salvador 5. Afghanistan
English-Language Fluency	• English only (EO) • Initial fluent English proficient (IFEP) • English learner (EL) • Redesignated fluent English proficient (RFEP) • Ever-ELs (EL or RFEP) • TBD • English proficiency unknown
Economic Status	• Not economically disadvantaged • Economically disadvantaged
Special Education Services Status	• No special education services • Special education services
Enrollment in U.S. Schools	• Less than 12 months • 12 months or more • Duration unknown

3.6. Training Test

The training tests were provided to LEAs to prepare students and LEA staff for the computer-based 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 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 test delivery system as well as 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://bit.ly/3gZ0sO3

California Department of Education. (2019b). *Matrix Four: Universal tools, designated supports, and accommodations for the English Language Proficiency Assessments for California*. Retrieved from https://bit.ly/2RtFmws

California Department of Education. (2020a). *Accessibility and usability for the English Language Proficiency Assessments for California: A cognitive lab study with students who are deaf or hard of hearing and students who are blind or have low vision*. Sacramento, CA: California Department of Education. https://www.cde.ca.gov/ta/tg/ep/documents/elpaccognitiverpt.pdf

California Department of Education. (2020b). *Summative English Language Proficiency Assessments for California technical report 2018–2019 administration*. [Draft 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

Educational Testing Service. (2019). *English Language Proficiency Assessments for California usability pilot: A final report.* [Unpublished document]. Sacramento, CA: California Department of Education.

Chapter 4: Scoring

This chapter summarizes scoring at the item level, including the scoring approaches that are used for each item type in the computer-based Summative English Language Proficiency Assessments for California (ELPAC) field test and the approach implemented to produce student scores.

4.1. 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 received in-depth training and were 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. The quality-control processes for CR scoring are explained further in *Chapter 8: Quality Control*.

4.1.1. 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.

4.1.2. Scoring Rubric Development

For the previous paper–pencil ELPAC, ETS' Assessment & Learning Technology Development (ALTD) group developed 9 rubrics for the scoring of Speaking CR task types and 10 rubrics for the scoring of Writing CRs (California Department of Education [CDE], 2019a and 2019b). For the computer-based assessment of the Summative ELPAC field test, a new Writing task type was introduced at grade one and grade two; otherwise, task types remained the same as on the paper–pencil ELPAC.

During range finding for the computer-based field test, paper–pencil ELPAC rubrics were evaluated and used for computer-based items. The rubric evaluated for the new Writing task type at grade one was the rubric used for similar tasks at grade one, and the rubric evaluated for the new Writing task type at grade two was the rubric used for the same task type at grade span three through five on the paper–pencil test (PPT).

Rubrics were edited as needed on the basis of feedback from the CDE and California educators during the range finding process for the computer-based field test. Changes from the PPT rubrics were made for clarification and to address keyboarding errors in grades three through twelve.

When reviewing student responses during the Writing range finding meetings, educators decided that keyboarding errors on the computer-based ELPAC should be treated the same as spelling errors. As a result, in each case where the rubrics had descriptors about spelling errors, keyboarding errors were added to the descriptor. For example, the highest score point for *Writing—Write About an Experience* was updated to state, "Minor errors in spelling/ keyboarding and punctuation may be present, but they do not impede meaning" (CDE, 2019b).

No substantial revisions were made that would change the similarity of how the paper–pencil responses and computer-based responses are scored. Proposed rubric revisions underwent internal ETS ALTD review and CDE review, resulting in the acceptance of rubrics for the two new Writing task types as well as minor revisions to one Speaking rubric and several Writing rubrics.

4.1.3. Range Finding

Soon after receiving Writing responses from California schools, ETS and the Sacramento County Office of Education facilitated a range finding event. The goal of the range finding was to enlist 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 prescored responses representing each score point on the rubric for each item. 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

Table 4.1 *(continuation)*

Sample Type	Purpose	Number of Sets and Samples in Sets	Configuration of Sets
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 be used 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.

4.1.4. Rater Recruitment and Certification Process

Several weeks prior to the start of the Summative ELPAC field test Writing CR scoring, 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.

4.1.5. 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.

4.1.5.1. 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.

4.1.5.2. 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.

4.1.5.2.1. 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. In the ONE system, the rubric was accessed through the [**Scoring Guide**] tab. The benchmarks, also called anchors, were accessed in ONE through the [**Benchmarks**] tab. The benchmarks had annotations associated with them to call the rater's attention to specific content in the sample responses.

4.1.5.2.2. 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.

4.1.5.2.3. 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 could review training sets and attempt 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.

4.1.5.3. Scoring Rules and Processes

For computer-based ELPAC field test scoring, approximately 10 percent of responses were double-scored as a check for accuracy. Raters were not aware when a second scoring was occurring and so did not have access to the first score.

4.1.6. Scoring Monitoring and Quality Management

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.

4.1.7. 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 indicated how many responses a rater scored during a shift.

4.2. Automated Scoring for Selected Response Items

During the field test administration, one goal was to evaluate the performance of the CR items in computer-based form compared to the CR items in the paper–pencil format. Consistency of scores and resulting outcomes are important considerations when changing mode of delivery. This was the first step toward assessing mode comparability and investigating processes to fully utilize the advantages of computer-based testing.

One element of the transition to computer-based delivery is the ability to seamlessly implement automated scoring. Data to investigate potential scoring models and algorithms needed for this process were collected during the spring 2020 Summative computer-based operational administration. This data will inform the ability to implement automated scoring for future administrations.

References

California Department of Education. (2019a). *Speaking Rubrics for the English Language Proficiency Assessments for California.* [Unpublished document]. Sacramento, CA: California Department of Education.

California Department of Education. (2019b). *Writing Rubrics for the English Language Proficiency Assessments for California.* [Unpublished document]. Sacramento, CA: California Department of Education.

Chapter 5: Analysis Plans

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

5.1. Data Collection Plan

Four test forms were included in the fall 2019 computer-based Summative ELPAC field test. Forms C1, C2, and C3 were earmarked for the mode comparability study, in which data for only existing English learner (EL) students was included (Educational Testing Service [ETS], 2020). Form F1 was a preassembled Summative ELPAC form, for which students targeted to take the Summative ELPAC—returning EL students and newcomer students within a school, regardless of their EL status at the time of the field test administration—were eligible to participate.

The mode comparability analyses were based on three versions of Summative ELPAC forms: C1, C2, and C3. Analyses to create 2020–2021 operational forms were conducted using data from the F1 field test form.

1. **Form C1** was a paper reprint of the 2018–2019 Summative ELPAC paper–pencil test that was developed according to the *Summative Assessment Test Blueprints for the ELPAC*, approved by the State Board of Education (SBE) on September 14, 2017 (California Department of Education [CDE], 2017).
2. **Form C2** was a computer-based form consisting of oral language items that were computer versions of all oral language items in the 2017–2018 Summative ELPAC as well as additional items to allow alignment to the adjusted test blueprints for the Summative ELPAC approved by the SBE on May 8, 2019 (CDE, 2019).
3. **Form C3** was a computer-based form comprising written language items that were computer renderings of all written language items in the 2017–2018 Summative ELPAC as well as additional items to allow alignment to the adjusted test blueprints for the Summative ELPAC approved by the SBE on May 8, 2019.
4. **Form F1** was a preassembled 2019–2020 computer-based Summative form aligned with the 2019–2020 adjusted Summative ELPAC blueprints. Additional oral language items were included to serve as vertical and horizontal linking items.

5.1.1. Form Assignment

The target population for Summative ELPAC was returning students who had been identified as EL students.

5.1.2. Challenges in Sample Recruitment

Local educational agencies (LEAs) were encouraged to enroll multiple schools to participate in the computer-based ELPAC field test. 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 or grade span. Eligible LEAs were asked to provide SCOE with voluntary sample N counts, by school and grade or grade span, for ETS to assign test forms. 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 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.
- Because analyses to create 2020–2021 operational forms were essential, students at more LEAs were targeted to respond to F1, which meant that students at fewer LEAs were targeted to respond to the other forms. This introduced a risk for the paper–pencil forms.
- It was difficult to monitor paper–pencil submissions, which made it difficult to know when samples obtained were smaller than anticipated for paper–pencil forms.

Table 5.1 shows the case counts targeted for the fall 2019 computer-based ELPAC field test.

