English Language Proficiency Assessments for California

CONSIDERATIONS IN THE TRANSITION OF THE ENGLISH LANGUAGE PROFICIENCY ASSESSMENTS FOR CALIFORNIA (ELPAC) PAPER-PENCIL TESTS TO COMPUTER-BASED ASSESSMENTS

Version 4

April 7, 2017

Prepared by:

Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541

TABLE OF CONTENTS

TABLE OF TABLES

1. INTRODUCTION

The use of computer technology continues to expand in educational settings, for both instruction and assessment. Computers[1] are commonly used during instruction, and students use various computer programs for schoolwork inside and outside of the classroom. Since 2014, California students in grades 3 through 8 and grade 11 have also been taking statewide summative assessments in English language arts (ELA) and mathematics on computer, using the Smarter Balanced Summative Assessments (SBAC) for accountability purposes. In 2014, California established the California Assessments of Student Performance and Progress (CAASPP) assessment system where all content-area assessments are delivered on computer. A web-based data management system, the California Longitudinal Pupil Achievement Data System (CALPADS) which houses student demographic and achievement data through their kindergarten through 12th grade (K–12) career, has also been deployed, establishing consistent centralized data for the state and for local educational agencies (LEAs). As a result, schools are increasingly well equipped to administer and use the results from computer-based assessments (CBAs), and students in grades 3 and above are becoming increasingly familiar with CBAs.

While paper-pencil tests (PPTs) can remain valid for appropriate uses, a substantial body of theoretical and empirical literature in the field of language testing research has documented the advantages of CBAs over PPTs to assess English language proficiency (ELP) (e.g., Alderson & Huhta, 2005; Bachman, 2000; Chalhoub-Deville, 2001; Fulcher, 2003; Hauck, Wolf, & Mislevy, 2016; Roever, 2001). The advantages largely lie in the enhanced efficiency of a CBA's standardized administration, faster turn-around of scoring and reporting, centralized data management, and opportunities to better measure the ELP construct. Large-scale, standardized ELP PPTs have limitations in simulating authentic language-use contexts. Computer technology enables test developers to include more contextualized and interactive contexts in language assessment tasks. These improvements to task design can increase students' engagement in assessments and capture more accurate evidence about students' ELP.

In order to make an informed decision on transitioning the English Language Proficiency Assessments for California (ELPAC) to a CBA, the California Department of Education (CDE) requested that the Educational Testing Service (ETS) examine recent research findings and current practices on CBA, including the CAASPP system, to make practical

[1] In this report, the term *computer* is used to encompass a range of devices including desktop computers, laptop computers, and tablets, which are currently used in K-12 academic contexts for instruction and assessment.

recommendations and solutions for creating the ELPAC CBA. The purpose of this report is to provide the CDE with useful information and recommendations regarding considerations and issues involved in a potential move of the ELPAC from a PPT to a CBA.

ETS formed the ELPAC CBA study team, consisting of selected experts from the ETS divisions of Research, Assessment Development, Statistical Analysis, Information Technology, and Program Management. The team members, all of whom have extensive research or development experience in CBA, engaged in a series of in-depth discussions to develop this document. Accordingly, this document is organized with multiple sections contributed by relevant experts from specific areas. A brief overview of each section is provided below.

- Section 2 presents general, high-level recommendations concerning the appropriateness of a CBA model for the ELPAC Initial Assessment (IA) and for the ELPAC Summative Assessment (SA).
- Section 3 provides a research background for the ELPAC's transition to CBA, including:
    - A review of research and practice related to general considerations and validity considerations in CBAs assessing ELP
    - A discussion of accessibility and accommodations considerations for the ELPAC CBA
    - A review of automated scoring capabilities for spoken and written responses from kindergarten through 12th grade (K–12) English learners (ELs), as applicable to the ELPAC CBA
- Section 4 provides an outline of possible task types and a proposed test design for the ELPAC CBA including an analysis of existing and potential new task types by domain (Listening, Speaking, Reading, and Writing), as well as a draft of a proposed test blueprint for the ELPAC CBA, and estimates regarding the number and types of items that might need to be developed to support the transition to CBA.
- Section 5 discusses psychometric and other measurement considerations related to the transition to the ELPAC CBA and provides recommendations regarding psychometric work to be done based on the information contained in Section 3 and Section 4.
- Section 6 provides an outline of systems and Information Technology (IT) work needed to support the transition to CBA and provides a high-level systems solution

for the ELPAC CBA, showing what systems would be needed and how they would work together to support the ELPAC CBA.

- Section 7 presents the program management approach to be followed in ensuring the success of the transition of the ELPAC to CBA, including a high-level timeline consisting of key tasks and milestones.

Each section follows a similar overall structure: first, there is an introduction to the section, which provides a general discussion of considerations relevant to the transition to CBA; then, specific recommendations are provided, with explanation or rationales based on prior literature and/or ETS's experience on other assessment programs, as appropriate. Within this general structure, the organization of sections varies to some degree based on the information to be communicated. The document also contains appendices, providing additional information to support discussion and recommendations in some sections.

There are, of course, a large number of decisions and alternative paths that can be taken in such a large undertaking of transitioning ELPAC to CBA. The approach that ETS has taken is to offer what we consider to be one reasonable path based on a review of the research literature and existing practice, our experience in the development of CBA (for K–12 ELP assessments and for related assessments), and our understanding of the CDE's values and priorities for the ELPAC. The report is intended to support the following goals for the development and deployment of the ELPAC CBA.

- The ELPAC assessments will remain valid, fair, and technically sound as CBAs.
- The CBA will take appropriate advantage of the computer platform for improvement in coverage of the standards, improvements in student engagement, and other available improvements.
- The transition to CBA will be efficiently managed, making good use of time, budget, and resources, including the utilization of the existing CAASPP platform and features as appropriate.
- The planning and execution of the transition to CBA will be conducted with strong CDE and stakeholder engagement, ensuring that informed decisions are made.

While ETS hopes this document contains information that will be of substantial use to the CDE in moving forward with the work of transitioning the ELPAC to a CBA format, we note that this document presents considerations and issues at a relatively high level intended to inform conceptual planning and provide a basis for further discussions. Details of the proposed test design for the ELPAC CBA and other features of this report are not intended to be sufficient to serve as the basis of operational work without further

analysis and documentation. While the timeline and the high-level information related to estimated costs are intended to be sufficient to inform CDE planning, they do not constitute any commitment to perform this work.

In addition, it is worth adding a word of caution about what advantages can and cannot be gained by transitioning the ELPAC to a CBA format. As detailed in the following sections, there are important advantages to a CBA. At the same time, it should be noted that transitioning the ELPAC to a CBA format will not significantly expand the test's scope by, for example, providing more detailed diagnostic information within the same testing window or providing information about student abilities unrelated to English language proficiency, such as disability evaluation and diagnosis. Rather, the CBA ELPAC will continue to serve the same functions as the PPT ELPAC—that is, providing information about student English language proficiency (ELP) based on standards.

2. GENERAL RECOMMENDATIONS FOR THE TRANSITION OF THE INITIAL ASSESSMENT AND THE SUMMATIVE ASSESSMENT TO CBA

A fundamental consideration to be made in the planning of the transition of the ELPAC to a CBA is the different status of students who take the ELPAC IA and SA. The IA and the SA have distinct purposes and specific target populations.

The IA is used to identify students who should be classified as ELs and provided with specific instructional support. Since the target population for the IA is newly arrived students who are entering the district and school for the first time, a substantial majority of IA test takers are young learners, including pre-kindergarten (pre-K) or transitional kindergarten students. IA test takers at the upper grades are far fewer in number compared to those in kindergarten and grade 1 (K–1), and they tend to be newly arrived students from outside the U.S. It is also important to note that students take the IA only once, when they first arrive in school. As the IA is given to students within 30 days of their arrival (or even before they enter the school in some cases), there is an extremely limited opportunity to provide appropriate instruction in advance to prepare students to take the test or to identify students who may need additional support to take the assessment.

The SA, in contrast, is used to fulfill the accountability requirements of measuring ELs' annual progress toward, and attainment of, ELP. The SA is administered toward the end of the school year when students typically have had several months of instruction. ELs also typically gain familiarity with the SA through repeated exposure, as they must take it annually until they meet the criteria to exit from EL services. Teachers and school administrators tend to know students taking the SA well, which means they have ample time to ensure that all students are prepared for the test, have computer skills sufficient to navigate a CBA format, have familiarity with the interface and navigation features of the CBA, and to ensure that students who may be in need of accommodations have been identified and provided for well in advance of testing.

These factors lead ETS to make the following general recommendations regarding the appropriateness of a CBA IA.

- The administration mode for the ELPAC IA should retain the PPT approach throughout the transition instead of launching a CBA IA and SA simultaneously.
    - IA test takers may not have previous formal schooling and assessment experience (e.g., students entering kindergarten), or they may have various kinds of formal schooling experience outside the United States. This means that IA test takers may have a wide range of computer familiarity.

  - The PPT approach reduces the risk of IA test takers having difficulty in demonstrating their abilities as a result of their unfamiliarity with computers in general and/or CBA.
  - Accessibility and accommodations for the ELPAC IA could follow what is available for the existing ELPAC PPT. Additionally, if desired, part of the accessibility documentation for the ELPAC could include a crosswalk that links the available PPT supports with parallel supports available in a CBA (for an example of this, see WIDA, 2015a).
  - As a practical matter, keeping the IA as a PPT will significantly reduce the expense of the overall transition to the CBA. For example, no CBA IA field test will be required, and the existing forms of the PPT IA can continue to be used when the CBA SA is administered.

- The possible later development of a CBA IA should be considered after the successful implementation of the CBA SA. The CBA IA development should consider the possible use of digital devices, including tablet computers and touch-screen computers.
  - The CBA SA experience will provide useful information regarding the ELPAC CBA format, features, and students' interactions with them. This staged-development approach will make it possible to ensure that a sound IA can be developed in a cost-effective fashion.
  - Given the wide range of variability of IA test takers' computer familiarity and the lack of time for test preparation for these students, tablets and touch-screen computers (instead of those with a keyboard and a mouse) may mitigate some computer skill issues. Wider access to these devices may be available and more affordable in the longer term.
- All data management for the PPT IA should be done using the same systems as the CBA SA.
  - Although the IA is delivered on paper, ensuring that there is one data management system for the ELPAC program (including student responses and scores) will reduce complexity and potential confusion for users in California schools. This will also ensure that all ELPAC data can be accessed easily.
  - Use of a data portal for the IA would provide local educational agency (LEA) staff with a tool for data entry and accurate derived score conversion.

With respect to the SA, our general recommendation is to use a CBA delivery and response modes for all domains at all grades except those in which there is a compelling construct and/or developmental motivation for retaining a PPT approach.

The following two tables provide details related to these general recommendations. Table 2.1 presents information on the PPT model, which is being employed in the current development effort for both the IA and the SA, and which is recommended for continued use in the IA. Table 2.2 summarizes the general recommendation for the CBA SA.

Both tables follow the same structure and are organized by two grade spans kindergarten through grade 2 (K–2) and third grade through 12th grade (3–12). Within each grade span, the following information is detailed for each of the four domains (Listening, Reading, Speaking, and Writing):

- The **delivery mode**, or the format in which the student receives test content (directions, stimuli, test questions, and prompts)
- The **response mode**, or the format in which the student provides the answer or response to each test item
- The **response capture**, or the technology by which the student's response moves from the student to the appropriate systems for scoring
- The **response format**, or the type of response which the student is asked to give, including the following:
    - Selected response (SR), which is the traditional multiple-choice format
    - Technology-enhanced Items (TEI), as described in detail in Section 4.1, in which the student responds by some means other than traditional multiple choice (e.g., by clicking on part of a picture or by marking cells in a table grid)
    - Constructed response (CR), in which the student produces an original spoken or written response
- The **administration model**, or the organizational structure within which a Test Examiner interacts with either a single student, a small group of students, or a class-sized group of students as the test is administered
- The **scoring approach,** which indicates how item-level scores are generated for each student response. For both SR items and TEIs, item-level scores are generated mechanically as the student's response is compared to either the correct response or to a set of possible correct responses. For CR items, the student's

unique response must be scored either via a judgment made by a trained human or by an artificial intelligence (AI) scoring model (see section 3.3 for more details).

Table 2.1: Overview of Current PPT Initial Assessment and PPT Summative Assessment

Grades Domains	Delivery Mode	Response Mode	Response Capture	Response Format	Administration Model	Scoring Approach
K–2						
Listening	PPT + Read-Aloud Audio	PPT	PPT	SR	K–1: one-on-one Grade 2: small group	IA: Locally scored (scores entered into local scoring tool) SA: Machine-scored (paper answer books are scanned)
Reading	PPT	PPT	PPT	SR	K–1: one-on-one Grade 2: small group	IA: Locally scored (scores entered into local scoring tool) SA: Machine-scored (paper answer books are scanned)
Writing	PPT	PPT (handwriting)	PPT (handwriting)	CR	K–1: one-on-one Grade 2: small group	IA: Locally human-scored (scores entered into local scoring tool) SA: Human-scored via distributed network (paper answer books are scanned)
Speaking	PPT	Spoken Aloud	None	CR	One-on-one	IA and SA: Locally human-scored “in the moment” IA: Scores entered into local scoring tool SA: Scores gridded onto answer book

Table 2.1 (Continued)

Grades Domains	Delivery Mode	Response Mode	Response Capture	Response Format	Administration Model	Scoring Approach
3–12						
Listening	Paper + Recorded Audio	PPT	PPT	Selected Response (SR)	Group	IA: Locally scored (scores entered into local scoring tool) SA: Machine-scored (paper answer books are scanned)
Reading	PPT	PPT	PPT	SR	Group	IA: Locally scored (scores entered into local scoring tool) SA: Machine-scored (paper answer books are scanned)
Writing	PPT	PPT (handwriting)	PPT (handwriting)	Constructed Response (CR)	Group	IA: Locally human-scored (scores entered into local scoring tool) SA: Human-scored via distributed network (paper answer books are scanned)
Speaking	PPT	Spoken Aloud	None	CR	One-on-one	IA and SA: Locally human-scored “in the moment” IA: Scores entered into local scoring tool SA: Scores gridded onto answer book

Table 2.2: Overview of Recommendations for CBA Summative Assessment

Grades Domains	Delivery Mode	Response Mode	Response Capture	Response Format	Administration Model	Scoring Approach
K–2						
Listening	CBA	CBA	Test Delivery System (TDS)	Selected Response (SR) Technology-enhanced Item (TEI)	K–1: one-on-one Grade 2: small group	Machine-scored
Reading	CBA	CBA	TDS	SR TEI	K–1: one-on-one Grade 2: small group	Machine-scored
Writing	PPT	PPT (handwriting)	PPT (handwriting)	Constructed Response (CR)	K–1: one-on-one Grade 2: small group	Human-scored via distributed network
Speaking	CBA	CBA (digital voice capture)	TDS (audio responses sent downstream with test results)	CR	One-on-one	Human raters (via distributed network) AND/OR AI scoring
3–12						
Listening	CBA	CBA	TDS	SR TEI	Group	Machine scored
Reading	CBA	CBA	TDS	SR TEI	Group	Machine scored
Writing	CBA	CBA (keyboarding) AND/OR PPT (handwriting)	TDS AND/OR PPT (handwriting)	CR	Group	Human raters (via distributed network) AND/OR AI scoring
Speaking	CBA	CBA (digital voice capture)	TDS (audio responses sent downstream with test results)	CR	Group	Human raters (via distributed network) AND/OR AI scoring

Considerable differences between the CBA and the PPT models can be seen in comparing the Delivery Mode, Response Mode, and Response Capture columns. In the CBA model, each of these functions occurs within the computer environment, while in the PPT model each occurs on paper. One exception is the recommendation for K–2 Writing. As detailed in Section 4.2, ETS is recommending that the Writing domain for the SA remain in PPT format for the K–2 grade span, even as the rest of the SA transitions to a CBA format.

The Response Mode and Response Capture columns reflect a notable change in how the CBA will capture students' spoken and written responses. For Speaking tasks at all grades, the CBA will employ digital voice capture technology in which students speak into a microphone and their response is recorded digitally. For Writing tasks at grades 3–12, students will keyboard their responses directly into the test delivery system. As discussed in Section 3.2, ETS is also recommending that a handwriting option be made available in the Writing domain for some students in grades 3–12. To avoid requiring local scoring of responses, these responses will be processed for scoring as in the current PPT SA.

As indicated in the Response Format column, the CBA contains both the SR and CR response formats included in the PPT while also adding TEI, a response format that is made possible by the computer delivery model.

While the Administration Model appears the same in the CBA and PPT charts in terms of the examiner-to-student ratio (one-on-one, small group, or classroom-sized group), the interactions between the student and the Test Examiner may change in important ways in the CBA model. At K–1, the Test Examiner will likely play a significant role by supporting students to make sure they understand directions, and also helping students indicate their answers and interact with the test-delivery system. Small-scale tryouts, such as cognitive laboratories, will play a crucial role in determining exactly how these interactions are structured, and how the test-delivery system must be designed to support them. At grade two, the Test Examiner will likely play only a supporting role, with the student responsible for interacting with the test-delivery system and indicating responses; however, the small group size is recommended to allow the Test Examiner to be active and available in providing support.

At grades 3–12, a significant change for the CBA is that the Speaking tasks can be amenable to group administration. Rather than requiring Test Examiners to individually administer each Speaking prompt and score the student response "in the moment," the CBA will allow the student to interact directly with the test delivery system with only minimal support needed from the Test Examiner. Depending on the facilities available in a given school, it may be advisable that group sizes for the Speaking administration be

somewhat smaller than for Listening, Reading, and Writing. For example, there will need to be sufficient space between students to ensure that the recordings of their responses are of sufficient audio quality. However, the ability for Speaking to be administered in a group setting at grades 3–12 should provide a significant administrative relief over the individually administered PPT model.

As noted above, the CBA calls for students' spoken responses to be recorded digitally and for students' written responses at grades 3–12 to be keyboarded. This change in Response Mode and Response Capture also has important impacts on the Scoring Approach. For CBA spoken responses, the Test Examiner will not be responsible for scoring student responses "in the moment." Instead, the digitally recorded student responses will be sent downstream in the test-delivery system along with the test results and will be routed to human raters (or, potentially, to an AI scoring system) for item-level scoring via a distributed scoring network. The impacts of this transition of Speaking scoring responsibility from Test Examiners to a distributed network are discussed in Section 7. The keyboarding of student responses at grades 3–12 has less direct impact on local examiners, but it does have the benefit of making AI scoring possible, as AI scoring requires keyboarded responses rather than handwritten responses.

Not included in Table 2.2 is an explanation of how testing sessions should be organized in the CBA. To ensure that CBA testing sessions are standardized with consideration given to CBA-specific features (e.g., using headsets, speaking into a computer), guidelines about administration conditions should be provided to LEAs. These guidelines, which can be embedded in the Test Examiner's manual, should describe appropriate conditions regarding testing location, seating arrangements (e.g., students seated far enough apart not to interfere with each other's testing experience), number of students to be tested at the same time, and break time. The guidelines should also describe the types of flexibility that LEAs have, given the variability of resources across LEAs. For instance, the CBA delivery system should be structured in such a manner that schools can choose to administer the domains in whatever order meets their administrative needs, and it should be possible to administer each domain on a separate day or multiple domains on the same day, as long as the plan for administration does not lead to students being overly fatigued.

The subsequent sections of this document detail the recommendations for the transition of the SA to a CBA, accompanied by specific background and rationale.

3. RESEARCH BACKGROUND ON CBA, ACCESSIBILITY, ACCOMMODATIONS, AND AUTOMATED SCORING

3.1. GENERAL VALIDITY CONSIDERATIONS IN CBAS OF ENGLISH LANGUAGE PROFICIENCY

In planning and managing the transition of the ELPAC to a CBA format, it is of critical importance to identify those CBA features which can enhance the quality of measurement. It is of equal importance to identify potential threats to the validity of score interpretations and inferences about the ELP for ELs.

This subsection describes general validity issues to consider in transitioning to CBA ELP assessments based on a review of prior literature and of current CBA practices. It includes practical recommendations for the development of the ELPAC CBA in order to sustain and even augment the validity argument that the ELPAC adequately measures ELs' English language abilities based on the California English Language Development Standards: Kindergarten Through Grade 12 (2014; hereafter, California ELD Standards). This subsection focuses on two major validity issues: (1) construct representation in CBA, and (2) students' computer familiarity.

In the area of computer familiarity, a CBA should be designed in such a way that any limitations in students' computer proficiency do not create systematic construct-irrelevant variance in measuring students' English language abilities. That is, students' abilities to demonstrate their knowledge and skills should not be inadvertently impeded by a lack of familiarity with computers. Moreover, the ELPAC CBA should be developed, to the extent practical, in a manner that utilizes current technology to allow students to engage in more authentic language-use tasks than is possible on a traditional PPT. Authentic language-use tasks can be expected to lead to improvements in the assessment of students' English-language skills (i.e., the target construct) and a stronger validity argument for the ELPAC.

The recommendations for the development of an ELPAC CBA in this subsection are based on a review of the recent literature and practice in the following areas:

- The mode effect (i.e., computer-based vs. paper-based), mainly for K–12 students
- Computer interface features in language assessments
- Field test results of several CBAs used for K–12 students, including the National Assessment of Educational Progress (NAEP) digitally based assessment, Partnership for Assessment of Readiness for College and Careers (PARCC), Smarter Balanced Assessment Consortium (SBAC) student assessments, World-class Instructional Design and Assessment (WIDA)'s ACCESS 2.0 for ELLs, and the

English Language Proficiency Assessment for the 21st Century (ELPA21)'s ELP assessments

The findings from empirical research on computer familiarity and mode effect are briefly summarized in Appendix A. The main computer interface features employed in recent computer-based ELP assessments for K–12 ELs are also presented in Appendix B. Additional materials reviewed included the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; hereafter, *AERA/APA/NCME Standards*) and the Assessment Peer Review Guidance (U.S. Department of Education, 2015; hereafter, *Peer Review Guidance*), which were used to identify key validity issues and evidence supporting the appropriate use of technology in assessing students' ELP. A rationale for each recommendation with relevant literature is also presented.

Recommendations Concerning Task Design and Administration to Mitigate Computer Familiarity Issues

- The task design effort should reflect best practices for minimizing the impact of computer maneuvering skills. Excessive demands for computer maneuvering skills should be avoided in designing the CBA tasks.
    - The previous literature indicates that students' computer proficiency impacts their performance on CBAs of content-area and ELP assessments (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008; ELPA21, 2015; Horkay, Bennett, Allen, Kaplan, & Yan, 2006; Ling & Bridgeman, 2013; Odo, 2012; Tate, Warschauer, & Abedi, 2016; White, Kim, Chen, & Liu, 2015). That is, students who had more experience with computers outperformed students who had less experience on CBA.
- Tasks for K–2, in particular, should be designed to minimize the demand for computer maneuvering skills that these young students may not commonly possess, including scrolling, paging up/down, and using a keyboard/mouse for navigation. While the ELPAC CBA may include features of the CAASPP interface so as to maintain a uniform interface system for ELs who take both the ELPAC and CAASPP assessments, simpler interface features should be considered for K–2 ELs who would take the ELPAC assessment only.[2]

[2] We also note that if consideration of the needs of the ELPAC should provide opportunities to adjust the CAASPP delivery system in a manner that would be beneficial to all students, those opportunities should be considered as part of the ongoing improvement of the CAASSP system.

  - Special attention should be paid to K–2 students, who are likely to have less experience with CBA than students at higher grades. While young students adapt readily to technology, it is important to consider various background characteristics of young students whose computer use outside of school may be limited. Students at these grade levels are also still developing the fine motor skills needed to use a mouse and keyboard (Choi & Tinkler, 2002; Wolf, Guzman-Orth, & Wain, 2014).

- An explicit and systematic process for test familiarization, particularly for the CBA interface features, should be developed for students. An interactive test demo with sample practice tasks should be made available for students to familiarize themselves with the CBA task types (e.g., TEI features), directions, and key features (e.g., use of CBA interface features, accessibility tools, and equipment such as headphones/ear buds and microphones). Students should have ample opportunity to take this interactive demo at a time comfortably in advance of the test administration. An explicit guide for test coordinators to implement the interactive demos should be created, and practice time for students should be put in place.
  - Prior field test results of an ELP CBA (e.g., WIDA, ELPA21) explicitly underscore the importance of students' having hands-on experience with interactive demos of CBA tasks (Mitchell, 2015; ELPA21, 2015). Equal familiarity with the features of a CBA format is essential to ensuring that all students have a fair opportunity to demonstrate their knowledge and skills.

- Interactive Test Examiner guides and training materials should be made available for Test Examiners and educators to familiarize themselves with the CBA system, the task types, and their roles and responsibilities in administering the ELPAC. The training materials should include standardized guidelines about how to actively facilitate students' engagement in CBA (e.g., appropriate prompting, monitoring students' use of general computer features and TEI features for task completion).
  - Students may require assistance from Test Examiners both to interact with some task types (particularly in K–2) and in case of any technical difficulties. Test Examiners and educators must have ample time to familiarize themselves with CBA features using the interactive demos and training materials (ELPA21, 2015; SBAC, 2014).

- Appropriate guidelines for student-to-Test Examiner ratios, such as those shown in Table 3.1 below, should be implemented for the CBA in order to allow for close monitoring of any technical issues or other issues students may have during administration. (Note that ratios for students with individualized education programs

[IEPs] or with any specific needs should be determined by accommodation guidelines.)

- As emphasized earlier, younger students need individual attention due to their likely lack of experience with a CBA and their developmental progression (e.g., attention span; Bailey, 2008; Bailey, Heritage, & Butler, 2014; McKay, 2006; Wolf & Butler, in press; Choi & Tinkler, 2002; White et al., 2015).

Table 3.1: Suggested Student-Test Examiner Ratios by Grade Span

Grade(s)	Ratio
K–1	One-on-one
2	1–10 students per Test Examiner
3–12	1–15 students per Test Examiner

Recommendations to Mitigate Computer Familiarity Issues by Domain

Listening

- An option to listen to the Listening stimuli more than once with clear directions provided to students should be considered. The decision regarding allowing listening more than once should be made prior to finalizing the CBA format design and should be based on empirical data (e.g., prototyping or pilot-testing data) and input from relevant experts, including experienced educators, test developers, and language testing experts.
  - For example, PPTs tend to enable students to review a set of items associated with a stimulus (i.e., a Listening passage) prior to listening, allowing students to listen for specific information and purposes. As CBAs generally present items on the screen one at a time, students are not informed of the specific information to listen for in a given stimulus, thus increasing the burden on their working memory.

Speaking

- Tasks should be designed to provide clear visual contexts in order to lessen the contextual inauthenticity of students' speaking into a computer and/or microphone.

- Students may not be accustomed to speaking into a computer, as prior Speaking tests were delivered by a human interlocutor (Douglas, 2013). However, technology can generate simulated situations where students can employ the target language with the aid of multimedia and visuals (Ginther, 2002). This can improve authenticity and increase students' engagement in the ELPAC CBA Speaking tasks.

Reading

- Scrolling and paging Reading passages should be avoided for younger students (i.e., K–2) and minimized for all other students.
  - Prior studies suggest that scrolling up and down in Reading passages negatively impacted students' performance on Reading tests (e.g., Bridgeman et al., 2003; Choi & Tinkler, 2002; Higgins, Russell, & Hoffmann, 2005; Poggio, Glasnapp, Yang, & Poggio, 2005; Pommerich, 2004). However, it is important to note that these comparability studies were conducted some years ago when student reading on computers was less prevalent. Students in the upper grades are becoming ever more exposed to computer- and internet-based Reading content. Younger students' daily Reading practices, however, remain predominantly paper-based.

Writing

- The response mode for Writing tasks should be paper-based for students at K–2.
  - Prior studies suggest that students demonstrate their Writing skills best when an assessment's Writing mode matches the mode used for instruction (Texas Education Agency, 2008; White et al., 2015). While computer-based Writing is increasing in upper grades, writing by hand remains at the core of Writing curricula and instruction for K–2 students. Additionally, some ELPAC task types designed to assess the K–2 standards call for handwriting of letters and/or words and are not suitable for keyboarding.

Task Design Recommendations for Improved Construct Representation and Student Engagement

- Across all domains, features of digital technology should be utilized to make the tasks more contextualized and interactive for students (e.g., color visuals, audio, and technology-enhanced features).
    - For the past two decades, the language testing field has advocated for the use of technology to develop authentic tasks in order to elicit real-life language from test takers (e.g., Bachman, 2000; Bailey, Heritage, & Butler, 2014; Douglas, 2013; Gruba, 2014; Martin, 2009; Roever, 2014; Wolf, Guzman-Orth, Lopez, Castellano, Himelfarb, & Tsutagawa, 2015). In particular, the provision of contextual information with integrated language skills can be more effectively implemented with technology (Ockey, 2009).
- The use of multimedia in task stimuli (e.g., video or animation) should be carefully considered in conjunction with cognitive load as well as the practicality of producing and implementing such features and content.
    - While real-life language-use situations involve various multimedia, it is critical to consider the cognitive load for students in assessment settings (Clark & Mayer, 2016; Mayer, 2003). The use of multimedia, including animations, should thus be embedded with considerations to construct representation and practicality. Given the large-scale administration of the ELPAC, it may be appropriate to include only a modest number of multimedia elements in each test form due to cost and bandwidth constraints.
- The use of technology-enhanced task features should be considered for the purposes of improving construct representation and student engagement. Where applicable, commonly known technology-enhanced features (e.g., zone, text selection), particularly those found on CAASPP, should be used rather than completely new features that may be less familiar to students.
    - Technology-enhanced task features boost students' interaction with given tasks, leading to increased student engagement (Martin, 2009; National Association for the Education of Young Children, 2012; Ockey, 2009; PARCC, 2014; Scalise, 2012). Additionally, the California ELD Standards call for the use of technology, where appropriate (e.g., standard PI.A.2). However, new or unfamiliar technology-enhanced task features may cause barriers to accessibility, particularly for young learners (Choi & Tinkler, 2002; SBAC, 2014; Wolf et al., 2014), or if they require special computer skills that students need to learn for the sake of the assessment (ELPA21, 2015; SBAC, 2013).

- User interface features for each language domain should be designed in a principled way so as to enhance construct representation and accessibility (see Section 3.2 for more specific information about accessibility). For example, user interface features such as volume control, replaying the directions and questions, replaying the recorded spoken responses, and word processing features (e.g., copy, cut, paste) should be considered in conjunction with real-life use situations. This will ensure that the user interface features enhance the representation of the construct of interest.
    - As mentioned earlier, the field of language testing emphasizes the authenticity of tasks in order to elicit evidence of students' language abilities appropriately. As CBA reduces the degree of human interaction in assessment, it is important to embed appropriate user interface features in order to better approximate real-life language use situations (Douglas, 2013; ELPA21, 2015).

Recommendations for Collecting Validity Evidence during the Development Stage

- The test design and task type design for a CBA should be formally documented via such work products as assessment claims, a construct definition, and item writing guidelines. The user interface and other aspects of the test delivery system should also be appropriately documented and reviewed to ensure that the target constructs are measured in valid, fair, and accessible ways.
    - The *AERA/APA/NCME Standards* (AERA, APA, & NCME, 2014) and the *Peer Review Guidance* (U.S. Department of Education, 2015) suggest that test design documents and administration manuals offer critical evidence to support validity arguments about the assessment's appropriate measurement of the target construct. Particularly, the *Peer Review Guidance* requires states to submit specific evidence to support the quality of a new technology-based assessment or the change of the administration mode of an existing assessment, a requirement which applies to the ELPAC CBAs. Thus, all the documents listed above, as well as guide/manual materials developed for CBA users, are critical to evaluate the quality and validity of the ELPAC CBAs.
- A usability study (or studies, as needed) should be conducted to examine the interaction of CBA features from the perspective of various users (students and educators including LEA ELPAC Coordinators and Test Examiners) in order to inform the platform design and selection of interface features prior to finalizing the operational test development. ETS recommends that cognitive laboratories be conducted at a relatively early stage of CBA development, once an initial CBA delivery platform with a representative set of TEIs and accessibility features is

available. It is also possible that other research methods (e.g., user surveys and/or interviews) could be used to address this need. Additionally, it is desirable to include a mode effect study (particularly for Writing and Speaking) examining the PPT and CBA modes during field-testing in order to provide empirical data to inform the creation of the administration and scoring guides.

- The *Peer Review Guidance* (U.S. Department of Education, 2015) explicitly mentions the implementation of usability studies to provide validity evidence for states' technology-based assessments. Additionally, an empirical investigation of targeted test takers' use of CBA features during the development stage would facilitate informed decision-making concerning the ELPAC CBA's design and administration features. Further, the initial usability study will offer useful insights and guidance for further empirical studies to be conducted.