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

Grade	C1–C3 and F1** Summative Grade	N C1/ C2	N C2/ C1	N C1/ C3	N C3/ C1	N F1**	Total
TK*	N/A	0	0	0	0	0	0
K	TK and K	400	400	400	400	1,200	2,400
1	TK and K	400	400	400	400	1,200	2,800
2	Grade 1	800	800	800	800	2,400	5,600
3	Grade 2	800	800	800	800	2,400	5,600
4	Grade span 3–5	275	275	275	275	800	1,900
5	Grade span 3–5	275	275	275	275	800	1,900
6	Grade span 3–5	275	275	275	275	800	1,900
7	Grade span 6–8	275	275	275	275	800	1,900
8	Grade span 6–8	275	275	275	275	800	1,900
9	Grade span 6–8	275	275	275	275	800	1,900
10	Grade span 9–12	500	500	500	500	1500	3,500
11	Grade span 9–12	500	500	500	500	1500	3,500
12	Grade span 9–12	500	500	500	500	1500	3,500
Total	N/A	5,550	5,550	5,550	5,550	16,500	38,700

*Transitional kindergarten (TK)/Kindergarten (K)

**Includes both returning students and newcomers

5.1.3. Form Assignment Principles

ETS and SCOE used a two-step process to assign samples. Sample N counts from LEAs were used for test form assignment. During this process, LEAs and schools within LEAs were informed of the number of students who were assigned for testing and the exact test forms that were assigned.

Statewide 2018–2019 Summative ELPAC data aggregated at the school level was used to inform form assignment such that forms were assigned to schools that had similar levels of achievement. Sample rosters were provided to the Test Operations Management System (TOMS) to preregister students in the system at the end of August 2019, when California Longitudinal Pupil Achievement Data System (CALPADS) data accurately reflected the grade level for students in the 2019–2020 school year. The following principles for form assignment were developed to overcome challenges related to sample recruitment:

- Multiple schools within an LEA were encouraged to participate. At least one school within a given LEA and at least one grade or grade span within the school was assigned F1 forms, which were the full-length, preassembled, computer-based ELPAC assessments. The second, third, or fourth school within an LEA may have been assigned to take the mode comparability forms C1 through C3.
- Since the C1/C3 combination for the mode comparability study did not allow schools to preview the Speaking assessment, it may have been viewed by LEAs as the less-preferable form assignment. The goal was for this condition to be assigned to only one school per LEA.
- Each school within an LEA was targeted for testing, which could have included test content from more than two grades or grade spans. For example, a school might have participated in the grade span three through five test that was administered to students in grades four, five, and six in fall 2019 and also the kindergarten test that was administered to grade one students.
- Within each school, the same test form conditions were assigned. For example, within a school, if grade one students were assigned to take the kindergarten test for the C1/C3 combination, students in grades four through six were assigned to take the grade span three through five test for the C1/C3 combination. Note that this was a TOMS design restriction.
- In the form assignment process, each LEA was grouped into northern, central, or southern districts according to geographical location. The goal of grouping was to have target numbers of students assigned each form with representative proportions for the northern, central, and southern parts of California.
- Because recruitment counts were lower than expected, the focus was on obtaining at least 1,500 students for the F1 form for each grade or grade span test before targeting students to take the mode comparability forms. Forms C2 and C3 contained potential backup items for F1, which made it desirable to have approximately 1,000 students taking C2 or C3 for each grade or grade span test.

- For LEAs and schools that signed up for participation after the recruitment window closed (approximately mid-May), a form was assigned to each LEA on a rotating basis, going by the order of F1, C1/C2, and C1/C3, with the aforementioned goal of giving priority to obtaining sufficient students for the F1 forms.

5.1.4. Student Roster Selection

Student rosters were developed for the computer-based ELPAC field test to provide structure for schools participating in the field test. Form assignment information was developed and communicated to LEAs in late May and early June and was later updated to include all LEAs and schools participating in the field test.

For each school, up to 50 percent more students per grade or grade span were included in the roster than were pledged by individual schools at each specific grade 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 CALPADS in August 2019 were used for roster selection. Previous Summative 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. 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. For these students, their previous ELPAC performance, as well as gender, home language, and other demographic information available in CALPADS, were used to evaluate if the roster of students were representative of the state EL population.

At the request of the CDE, 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 each form and the number of students included in analyses after data cleaning rules were applied.

It was especially important to select an equivalent roster of students taking the C1/C2 and C2/C1 combinations. Similarly, students taking C1/C3 and C3/C1 also needed to be as equivalent as possible, since assumptions of their equivalency needed to be made to understand the mode effect using classical statistics.

5.2. Data Analysis Plan for the Summative ELPAC

5.2.1. Data Cleaning

The field test sample was created by performing the following steps for each domain and grade or grade span:

1. Remove all test takers who are not EL students (i.e., English Language Acquisition Status is reclassified fluent English proficient, initial fluent English proficient, English only, or blank).
2. Remove all test takers with test irregularities as defined in the *Field Test Administration Manual* (CDE, 2019b).
3. Remove all test takers with fewer than at least four, three, five, and two item scores for Listening, Speaking, Reading, and Writing, respectively.
4. For mode comparability analyses, require that at least half of the items had responses in both C1 and C2, or C1 and C3, to be included.

Omitted or not-reached responses were handled in the same way in all statistical analyses (item analysis, differential item functioning [DIF], item response theory [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 item analyses were conducted for each item across all domains. However, the results for students who took the braille version of test forms were excluded from these item analyses.

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 the item difficulty indices (i.e., *p*-values) and the 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 items are also included in the classical item analyses.

5.2.2.1.1. Item Difficulty

For 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 constructed-response (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. The formula for AIS is the same as the *p*-value formula used for multiple choice items, but the x_{ij} values range from zero to the maximum possible points for the item.

$$AIS = \frac{\sum X_{ij}}{N_i} \tag{5.2}$$

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

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

$$p\text{–}value_{CR} = \frac{\sum X_{ij}}{N_i \times Max(X_i)} \tag{5.3}$$

Refer to the Alternative Text for Equation 5.3 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 polytomous 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 item response theory (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.

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 average item score (less than .25)
D	MC items with proportionally higher ability students selecting a distractor over the key
H	High average item score (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 Summative 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, ethnicity, and mode—paper-based compared to computer-based assessments—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, ethnicity, or linguistic background—when students are matched on ability, the item may be measuring something else 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 Summative ELPAC DIF procedures used were the Mantel-Haenszel (MH) procedure (1959) for MC items and the 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 on the basis of total test raw score. That is, all examinees obtaining a raw score of 10 represent one matched ability group, for example. 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:

$$\alpha_{MH} = \frac{\sum_k \frac{A_k D_k}{N_{Tk}}}{\sum_k \frac{B_k C_k}{N_{Tk}}} \tag{5.4}$$

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

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

$$\Delta_{MH} = -2.35 \log_e(\alpha_{MH}) \tag{5.5}$$

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

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 standardized mean difference (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:

$$SMD = \sum_{k=1}^{K} \frac{n_{k,\,focal}}{n_{focal}} \left(\frac{\sum_{s=0}^{S} s \cdot n_{ks,\,focal}}{n_{k,\,focal}} - \frac{\sum_{s=0}^{S} s \cdot n_{ks,reference}}{n_{k,reference}} \right) \tag{5.6}$$

Refer to the Alternative Text for Equation 5.6 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 .05 level (i.e., the *p*-value of MH_Chi_Sq > .05), or |MH D-DIF| ≤ 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 1.5.
C (moderate to large)	• MH D-DIF is significantly different from 1 at the .05 level and is at least 1.5.

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| is rounded to two decimal places before it is 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" 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 F1 form for the Summative ELPAC. Because F1 forms vary somewhat in length from the spring 2020 Summative ELPAC forms, the F1 total form durations were projected to the spring 2020 operational lengths; this was done based on the number of items in the field test and operational forms.

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 for the F1 forms, at the first, tenth, twenty-fifth, fiftieth, seventy-fifth, ninetieth, and ninety-ninth percentiles. Total test response times calculated for the 50th and 90th 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

IRT is based upon the item response function, which describes the probability of a given response as a function of a test-taker's true ability. IRT can be used to implement item calibrations, link item parameters, scale test scores across different forms or test administrations, evaluate item performance, build an item bank, and assemble test forms.

The two-parameter logistic (2PL) IRT model was used for the Summative ELPAC item calibration. Specifically, the generalized partial credit (GPC) model (Muraki, 1992) was applied to both dichotomous and polytomous items. The mathematical formula of the GPC model is the following:

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

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,

a_i is the discrimination parameter for item i,

b_i is the location parameter for item i,

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

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

5.2.6. Linking Procedures for the Summative ELPAC

As part of the Summative 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 not preferred initially to conduct this linking, because common items cannot be assumed to be equivalent across delivery modes. Two alternative methods were evaluated to provide additional sources of evidence with regard to mode comparability: equivalent groups and single group analyses.

The methodology, analyses, and results of this study are 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, 2020). 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.

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

References

California Department of Education. (2017). Summative assessment test blueprints for the English Language Proficiency Assessments for California. In *California State Board of Education September 2017 agenda; Subject English Language Proficiency Assessments for California: Approve the Revised Test Blueprints, the Revised General Performance Level Descriptors, and the Reporting Hierarchy* (pp. 9–21). Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/be/ag/ag/yr17/documents/sep17item18.doc

California Department of Education. (2019b). *Computer-based ELPAC field test administration manual.* Sacramento, CA: California Department of Education. Retrieved from https://bit.ly/3gZ0sO3

California Department of Education. (2019a). *Proposed adjustments to the test blueprints for the Summative English Language Proficiency Assessments for California*. Approved by the California State Board of Education in May 2019. Sacramento, CA: California Department of Education.

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.

Educational Testing Service. (2020). *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.

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.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2): 159–76.

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 average item score 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 the total number of students who were presented with item i.

Alternative Text for Equation 5.3

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 5.4

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 5.5

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 5.6

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 5.7

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. Summative ELPAC Results

This chapter summarizes the results of the item- and test-level statistical and psychometric analyses for the administration of the fall 2019 computer-based Summative English Language Proficiency Assessments for California (ELPAC) field test.

6.1.1. Overview

The descriptions for these analyses are provided in chapter 5. They include classical item analyses, response time analyses, test completion analyses, differential item functioning (DIF) analyses, item response theory (IRT) analyses, 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 this technical report were based on all valid student scores from the field test samples. An exception occurred in 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.

Table 6.1 through table 6.3 show the number of students tested, by grade or grade span, for each field test form (i.e., forms C1, C2, and C3 for the mode comparability analyses and form F1 for the Summative ELPAC field test analyses). As expected, the N counts for the C forms were lower than for the F1 forms. The decision was made to prioritize the latter in the field test form assignment process to ensure sufficient data for linking to the paper-based scale.

The N counts for the C1 form of the written composite for kindergarten through grade two (K–2) were very low due to the small number of Answer Books returned to ETS. Among the C forms, the percentage of students who responded to all items was also the smallest for these lower grades and C1 written forms. With the exception of C1 written forms for grades K–2, the percentage of students who answered all items varied from 64 percent to 99 percent.

The N counts presented in table 6.1 to table 6.3 may not always match those shown in other tables and appendices of this report, due to different reporting specifications requiring demographic information that may be missing from some student records.

Table 6.1 Summary of Completion of the Field Test—C Forms

Grade Level or Grade Span	Test Form	Total Number of Students Responding to 75% of Items	Percent of Students Responding to 75% of Items	Total Number of Students Responding to 90% of Items	Percent of Students Responding to 90% of Items	Total Number of Students Responding to All Items	Percent of Students Responding to All Items	Total Number of Registered Test Takers
K	C1_Oral	357	98.89	354	98.06	345	95.57	361
1	C1_Oral	436	99.32	436	99.32	418	95.22	439
2	C1_Oral	391	98.49	390	98.24	378	95.21	397
3–5	C1_Oral	755	99.60	748	98.68	730	96.31	758
6–8	C1_Oral	625	99.68	620	98.88	597	95.22	627
9–10	C1_Oral	341	99.42	322	93.88	307	89.50	343
11–12	C1_Oral	108	99.08	99	90.83	90	82.57	109
K	C1_Written	49	81.67	46	76.67	43	71.67	60
1	C1_Written	46	100	37	80.43	36	78.26	46
2	C1_Written	82	100	58	70.73	45	54.88	82
3–5	C1_Written	1,281	99.38	1,260	97.75	1,147	88.98	1,289
6–8	C1_Written	1,292	99.61	1,277	98.46	1,176	90.67	1,297
9–10	C1_Written	538	99.26	529	97.60	481	88.75	542
11–12	C1_Written	111	100	110	99.10	100	90.09	111
K	C2_Oral	353	98.60	352	98.32	352	98.32	358
1	C2_Oral	487	99.59	486	99.39	486	99.39	489
2	C2_Oral	390	99.74	388	99.23	387	98.98	391
3–5	C2_Oral	770	97.59	764	96.83	756	95.82	789
6–8	C2_Oral	722	100	713	98.75	708	98.06	722
9–10	C2_Oral	404	99.75	387	95.56	382	94.32	405
11–12	C2_Oral	125	100	121	96.80	120	96	125
K	C3_Written	1,304	99.85	1,288	98.62	1,274	97.55	1,306
1	C3_Written	1,251	99.44	1,235	98.17	1,224	97.30	1,258
2	C3_Written	527	99.81	521	98.67	511	96.78	528
3–5	C3_Written	112	100	112	100	111	99.11	112
6–8	C3_Written	1,120	99.12	1,120	99.12	1,120	99.12	1,130
9–10	C3_Written	706	98.74	706	98.74	706	98.74	715
11–12	C3_Written	630	99.68	630	99.68	630	99.68	632

Table 6.2 Summary of Completion of the Field Test—C1 Written (Reported Separately for Reading and Writing)

Grade Level or Grade Span	Test Form	Total Number of Students Responding to 75% of Items	Percent of Students Responding to 75% of Items	Total Number of Students Responding to 90% of Items	Percent of Students Responding to 90% of Items	Total Number of Students Responding to All Items	Percent of Students Responding to All Items	Total Number of Registered Test Takers
K	C1_Reading	1,071	99.44	1,071	99.44	1,060	98.42	1,077
1	C1_Reading	651	94.08	650	93.93	643	92.92	692
2	C1_Reading	634	98.14	610	94.43	540	83.59	646
K	C1_Writing	71	79.78	68	76.40	68	76.40	89
1	C1_Writing	66	85.71	66	85.71	66	85.71	77
2	C1_Writing	74	67.89	70	64.22	70	64.22	109

Table 6.3 Summary of Completion of the Field Test—F Form

Grade Level or Grade Span	Test Form	Total Number of Students Responding to 75% of Items	Percent of Students Responding to 75% of Items	Total Number of Students Responding to 90% of Items	Percent of Students Responding to 90% of Items	Total Number of Students Responding to All Items	Percent of Students Responding to All Items	Total Number of Registered Test Takers
K	F1	980	98.69	901	90.74	808	81.37	993
1	F1	728	96.94	709	94.41	666	88.68	751
2	F1	906	99.67	902	99.23	856	94.17	909
3–5	F1	1,642	100	1,635	99.57	1,576	95.98	1,642
6–8	F1	1,299	99.69	1,296	99.46	1,252	96.09	1,303
9–10	F1	678	96.17	673	95.46	625	88.65	705
11–12	F1	737	99.86	734	99.46	698	94.58	738

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 the hand-scored, constructed-response (CR) items. Appendix 6.A contains the raw score frequency distributions and summary statistics tables by form and by grade or grade span (table 6.A.1 through table 6.A.64).

The C forms appeared to be easy for the students, as more students received higher raw scores than lower scores; few students received scores of zero; and many of the forms had students receiving the maximum possible raw score. A possible explanation for this finding, particularly for C1, is that many students would have already responded to these items during the 2018–2019 operational Summative paper–pencil testing (PPT) administration. For all grades and grade spans, the mean raw scores from these forms accounted for about 55 percent to 84 percent of the total possible raw score (tables 6.A.8, 6.A.16, 6.A.20, 6.A.24, 6.A.32, 6.A.40, 6.A.48, 6.A.56, and 6.A.64). Students in kindergarten through grade two performed better on the Reading items than the Writing items (refer to table 6.A.20 and table 6.A.24).

For the F1 forms, the distribution of the scores had more students in the middle of the score range. The mean percentage correct scores for these forms was about 42 percent to 82 percent of the total possible scores (table 6.A.48, table 6.A.56, and table 6.A.64); the Writing domain tended to have the lowest mean percentage correct scores of the total possible scores.

6.1.4. Results of Classical Item Analyses

ETS psychometric and content assessment staff carefully reviewed each of the items flagged after the 2018–2019 Summative ELPAC administration. These results were summarized and submitted to the California Department of Education (CDE) 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 2018–2019 test items. Table 6.A.65 through table 6.A.68 in appendix 6.A present the *p*-value and item-total correlation information by grade or grade span as well as the number of unique items in each test.

Overall, the classical item analysis results were within acceptable ranges (i.e., *p*-values ranged from 0.25 and 0.95; item-total correlations were greater than 0.20). Across the grades, grade spans, and domains there was a wide range of item difficulty values and item-total correlations. The item difficulties ranged from 0.06 to 1.00 and item-total correlations ranged from -0.16 to 1.00. Items with item difficulty values less than 0.25 or greater than 0.95 and items with item-total correlations less than 0.20 were flagged for psychometric and content review. ETS content experts looked at each of the flagged items and collaborated with the psychometrics team to make decisions about the items.

The average item difficulties across the forms ranged from 0.41 to 0.91, with most forms having average item difficulties in the range of 0.60 to 0.80. Some forms were relatively difficult (e.g., grade span six through eight Reading in form F1) and other forms were relatively easy (e.g., grade span three through five Speaking in form C1).