3.2. ACCESSIBILITY AND ACCOMMODATIONS CONSIDERATIONS FOR THE ELPAC CBA

The introduction of the ELPAC CBA presents a significant opportunity to reevaluate and enhance the approach to ensuring that the ELPAC is as accessible as possible to all students in the target population, including ELs with disabilities (ELSWDs).

This subsection illustrates considerations for accessibility and accommodations associated with various stages of the ongoing test development process for the ELPAC CBA, particularly regarding ELSWDs. It is important to note that this subsection focuses only on students who would take the general ELPAC assessments, with or without accommodations. Students with significant cognitive disabilities who would need an alternate assessment (the one percent) are beyond the scope of this report. Based on information reviewed from both research literature and CBA practice, key considerations are raised as they pertain to the intersections between the target population, the test design process, the ELP construct, and the CBA context. This subsection also includes recommendations regarding accessibility and accommodations for the ELPAC CBA, including a rationale for each recommendation based on findings from research and practice.

Accessibility is a critical component to keep in mind at the start of the test design process. An early and consistent focus on accessibility provides opportunities to foster innovative solutions, to ensure that any challenges related to the use of technology-enabled features are addressed in a principled and efficient manner, and to improve accessibility planning for the various stages of the ELPAC CBA test development. This

process serves to make a testing program more effective by ensuring that meaningful scores are reported for a broader population of students.

Guidelines exist to help ensure the quality of test development processes, such as the *AERA/APA/NCME Standards* (AERA, APA, & NCME, 2014), universal design for learning and assessment, and best practices for ELs (CAST, 2011; Liu & Anderson, 2008; Pitoniak et al., 2008; Thompson, Johnstone, & Thurlow, 2002). Previous generations of assessments included reference to variations, accommodations, and modifications (e.g., California Department of Education, 2016) or used a five-category breakdown for accommodations: presentation, response, equipment and material, scheduling or timing, and setting.

The current generation of CBA has moved forward by structuring multi-tiered models of accessibility (e.g., *universal tools* available for all students; *designated supports* available for some students; and *accommodations* available for few students). The ELPAC CBA has an opportunity to build upon the existing CAASPP system of CBA, including the Smarter Balanced English language arts and mathematics assessments and other content-area CBAs, in addition to existing ELP CBAs (e.g., ELPA21, WIDA). This wide view can inform the accessibility and accommodations practices implemented for the ELPAC CBA. Because CBAs and the research and technology behind them are evolving, there are great benefits to involving a multidisciplinary team of experts to employ a holistic view to develop a multi-faceted accessibility solution for the ELPAC CBA.

Given the complexity of the interactions between the target population and test design decisions, the designers of the ELPAC CBA also have an important opportunity to take both of these elements into account, individually and in combination, when designing and implementing a multi-faceted accessibility solution for the ELPAC CBA. This practice includes not only looking at the high-level categories (e.g., ELSWDs, accessibility features) but also involves breaking down each category (i.e., the CALPADs disability categories) with a team of experts to identify potential challenges and possible solutions to move forward in the ELPAC CBA test development process. Although it is anticipated that the solutions for the ELPAC CBA will adopt the systems and procedures used for the CAASPP, the multidisciplinary team should provide guidance regarding instances in which the ELPAC CBA test development process should differ from CAASPP due to the target population, the construct being assessed, item types, or other considerations. As noted in Section 3.1, these reviews may also generate opportunities to consider potential enhancements to the CAASPP assessment system that will be appropriate for all CAASPP test takers.

Specific recommendations and rationales are included below.

Recommendations for Accessibility Features and Accommodations for the ELPAC CBA

- From the beginning of the test design effort, a formal process should be instated for reviewing and evaluating accessibility tools, supports, and accommodations that will be made available on the ELPAC CBA. This review process should include ongoing advisory work by a multi-disciplinary team of experts (e.g., experts in ELs, ELSWDs, SWDs, technology, assessment validity, accessibility) who will take into consideration accessibility features and accommodations used by existing content-area and ELP CBAs, as well as those commonly used in an ELP PPT (e.g., dropping test items/domains). In addition, the team of experts should systematically review their accessibility and accommodations recommendations with considerations for the unique features of the ELPAC test design (e.g., K–2, ELPAC task types by domain, integrated task types).
    - Current practice for accessibility and accommodations involves input and collaboration from experts to ensure that the practices meet the needs of the target population taking the CBA, either through reviews of existing research or specially targeted studies (Abedi & Ewers, 2013; Guzman-Orth, 2014; Guzman-Orth, Wolf, King, & Tolentino, 2016; King, 2014; Laitusis et al., 2012; Solano-Flores, Shade, & Chrzanowski, 2014; Wolf et al.,, 2014). Variations in the multi-tiered structure (e.g., universal tools, designated supports, accommodations, domain differences, delivery mode [e.g., CBA version and PPT version, WIDA, 2015a]) and in design for allowed accessibility features and accommodations (e.g., digital notepad) implemented across consortia (see Table C2 in Appendix C for an example) can have meaningful impact for the overall design of the ELPAC CBA accessibility and accommodations (see Table C1 in Appendix C for a list of possible accessibility features and accommodations that can be used for the ELPAC CBA).
- An accessibility framework should be created to guide the accessibility work from the initial test design stages throughout the test development cycle for the ELPAC CBA. At minimum, the framework should include sections related to accommodations policy, accommodations guidelines, and an accommodations assignment process. The framework should be considered a living document that is consulted and updated as appropriate at each stage of the design process for the ELPAC CBA. For example, certain components of the framework may not be available to update if they have been finalized in policy or other regulations. However, other components of the framework—such as ongoing research, validity evidence, stakeholder meetings, or documentation of accessibility procedures for design and development—could be updated during the length of the test development contract.

  - The accessibility framework should be the ongoing and evolving foundation for the conceptual and empirical accessibility information to guide the test development process through operational delivery, similar to the ongoing research and development in the existing accessibility frameworks for content-area and ELP CBAs (ELPA21, 2016a; PARCC, 2016; Shyyann et al., 2016; SBAC, 2014, 2016; WIDA, 2015b).

- Early in the initial ELPAC CBA design work, a formal process should be implemented to systematically evaluate the task design decisions and corresponding test design documents for potential challenges to accessibility. These challenges should be linked to multiple accessibility solutions (e.g., accessible item writing guidelines, accessibility features or accommodations, alternative formats). These accessibility solutions should be continuously evaluated and updated as the design decisions and test design documents are updated during the test development process excepting any considerations for documentation that may have already been approved and finalized in policy or other regulations.
  - WIDA specifies their allowed accessibility features and accommodations with attention to differences in the target construct by domain and allowed supports (WIDA, 2015a). In contrast, the use of digitally delivered "twins" (alternate versions of specific items presented in accessible formats) to measure ELP for students with visual impairments (VI) is an example of a unique alternate format promoted by ELPA21 (ELPA21, 2016b). Various accommodation supports work together to create a customized, yet standardized test setting via accessibility standards (e.g., Accessible Portable Item Protocol [APIP] interoperability standards; IMS, 2016), which digitally link the item authoring process to the item delivery via the item metadata. However, some accessibility features (e.g., highlighter) are not applicable for all ELP task types (e.g., tasks with picture based stimulus and response options), and cannot be used appropriately in those contexts (Wolf et al., 2014).

- The ELPAC CBA should develop an accommodations assignment process and mechanism to determine *when* (IA, SA), *where* (domain level), and *which* accessibility features and accommodations are allowed, and *for whom* (e.g., all students, some students, ELSWDs, new arrivals). This accommodations assignment mechanism should be based on the CAASPP system of assessments in consideration of other CBAs, particularly those assessing English language proficiency. The accommodations assignment process should be developed in accordance with the allowed accessibility features and accommodations and corresponding organizational structure (e.g., universal tools, designated supports, and accommodations). The process should take into account unique characteristics

of ELSWDs who may be new arrivals and have had limited time for evaluation or limited access to their assigned accommodation. A comparable process should be developed for accommodations of students qualifying for the PPT IA, or those who are allowed to handwrite rather than keyboard responses on the ELPAC CBA. Because of the nature of the differences between the PPT and CBA formats, it is reasonable to expect that only select accommodations for the CBA will apply to the PPT context. In this aspect, the multi-disciplinary team of experts should provide guidance and justification for which administration guidelines and accommodations are appropriate for students qualifying for the paper-based IA.

  - Both content-area and ELP CBA use an accommodations assignment process and underlying mechanism in practice (e.g., a Personal Needs Profile [PNP; see ELPA21, 2016; PARCC, 2016], an Individual Student Assessment Accessibility Profile tool [ISAAP; see SBAC, 2015], or the WIDA Assessment Management system [see WIDA, 2015a]). ELs who are new arrivals (with a true disability or suspected of a disability) may not have adequate time to go through the referral and evaluation process to have appropriate documentation (i.e., IEP/504 Plan documentation) or may require alternate accommodations like item- or domain-level exemption (Liu et al., 2013; Christensen et al., 2013; 2014a; 2014b). For students with a diagnosed disability upon enrollment (e.g., profoundly hearing-impaired [HI] or VI students), they may not be proficient with the allowed accommodations or communication system (e.g., Unified English Braille; American Sign Language) at the time of the assessment, which further emphasizes the need for the assignment process to take into account multiple solutions for the ELPAC CBA (Guzman-Orth, Laitusis, Thurlow, & Christensen, 2016).

- The ELPAC CBA test development effort should include the creation of accessible forms created specifically for the needs of students with VI (low vision, blindness). The accessible form design should take into account the task type presentation, alternate means of representation, TEI usability and interactions with technology-enhanced accommodations (e.g., text to speech software), and accommodation familiarity. Additionally, characteristics of the target population (e.g., grade level, emergence of literacy skills, new arrival status, migrant status, refugee status, and presence of students with interrupted formal education [SIFE]) should be taken into account in the accessible form design since they may impact students' performances and accommodation familiarity. Although practices for VI accessibility should follow procedures for the CAASPP, guidance from the multi-disciplinary team of experts, as well as input from stakeholders who have experience with ELs with VI, should provide input on how these considerations could interact in the ELP context.

These interactions should be taken into account to develop principled solutions where accessible content for ELs who have VI may differ from those solutions developed for the CAASPP. Accessible item formats should be prototyped with a small number of students in K–2 (non-Smarter Balanced grades). ELSWDs should be included in the field test process, either via access to the content through accessible test forms, or through other means, like a cognitive laboratories study for students with VI.

  - The VI disability can impact access across domains, but the impact may be less obvious to recognize (Guzman-Orth et al., 2016). Full-color art (regardless of construct relevance) or animations pose challenges for the VI population (Albus & Thurlow, 2008; Christensen et al., 2014a; Guzman-Orth et al., 2016). Select TEIs also pose some challenges in being made accessible via certain technology-based accommodations, e.g., text-to-speech software or zoom and magnification supports (Guzman-Orth et al., 2016; Wolf et al., 2014). Accessibility standards (e.g., Accessible Portable Item Protocol [APIP] interoperability standards; IMS, 2016) are evolving, and current limitations in accessibility technology call for multifaceted accessibility solutions. The ELPA21 Consortium has implemented accessible item twins as one example of an alternate design format (ELPA21, 2016b).

- The ELPAC CBA should include interactive demos and tutorials to guide students through the use and functionality of each accessibility feature and accommodation allowed to them before taking the ELPAC CBA. These interactive demos and tutorials should include general and accessible formats of the ELPAC CBA. The training features and help features in the interactive demos, as well as tutorials, should follow those implemented in the CAASPP, with possible modifications as suggested by accessibility or content experts who may suggest key areas for differentiation based on the ELPAC target population, construct, and item types.
  - Opportunities to practice with the test and accessibility features and accommodations are critical components for any ELP CBA (Guzman-Orth et al., 2016). When taught through a video and interactive tutorial, ELs and ELSWDs demonstrated the ability to familiarize themselves with digital accessibility features and to use these features in their test experience (Wolf et al., 2014).

3.3. AUTOMATED SCORING CAPABILITIES FOR WRITTEN AND SPOKEN RESPONSES FROM K–12 ENGLISH LEARNERS

The transition of the ELPAC to a CBA model provides an opportunity to introduce the

use of Artificial Intelligence technology (automated scoring or AI scoring) to evaluate students' written responses and spoken responses. AI scoring technology is a field that continues to evolve as the capabilities of computers expand and as our understanding of the constructs in language assessments advances. This subsection first briefly discusses advantages and disadvantages of AI scoring in general. Then AI scoring of written responses and spoken responses by K–12 ELs, respectively, are discussed.

From a practical perspective, the most obvious advantage of AI scoring for a large-scale assessment is that it can eliminate, or at least dramatically reduce, the human labor required for scoring student responses. If the responses are currently being scored by hired raters (as is the plan for the ELPAC PPT written responses), this can create savings in the recurring costs associated with compensating raters. If the responses are currently being scored by local educators (as is the plan for the ELPAC PPT spoken responses), AI scoring can free up staff time that had been devoted to training and scoring. Another advantage of AI scoring is consistency; AI scoring engines are not affected by subjective factors that can negatively influence human rater reliability such as fatigue, distraction, or bias.

While the amount of teacher time devoted to training and scoring would be reduced with the introduction of automated scoring, there would still be opportunities for professional development for teachers through sample selection, range finding, and providing scores to a subset of the responses as a reliability check.

An important limitation of AI scoring is that AI scoring systems are not able to adequately assess all aspects of the targeted construct for certain CR task types due to limitations in the system's ability to fully understand the content of the response. For example, in AI scoring of written responses, the technology for evaluating writing features such as conventions, structure, vocabulary, and correctness of content as defined in the rubrics is more advanced than that for evaluating argumentation or authorial voice (Weigle, 2013). AI scoring of responses to more open-ended task types can be more challenging. In assessments where these task types are required to provide a full assessment of the construct, it would be necessary to conduct additional research to develop new AI scoring capabilities to address the targeted construct or to use a hybrid human-machine scoring approach (i.e., to use human raters to evaluate those task types, or aspects of task types, for which AI scoring is not sufficiently effective). This, of course, would offset some of the potential cost savings of AI scoring.

It is worth noting that AI scoring requires an appropriate infrastructure for the capture of student responses. In Writing, AI scoring requires that student responses be keyboarded. In Speaking, AI scoring requires digital voice capture of student responses. More details on the required infrastructure appear in the recommendations below.

Another practical disadvantage of AI scoring is that it requires substantial up-front investment. An AI "model" needs to be built for each constructed response task type, and a prerequisite for building and evaluating such models is a relatively large number of double-human-scored responses (typically at least 1,500 per item). These responses need to be collected in a pilot or field test far enough in advance to allow for model building, evaluation, and deployment. This necessitates careful planning for the timing of every stage of test development and deployment. In addition, the up-front development costs associated with AI model building are only justified for an assessment with relatively large operational volumes (at least tens of thousands of tests administered annually). A full analysis of the development costs would be necessary.

Finally, it is important to note that AI scoring models have been explored more fully for adult students than for K–12 students. There are some challenges related to the assessment of younger students that are discussed in the Automated Scoring of Written Responses and Automated Scoring of Spoken Responses sections below.

Automated Scoring of Written Responses

Research on automated evaluation of student written responses is approximately half a century old (Page, 1966), and the technology itself has been in use for high-stakes applications for adult ELP assessments for several years, such as in the TOEFL iBT assessment of English for academic purposes (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012). There is a large body of work discussing the linguistic features employed, algorithms used, its construct coverage, and validity. In terms of reliability, the summary of results of a study comparing nine existing commercial AI scoring engines by Shermis and Hamner (2013) shows that all engines that participated in the study generally correlated well with human raters in scoring essay-length responses.

Notably, there has been relatively little research that focuses specifically on AI scoring of Writing in K–12 ELP assessments. To our knowledge, Pearson's Test of English Language Learning (TELL) is currently the only assessment in the K–12 EL domain which uses AI scoring of Writing (Bonk, 2016). TELL was introduced in 2015, and we have little information on its performance to date. The K–12 ELP assessments offered by the WIDA and ELPA21 consortia do not currently make use of AI scoring.