Table 6.A.71 through table 6.A.75 present the item difficulties and item-total correlations by task type for each domain and item type. The mean item difficulties by task type ranged from 0.34 to 0.97; many of the task types had average item difficulties between 0.70 and

0.90. Thus, most of the task types were easier than average for the students, even though there were some task types that were difficult. Detailed results of the item analyses for each item by grade or grade span are presented in table 6.A.76 through table 6.A.89. The item statistics, including *p*-value, item-total correlation, and item type, are presented in those tables.

6.1.5. Differential Item Functioning (DIF) Results

DIF analyses were conducted for 2018–2019 ELPAC 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 6.4.

Table 6.4 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 analyses provided in appendix 6.C in form F1, no items were identified as having significant DIF for gender. For the Hispanic versus non-Hispanic analyses, two items were flagged for possible DIF. These two items are identified in table 6.C.3 of appendix 6.C. One item was flagged at grade span six through eight for Speaking and favored the Hispanic or Latino students. The content review suggested that Hispanic or Latino students *may* have performed better on this item because one important word in the prompt is similar to the corresponding Spanish word. This item was retained in the item pool for future operational use.

The other item was flagged at grade span eleven and twelve for Listening and favored the non-Hispanic or non-Latino students. The content review identified no obvious reason as to why non-Hispanic or non-Latino students performed better on this item. DIF analyses could not be performed at grade span nine and ten for Hispanic or Latino versus non-Hispanic or non-Latino comparisons in the Speaking, Reading, and Writing domains because there were fewer than 100 non-Hispanic or non-Latino students.

No items were flagged for significant DIF when comparing the female and male groups.

6.1.6. Response Time Analyses

Response times for computer-based Summative ELPAC forms were collected and are summarized in table 6.B.1 of appendix 6.B. Because the C1 form was paper-based, the table does not contain the statistics for this form. Instead, the table shows descriptive statistics of response time by grade or grade span, form, 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 was zero minutes for low-performing students in kindergarten. The maximum testing time was about seven hours for grade span six through eight students taking the F1 form. The average testing time for students in each quartile varied from about one minute to two hours. The most extreme average testing time of 0.69 minutes belonged to one student in the first quartile who took the C3 form in grade two. The average testing time for the remaining student groups was at least eight minutes. It took students more time to complete the F1 form than the C forms because the former had items across all four domains and the latter only had items for two out of the four domains. With a few exceptions (e.g., forms C3 and F1 of grade span eleven and twelve), students with higher total raw scores spent more time on the assessments than their peers with lower scores. As grades and grade spans increased, students tended to take more time to complete the test.

Note that the ELPAC is an untimed test.

6.1.7. IRT Results for the Summative ELPAC

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

1. The composite oral language scale was comprised of the Listening and Speaking domain assessments.
2. The composite written language scale was comprised of the Reading and Writing assessments.

The two-parameter logistic (2PL) model was used to calibrate dichotomous items. The generalized partial credit model was used to calibrate polytomous items. Appendix 6.D contains all the tables summarizing the IRT results for the Summative ELPAC forms.

Table 6.5 shows the number of items, score points, and students included in the IRT analyses for the Summative ELPAC forms. Because ETS prioritized assigning more students to the F1 form than the C forms, sample sizes for all grades and grade spans for the F1 form were greater than 700 students, with the exception of the written composite in grade span nine and ten. These N counts for F1 were sufficient to fit a 2PL IRT model to the corresponding field test data to investigate the IRT characteristics of the items to be used in the 2019–2020 computer-based ELPAC operational forms. Note that C2 and C3 forms did not have sufficient sample sizes to be included in the IRT calibrations (refer to table 3.A.21 through table 3.A.34 in appendix 3.A).

Table 6.5 Number of Items, Score Points, and Students for IRT Analyses of Summative ELPAC Forms

Language Composite	Form	K	Grade 1	Grade 2	Grade Span 3–5	Grade Span 6–8	Grade Span 9–10	Grade Span 11–12
Number of Oral Items	F1	32	38	41	42	43	45	45
Number of Written Items	F1	32	49	43	47	47	48	48
Maximum Score Points, Oral	F1	49	55	61	63	62	64	67
Maximum Score Points, Written	F1	42	66	58	63	63	64	64
Number of Oral Students	F1	960	791	965	1,644	1,311	700	735
Number of Written Students	F1	874	716	851	1,607	1,281	683	723
Number of Oral Items	C1	30	32	35	35	35	35	35
Number of Written Items	C1	22	27	33	32	32	32	32
Maximum Score Points, Oral	C1	41	43	49	51	51	51	51
Maximum Score Points, Written	C1	29	35	41	43	43	43	43
Number of Oral Students	C1	151,356	126,367	113,897	285,815	199,638	97,852	72,801
Number of Written Students	C1	148,388	124,736	113,587	284,284	198,807	97,017	72,601

Distributions of the IRT *a*-value parameter estimates by domain and grade or grade span for the oral composite and written composites are shown in table 6.6 and table 6.8. The discrimination estimates for F1 and C1 oral composite were within the range of .20 to 1.07, and the range for the written composites was 0.34 to 1.62.

Table 6.7 and table 6.9 show the summary statistics of the *b*-value parameter estimates for the oral and written composites. The average values ranged from -3.11 to -0.58 for the oral composites and -4.4 to +1.53 for the written composites. The *b*-parameter estimates were relatively low for many items across the grades and grade spans, for both the oral and written composites. The lowest values (i.e., easiest items) were observed for kindergarten in both the Reading and Writing domains.

Table 6.D.2 through table 6.D.8 in appendix 6.D contain parameter estimates for each item in the oral composite from kindergarten through grade span eleven and twelve. For dichotomous items, each table shows item discrimination and difficulty parameter estimates. For polytomous items, category thresholds (i.e., the *d*-parameter) are also presented.

Table 6.6 reports summary statistics of *a*-parameter estimates for the oral composite across the grades and grade spans. Almost all of the statistics were within an acceptable range with a few exceptions, such as the Listening section of the F1 form for grade spans nine and ten and eleven and twelve. However, the sample sizes for these grade spans were around 700, which was the lowest among all the samples for F1. Thus, the *a*-parameter estimates for these grade spans might be less robust than for the other samples. Between the two domains of the oral composite, Speaking tended to have higher *a*-parameter estimates than Listening.

Table 6.6 IRT *a*-values for Oral Language Skill by Grade or Grade Span for F1 and C1

Grade Level or Grade Span	Form	Domain	N Items	Mean	SD	Minimum	Maximum
K	F1	Listening	20	0.61	0.22	0.24	1.04
K	F1	Speaking	12	0.99	0.11	0.83	1.14
1	F1	Listening	26	0.59	0.21	0.09	0.98
1	F1	Speaking	12	1.05	0.20	0.73	1.37
2	F1	Listening	26	0.58	0.25	0.11	1.20
2	F1	Speaking	15	0.87	0.21	0.54	1.23
3–5	F1	Listening	27	0.34	0.15	0.09	0.59
3–5	F1	Speaking	15	0.92	0.32	0.42	1.60
6–8	F1	Listening	29	0.27	0.11	0.06	0.57
6–8	F1	Speaking	14	0.74	0.21	0.50	1.15
9–10	F1	Listening	31	0.23	0.08	0.02	0.36
9–10	F1	Speaking	14	0.59	0.14	0.40	0.84
11–12	F1	Listening	30	0.25	0.08	0.11	0.37
11–12	F1	Speaking	15	0.70	0.15	0.48	0.92

Table 6.6 *(continuation)*

Grade Level or Grade Span	Form	Domain	N Items	Mean	SD	Minimum	Maximum
K	C1	Listening	20	0.58	0.17	0.29	0.92
K	C1	Speaking	10	1.07	0.14	0.88	1.26
1	C1	Listening	22	0.66	0.19	0.26	1.08
1	C1	Speaking	10	0.99	0.28	0.73	1.63
2	C1	Listening	22	0.53	0.21	0.15	0.90
2	C1	Speaking	13	0.78	0.16	0.58	1.09
3–5	C1	Listening	22	0.37	0.11	0.19	0.59
3–5	C1	Speaking	13	0.84	0.19	0.58	1.22
6–8	C1	Listening	22	0.25	0.12	0.04	0.51
6–8	C1	Speaking	13	0.75	0.21	0.54	1.16
9–10	C1	Listening	22	0.24	0.08	0.06	0.36
9–10	C1	Speaking	13	0.68	0.13	0.43	0.86
11–12	C1	Listening	22	0.20	0.06	0.07	0.34
11–12	C1	Speaking	13	0.68	0.12	0.51	0.84

For *b*-parameter estimates of the oral domains, table 6.7 presents the summary statistics. In general, the means and minimum and maximum values of these estimates appeared to be higher for the higher grade spans. This trend was more salient for the Speaking items compared to the Listening items. For example, the mean, minimum, and maximum values of *b*-parameter estimates for Listening in grade one were smaller than their counterparts in kindergarten. These results will be reexamined when analyses are conducted on the results from the operational computer-based ELPAC data from the 2019–2020 Summative assessment.

Table 6.7 IRT *b*-values for Oral Language Skill by Grade or Grade Span for F1 and C1

Grade Level or Grade Span	Form	Domain	N Items	Mean	SD	Minimum	Maximum
K	F1	Listening	20	-2.91	1.17	-4.69	0.11
K	F1	Speaking	12	-2.69	0.78	-3.93	-1.61
1	F1	Listening	26	-3.11	1.30	-7.15	-0.65
1	F1	Speaking	12	-2.44	0.82	-3.51	-0.83
2	F1	Listening	26	-2.35	1.50	-3.91	4.00
2	F1	Speaking	15	-2.34	0.79	-3.29	-0.85
3–5	F1	Listening	27	-1.41	2.67	-3.80	9.40
3–5	F1	Speaking	15	-2.00	0.61	-2.86	-0.82
6–8	F1	Listening	29	-1.78	1.94	-5.44	3.26
6–8	F1	Speaking	14	-1.85	0.98	-3.29	-0.34

Table 6.7 *(continuation)*

Grade Level or Grade Span	Form	Domain	N Items	Mean	SD	Minimum	Maximum
9–10	F1	Listening	31	-0.58	6.54	-5.61	32.07
9–10	F1	Speaking	14	-1.63	1.10	-3.36	0.22
11–12	F1	Listening	30	-1.15	1.18	-3.85	1.04
11–12	F1	Speaking	15	-1.88	1.11	-3.87	-0.41
K	C1	Listening	20	-3.02	1.27	-4.44	0.13
K	C1	Speaking	10	-2.71	0.57	-3.38	-1.79
1	C1	Listening	22	-2.69	0.67	-3.94	-1.59
1	C1	Speaking	10	-2.39	0.75	-3.36	-0.94
2	C1	Listening	22	-2.55	1.30	-4.11	1.46
2	C1	Speaking	13	-2.23	0.55	-3.13	-1.03
3–5	C1	Listening	22	-1.84	1.09	-3.33	0.39
3–5	C1	Speaking	13	-2.15	0.61	-3.02	-0.96
6–8	C1	Listening	22	-2.20	1.65	-4.32	2.23
6–8	C1	Speaking	13	-1.87	0.84	-3.44	-0.44
9–10	C1	Listening	22	-1.94	1.27	-4.04	0.53
9–10	C1	Speaking	13	-1.57	0.82	-3.22	0.34
11–12	C1	Listening	22	-1.41	1.11	-3.21	1.44
11–12	C1	Speaking	13	-1.64	0.75	-2.95	-0.51

For the written composite, table 6.D.9 through table 6.D.15 in appendix 6 present parameter estimates for all items from kindergarten through grade span eleven and twelve. The structure of these tables is similar to table 6.D.2 through 6.D.8.

Parameter estimates at both the item level and corresponding descriptive statistics at the form and domain levels were within an acceptable range with a few exceptions such as the Reading section of the F1 form for grade spans three through five and nine and ten. The *a*-parameter estimates for F1 items varied from 0.04 to 2.27. Mean values were in the range of 0.34 for grade span six through eight for Reading to 1.61 for kindergarten Writing. The *a*-parameter estimates for C1 were slightly smaller than F1, which is not surprising since their estimates were based on less data.

For the difficulty parameter, item-level estimates and corresponding descriptive statistics were within acceptable ranges and followed expected trends. For example, the minimum, maximum, and mean *b*-parameter estimates generally increased from lower to higher grades or grade spans. There were a few exceptions to this finding. For instance, Writing items for kindergarten F1 demonstrated markedly more difficulty than Writing items for grade one. This result might be due to the small sample size available for the calibration of the F1 written composite. This finding will be further investigated when the computer-based ELPAC 2019–2020 operational data becomes available.

Table 6.8 and table 6.9 present the descriptive statistics for item parameter estimates for the Reading and Writing domains.

Table 6.8 IRT *a*-values for Written Language Domains by Grade or Grade Span for F1 and C1

Grade Level or Grade Span	Form	Domain	N Items	Mean	SD	Minimum	Maximum
K	F1	Reading	20	0.61	0.25	0.36	1.31
K	F1	Writing	12	1.61	0.42	0.96	2.27
1	F1	Reading	32	0.83	0.25	0.34	1.30
1	F1	Writing	17	0.96	0.34	0.41	1.87
2	F1	Reading	35	0.80	0.37	0.17	1.79
2	F1	Writing	8	0.80	0.18	0.44	1.00
3–5	F1	Reading	38	0.41	0.25	0.04	1.27
3–5	F1	Writing	9	0.62	0.10	0.49	0.80
6–8	F1	Reading	38	0.34	0.14	0.09	0.65
6–8	F1	Writing	9	0.65	0.17	0.40	0.91
9–10	F1	Reading	39	0.38	0.18	0.06	1.01
9–10	F1	Writing	9	0.51	0.10	0.35	0.64
11–12	F1	Reading	39	0.42	0.17	0.11	0.83
11–12	F1	Writing	9	0.53	0.10	0.30	0.64
K	C1	Reading	14	0.53	0.34	0.27	1.40
K	C1	Writing	8	1.62	0.37	1.08	2.01
1	C1	Reading	20	0.99	0.21	0.53	1.40
1	C1	Writing	7	0.70	0.16	0.51	0.91
2	C1	Reading	26	0.83	0.32	0.34	1.66
2	C1	Writing	7	0.87	0.17	0.60	1.11
3–5	C1	Reading	26	0.56	0.22	0.14	1.03
3–5	C1	Writing	6	0.66	0.08	0.59	0.79
6–8	C1	Reading	26	0.38	0.14	0.11	0.69
6–8	C1	Writing	6	0.62	0.07	0.54	0.74
9–10	C1	Reading	26	0.42	0.14	0.16	0.69
9–10	C1	Writing	6	0.47	0.08	0.39	0.59
11–12	C1	Reading	26	0.37	0.17	0.04	0.66
11–12	C1	Writing	6	0.49	0.07	0.41	0.56

Table 6.9 IRT *b*-values for Written Language Domains by Grade or Grade Span for F1 and C1

Grade Level or Grade Span	Form	Domain	N Items	Mean	SD	Minimum	Maximum
K	F1	Reading	20	-4.40	0.83	-6.07	-2.81
K	F1	Writing	12	-3.77	0.42	-4.86	-3.30
1	F1	Reading	32	-2.70	0.93	-4.86	-1.23
1	F1	Writing	17	-3.11	1.20	-6.83	-1.69
2	F1	Reading	35	-1.36	1.19	-3.23	2.47
2	F1	Writing	8	-1.86	0.26	-2.11	-1.29
3–5	F1	Reading	38	1.15	2.31	-2.29	9.23
3–5	F1	Writing	9	-1.02	0.55	-1.72	-0.15
6–8	F1	Reading	38	1.25	1.80	-3.35	7.03
6–8	F1	Writing	9	-0.75	0.58	-1.67	0.03
9–10	F1	Reading	39	0.89	1.81	-5.30	8.22
9–10	F1	Writing	9	-0.42	0.83	-1.52	0.90
11–12	F1	Reading	39	1.21	1.62	-0.98	7.72
11–12	F1	Writing	9	-0.34	0.67	-1.31	0.75
K	C1	Reading	14	-4.24	0.92	-5.65	-2.31
K	C1	Writing	8	-3.95	0.42	-4.81	-3.54
1	C1	Reading	20	-2.45	0.61	-3.51	-1.41
1	C1	Writing	7	-3.04	1.04	-4.61	-2.15
2	C1	Reading	26	-2.04	0.56	-2.99	-0.98
2	C1	Writing	7	-2.18	0.63	-3.14	-1.67
3–5	C1	Reading	26	-0.44	0.97	-2.66	1.85
3–5	C1	Writing	6	-1.08	0.54	-1.61	-0.40
6–8	C1	Reading	26	1.37	1.31	-0.46	5.25
6–8	C1	Writing	6	-0.66	0.63	-1.40	0.18
9–10	C1	Reading	26	0.87	0.88	-0.68	2.85
9–10	C1	Writing	6	-0.63	0.82	-1.53	0.58
11–12	C1	Reading	26	1.53	2.38	-0.40	12.12
11–12	C1	Writing	6	-0.37	0.72	-1.39	0.63

6.1.7.1. Horizontal Linking to Summative Paper-based Scale

As described in the Summative ELPAC mode comparability report (ETS, 2020), 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 PPT scale.

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 assessment across the grades 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 assessments and

grades and grade spans, 17 common items were removed. The written composite for grade span eleven and twelve had the largest number of linking items excluded: 6 out of 25 items were excluded.