However, based on previous research into AI scoring of non-native written responses for other assessments, the prospects for developing successful AI scoring capabilities for current ELPAC Writing task types in grades 3–12 are promising. The content of test taker responses in the Describe a Picture and Write about Academic Information task types is substantially constrained based on information presented in the stimulus

materials, and it is likely that valid and reliable AI scoring models could be developed for these task types. The Justify an Opinion task type is the most open-ended task type in the current ELPAC Writing domain, since the test takers are likely to support their opinions using a range of different content based on their personal experiences. However, this task type is similar to Writing tasks included in assessments from other areas, such as the GRE Analyze an Issue task, for which AI scoring systems have been successfully developed and deployed; therefore, with sufficient training data for the AI scoring models from a field test, it may also be possible to develop valid and reliable AI scoring for this task type.

Automated Scoring of Spoken Responses

Research into AI scoring of non-native speakers' spoken responses has been conducted since around 1990; however, most of the early research focused on task types that were highly restricted (e.g., reading printed content out loud). Responses to these types of tasks can be processed accurately by automatic speech recognition (ASR) systems, which is a requirement for successful AI scoring of speech. AI scoring systems for these task types are mature and produce scores that correlate highly with human scores, typically matching human-human agreement levels (Neumeyer, Franco, Digalakis, & Weintraub, 2000; Bernstein, Van Moere, & Cheng, 2010; Evanini, Heilman, Wang, & Blanchard, 2015). The proficiency features extracted by AI scoring systems for these task types primarily address the delivery aspect of the Speaking construct, including pronunciation, fluency, rhythm, and intonation.

More recent research has investigated the use of AI scoring for Speaking tasks that elicit spontaneous speech, such as describing a picture, retelling a story, and summarizing a lecture, and has expanded the construct coverage to include vocabulary complexity, grammatical accuracy, grammatical complexity, discourse coherence, and content appropriateness. While AI scoring systems for these Speaking tasks types are still under active development, strong empirical results have been obtained in some cases. For example, Hassanali et al. (2015) report on an AI scoring system for a narrative retelling task for ELs in primary school, and Xiong et al. (2013) describe how features assessing content accuracy contribute to the performance of an AI scoring system for an academic summarization task for ELs in middle school.

AI scoring of speech from young ELs is especially challenging due to the difficulty of obtaining accurate results from the ASR system for this target population. For that reason, it is important to pay attention to the methodology used for digital speech capture in an assessment to ensure that the audio is of high quality; this will, in turn, lead to better ASR performance.

Similar to Writing, AI scoring for Speaking has been used sparingly to date in operational K–12 ELP assessments; the two instances that we know of are the Arizona English Language Learner Assessment (AZELLA; Cheng, D'Antilio, Chen, & Bernstein, 2014) and the TELL (Bonk, 2016), both administered by Pearson. These assessments include a combination of task types that elicit both restricted speech and spontaneous speech. While the empirical performance in terms of the correlation with human scores for the Speaking domain of the assessment can be quite high (for TELL these correlations range from 0.59 to 0.85; for AZELLA they range from 0.88 to 0.95), compelling evidence that the AI scores for these task types fully cover all aspects of the targeted Speaking construct has not been made publicly available; it seems possible that the high correlations are due primarily to emphasis on a smaller number of reliable features assessing delivery, with limited coverage of vocabulary, grammar, and content. Best practice in planning the use of AI scoring for the ELPAC Speaking task types would be to ensure appropriate construct coverage before operationally implementing AI scoring (Evanini, Hauck, & Hakuta, in press).

Based on this prior research into AI scoring of non-native spoken responses for other assessments, the prospects for developing successful AI scoring capabilities for current ELPAC Speaking task types are promising. The Retell a Narrative, Present and Discuss Information, and Summarize and Academic Presentation task types are very similar to task types that have been previously investigated in the context of AI scoring. Since the expected content in the test takers' responses for these task types is based primarily on information that is contained in the stimulus materials, it is likely that AI scoring technology will be able to cover aspects of the scoring rubric related to content (for example, "The response is appropriate to the task and conveys sufficient and relevant details"). Less prior research has been conducted on task types that are similar to the Talk About a Scene and Speech Functions tasks; however, the content of these responses is also substantially constrained based on the stimulus materials, and it is likely that valid and reliable AI models can be developed based on a set of scored responses from the field test. The Support an Opinion task type in the current ELPAC Speaking domain is the most likely to elicit unpredictable content since the test takers' responses are based on personal experiences and opinions; for this task type, it may therefore be difficult to develop AI scoring models that fully address all aspects of the targeted Speaking construct.

Recommendations for Maximizing Automated Scoring Capability for the ELPAC CBA

Overall Recommendations Regarding Automated Scoring

- The ELPAC CBA should be developed with the goal of supporting operational AI scoring of written responses at grades 3–12 and of spoken responses at K–12, with details of implementation to be based on an analysis of field test data. AI scoring of written responses at K–2 is not possible because of the recommendation that written responses be handwritten.
- The CDE should make decisions regarding the pace of adoption for AI scoring through a process that considers input from appropriate stakeholders as well as empirical data. Because these technologies will continue to evolve, periodic reviews of capabilities and potential applications to the scoring of the ELPAC responses should be conducted with an eye to further enhancing the use of AI scoring.

Specific Recommendations Regarding Infrastructure and Technical Requirements for Automated Scoring

- The ELPAC CBA should be implemented with appropriate infrastructure to support potential AI scoring of written responses and of spoken responses.
    - Technical requirements for AI scoring of written responses are as follows:
        - Written responses should be provided via keyboard input at grades 3–12. Handwriting recognition, whether on paper and scanned or using a tablet, is not reliable enough for a high-stakes assessment context.
        - Keyboard input should be captured as plain text. AI scoring engines require plain text as input. Although HTML outputs may be cleaned, there is no guarantee that the resulting text will exactly match what the student entered.
    - Technical requirements for AI scoring of spoken responses:
        - Spoken responses should be collected using a digital voice mechanism that includes an external microphone (e.g., a headset microphone or earbuds with an in-line microphone instead of the computer's built-in microphone) to ensure the highest quality audio input. The digital voice capture procedure (including the system hardware and microphones) should be standardized across all test takers in the field test, and the same equipment and procedures should be used in the operational administration. Prior to making a final decision

about the procedure to be used for digital voice capture, a small number of student responses should be collected under simulated testing conditions to ensure that the quality of the audio is sufficient for AI scoring.

- Specifications and requirements for headsets/earbuds/microphones should consider the range of sizes of children across grades and grade spans (i.e., while adult sizes may work for students in grades 9–10 and 11–12, smaller equipment is necessary for students in lower grades).
- The interface for capturing spoken responses should include functionality for verifying that high quality audio is being captured for each test taker; for example, the test taker should be asked to record a few seconds of speech and then listen to verify that the recording was clearly audible. Additionally, guidelines for establishing an appropriate environment (e.g., establishing a quiet location, minimizing background noise, seating students an appropriate distance apart) and ensuring appropriate use of the digital speech capture interface (e.g., optimal placement of the microphone) should be established.
- For Speaking tasks types that involve interaction between a student and Test Examiner, the student's spoken response should be captured separately by the digital voice capture platform. (If the Test Examiner's voice is also captured, it can interfere with AI scoring of the student's response.)

Recommendations Regarding Task Design and Data Collection for Automated Scoring

- The task type design process, including evaluation of the prototype/pilot responses, should include ongoing consulting by natural language processing and AI scoring staff to ensure that all tasks for both the ELPAC Writing domain and Speaking domain are designed to maximize their suitability for AI scoring (while keeping construct representation as the highest value).
- The CBA field test should include a data collection effort to support the training and evaluation of AI scoring models for both written and spoken responses.
    - A sufficient number of spoken and written student responses should be collected for each item in order to train and evaluate AI scoring models so that a recommendation can be made about the potential use of AI scoring in the CBA (for most task types, a minimum of 1,500 responses per item is recommended).
    - The responses should be sampled from test takers that represent all major primary language backgrounds and the full range of ELP in the target population.

A skewed sample has negative impact on AI scoring model performance.

- A portion of the spoken responses from the field test should be used for adapting the automatic speech recognizer to improve its performance on these tasks and this population of ELs; a minimum of 200 hours of spoken responses should be transcribed and used for ASR adaptation. An adapted ASR system will improve the accuracy of the speech recognizer, thus leading to more valid and reliable AI scoring models.
- The field test responses that will be used for training and evaluating AI scoring models should receive double human scores so that human-machine agreement results can be compared to human-human agreement results.

Recommendations Regarding Deployment of Automated Scoring Models

- After the field test results have been analyzed, a formal set of recommendations should be made regarding how and when AI scoring should be implemented for operational use (e.g., for which task types AI scoring is performing well enough to be implemented and for which human scoring or a hybrid model is most appropriate). This set of recommendations should be reviewed and approved by an appropriate set of stakeholders. These are crucial decisions that will have significant impact on the ELPAC CBA testing program. It is important that these decisions be made on appropriate data and be thoroughly vetted before implementation.
  - AI scoring should be deployed operationally only for task types for which the models contain a sufficient number of construct-relevant features. In the event that valid AI scoring models do not exist for all task types in an assessment form, a hybrid approach to scoring is recommended in which humans score responses to some tasks and the AI scoring engine scores responses to other tasks. Lack of sufficient construct-relevant features in AI scoring models negatively affects the validity of AI scores.
  - AI scoring should be deployed operationally only for task types for which the performance on the field test responses meets pre-defined reliability standards in terms of raw correlations, as well as comparisons to human-human agreement levels.
  - Human raters should be made available as a backup for all task types in order to provide scores for any responses that are filtered out as non-scorable by the AI scoring engine (e.g., responses that are off-topic or spoken responses that have

too much background noise). This is to ensure that every response receives a valid score.

- If the AI engine score is the sole score for a given task type, a percentage of randomly sampled responses should be scored by human raters in order to monitor the AI scoring model performance during operational administrations, as substantial changes in test-taker population may affect the AI scoring model performance.

Recommendation Regarding Automated Scoring for ELs with Disabilities

- Speaking tasks that may involve AI scoring should have additional procedures in place for ELSWDs who may need assistance with the equipment, procedures, or other components. Additionally, guidelines for the appropriate use of AI scoring should be developed for ELSWDs with speech/language impairments.
  - To date, limited research exists to determine the impact that speech/language impairments may have on the AI scoring process, especially for children who are non-native speakers with speech/language impairments.

4. TASK TYPE AND TEST DESIGN CONSIDERATIONS FOR THE ELPAC CBA

Based on the research background information and recommendations presented in the previous sections (e.g., CBA validity considerations, CBA enhanced features, accessibility and accommodations, and potential use of AI scoring), the present section aims to provide more detailed recommendations regarding the design and development of task types for the ELPAC CBA. To achieve this aim, the ELPAC CBA study team members (who have been centrally involved in the development of the ELPAC PPT task types) conducted a close, qualitative examination of the current ELPAC PPT task types and discussed the extent to which the current task types were suitable for use in the CBA, with modifications as appropriate. Then, the team created a draft of a proposed test blueprint for the ELPAC CBA summative assessment based on this analysis. This draft of a proposed test blueprint is presented in Appendix D.

The task types and blueprints that will form the basis of the test design will be a crucial element of the ELPAC CBA, serving as key guiding documents outlining how information about student English language proficiency will be captured as well as establishing the framework for other aspects of the ELPAC CBA transition and implementation, including the psychometric requirements and the systems requirements.

This section first presents ETS's overall recommendations for the task types and blueprint for the ELPAC CBA summative assessment, including features to be added for the CBA that are applicable to all grades and across multiple domains. Then, more specific recommendations for enhancing task types are presented by grade span and by domain. The final subsections describe the draft of a proposed test blueprint for the ELPAC CBA summative assessment and offer concrete recommendations for test item development to support an efficient transition to the CBA.

4.1. OVERALL RECOMMENDATIONS FOR TASK TYPES FOR THE ELPAC CBA SUMMATIVE ASSESSMENT

- The task types of the current ELPAC should be maintained for the CBA with technology-enhanced features added in a principled way. The principles for enhancements include better measurement, improved student engagement, and efficient standardized administration.
    - The conceptual analysis performed as part of the PPT test design effort, reflecting decisions about how the California ELD Standards can be best

assessed within the constraints of a standardized assessment, remains valid. At the CDE's direction, this conceptual analysis was conducted with consideration that the ELPAC program would likely transition to CBA in the future. As a result, the task types identified for the PPT were developed with the goal of selecting tasks that could be readily repurposed and enhanced for the CBA. The CBA should therefore include a high proportion of the same task types that were designed for the PPT, but the items delivered via computer will feature a number of enhancements that will result in an improved assessment.

- In general, the PPT blueprint should be similar for the CBA in terms of the number of task types, items, and raw score points.
    - As described above, future-looking efforts were made in developing the current blueprint for the ELPAC. While minor differences in the blueprint may emerge over the course of developing the CBA (reflecting, for example, revisions that are made to task types based on prototype/pilot results or differences in score points as rubrics are revised), the general approach and contents of the PPT blueprint are expected to be largely appropriate for the CBA. This will provide a useful element of stability in the transition.
- As a general CBA enhancement feature, ETS recommends the inclusion of an avatar or guide in the form of an illustration or animation of a student and/or teacher who guides the student through the assessment. The design of the avatar should be carefully considered and its usability should be empirically examined as part of the development process.
    - The avatar has the potential to increase engagement in the assessment, reinforce and clarify directions, model responses, and create more authentic contexts for students to respond to. To realize the benefits of including avatars as guides, an empirical investigation of students' interaction with avatars is an important part of the pilot/prototype stage of the development effort.
- All graphics to be included in ELPAC CBA items should be presented in full color. New graphics developed for the CBA should be developed in color; if existing PPT items are to be used on the CBA, their graphics should be updated to be in color.
    - Full-color graphics are a standard feature of most CBAs and can be expected to help increase student engagement (Douglas, 2013; ELPA21, 2015).
- Context-setting images should be added to several task types as detailed in Appendix D, increasing the amount of context for the student and improving engagement.

- New TEI formats should be considered for use in several task types. They make the assessment more interactive, authentic, and potentially more engaging, and help to meet the directive of the California ELD Standards for the use of technology where appropriate (see standard PI.A.2). However, care must be taken to ensure that TEIs are developmentally appropriate for young test takers, especially those in grades K–2. TEIs may require a certain level of motor and/or computer maneuvering skill, or they may be cognitively demanding due to the complexity of the response mode. Consideration of the suitability of TEIs for all students, especially those in grades K–2, should be an important part of the pilot/prototype stage of the development effort. Recommended TEI formats are listed below; appropriate task types for the use of each TEI format are provided in Appendix D.
    - **Match** (Drag and Drop/Click and Click): The test taker responds by dragging and dropping answer choices ("sources") into the appropriate locations ("targets"). Alternately, test takers click on answer choices to select the item and click again to place or drop the item into the appropriate location. The latter requires minimal fine motor skills which is of special concern for K–2 test takers.
    - **Zone** (Hotspot): Items where the answer choices are pre-defined "hotspots" on an image. When the test taker clicks on the hotspot, the selection is highlighted, shaded, or outlined. Test takers select one zone as a response.
    - **Grid** (Matching Tables): The test taker responds by marking two or more cells in the table grid. The response can be restricted to one selection per row, column, or table.
    - **Select in Passage:** Items where the answer choices are a pre-defined set of words, phrases, sentences, or paragraphs within a set leader. When the test taker makes a selection, the word/sentence is highlighted in the passage. The candidate can select one or more choices as a response.
- The CBA test delivery system should support seamless delivery of professionally recorded audio of aural elements of the test materials. This includes directions for all domains, any audio stimulus material, and either the question and response options (for SR and TEIs) or the prompt (for CR items). This will help to ensure an intuitive and standardized delivery of all domains of the CBA.

4.2. ANALYSIS OF EXISTING AND POTENTIAL TASK TYPES BY DOMAIN

It is assumed that the ELPAC CBA will report scores in four domains, as the PPT does. The following subsection discusses the suitability of ELPAC PPT task types for use on

the CBA and provides recommendations for modifications and enhancements for the design of task types for each language domain by grade levels.

Grades K–2

Listening Domain

- Instead of having Test Examiners read the Listening stimuli, questions, and response options aloud, the K–2 Listening content should be recorded by professional voice actors (as is the ELPAC PPT at grades 3–12). The students should receive the CBA audio through a headset or earbuds. The use of speakers is unlikely to be appropriate if there are multiple students in the testing room because students may move through the Listening domain at different speeds.
- The CBA test delivery system should be designed to allow for user control of the replay of audio inputs (i.e., directions, questions, response options). This will improve engagement (and will potentially improve measurement) by allowing students to have more control over the testing experience. For K–2, the stimuli should be replayable by the Test Examiner, if needed, to align with the prompting and replay guidelines that are to be established (e.g., if the student did not hear the stimulus originally due to a technical or other issue).
- A decision regarding the permissibility of replay of stimuli, as discussed in Section 3.1, must be made during the task design process, as stimuli will be written differently if they are intended to support replay.
- The Listening domain should be administered individually to students in K–1 in the CBA (as it is for those grades on the PPT) in order to provide adequate support and guidance to test takers.