Following exclusion of problematic linking items, the final linking item sets accounted for approximately 46 to 75 percent of the total items on the Summative ELPAC forms. These percentages were considered reasonable 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 F1 field test forms. They also had PPT item parameter estimates from the Summative ELPAC item bank.

The software STUIRT (Kim & Kolen, 2004) was used to find the Stocking-Lord constants necessary to perform the required linear transformations. Table 6.10 presents the resulting scaling constants. Because the vertical scale for the Summative ELPAC was created in 2017 using grade span three through five as the base group, the slope tends to be closer to one for grades and grade spans closest to grade span three through five. The intercepts are also closer to zero for these grades and grade spans.

The 2017 vertical scaling procedure applied scaling constants to each grade or grade span. Consequently, a scale transformation of data from the 2019 F1 form back to the 2017 scale would not be expected to have a slope equal to one and an intercept equal to zero. Instead, if it is presumed that the F1 and the 2017 test takers are similar in ability, the F1 slope and intercept should have values consistent with the original vertical scaling values. The slopes and intercepts resulting from the vertical scale can be found in table 8 of the *English Language Proficiency Assessments for California Summative Assessment Field Test Item Response Theory Calibration and Vertical Scaling Review* (ETS, 2017).

A comparison between the 2017 vertical scaling results and the 2019 field test results indicates that most differences between the slopes and the intercepts were near zero. An exception was seen in the intercepts for Written kindergarten, grade one, and grade two; the intercepts for the vertical scaling results were larger than those for the field test results by approximately 1.3 to 1.5.

Based on the overall results, it was determined by the CDE and ETS that the analysis to transition computer-based scores to the paper-based score scale produced reasonable results. Therefore, the paper-based scale could continue to be used for scoring of future operational ELPAC test forms.

Table 6.10 Linking Constants for Summative Form F1 Using Common Item Equating

Grade Level or Grade Span	Composite	Slope	Intercept
K	Oral	0.90	-1.45
1	Oral	0.84	-0.99
2	Oral	1.20	-0.09
3–5	Oral	1.24	0.35
6–8	Oral	1.96	0.78
9–10	Oral	2.16	0.27
11–12	Oral	2.22	0.15

Table 6.10 *(continuation)*

Grade Level or Grade Span	Composite	Slope	Intercept
K	Written	0.90	-1.56
1	Written	0.84	-0.18
2	Written	1.24	0.66
3–5	Written	1.11	1.13
6–8	Written	1.11	0.39
9–10	Written	1.37	1.34
11–12	Written	1.32	1.68

6.1.7.2. Investigation and Modification of the Vertical Scales

The reasonableness of the vertical scales for the Summative ELPAC scores can be evaluated by means of the test characteristic curves (TCCs) for the F1 oral and written composites, as shown in figure 6.1 and figure 6.2. Data used to create these figures can be found in appendix 6, in table 6.D.20 and table 6.D.21.

The TCCs for the lower grades tend to be on the left of those for the higher grade spans. This ordering is the most distinct for kindergarten, grade one, and grade two; and grade spans three through five and six through eight. The TCCs for grade spans six through eight, nine and ten, and eleven and twelve are much closer to each other when compared to the lower grades and grade span. This finding has been observed in previous years (CDE, 2020). These TCCs were created using limited data resulting from the field test phase; TCCs will be recreated using computer-based Summative ELPAC 2019–2020 operational data.

Figure 6.1 and figure 6.2 use the following abbreviations:

- G/GS_KN = kindergarten
- G/GS_01 = grade one
- G/GS_02 = grade two
- G/GS_03 = grade span three through five
- G/GS_06 = grade span six through eight
- G/GS_09 = grade span nine and ten
- G/GS_11 = grade span eleven and twelve.