Speaking Domain

- The CBA test delivery system should support digital capture of speech (digital voice capture), allowing all student spoken responses to be scored by certified raters via a distributed network, thereby increasing the reliability of scoring and reducing the burden on Test Examiners.
- As with Listening, the CBA test delivery system for the Speaking domain should be designed to allow directions and the prompt (though not stimuli) to be controlled by the student and to be played or replayed as the student chooses. For K–2, the stimuli should be replayable by the Test Examiner, if needed, following prompting

and replay guidelines that should be established (e.g., if the student did not hear the stimulus originally due to a technical or other issue).

- Regarding the recommendation above to employ an avatar for the Speaking domain, the avatar would pose the questions to the student, and the student would respond to the avatar. Creating a realistic context for speech (student-student or student-teacher) is an important part of effectively measuring the Speaking construct.
- The Speaking domain should be administered individually to students in K–1 (as it is on the ELPAC PPT) in order to provide adequate support and guidance to test takers. Prompting guidelines should be established to indicate what level of support Test Examiners can provide for students who offer no or incomplete responses, or who provide responses in a language other than English.
    - An appropriate policy regarding the capture of prompting on the part of the Test Examiner should be established. As noted in Section 3.3, recording the speech of the Test Examiner can create challenges for AI scoring; however, for human raters it can be important to listen to the entire interaction in order to evaluate the student's response.

Reading Domain

- The Reading domain should still be administered individually to students in K–1 in order to provide adequate support and guidance to test takers.
- Where applicable for K–2, the general design of the test delivery system for the presentation of the CBA Reading domain (layout, display, available tools, etc.) should be parallel to that of CAASPP unless there is a principled reason for the ELPAC CBA to be different (i.e., a reason rooted in the construct or target population of the ELPAC). Recommendations about appropriate font size, layout, amount of text visible per page, and screen navigation for K–2 students should be made and documented early in the test design process.
- For "read-along" tasks, where the student hears and sees text, the text should be highlighted word-by-word and synchronized with an audio recording of the text.

Note that the general recommendation for the inclusion of color graphics (in Section 4.1) has particular value for the K–2 Reading domain. The use of color in Reading task types will more closely resemble the classroom materials that students are exposed to.

Writing Domain

- As noted in Table 2.1, the ELPAC Writing domain should remain a fully PPT for K–2. Depending on the outcome of the PPT field test and the proposed CBA blueprint, it may appropriate for the blueprint for the K–2 Writing domain to remain identical to that of the PPT.
    - In contrast to the other three domains, the Writing domain at K–2 is not suitable for assessment via CBA, as described in Section 3.1. Although there are models in which aspects of a CBA approach could be used for K–2 Writing (e.g., the stimulus material could be presented on a computer screen and the student could be asked to respond on paper), such models contain significant risks (e.g., greatly increasing the cognitive load for these young students) without compelling benefits.

Grades 3–12

Listening Domain

- The CBA test delivery system should be designed to allow both directions and also questions and response options to be controlled by the student and to be played or replayed as the student chooses. This improves engagement (and potentially improves measurement) by allowing student more control over the testing experience. (Enabling students to replay directions provides a higher level of confidence that students have understood the tasks.)
- As noted for grades K–2, a decision regarding the possible replay of stimuli must be made during the task design process, as stimuli will be written differently if they are intended to support replay.

Speaking Domain

- The CBA test delivery system should support digital capture of speech (digital voice capture), allowing all student spoken responses to be scored by certified raters via a distributed network, thereby increasing the reliability of scoring and reducing the burden on Test Examiners.
- As with Listening, the CBA test delivery system for Speaking should be designed to allow directions and the prompt (though not stimuli) to be controlled by the student, and to be played or replayed as the student chooses.
- Additional recommendation regarding the use of an avatar: in the Speaking domain, the avatar should pose the questions to the student and the student should respond to the avatar. This will help to create a realistic context for speech (student-student or student-teacher), which is an important part of the Speaking construct.
- The use of multimedia, animation, or video could be of value throughout the CBA, but the Speaking domain is a particularly promising area in which to use such features. As noted in Appendix D, the Summarize an Academic Presentation task type could be modified from the PPT format, in which the student hears an audio recording, to a CBA format, in which the student would view an animation of a teacher in a classroom giving a presentation, providing a much richer stimulus for the student's summary.

Note that the recommendations made for the Speaking domain have been limited to those which are based on existing capabilities that can be implemented on a known schedule and with a predictable budget. It is also the case that, as technology and

innovation in the field of language proficiency advance, new capabilities will be developed. For example, ETS currently has in development a capability known as Spoken Dialog Systems (SDS) that has the potential to make Speaking tasks significantly more interactive by providing real-time, computer-generated responses based on in-the-moment AI recognition of student speech. While the SDS approach is not currently mature enough to meet the criteria for inclusion in this report (e.g., the readiness of capabilities for large-scale, high-stakes standardized assessments), ETS welcomes the opportunity to discuss when and how such cutting-edge capabilities might be appropriate for use in the ELPAC.

Reading Domain

- The general design of the test delivery system for the presentation of the CBA Reading domain (layout, display, available tools, etc.) should be parallel to that of CAASPP unless there is a principled reason for the ELPAC CBA to be different (i.e., a reason rooted in the construct or population of the ELPAC).

Writing Domain

- The CBA test delivery system should be designed to allow students in grades 3–12 to complete the Writing domain at their own pace, rather than having to wait for all students in the class to finish a task type before all can move on to the next task type. This has the potential to be a significant improvement to the student experience in the Writing domain, offering self-pacing results for less student downtime during the test and a more active testing experience.
- The CBA test delivery system should be designed so that the prompt is presented on screen, and students keyboard their responses to all Writing tasks. The keyboarding functionality should include the same features as provided for CAASPP (e.g., cut and paste, copy, delete, but not spell check or grammar check), while also ensuring that text is captured in a mode that supports potential use of AI scoring.
- A new task type, an integrated Writing with Listening task, should be included in the prototype/pilot for the ELPAC CBA and considered for inclusion on the CBA blueprint. This task type was successfully piloted for the ELPAC PPT but was not included because of logistical challenges related to administration. These logistical requirements (i.e., the inclusion of an audio stimulus within the Writing domain) are not a challenge for the CBA delivery model.

4.3. DRAFT OF A PROPOSED CBA TEST BLUEPRINT

A draft of a possible test blueprint for the ELPAC CBA is included in Appendix D. This blueprint takes the form of several tables based on the current PPT blueprints.

Each table presents a brief description of the existing PPT task type, the primary California ELD Standard(s) that the task type is designed to assess, information regarding whether the task type is a discrete or set-based item and the number of raw score points assigned to the each item, and the number of items that appear on a test form at each grade or grade span.

The new information in each table is the right-most column, “Recommended CBA Enhancements.” This column provides, for each task type, information regarding how ETS recommends enhancing the task type to take advantage of the capabilities of the CBA.

As noted in the beginning of Section 4, the PPT task types were developed with the goal of selecting tasks that could be readily repurposed and *enhanced* for CBA. Therefore, the proposed CBA blueprint closely resembles the current PPT blueprint. The main difference is that the draft CBA blueprint includes a number of newly designed TEI types and additional enhancements to redesign the assessment experience, as described above.

The blueprint contains task types for each assessment domain. In order to provide information about coverage of the California ELD Standards, aligned standards have been listed for each task type. In addition, the number of items and point value for each item is listed. Finally, the last column lists the redesigned enhancements to be implemented for existing ELPAC task types.

4.4. ITEM DEVELOPMENT CONSIDERATIONS

A key task in the development and implementation of the ELPAC CBA will be the creation of high-quality test items sufficient to support the initial launch and the ongoing refresh of the CBA. The following recommendations are designed to support that work.

- An item pool robust enough to support the launch of the ELPAC CBA and to support pool refreshment for the second operational development should be developed for the CBA field test.

  - ETS estimates that developing approximately 1,400 items, of which approximately 70% (or 980) would be multiple-choice or TEIs and approximately 30% (or 420) would be CR items, would comfortably support this requirement, as well as allowing for use of CBA items for public release purposes such as inclusion in tutorials and demos.
  - ETS estimates that approximately 80–90% of these CBA items would be new development and approximately 10–20% would be adapted ELPAC PPT SA items (with art revised to be in color and other adjustments as appropriate).
  - The ELPAC PPT items can be entered into the CBA item pool in various stages of the CBA item development process, prior to large-scale field testing.
  - In order to ensure that an adequate pool of operationally-ready items are available after the standalone field test, an overage of 10–15% should be developed in order to allow for attrition through the review process and/or through field testing.

- The evaluation process for the prototype/pilot should be used to make determinations about which task types should be included in the field test (and thus eligible for inclusion in the operational CBA).
  - The prototype/pilot represents a crucial opportunity to determine which CBA task types are likely to be most effective in capturing evidence about student language proficiency. For that reason, it would be appropriate to include in the prototype/pilot both task types that are expected to be successful and more ambitious task types that may or may not be successful. As part of the evaluation process for the prototype/pilot, decisions should be made about which task types should and should not be included in the blueprint and field test development effort.
- Before item development for the CBA field test begins, all documents supporting expected criteria for items for content, accessibility, and presentation on the test delivery system should be reviewed and confirmed.
  - The item development process represents a major investment of time and resources. In the case of the ELPAC CBA, stakes for this effort are particularly high because it is anticipated that new features related to construct coverage, accessibility, and presentation of items on the test delivery system will be implemented. For this reason, it is important that key documents such as the assessment claims, test blueprint, item writing guidelines, metadata definitions, accessibility framework, item authoring guidelines (including item style guide, graphics style guide, and display guide), and test delivery system configuration

(including all specifications for item presentation and user interface) are reviewed and confirmed before the item development effort is launched.

- Limits on embedded media file sizes for individual items, for test domains, and for the test as a whole should be in alignment with the CAASPP and should be documented in the ELPAC item authoring guidelines. The authoring guidelines are critical in achieving the optimal student testing experience by developing content that can be delivered over the Internet.

5. PSYCHOMETRIC CONSIDERATIONS

This section describes key psychometric issues and our recommended solutions to those issues in transitioning from the current ELPAC PPT to a CBA model. In developing psychometric recommendations, a major consideration was the accountability requirements based on the Every Student Succeeds Act (ESSA, 2015). It is critical to carefully plan a transition process that will support valid reporting of ELs' progress during the transition period when both the PPT and the CBA will be in use, and beyond. Thus, our recommendations largely concern the creation of a technically sound linkage between ELPAC PPT scores and ELPAC CBA scores.

In this section, we first present our overall recommendation on the use of linear vs. adaptive format for the ELPAC CBA. One of the prominent features of CBAs is that they support adaptive testing models. However, a number of factors need to be carefully reviewed to make appropriate decisions on a CBA format for the ELPAC, as described below. In addition, we present key psychometric-related recommendations to be considered for the ELPAC CBA development stage, organized around field testing, the first operational administration stage, and accountability reporting.

Recommendation Regarding the Linear vs. Adaptive CBA Format for the ELPAC

- The ELPAC CBA should use linear forms, rather than employing an adaptive model (such as a multi-stage test [MST] or a computer-adaptive test [CAT]). While an adaptive model can have important advantages (e.g., shorter administration times and immediate scoring and reporting), these advantages are not necessarily realized across all contexts. There are several reasons that an adaptive model is not a good fit for the ELPAC.
    - The ELPAC task types developed to assess the California ELD Standards are not well suited for adaptive testing, and the time savings from an adaptive ELPAC CBA would be modest.
        - The ELPAC consists of four separate domains (Listening, Speaking, Reading, and Writing), each of which would require, in effect, a separate adaptive model, involving considerable complexity and cost.
        - The ELPAC Listening and Reading domains contain a high proportion of set-based items (those based on a common stimulus), which increases the complexity of appropriate assignment of specific items.

    - The ELPAC Speaking and Writing domains consist entirely of CR items. Immediate scoring is prerequisite to apply an adaptive model during the assessment. However, the feasibility of implementing AI scoring models for all CR task types is, as described in Section 3.3, not yet known.
    - An adaptive model requires that highly reliable item parameter estimates be obtained well in advance of the operational testing. Considering the heterogeneous characteristics of the target EL population, all items would need to be piloted with a representative, large-scale sample with a careful sampling plan. The acceptable accuracy level of a priori item parameters is challenging to obtain.
  - An adaptive test design would require a considerably larger item pool in order to provide highly discriminating items in all domains at all difficulty levels, which would considerably increase the item development volume and costs.
  - Introducing an adaptive model for the ELPAC SA CBA would complicate the recommendation of maintaining score relationships between the ELPAC PPT and the ELPAC CBA.

Recommendations for the ELPAC CBA Field Testing Stage

The major psychometric goals to achieve during the large-scale field testing for the transition to CBA include:

1. Ensuring items in the CBA pool are of reasonable statistical quality;
2. Establishing the dimensionality of the ELPAC CBA; and
3. Creating a linkage between the ELPAC CBA and the ELPAC PPT.

Specific recommendations to support these goals are described below, with rationales provided as appropriate.

- The ongoing embedded pretesting model to support refresh of future editions of the CBA should support a relatively large number of unique field test versions for the CBA (e.g., twelve unique CBA field test forms compared to six for the PPT).
  - A CBA format allows a much more flexible assembly of field test forms than does a PPT. Spreading the embedded pretest items over more forms lessens the burden of responding to pretest items for each individual student (as each student will take about half the number of field test items compared to the ELPAC PPT), and reduces the incremental time needed to support field test items in

each participating school. In addition, this process is likely to gather more robust data by obtaining responses from a broader pool of students.

- It is not necessary to develop an independent breach form for the ELPAC CBA.
    - Security breaches for CBA are considered very differently compared to PPT because CBAs have no test materials to distribute and collect physically. It is also much easier to allow for a dynamic plan to lengthen or shorten a field test domain(s) in an operational administration when needs arise due to item breach and other unforeseen circumstances.
- A detailed sampling plan for the ELPAC CBA field test should be developed, calling for recruitment of a student sample that reflects California's diverse background of ELs (e.g., urban to rural, varying levels of computer familiarity), including students with disabilities. This sample should consist of students who will be taking the ELPAC PPT for accountability purposes.
    - A sampling plan is necessary to support adequate planning for the field test. Including a sample that reflects the diversity of California's EL population, including students with disabilities, will help to ensure the validity of the sample and ensure the implementation of several of the recommendations made in Section 3.2.
- The majority of the items used in the ELPAC CBA field test should be newly developed CBA items. PPT items that have been adapted to the CBA format should be field tested during the CBA field test study with caution due to risks related to item exposure.
    - It is anticipated that the ELPAC CBA field test will overlap with the operational ELPAC PPT administration. In order to maintain item security and the validity of the PPT operational scores, an item use plan should be drafted to clearly delineate the possible role of repurposed PPT items in the CBA field test. The repurposed PPT items in the CBA field test may be able to provide information about the mode effect between PPT and CBA, so their inclusion might be helpful. It is sensible to require that the CBA field test be administered after the administration of the operational PPT to ensure that the validity of scores from the operational administration is not compromised.
- Data from the ELPAC CBA field test should be used to conduct item analyses and dimensionality analyses to compare PPT and CBA item performance.
    - Item analyses are basic analyses that need to be conducted to ensure that items are of reasonable difficulty and quality. Following recommendations in Section 4,

CBA task types are expected to feature several enhancements, taking advantage of the capabilities of a digital test delivery system. As a result, it is necessary to replicate all the psychometric analyses conducted during the PPT field test in evaluating the CBA items. Although integrated-skills items (those that combine multiple domains) are included in the ELPAC PPT, improvements in the presentation of integrated items on the CBA may impact how students respond to them. Therefore, ETS recommends repeating the dimensionality study for the ELPAC CBA.

- A concordance study should be conducted to link the ELPAC CBA to the ELPAC PPT, with students taking both the ELPAC CBA field test and the ELPAC PPT to establish the relationship between the two tests in terms of their score scales. Creating a linkage between ELPAC PPT and CBA will also allow the CDE to report student progress over time during the CBA transition period.
    - It is anticipated that a small part of the target population will not be able to access ELPAC online for various reasons, as is the case for the CAASPP program. It is desirable to have the ability to offer the ELPAC PPT to an appropriate subset of students even after the ELPAC CBA becomes operational (e.g., for newly arrived students who have no familiarity with computers). This could be done by keeping one or more existing editions of the ELPAC PPT in print after the launch of the ELPAC CBA. A concordance table to link ELPAC PPT and CBA will be instrumental for schools or students who have difficulty taking the ELPAC CBA.

      It is assumed that ELPAC PPT operational administration will take place around the same time that the voluntary ELPAC CBA field test study is being conducted. The field testing stage is the best time to collect data for the concordance study because the ELPAC PPT will still be mandatory at that point. Note that the final linkage between ELPAC CBA and PPT will be established when the ELPAC CBA scales are finalized, after the first ELPAC CBA operational administration.

Recommendations for the First Operational Administration Stage

The chief psychometric goals to accomplish around the first operational administration for the CBA include:

1. To construct the ELPAC CBA vertical scales;
2. To use target population data to derive the ELPAC CBA reporting scales;
3. To support a standard setting/threshold score review for the ELPAC CBA; and

4. To establish a linkage between the ELPAC PPT IA and CBA SA.

Specific recommendations to accomplish these goals are described below, with rationales provided as appropriate.

- The ELPAC CBA test blueprint should be reviewed and confirmed or revised after the field test results have been analyzed. Items/item types that are not performing as expected would be revised and field tested or discarded.
- Items surviving from the field test should be used to construct operational CBA test forms that are aligned to the test blueprint. Only items exhibiting reasonable statistical quality should be included in the operational test.
- Post-equating analyses based on the full target population should be conducted for the first operational administration of the CBA. A post-equating design for the first census operational administration is recommended to set the ELPAC CBA reporting scale, including ELPAC common (vertical) scales. It should be noted that the post-equating design will likely introduce delays in score reporting for the first operational administration.
    - It is best practice to derive the reporting scale based on the full population of students with a test form that reflects the operational test blueprint. Biases might be introduced if item parameter estimates are based on a sample of student responses to field test forms that might not fully represent the test blueprint.
- In the first operational assessment and each subsequent operational administration, a sample of students should be given additional field-test items to support refreshing of the item pool for future ELPAC CBA test editions.
    - For security and item exposure/student experience reasons, the ELPAC item pool should be refreshed annually. Given the ELPAC grade span organization, some students will take the same test two—or even three—years in a row; refreshing forms year to year will provide a more valid and engaging testing experience for these students. While the historical CELDT and ELPAC refresh rate of 30% per year may work as a minimum, a higher refresh rate (perhaps 50% or 60%) would be desirable in order to decrease the possible impact of student familiarity with test items, if budgets can support it. The familiarity impact is only likely to be an issue in cases where a single test edition is used across multiple grades: 3–5, 6–8, 9–10, and 11–12.
- In addition to field testing to support item pool expansion, a separate sample of students should be given additional items from the grade/grade span below to construct the common (vertical) scale for the ELPAC CBA.

  - Constructing common (vertical) scales with operational data allows population-level student achievement data at each grade/grade span to be used in evaluating such scales.

- Once the ELPAC CBA reporting scale has been finalized, the ELPAC CBA and the ELPAC PPT linking should be completed, reflecting the final CBA scales.
  - The link between CBA and PPT is established using field test data. The linkage can be updated by placing the preliminary Item Response Theory (IRT) parameters obtained in the CBA field test onto the final ELPAC CBA scale, therefore placing the whole pool of items developed in the field test stage onto the final reporting scales.
  - A study should be conducted to establish a linkage between the ELPAC PPT IA and the ELPAC CBA SA by using data available in the CALPADS system. The results of this study can be used as a piece of evidence for the convergent validity of the ELPAC.
    - Students who have taken the IA during the same school year as the first operational CBA SA will be used to establish a predictive relationship between the two assessments. The results of this study will allow stakeholders to understand the relationship between the IA and SA. It will serve as a piece of validity evidence for the ELPAC assessment system (both IA and SA).

Recommendations Regarding Score and Accountability Reporting

One of the important goals for ELPAC is to provide stakeholders meaningful scores for both local and state level accountability. To ensure that the score reporting information for the ELPAC CBA (including the Performance Level Descriptors [PLDs], the score scales, and the reported scores) are meaningful and that the relationship between all versions of the ELPAC (PPT IA, PPT SA, CBA SA) are understood and can be communicated effectively to the public, the following are recommended.

- The existing ELPAC general (policy level) PLDs should be remain the same.
  - Both the ELPAC PPT and CBA are designed to align with the California ELD Standards. The assumption follows that there is no change in overall expectations for students' performance.

- The existing ELPAC grade- and grade-span domain-specific PLDs should be reviewed and updated as appropriate in light of the test blueprint for the ELPAC CBA SA.
    - The original specific PLDs established with the ELPAC PPT considered both the IA and SA designs. Given the possible changes to task types and to the test blueprint for the CBA, it is appropriate to examine whether the PLDs are still applicable or should be revised to reflect differences in evidence provided by the CBA compared to the PPT. This decision should be made after the ELPAC CBA field test.
- Engage stakeholders in reviewing or updating the ELPAC CBA SA performance threshold scores.
    - If the PLDs remain the same, a review of threshold scores should be conducted. The CDE and selected stakeholders or advisors could review threshold scores, using the statistical linkage between the ELPAC PPT SA and ELPAC CBA SA. The review may include, but not be limited to, percentages of students classified in each performance level for the total group and in selected subgroups. Analyses should take into account standard errors of measurement and sample sizes, and may potentially lead to a change in threshold scores being recommended.
    - If the ELPAC-specific PLDs are updated substantially to reflect the CBA, a standard setting study should be conducted after the first operational administration, when the ELPAC CBA scales have been finalized, to provide threshold recommendations aligned with the revised grade- and grade-span domain-specific PLDs.
- Samples of reference groups for English Only (EO) and Fluent English Proficient (FEP) students should be recruited to participate in the first operational ELPAC CBA administration in order to establish a data source capable of informing the scaling and standard-setting.
    - It is important to understand the performance of both EL and EO/FEP (reference group students) when setting threshold scores such that expectations for ELs not exceed those for EO/FEP students.
- The ELPAC CBA and PPT linking study results will also be included in the standard-setting design.
    - Standard setting/cut-score review panelists should review the CBA threshold scores in light of the previously established PPT threshold scores. Depending on

the latest requirement of ESSA and California's accountability policies, it may also be appropriate to update the ELPAC CBA reporting scores to facilitate various requirements for both federal and state reporting.

6. INFORMATION TECHNOLOGY CONSIDERATIONS FOR THE ELPAC CBA

As described in the previous sections, the transition of ELPAC to a CBA model provides an opportunity to take advantage of the benefits that a computer-based platform allows, including faster score turnaround times, improved monitoring and reporting, improved test security, broader accessibility and accommodations support, and the ability to support TEIs.

This section elaborates more on several key areas in which the ELPAC can take advantage of a CBA format from an IT perspective, including recommendations for systems solutions to support the ELPAC CBA.

Faster Score Turnaround Time

With a CBA, days or even weeks can be saved in providing the official scores in cases where all test parts are administered online since student responses are sent immediately upon test completion to the scoring engines by the online test delivery system. The scoring system is expected to manage test parts and produce the final score once all test parts are received. These test parts include responses from the test delivery system and human scores if a particular test part contains constructed responses.

For the CBA, the ELPAC can be aligned with the CAASPP results processing chain and will benefit from established quality control checks that verify no student response is lost in transmission or during scoring.

State-level Service Level Agreements: Monitoring and Reporting

CBA test results are tracked and reconciled between test delivery system and back-end scoring processes. The delivery/processing system should track individual student responses/results progression and ensure that results for each student meet scoring state-level service agreements (SLAs). SLA monitoring and reporting is available at the start of the administration window and until all student responses are scored and reported. Distinct scoring and reporting SLAs should be established for those students who take the Writing domain on paper.

Online Reporting System

For the ELPAC CBA, individual student scores can be made viewable online via an online reporting system (ORS). All student scores can be sent to the ORS to be available for aggregate or student-level reporting within the Test Examiner's jurisdiction. Further disaggregation of student results is possible via a predefined list of demographic data definitions. Student demographics data received daily from CALPADS can be updated throughout student active enrollment. Educators with appropriate access may exercise the ORS roster upload functions to create classroom-level aggregation.

Unified ORS can serve as a one-stop reporting application, in which both CAASPP and ELPAC test scores can be viewed by LEAs that have permission to access both programs. ORS configuration will need to expand to accommodate ELPAC-specific reporting requirements.

Test Security

The ELPAC CBA can take advantage of the security protocols currently used in the CAASPP assessment platform, including the unified test registration system, test delivery system, and ORS applications, to ensure a highly secure testing experience that employs industry-standard measures both during the test delivery and transmission of content or results.

The test delivery system currently used in California provides a secure browser that locks down the student's desktop by preventing the student from navigating away from the test, blocking certain external applications, and disabling keystrokes that can threaten the security of the test. For example, the secure browser disables screen shots and navigation and prevents test takers from viewing the source and opening the "taskbar." Any student or item data communicated to and from the test delivery system uses industry-standard encryption to enable secure content delivery. The secure browser software continuously monitors other activity on the computer for possible threats and terminates testing if it detects a threat.

All interfaces for the ELPAC CBA should ensure that data is encrypted while at rest and in transit. Encryption at rest primarily applies to any data files that reside on a server that uses the secure file transfer protocol (SFTP) waiting to be retrieved. The test delivery system must use encryption (in transit and at rest) using a Federal Information Processing Standard (FIPS) 140-2 validated solution (128-bit AES encryption or better) to protect confidential information handled by the system, including student registration information, student identifiable results information, test items and packages, and other information as identified by the CDE Information Security Officer (ISO).

The test delivery system should provide a way to identify item content when a photo of a screen is taken during test administration.

Common Application of Student Accessibility Requirements

Moving to an online delivery system allows for broader support of accessibility requirements. All online accommodations currently defined by the Smarter Balanced Assessment Consortium, and used in the online CAASPP assessments, can be supported in an online ELPAC test where task types/constructs are the same, allowing for consistent support across programs. In addition, the Individual Student Assessment Accessibility Profile (ISAAP) tool, currently available with the CAASPP system, could be leveraged for the ELPAC. As discussed in Section 3.2, a process should be established to provide guidance in cases where it is appropriate for the ELPAC process to differ from that of the CAASPP.

Section 3.2 recommends a process for evaluating which of the CAASPP tools, supports, and accommodations should be made available to students taking the ELPAC CBA, and what additions or modifications to the CAASPP accessibility requirements may be appropriate for the ELPAC.

Response Capture for Writing Domain

Written responses for grades 3–12 will be collected using the test delivery system and sent for scoring in electronic format. The K–2 Writing domain will continue to be administered on paper and can be available as a paper alternative for grades 3–12. LEAs will order test materials via the CAASPP order management system and ship test materials back to the vendor for scanning and human scoring. Final scoring and reporting will be done once all required test parts are completed by the student and received in the scoring system. Reconciliation and scoring rules will need to be clearly defined for those students that are missing one or more parts of the test.

In addition, resolution processes will be required to address issues like incorrect SSID gridded on the answer document. Since scores are reported only after all test parts have been received, scoring and reporting SLAs will need to be adjusted accordingly when the paper test part is administered for grades K–2 or optionally selected for grades 3–12.

Digital Voice Capture for Speaking Domain

The ELPAC CBA can support the capture of digital voice responses, which can then be transmitted to raters over a distributed network for human scoring. The technical challenges involved in capturing and transferring the digital voice files can be addressed by transferring these audio files from the test delivery system to the scoring vendor as part of the test results.

See Section 3.3 for a discussion of the advantages of the digital voice capture model in terms of reliability of scoring. Additionally, digitally captured spoken responses can be used to assess the feasibility of AI scoring as discussed in Section 3.3.

Use of Technology-Enhanced Items

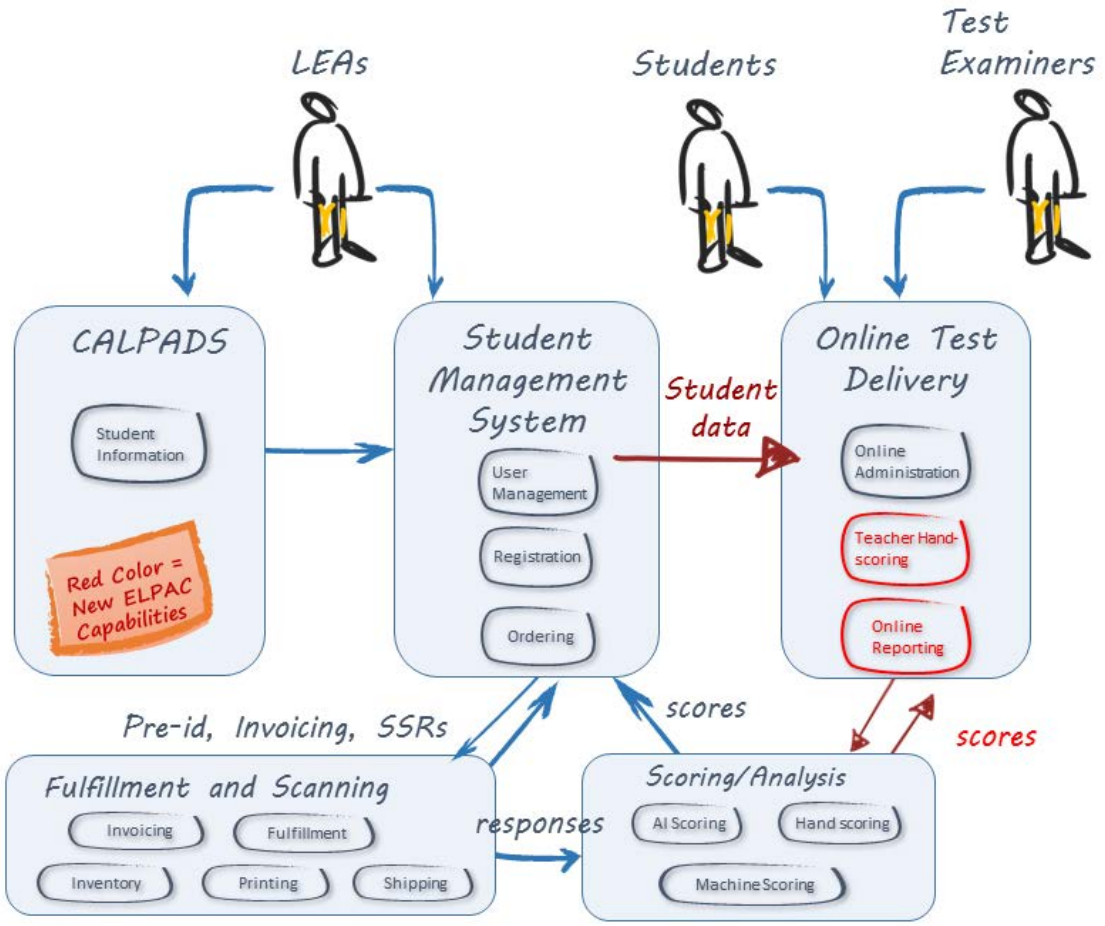
As described previously in Section 3.1 and Section 4, the CBA test delivery system can support a range of innovative item types not available in a PPT. The TEI types/response formats recommended in Section 4 and in Appendix D have been selected as appropriate to the assessment of the California ELD Standards and are also expected to be supported by the current CAASPP platform. However, confirmation that all item types/response formats as configured for the ELPAC CBA will be supported by the CAASPP platform should be a key step in the test development process. Item type, response collection, scoring gap analysis and remediation effort should dictate which items are supported by the CAASPP platform.

In order to support development of TEIs and of all ELPAC CBA item types, the systems provider should make available an item previewer integrated with the ELPAC item bank supporting all ELPAC CBA item types. This previewer should be made available before item development for the CBA field test begins (in order to support review and sign off of the planned presentation of the item type) and during the item review process so that reviewers are able to evaluate item content and item functionality simultaneously.

Recommendations for Systems Solution for the ELPAC CBA

In consideration of the major benefits described above, the systems solution for the ELPAC CBA should be practical and efficient, utilizing existing California systems (e.g., CAASPP, CALPADS) wherever appropriate. An overall concept of a recommended system can be seen in Figure 6.1. Some additional recommendations pertaining to specific elements of the solution depicted is described below Figure 6.1.