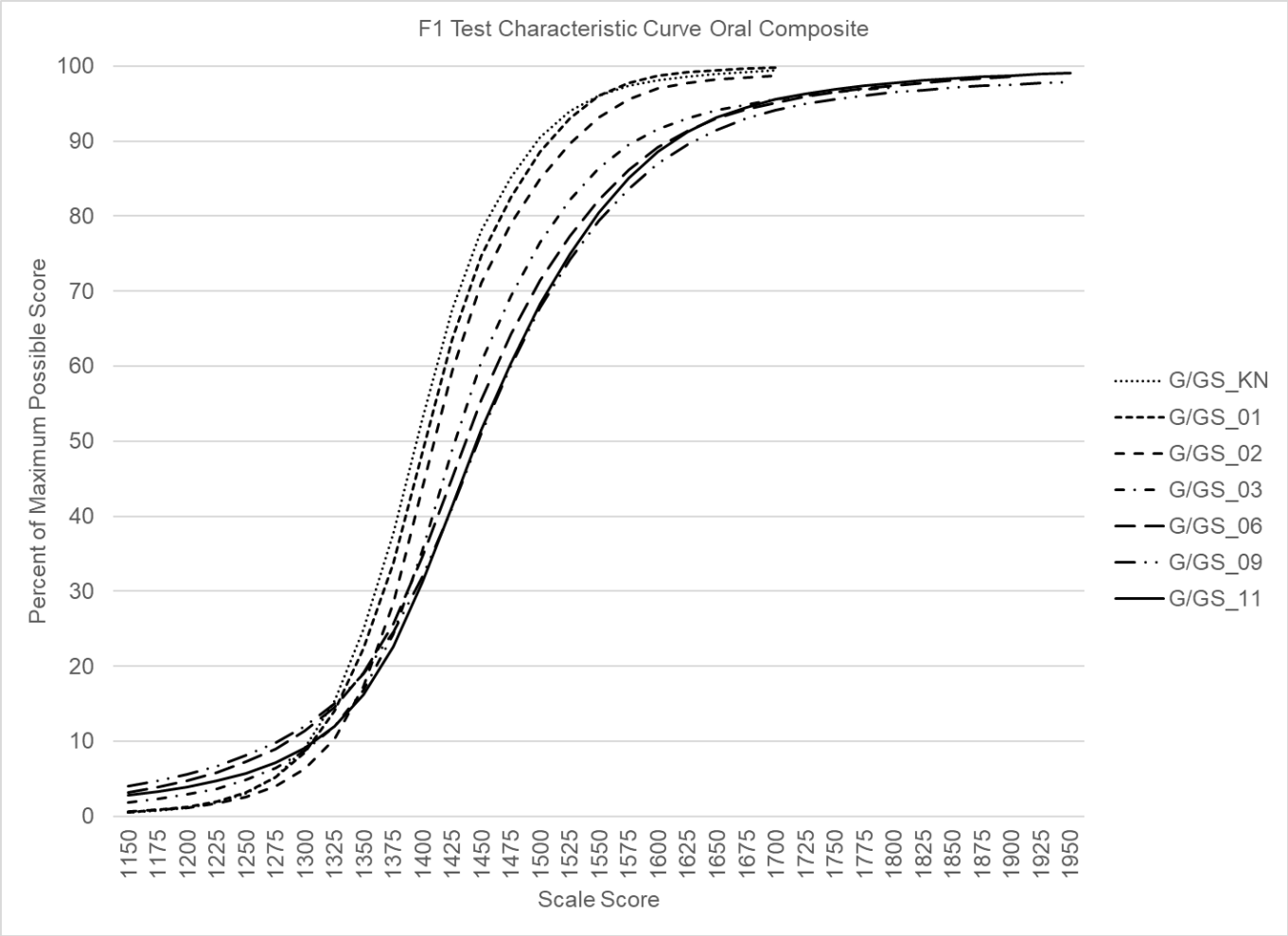

Figure 6.1 F1 oral test characteristic curves

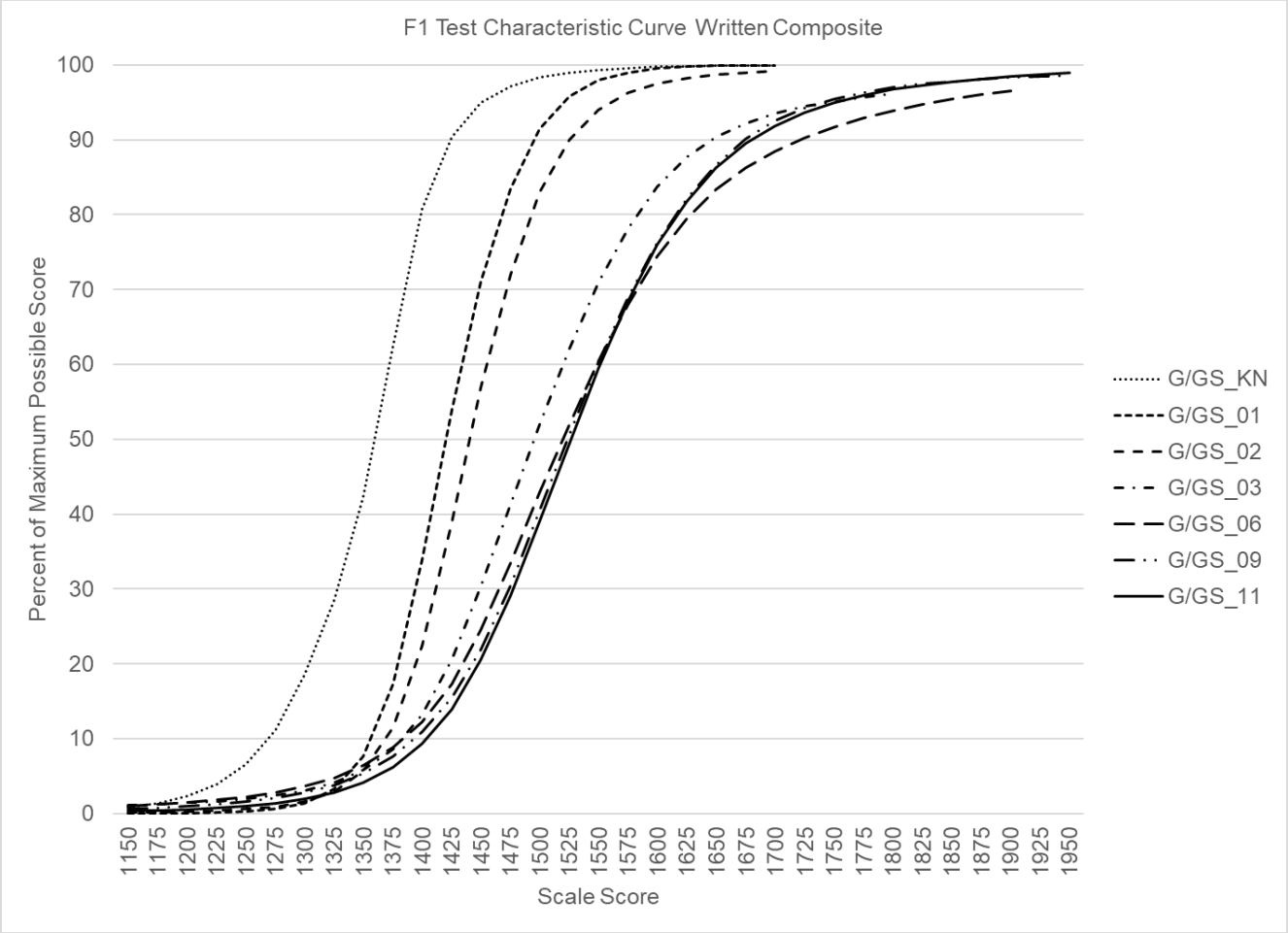

Figure 6.2 F1 written test characteristic curves

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 (or reliability) of the ratings. This interrater consistency 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 stringent criterion, which tends to decrease with increasing numbers of item score points. The fewer the item score points, the fewer degrees of freedom on which two raters can vary and the higher the percentage of exact agreement.

Table 6.11 shows, for all Writing items, the average percent exact, adjacent, and discrepant for each grade level, by the number of maximum score points. With only a few exceptions, the percent exact across all grade levels and 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]).

Table 6.11 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	97.60	2.33	0.07
K	1-pt Score Items	98.13	1.88	0.00
K	2-pt Score Items	97.07	2.79	0.14
1	All Writing Items	88.59	11.14	0.27
1	1-pt Score Items	99.03	0.97	0.00
1	2-pt Score Items	94.26	5.63	0.11
1	3-pt Score Items	77.32	22.11	0.57
2	All Writing Items	83.68	15.99	0.33
2	1-pt Score Items	99.13	0.88	0.00
2	2-pt Score Items	93.96	5.95	0.09
2	3-pt Score Items	80.06	19.49	0.45
2	4-pt Score Items	59.16	40.14	0.70
3–5	All Writing Items	72.34	27.28	0.39
3–5	2-pt Score Items	76.47	23.29	0.24
3–5	3-pt Score Items	75.84	23.89	0.27
3–5	4-pt Score Items	63.55	35.75	0.70
6–8	All Writing Items	72.74	26.89	0.38
6–8	2-pt Score Items	80.58	19.27	0.15
6–8	3-pt Score Items	70.41	29.18	0.41
6–8	4-pt Score Items	61.75	37.55	0.71
9–10	All Writing Items	72.27	27.38	0.36
9–10	2-pt Score Items	78.24	21.60	0.15
9–10	3-pt Score Items	72.24	27.49	0.27
9–10	4-pt Score Items	62.90	36.38	0.72
11–12	All Writing Items	71.02	28.55	0.43
11–12	2-pt Score Items	77.51	22.06	0.43
11–12	3-pt Score Items	67.59	32.12	0.30
11–12	4-pt Score Items	62.15	37.37	0.48

6.3. Limitations and Caveats for Data Interpretation

As discussed in chapter 3 and section *6.1 Summative 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 by several months compared to the typical ELPAC testing window. The typical Summative ELPAC testing window occurs from February 1 through May 31, 2020, while the fall 2019 field test window occurred from October 1 through November 8, 2019. The lag in testing windows likely resulted in some differences in the English proficiency of students who took the 2018–2019 Summative ELPAC forms and the field test forms. In addition, some C1 analyses—as well as the IRT calibrations for kindergarten, grade one, grade two, and grade span eleven and twelve—were based on relatively small numbers of students (refer to table 6.1 through table 6.3).

Another factor that might impact interpretation of the field test results is that forms were repeated, and there were many common items between some forms. As described earlier, C1 forms were the same as the 2018–2019 Summative ELPAC operational forms. C2 and C3 forms were the computer-based version of the 2017–2018 forms. Notably, about 70 percent of the items were common between the C1 and C2 forms. The same condition was true for the C1 and C3 forms. The percentage of common items between F1 and the C forms was smaller, but at least 20 percent of items were common.

The repeated use of forms and items might have resulted in improved student performance during the field test phase and could impact the interpretation of results. To evaluate the effect of these potential limitations, follow-up analyses will be conducted using 2019–2020 computer-based Summative ELPAC operational data to evaluate the psychometric quality of the items, test forms, and linking sets.

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. (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. (2017). English Language Proficiency Assessments for California summative assessment field test item response theory calibration and vertical scaling review. [Unpublished 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 Summative ELPAC. 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, 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. 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 dimensionality and correlational analyses. These analyses were conducted using the 2017–2018 field test data. Results of these analyses are summarized in the *ELPAC Summative Dimensionality Report*, (CDE, 2019).

Evidence collected from the 2017–2018 field test data supported the oral and written composites that are currently used to report ELPAC scores. Correlations were calculated using data from the fall 2019 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 test scores.

7.2.1. Correlations Between Domains by Administration and Delivery Mode

The data collection design of the field test, described in chapter 5, was such that comparisons could be made between student performance on the 2018–2019 paper–pencil test (PPT) operational forms and the fall 2019 C1 PPT field test forms. Additionally, comparisons could be made between results from the 2017–2018 PPT operational forms and the fall 2019 C2 and C3 computer-based field test forms.

Using student raw scores from these forms, correlation coefficients between the four domain scores were calculated by administration delivery mode. Table 7.A.1 through table 7.A.18 in appendix 7.A present the correlation coefficients for the forms across grades and grade spans. Table 7.A.1 to table 7.A.7 provide domain correlations for the PPT forms by administration. Domain correlations between the fall 2019 computer-based field test scores and the 2018–2019 PPT operational scores are presented in table 7.A.8 through table 7.A.14 for the Listening and Speaking domains, and table 7.A.15 through table 7.A.18 for the Reading and Writing domains.

The results indicate that domains were moderately related between administrations and delivery modes. Most correlations were between 0.44 and 0.76. An exception was found in the correlations between the Reading and Writing domains for PPT forms in kindergarten, grade one, and grade two, as shown in tables 7.A.1 through 7.A.3. These correlations were markedly lower, with coefficients of 0.24, 0.34, and 0.37, respectively. This is likely due to the small sample sizes (less than 85 students) for those grades, as shown in table 6.1.

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 are presented in the last column of table 7.B.1. Corresponding results, aggregated by student groups, are presented in the last column of table 7.B.2 through table 7.B.14 in appendix 7.B. The results shown in table 7.B.1 indicate that the reliability estimates for all summative test total scores across grades are within acceptable ranges, from 0.87 to 0.94. Reliability estimates for 9 out of 13 grades were 0.90 or higher.

When the analysis was conducted by student groups within each grade, the lowest reliability estimate observed was 0.83 for female students in grade four (table 7.B.6). The highest estimate was 0.95 for migrant students in grade two (table 7.B.4). 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 moderate to high, ranging from 0.59 to 0.94 across grades. Most of the estimates were in the range of 0.80 to 0.91.