Figure 6.1: Recommended Solution Concept for the ELPAC CBA



- In order to ease the transition to/administration of the online test delivery, the current CAASPP test delivery solution should be leveraged as much as possible. For instance, the current assessment platform is fairly configurable using test level color coding to prevent test examiner confusion when launching an assessment. Additional analysis should be performed to determine what other configurable features are missing to support the K–2 student population in particular. This may include additional configurations, such as more user-friendly test navigation buttons, which are currently not supported in the TDS. Streamlined TDS mode is available as an accommodation today but is generally documented in the student IEP and is designated on a student-by-student basis.

  - In addition to being able to take advantage of the current infrastructure already in place in schools, from a user experience perspective, it is beneficial for the LEAs and students to use a unified test delivery system. Many LEAs administer both CAASPP and the ELPAC assessments, and there are ELs who are also taking Smarter Balanced Math/ELA tests as well as the new California science assessment. Existing users are knowledgeable and comfortable with the current test delivery system and are familiar with the current test login and check-in processes.

    While the CAASPP system has been widely used at grade 3 and above, it is worth noting that the largest numbers of students taking the ELPAC are in grade 2 or below, grades at which the CAASPP system has not been deployed. In addition to ensuring that schools have the appropriate infrastructure and bandwidth to support the ELPAC for K–2, a thorough review of the proposed ELPAC task types for these grades should be conducted as part of the prototype/pilot effort to ensure that any complications are addressed before the large-scale item development effort begins.

- In addition to a common test delivery system, a unified test registration system should be used. Important considerations for a unified test registration system are the yearly data rollover and system outages. Aligning transition and downtime activities across the ELPAC and the CAASPP assessments will make unification of test registration system possible.
- The CALPADS daily enrollment file should be processed and students assigned test registrations based on demographics data received from CALPADS. Summative registrations should be derived based on the EL designation.
  - Users common to both the CAASPP and ELPAC programs will have access to the student population and be able to search, designate initial participation, add student accommodations, and view student score reports in the registration system.
- Using unified test delivery and registration systems will also allow for efficiencies during the development cycle. Existing system features and code base should be leveraged and reused wherever appropriate.
  - This approach will help to contain development costs and maintain the timeline. However, shared components require coordination between the programs in terms of feature development, outage periods, and transition timelines. Yearly student rollover can occur on different schedules. Additional discussion is

required to assess options for user transition between administration years.

- Deployment of the ELPAC CBA test delivery system and online reporting system should be implemented in coordination with the existing CAASPP test delivery system. Appropriate capacity analysis should be performed to ensure that the current infrastructure can support the additional testing volume.
    - The existing test delivery system already provides all of the above platform components as part of the CAASPP program. There will be additional configuration and onboarding steps required to launch the ELPAC online. The CBA testing vendor should perform all scoring activities and generate reporting deliverables for the SA.
- As with CAASPP, the test registration and management system should be used to provide LEAs with the official student score report (SSR). SSRs should be viewable and printable. The CBA testing vendor should be required to print and ship a copy of SSRs to the LEAs for the SA.
- It is recommended that the assessment platform provide student result files (i.e., LEA downloadable files) to the LEAs via downloadable files from the test registration and management system. Complete student and aggregate data files should be delivered to the CDE. ETS recommends creating a master file layout that incorporates both programs' data elements. The CDE may choose to receive a separate file specific to ELPAC or extend the current CAASPP layout with additional ELPAC record types.
- A data retention policy should be established for the ELPAC CBA. Given that the CAASPP data retention policy is currently three years, it seems appropriate that three years' worth of data for the ELPAC CBA should be made available on all reporting deliverables. Archiving student results beyond three years is possible if deemed necessary by the CDE.

7. PROGRAM MANAGEMENT CONSIDERATIONS

Program management considerations for the ELPAC CBA are based on several assumptions about the future suite of CDE assessment programs.

- The CDE envisions CBA as the preferred delivery mode except where there is a compelling case that students are best served through other modes of delivery.
- It will remain the CDE preference to deliver assessments on a common platform across programs.
- The CDE will continue to look for opportunities to reduce demands on LEA staff through use of technology and common processes.
- The intent of the transition to the ELPAC CBA is to take advantage of both the measurement and logistical advantages of CBA.
- The ELPAC PPT delivery will continue up to the point that the operational CBA will be delivered; there will be no skipped year or census field test.
- LEA technology assets will be adequate to support the ELPAC CBA.

General Approach to Managing the Transition to CBA

Effective development and deployment of the ELPAC CBA will best be served by being closely aligned with the existing ELPAC paper-based program to enhance a smooth and stable migration between the two modes. However, it will be important for the testing vendor to have a discrete group specifically focused on CBA development and implementation. Similarly, we would recommend that the CDE designate a CBA Lead from their ELPSA team.

Project Management best practices are critical. The Project Management practices for the ELPAC CBA (e.g., planning, meetings, deliverables tracking, and communications) should be consistent with those currently in place for the ELPAC PPT.

Importance of Communications with Local Educational Agencies

Moving ELPAC to CBA will mark a significant shift for the California English language proficiency exam which since the inception of CELDT has been paper-based. While a number of ELPAC test examiners will have experience with CAASPP, many will not.

Furthermore, ELPAC will potentially use different technologies to address the Speaking domain, which will be unfamiliar even to those who have administered CAASPP.

Education and training for LEA staff will be critical, both typical pretest activities but also earlier familiarization sessions to orient LEA staff to the pending changes and significant new aspects anticipated with the CBA deployment. Such orientation activities should extend beyond LEAs to the various stakeholder groups involved with this key California population. These latter audiences are most properly addressed by the CDE with support of the contractor while LEA sessions would best be joint ventures, blending CDE policy with contractor procedures.

Thus, it is important that a formal plan for communicating with LEAs and for stakeholder involvement should be developed early in the planning stages for the ELPAC CBA. Progress against this plan should be measured and reported on regularly (e.g., quarterly).

Scoring of the Speaking Domain

The most salient consideration for scoring the Speaking domain is whether to continue local scoring with LEA staff or use digital voice capture to support distributed online scoring, whether by humans or via AI scoring. A recommendation to support AI scoring is provided in Section 3.3; some additional notes from a program management perspective are provided here.

As with any design consideration, there will be tradeoffs involved in any decision. Table 7.1 presents advantages and disadvantages of each approach.

Table 7.1: Approaches to Scoring of Spoken Responses

Scoring Model	Advantages	Disadvantages
Local	• Real-time results • Lower CDE costs	• Lack of standardization • Potential errors in data entry of ratings • Higher LEA costs
Online	• Standardized scoring • Lower LEA costs • Less exposure to Test Examiner error during ratings entry	• Results delay (shorter if AI scored, longer if human scored) • Higher CDE costs • Potential data-capture failures not caught in real time

Considerations of whether or not to utilize digital voice capture include both efficacy and cost. California offers such a large EL population that questions of scale tend to lean towards the promises of AI scoring. As discussed in Section 3.3, AI scoring in the Speaking domain continues to evolve. AI scoring is likely to advance and become more broadly applicable and cost-effective over time. Since the CDE is now looking forward with the ELPAC program, we recommend utilizing audio capture to establish preconditions so that the program can benefit from future advances in AI capabilities.

Proposed Timeline

The proposed high-level timeline presented in Appendix E assumes a redesign leading to an ELPAC which takes advantage of the various capabilities described in Section 4, including TEIs. Precursor CDE activities are estimated to provide a potential start date. The timeline also assumes that the ELPAC CBA will deploy all affected grades simultaneously. We see little advantage to a grade-span-staged approach since it would extend the transition period, increase costs, and offer minimal advantages.

A high-level schedule for the development and deployment of the ELPAC CBA could be:

- 2017-18 Procurement process by the CDE
- 2018-19 Test and systems design, piloting
- 2019-20 Test production and large-scale field testing
- 2020-21 Operational deployment of the CBA

A timeline with detailed activities can be found in Appendix E.

Cost Implications

Cost considerations for the ELPAC CBA cover two aspects: development and operations. Previous sections of this report have described various changes to the ELPAC that are recommended in order to secure the full benefits of the CBA. While a degree of consistency with the current PPT is anticipated, the range of changes recommended for the CBA will require large-scale activities, including updating test design documentation, piloting/prototyping new or revised task types, field testing, standard setting, etc. These activities would overlap with the current ELPAC administration for two years and involve one-time costs on top of the continuing ELPAC operations.

Once the ELPAC CBA deploys, operational costs will still be split between online (for the SA) and paper (for the IA and for the SA Writing domain at K–2). If online scoring for Speaking is selected, the CDE will incur additional scoring costs. The CBA SA will represent the best opportunity for evolving savings due to efficiencies in leveraging a common CDE delivery platform and the developing benefits of AI scoring.

As the IA is recommended to remain on paper, ETS does not recommend that there be a new development effort for the IA while the CBA SA is being developed. Therefore, costs for the IA will likely be modest and quite similar to those for use of the IA in a typical operational year.

REFERENCES

Abedi, J., & Ewers, N. (2013, February). Smarter Balanced Assessment Consortium Accommodations for English language learners and students with disabilities: A research-based decision algorithm. Retrieved from https://portal.smarterbalanced.org/library/en/accommodations-for-english-language-learners-and-students-with-disabilities-a-research-based-decision-algorithm.pdf

Albus, D., & Thurlow, M. L. (2008). Accommodating students with disabilities on state English language proficiency assessments. *Assessment for Effective Intervention*, *33*(3), 156-166.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22(3),* 301-320.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing.* Washington, DC: AERA

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1)*,* 1-42.

Bailey, A. L. (2008). Assessing the language of young learners. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education, Vol. 7: Language Testing and Assessment* (pp. 379-398). New York, NY: Springer.

Bailey, A. L., Heritage, M., & Butler, F. A. (2014). Developmental considerations and curricular contexts in the assessment of young language learners. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., pp. 423-439). Chichester, UK: John Wiley & Sons, Inc.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *The Journal of Technology, Learning, and Assessment, 6*(9), 4-38.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests.

Language Testing 27(3), 355-377.

Bonk, W. J. (2016). *Scoring TELL (Test of English Language Learning).* Menlo Park, CA: Pearson. Retrieved from http://cdn2.hubspot.net/hubfs/559254/TELL/TELL_White_Paper_4_-_Scoring_TELL.pdf?t=1481317295078

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16*(3), 191–205.

California Department of Education. (2016). Matrix Two: Matrix of Variations Accommodations, and Modifications for Administration of the California High School Exit Examination, California English Language Development Test, and the Physical Fitness Test. Sacramento, CA: Author. Retrieved from https://www.cde.ca.gov/ta/tg/ai/pftaccessibilityresources.asp

California Department of Education. (2014). *California English Language Development Standards: Kindergarten Through Grade 12*. Sacramento, CA: Author. Retrieved from http://www.cde.ca.gov/sp/el/er/documents/eldstndspublication14.pdf

CAST. (2011). Universal design for learning guidelines version 2.0. Wakefield, MA: Author. Retrieved from http://www.udlcenter.org/aboutudl/udlguidelines

Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning & Technology, 5(2).* 95-98.

Cheng, J., D'Antilio, Y. Z., Chen, X., & Bernstein, J. (2014). Automatic assessment of the speech of young English learners. *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*. *Association for Computational Linguistics* (pp. 12-21). Baltimore, Maryland.

Choi, S. W., & Tinkler, T. (2002, April). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K–12 setting*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Christensen, L. L., Albus, D. A., Kincaid, A., Christian, E., Liu, K. K., & Thurlow, M. L. (2014a). *Including students who are blind or visually impaired in English language proficiency assessments: A review of state policies*. Minneapolis, MN: University of Minnesota, Improving the Validity of Assessment Results for English Language Learners with Disabilities (IVARED).

Christensen, L. L., Albus, D. A., Kincaid, A., Liu, K. K., Christian, E., & Thurlow, M. L. (2014b). *Including students who are deaf or hard-of-hearing in English language proficiency assessments: A review of state policies*. Minneapolis, MN: University of Minnesota, Improving the Validity of Assessment Results for English Language Learners with Disabilities (IVARED).

Christensen, L. L., Albus, D. A., Liu, K. K., Thurlow, M. L., & Kincaid, A. (2013) *Accommodations for students with disabilities on state English language proficiency assessments: A review of 2011 state policies*. Minneapolis, MN: University of Minnesota, Improving the Validity of Assessment Results for English Language Learners with Disabilities (IVARED).

Clark, R. C., & Mayer, R. E. (2016). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. San Francisco, CA: John Wiley & Sons, Inc.

Douglas, D. (2013). Technology and language testing. In C. A. Chappelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-7). Oxford, UK: Blackwell Publishing, Ltd.

English Language Proficiency Assessment for the 21st Century (ELPA21). (n.d.). ELPA21 assessment program interactive demo. Retrieved from ~~https://elpa-practice.nextera.questarai.com/student/WebClient/PracticeTest/~~ [Note: the preceding Web address is no longer valid.]

English Language Proficiency Assessment for the 21st Century (ELPA21). (2015). The ELPA21 field test implementation report. Washington, DC: Author. Retrieved from http://www.elpa21.org/sites/default/files/Field%20Test%20Implementation%20Report-September%202015%20(1).pdf

English Language Proficiency Assessment for the 21st Century (ELPA21). (2016a). ELPA21 Accessibility and accommodations manual, 2015-2016 school year. Washington, DC: Author. Retrieved from http://www.elpa21.org/sites/default/files/Accessibility%20and%20Accommodations%20Manual_SY15_16.pdf

English Language Proficiency Assessment for the 21st Century (ELPA21). (2016b). ELPA21 Blind and low vision information sheet. Washington, DC: Author. Retrieved from http://www.elpa21.org/sites/default/files/Blind%20and%20Low%20Vision%20Information%20Sheet.pdf

Evanini, K., Hauck, M. C., & Hakuta, K. (in press). Approaches to automated scoring of speaking for English language proficiency assessments. (Research Report). Princeton, NJ: Educational Testing Service.

Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). *Automated scoring for the TOEFL Junior® Comprehensive Writing and Speaking Test.* (Research Report No. RR-15-09). Princeton, NJ: Educational Testing Service.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and Applications of the Intelligent Essay Assessor. In M. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York, NY: Routledge.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language testing*, *20*(4), 384-408.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing, 19*(2), 133-167.

Gruba, P. (2014). New media in language assessments. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., pp. 997-1012). Chichester, UK: John Wiley & Sons, Inc.

Guzman-Orth, D. (2014, April). *PARCC Item Development: Usability of New Item Types for English Learners*. In C. Laitusis (Chair), *Research and Development to Ensure Accessibility for the Partnership for Assessment of Readiness for College and Careers Assessments.* Symposium conducted at the meeting of the National Conference on Student Assessment, Council of Chief State and School Officers Annual Conference, New Orleans, LA.

Guzman-Orth, D., Laitusis, C., Thurlow, M., & Christensen, L. (2016). Conceptualizing accessibility for English language proficiency assessments. (Research Report No. RR-16-07). Princeton, NJ: Educational Testing Service.

Guzman-Orth, D., Wolf, M., King, T., & Tolentino, F. (2016, April). *Illustrations Glossaries for English Learners: Findings from Cognitive Labs*. In R. Kachchaf (Chair), *Designing Illustrations Glossaries for Special Populations in a National Large-Scale Mathematics Assessment*. Symposium conducted at the Annual Meeting of the American Educational Research Association, San Antonio, TX.

Hauck, M. C., Wolf, M. K., & Mislevy, R. (2016). *Creating a next-generation system of K–12 English learner language proficiency assessments* (Research Report No. RR-16-06). Princeton, NJ: Educational Testing Service.

Hassanali, K.-N., Yoon, S.-Y., & Chen, L. (2015). Automatic scoring of non-native children's spoken language proficiency. *Proceedings of the Workshop on Speech and Language Technology for Education (SLaTE 2015)*, 13-18.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation of reading test performance. *The Journal of Technology, Learning and Assessment*, *3*(4). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1657/1499

Hidalgo, P., Sato, E., Suzuki, M. (2015). *Pearson TELL (Test of English Language Learning): New Technologies for Assessing Language Learners*. Menlo Park, CA: Pearson. Retrieved from http://www.pearsontell.com/hubfs/TELL/TELL_White_Paper_on_Test_Design_and_Automated_scoring-LW-5-22-15-1.pdf?t=1481317295078

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *The Journal of Technology, Learning, and Assessment, 5*(2), 4-49.