Speaking and Reading domains had higher reliability estimates than the Listening and Writing domains. For the oral and written composite scores, the reliability estimates were moderate to high, ranging from 0.77 to 0.94 across grades.

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 Summative 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.21 in appendix 7.B for all grades. Tables 7.B.15 through table 7.B.17 provide classification accuracy of overall scores, while table 7.B.18 through table 7.B.20 show classification consistency of overall scores.

At each threshold, the classification at adjacent performance levels appeared to be acceptably reliable and consistent. Classification accuracy ranged from 0.85 to 0.98, while classification consistency ranged from 0.79 to 0.98, with most values at or above 0.90.

These values are similar to the classification accuracy and consistency estimates reported in the *2018–2019 Summative ELPAC Technical Report* (CDE, 2020).

Table 7.B.21 presents the classification accuracy and consistency results for both the composite and overall scores. For both classification accuracy and consistency, the grade three oral composite and the grade six written composite had the lowest reliability estimates, while the grade six overall scores had the highest reliability estimates. Classification accuracy ranged from 0.71 for grade three oral composite scores and grade six written composite scores to 0.84 for grade two overall scores. Reliability estimates for classification consistency ranged from 0.61 grade three oral composite scores and grade six written composite scores to 0.78 for grade two overall scores. These values are similar to the classification accuracy and consistency estimates reported in the *2018–2019 Summative ELPAC Technical Report* (CDE, 2020).

7.3. Evidence Based on the Relationship Between ELPAC and California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced Test Scores

The relationship between scores from different tests is examined to support evidence of convergent and divergent validity. If the assessments measure similar constructs, their scores are expected to be closely associated. If the constructs are less similar, scores should have lower correlations.

Since many students from grade spans three through five and six through eight and grade eleven who participated in the fall 2019 ELPAC field test (i.e., F1, the preassembled 2019–2020 computer-based Summative form) also took CAASPP Smarter Balanced Summative Assessments for English language arts/literacy (ELA), overall ELPAC F1 scale scores were correlated with students' corresponding overall CAASPP Smarter Balanced scores.

Table 7.1 presents the number of students with scores from both CAASPP Smarter Balanced Summative Assessments and the ELPAC F1 assessments. Grade spans three through five and six through eight resulted in high percentages of matched pairs. The matched results for grade eleven are markedly lower. This may be because many grade eleven students taking CAASPP Smarter Balanced for ELA assessments in 2018–2019 were reclassified as English proficient and, therefore, not included in the fall 2019 ELPAC field test.

Table 7.1 Students with Scores from Both the Smarter Balanced for ELA and ELPAC F1 Test Forms

Grade or Grade Span	Total	Matched Percentage	Not Matched Percentage
3–5	1,642	96.35	3.65
6–8	1,303	92.71	7.29
11	738	24.80	75.20

Correlations between the overall Summative ELPAC F1 scores and CAASPP Smarter Balanced for ELA scores for students who completed both assessments are presented in table 7.2. The scores are moderately correlated, indicating that the Summative ELPAC and CAASPP Smarter Balanced for ELA assessments measure unique aspects of the English language. Additionally, the magnitude and direction of these correlations are similar to what was found between the 2018–2019 paper-based Summative ELPAC scores and corresponding CAASPP Smarter Balanced for ELA scores for grades three, six, and eleven. The 2018–2019 correlations between these assessments for each of the grades was 0.68, 0.61, and 0.56, respectively.

Table 7.2 Correlation of Overall and Smarter Balanced for ELA Scores

Grade or Grade Span	Total	Correlation
3–5	1,582	0.71
6–8	1,208	0.64
11	183	0.58

7.4. 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 Summative ELPAC goals may include analyses of students' opportunity to become proficient English language learners and thus reclassified as fluent English proficient (RFEP), as well as potential analyses to inform instruction. Ongoing collection of evidence of the validity of these test score interpretations is of critical importance, as these scores are one set of criteria used to determine whether individual students qualify for RFEP status. Results from the Summative ELPAC may also 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 Summative 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. (2019). *ELPAC summative dimensionality report.* [Unpublished report]. Sacramento, CA: California Department of Education.

California Department of Education. (2020). *Summative 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 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 and the set of approximately 360 new 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 and 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 Summative 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 (ALTD) 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 *2.10 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 item options were parallel in structure,
- language was grade-level appropriate,
- no more than three multiple-choice 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, LEA 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.

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, a scoring model is based on scoring one item at a time by a team of raters. Items are scored by a team of 10 to 15 raters under the supervision of a scoring leader. Scoring leaders are supervised by group scoring leaders. Each group scoring leader is responsible for multiple teams in a grade or grade span.

Responses to individual prompts are assigned to teams of 10 to 15 raters. Each rater must calibrate for an item type prior to scoring any response by passing the corresponding calibration test. The team scores multiple items of a similar type per shift. Once all responses of the same type are scored, each rater must calibrate for a new item type. Each rater works independently on the rater's own device to read each student response and enter a score for each item.

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 daily 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 each scheduled scoring session. 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 Summative 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 2018–2019	**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 team 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 (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.

The standards, provided in table 8.2, were qualification expectations for the various score point ranges and the qualification standard in terms of the percent of exact agreement. This qualification set, like the validity papers discussed in the next subsection (*Monitoring Raters*), had been scored previously by scoring experts. Raters scored the papers in the same manner according to the percentage of agreements listed in table 8.2.

Table 8.2 Rater Qualification Standard for Agreement with Correct Scores

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

The qualification process 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 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; this process is referred to as back-reading.

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. Supplemental samples were added as needed. The statistics for interrater reliability for all items at all grades are presented in table 6.11. These statistics include the percentage of perfect 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:

- 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 or grade span, and scoring method. Due to small sample sizes or large numbers of blank responses, which are scored automatically by the scoring system, only items with 50 or more responses scored by two human raters were evaluated using the flagging criteria; 88 out of 109 Writing items met the criterion of 50 or more responses. Of those 88 Writing items, 12 polytomous items were flagged across all grades and grade spans. No dichotomous items were flagged.

Table 8.3 Number of Constructed-Response Items Flagged, by Grade, Fall 2019 Computer-based Summative 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	12	0.0%
Human to Human	Writing	1	0	0	0	12	0.0%
Human to Human	Writing	2	1	0	1	10	10.0%
Human to Human	Writing	3–5	1	N/A	1	21	4.8%
Human to Human	Writing	6–8	4	N/A	4	21	19.0%
Human to Human	Writing	9–10	3	N/A	3	21	14.3%
Human to Human	Writing	11–12	3	N/A	3	21	14.3%

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.

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.

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*, 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 more than 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.

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 *Directions for Administration* 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 operational administration in February 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 very helpful or 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.

Twenty-eight percent of survey participants responded that the test directions were very clear and 56 percent responded that they were clear. From the 15 percent of respondents who indicated the test directions were somewhat clear and the 1 percent of respondents who indicated the directions were not clear, some respondents noted the directions were too wordy and lengthy, some noted students did not know when to select the [**Next**] button, and some noted the 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 Writing domains. About 41 percent of the respondents indicated students experienced no difficulty with typing responses in the Writing domain. Of the respondents who indicated students “always” or “sometimes” had difficulties with typing responses, the highest percentages were at 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.

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).

Reference

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.

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 Summative 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* (2019b) was created that covered guidelines for the use of accessibility resources. It was sent to the California State Board of Education as part of the June memorandum. The adoption of this document 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 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 *Directions for Administration*—**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. Post-Test 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:

- *DFAs* and the *Test Administration Manual* were updated with more concise wording and less repetition.
- The TRCS, training tests, and practice tests were promoted to help LEAs and students prepare for the operational assessment.
- The availability of the DEI demonstration video in Moodle was promoted.

- The use of universal tools, designated supports, and accommodations was clarified by providing information about and promoting the Student Accessibility Resources web page on the ELPAC website, at https://www.elpac.org/test-administration/accessibility-resources/.
- 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. These are posted on the ELPAC Student Accessibility Resources web page at https://www.elpac.org/test-administration/accessibility-resources/. 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

Recruitment, training, and communication will be focal points moving forward as ETS continues work on the computer-based Summative 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.

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 Summative ELPAC.

10.3. Human Scoring

Ten percent of responses are scored twice (i.e., “read behind”) to check agreement among raters. Second readings are scored independently from the first reading. Only scorable responses are selected for second readings. Nonscorable (i.e., condition code) responses are not eligible for second readings and so are not included in the calculation of interrater reliability. Second reading scores are used only for statistical analysis to obtain interrater reliability. They are not included in the calculation of the final item score.

10.4. Psychometric Analysis

As the computer-based Summative ELPAC transitions from a field test to operational administrations beginning in spring 2020, the PAR 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.