IMS Global Learning Consortium. (2016). *Accessible Portable Item Protocol® (APIP®)*. Retrieved from https://www.imsglobal.org/apip/index.html

King, T. (2014, April). *PARCC Item Development: Usability of New Item Types for Students with Disabilities*. In C. Laitusis (Chair), *Research and Development to Ensure Accessibility for the Partnership for Assessment of Readiness for College and Careers Assessments.* Symposium conducted at the meeting of the National Conference on Student Assessment, Council of Chief State and School Officers Annual Conference, New Orleans, LA.

Laitusis, C. C., Buzick, H., Stone, E., Hansen, E., & Hakkinen, M. (2012). Literature review of testing accommodations and accessibility tools. Commissioned report for the Smarter Balanced Assessment Consortium.

Ling, G., & Bridgeman, B. (2013). Writing essays on a laptop or a desktop computer: Does it matter? *International Journal of Testing, 13*, 105-122.

Liu, K. K., & Anderson, M. (2008). Universal design considerations for improving student achievement on English language proficiency tests. *Assessment for Effective Intervention, 33*, 167–176.

Liu, K. K., Goldstone, L. S., Thurlow, M. L., Ward, J. M., Hatten, J., & Christensen, L. L (2013). *Voices from the field: Making state assessment decisions for English language learners with disabilities*. Minneapolis, MN: University of Minnesota, Improving the Validity of Assessment Results for English Language Learners with Disabilities (IVARED).

Martin, R. (2009). Utilising the potential of computer-delivered surveys in assessing scientific literacy. In F. Scheuermann & J. Bjórnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 172-177). Luxembourg: Office for Official Publications of the European Communities.

Mayer, R. E. (2004). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13*, 125-139.

Mitchell, C. (2015, September 22). As ELL tests move online, educators hope for better gauge of skills. *Education Week*. Retrieved from http://www.edweek.org

National Association for the Education of Young Children. (2012). *T*echnology and interactive media as tools in early childhood programs serving children from birth through age 8 [Position Statement]. Retrieved from ~~http://www.naeyc.org/files/naeyc/PS_technology_WEB.pdf~~ [Note: the preceding Web address is no longer valid.]

Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication 30*, 83-93.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal, 93*, 836-847.

Odo, D. M. (2012). Computer familiarity and test performance on a computer-based cloze ESL reading assessment. *Teaching English with Technology, 12*(3), 18-35.

Page, E.B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, *48*, 238-243.

Partnership for the Assessment of Readiness for Careers and Colleges (PARCC). (n.d.). PARCC Practice Tests. Retrieved from https://parcc.pearson.com/practice-tests/

Partnership for the Assessment of Readiness for Careers and Colleges (PARCC). (2016). What's new in the fifth edition of the PARCC accessibility features and accommodations manual? Retrieved from http://avocet.pearson.com/PARCC/Home#10637

Partnership for the Assessment of Readiness for Careers and Colleges (PARCC) (2014). PARCC field test: Lesson learned. Retrieved from http://www.parcconline.org/files/81/Field%20Test/94/field-test-lessons-learned-final_0-2.pdf

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., & Ginsburgh, M. (2009). Guidelines for the assessment of English language learners. Princeton, NJ: Educational Testing Service.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematic testing in a large scale state assessment program. *The Journal of Technology, Learning, and Assessment, 3*(6), 4-29.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6). Retrieved from http://files.eric.ed.gov/fulltext/EJ905028.pdf

Ramineni, C., Trapani, C., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). Evaluation of the *e-rater*® Scoring Engine for the *TOEFL*® Independent and Integrated Prompts. (Research Report No: RR-12-06). Princeton, NJ: Educational Testing Service.

Roever, C. (2014). Assessing pragmatics. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., pp. 125-139). Chichester, UK: John Wiley & Sons, Inc.

Roever, C. (2001). Web-based language testing. *Language Learning & Technololgy, 5(2),* 84-94.

Scalise, K. (2012). Creating innovative assessment items and test forms. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history*

and predictions for the future (pp. 133-155). Charlotte, NC: Information Age Publishing.

Shermis, M. D. & Hamner, B. (2013). Contrasting State-of-the-Art Automated Scoring of Essays. In M. Shermis, & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York, NY: Routledge.

Shyyan, V., Thurlow, M., Christensen, L., Lazarus, S., Paul, J., & Touchette, B. (2016). *CCSSO accessibility manual: How to select, administer, and evaluate use of accessibility supports for instruction and assessment of all students.* Washington, DC: CCSSO.

Smarter Balanced Assessment Consortium (SBAC). (n.d.). Smarter Balanced Assessment Consortium: Practice and training tests. Retrieved from http://www.smarterbalanced.org/assessments/practice-and-training-tests/

Smarter Balanced Assessment Consortium (SBAC). (2015). Smarter Balanced Assessment Consortium: Accessibility and accommodations fact sheet. Los Angeles, CA: Author. Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/08/SmarterBalanced_Accessibility_Factsheet.pdf

Smarter Balanced Assessment Consortium (SBAC). (2016). Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines. Los Angeles, CA: Author. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

Smarter Balanced Assessment Consortium (SBAC). (2014). Smarter Balanced "Tests of the Test" Successful: Field Test Provides Clear Path Forward. Los Angeles, CA: Author. Retrieved from ~~https://www.smarterbalanced.org/wp-content/uploads/2015/08/FieldTest_Report.pdf~~ [Note: the preceding Web address is no longer valid.]

Smarter Balanced Assessment Consortium (SBAC). (2013). Smarter Balanced assessment consortium: Cognitive laboratories technical report. Lost Angeles, CA: Author. Retrieved from https://portal.smarterbalanced.org/library/en/cognitive-laboratories-technical-report.pdf

Smarter Balanced Assessment Consortium, Measured Progress, & National Center on Educational Outcomes. (2014). Smarter Balanced Assessment Consortium: Accessibility and accommodations framework. Los Angeles, CA: Author.

Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/09/Accessibility-and-Accommodations-Framework.pdf

Solano-Flores, G., Shade, C., & Chrzanowski, A. (2014). Smarter Balanced Assessment Consortium. Smarter Balanced Assessment Consortium: Item Accessibility and Language Variation Conceptual Framework. Retrieved from https://portal.smarterbalanced.org/library/en/item-accessibility-and-language-variation-conceptual-framework.pdf

Tate, T. P., Warschauer, M., & Abedi, J. (2016). Data on NAEP 2011 writing assessment prior computer use. *Data in Brief, 8*, 978-989.

Texas Education Agency (2008). A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests. Texas Education Agency Technical Report Series. Retrieved from http://tea.texas.gov/WorkArea/DownloadAsset.aspx?id=2147494120

Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report No. 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education. (2015). Assessment Peer Review Guidance. Retrieved from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf

Weigle, S. C. (2013). English as a Second Language Writing and Automated Essay Evaluation**.** In M. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York, NY: Routledge.

White, S., Kim, Y., Chen, J., & Liu, F. (2015). *Performance of fourth-grade students in the 2012 NAEP computer based writing pilot assessment: Scores, text length, and use of editing tools* (NCES 2015-119). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

WIDA. (n.d.). WIDA consortium test demo. Retrieved from https://wbte.drcedirect.com/WIDA/portals/wida

WIDA. (2015a). ACCESS for ELLs 2.0 Accessibility and Accommodation Descriptions. Madison, WI: Author. Retrieved from https://www.wida.us/assessment/WIDA_AccessAccDescriptions.pdf

WIDA. (2015b). ACCESS for ELLs 2.0 Accessibility and Accommodation Guidelines. Madison, WI: Author. Retrieved from https://www.wida.us/assessment/access20-prep.aspx

Wolf, M. K., & Butler, Y. G. (in press). An overview of English language proficiency assessments for young learners. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners.* New York, NY: Routledge.

Wolf, M. K., Guzman-Orth, D., & Hauck, M. C. (2016). *Next-generation summative English language proficiency assessments for English learners: Priorities for policy and research* (Research Report No. RR-16-08). Princeton, NJ: Educational Testing Service.

Wolf, M. K., Guzman-Orth, D., Lopez, A., Castellano, K., Himelfarb. I., & Tsutagawa, F. (2016). Integrating scaffolding strategies into technology-enhanced assessments of English learners: Task types and measurement models. *Educational Assessment*, *21*(3), 157-175.

Wolf, M. K., Guzman-Orth, D., & Wain, J. (2014). *Investigating the usability of technology-enhanced assessment items during the ELPA21 development process*. ELPA21 Cognitive lab study report. Final Deliverable to ELPA21.

Xiong, W., Evanini, K., Zechner, K., & Chen, L. (2013). Automated content scoring of spoken responses containing multiple parts with factual information. *Proceedings of the Workshop on Speech and Language Technology for Education (SLaTE 2013)*, 137-142.

APPENDIX A: A LIST OF RELEVANT STUDIES ON COMPUTER FAMILIARITY AND MODE EFFECT ON STUDENT ASSESSMENT PERFORMANCE

Table A1: Empirical Studies Related to Students' Computer Familiarity and Their Performance on Computer-Based Assessments

Author(s) and publication year	Data collection year	Grade level(s)	Test content	Did computer familiarity appear to affect CBA performance?
Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan (2008)	2001	8 (*N* = 1,970)	Mathematics (NAEP)	• Yes • Statistically significant • Small effect size
Horkay, Bennett, Allen, Kaplan, & Yan (2006)	2002	8 (*N* = 4,133)	ELA Writing (NAEP)	• Yes • Statistically significant • Small effect size
Kim & Huynh (2008)	2005	Middle, High School (*N* = 439)	ELA Reading, Writing (EOC)	• Yes • Statistically significant • Small effect size
Ling & Bridgeman (2013)	n/a	Undergraduate, Graduate (*N* = 367)	ELP Writing (TOEFL iBT)	• Yes • Statistically significant • Medium effect size
Odo (2012)	n/a	8–12 (*N* = 120)	ELP Reading (LOMERA)	• Yes • Statistically significant • Small effect size
PARCC (2015)	2014	3–11 (*N* = ~20,000)	ELA, Mathematics (PARCC field test)	• Yes • Negligible to small effect size
Pomplun, Frey, Beckers (2002)	n/a	High School, University (*N* = 215)	Reading (Nelson-Denny Reading Test)	• Yes • Statistically significant • Small effect size
Russell (1999)	n/a	8 (*N* = 229)	Mathematics, Science, Reading, Writing (NAEP, MCAS)	• Yes • Statistically significant • Medium effect size
Russell & Plati (2000)	1999	8, 10 (*N* = 290)	ELA Writing (MCAS)	• No

Author(s) and publication year	Data collection year	Grade level(s)	Test content	Did computer familiarity appear to affect CBA performance?
SBAC (2014)	2014	3–8, 11 (*N* = 19,600)	ELA, Mathematics (SBAC field-test)	• Yes (qualitative)
Tate, Warshauer, & Abedi (2016)	2011	8 (*N* = 24,600)	ELA Writing (NAEP)	• Yes • Statistically significant • Small effect size
Taylor, Jamieson, Eignor, & Kirsch (1998)	n/a	Undergraduate, Graduate (*N* = 1,169)	ELP (TOEFL CBT)	• No
White, Kim, Chen, & Liu (2015)	2012	4 (*N* = 10,400)	ELA Writing (NAEP)	• Yes

Note. ELA = English language arts; ELP = English language proficiency; EOC = end of course; LOMERA = Lower Mainland English Reading Assessment; MCAS = Massachusetts Comprehensive Assessment System; NAEP = National Assessment of Educational Progress; PARCC = Partnership for Assessment of Readiness for College and Careers; SBAC = Smarter Balanced Assessment Consortium; TOEFL = Test of English as a Foreign Language.

APPENDIX B: A REVIEW OF COMPUTER-BASED ELP ASSESSMENTS FOR K–12 STUDENTS

Table B1: Reviewed ELP Assessments and Their Grade-Level Clusters

Test name	Grade-level clusters
ACCESS for ELLs 2.0 (WIDA)*	K, 1, 2–3, 4–5, 6–8, 9–12
ELPA21	K, 1, 2–3, 4–5, 6–8, 9–12
LAS Links Online	K–1, 2–3, 4–5, 6–8, 9–12
Online IPT	Oral (Listening, Speaking): K, K–1, 2–6, 6–8, 9–12 Reading, Writing: K–1, 2–3, 4–6, 7–12
Test of English Language Learning (TELL)	K, 1–2, 3–5, 6–8, 9–12

*PPT is administered at the kindergarten level and computer-based Writing is administered starting at Grade 4.

Table B2: User Interface Features of Five Computer-based ELP Assessments for K–12 Students

	CBA test features across all domains	Speaking and Writing CBA features	Administration features	PPT options
WIDA	• Colored visuals • Color contrast, color overlay • Volume control • Progress indicator • Accessibility tools (“Help” button, highlighter, magnifier, line guide) • “Next” button activation upon the test taker’s response • Speech at adjusted (slower) speed • Listening questions played once • Technology-enhanced item (TEI) formats: hotspot, drag and drop	• Digital voice capture • Microphone testing • Virtual instructor and model peer student (avatars) in Speaking • The test taker controls the “Record” and “Stop” buttons • Recording limited to once • Cut/copy/paste, underline in writing space • On-screen notepad for writing • Directions in both written and spoken languages	• Online demo, practice tests available for students • Required online training course for Test Examiners • Kindergarten: one-on-one administration • Grades 1–12: group administration	• Kindergarten: only PPT • PPT Writing only for Grades K–3 • Grades 1–12: PPT available (the existing ACCESS for ELLs test) only for qualified students (pre-arranged)

	CBA test features across all domains	Speaking and Writing CBA features	Administration features	PPT options
ELPA21	• Colored visuals (both static and animated visuals) • Volume control • Progress indicator • Accessibility tools (selector, highlighter, answer choice eliminator, digital notepad, zoom in/out) • Flag items for review • Optional multiple audio playing for directions and prompts at the test taker's control • Listening questions can be played twice at the test taker's control (upper grades); automatic audio play for the lower grades • TEI formats: hotspot, drag and drop, drop-down menu, text selection	• Digital voice capture • The test taker controls the "Record" and "Stop" buttons • "Record," "Stop," "Play" buttons for recording responses • Recording limited to twice • Cut/copy/paste, underline, italicize, bold, undo/redo, bullets	• Online demo & practice tests available • Training site available for Test Examiners • K: 1–5 students, 1–2 proctors • Grade 1: 1–5 students, 1–2 proctors • Grades 2–3: 8 students, 1 proctor • Grades 4–5: 10 students, 1 proctor • Grades 6–12: 15–20 students, 1 proctor	• Print on request (designated support following each state policy)
LAS Links Online	• Colored visuals • Volume control • Progress indicator • Accessibility tools (highlighter, eraser) • Timer • "Go back"/"Go on"/ "Stop Test" buttons • Flag items for review • Optional audio for directions and prompts	• Digital voice capture • The test taker controls the "record" and "stop" buttons • "Record," "Stop," "Play" buttons for recording responses • Re-recording is allowed	NA	• PPT forms (the existing the LAS Links test) available for K–12
Online IPT	• Colored visuals • Volume control • Progress indicator • Accessibility tools (highlighter, eraser) • timer • "Go back" and "Go on" buttons	• Only the Test Examiner uses the computer, and the test takers see a paper-based Speaking booklet for Speaking • Digital voice capture	NA	• PPT forms (the existing IPT test) available for K–2

	CBA test features across all domains	Speaking and Writing CBA features	Administration features	PPT options
TELL	• Tablet touch screen • Colored visuals (both static and animated visuals) • Volume control • Progress indicator • timer • "Next" button activation upon the test taker's response • TEI formats: drag and drop, touching the screen	• Digital voice capture • Automated scoring • Fixed response time	NA	NA

Note. CBA = computer-based assessment; PPT = paper-based test.

APPENDIX C: ACCESSIBILITY FEATURES AND ACCOMMODATIONS FOR CBA

Table C1: A List of Common Accessibility Features and Accommodations Used in Consortia

Support	Sources	Delivery system location (embedded/non-embedded)	Tier
American Sign Language Also known as "Interpreter Signs Test Directions in ASL" [WIDA]	• CCSSO • ELPA21 (directions only) • PARCC • SBAC • WIDA (directions only)	• Embedded (CCSSO, ELPA21, PARCC, SBAC) • Non-embedded (WIDA)	Accommodation
Assistive Technology	• CCSSO • ELPA21 • PARCC • WIDA (Reading, Listening, Speaking, Writing)	• Non-embedded	Accommodation
Audio Amplification *Also known as Audio Aids [WIDA]; includes amplification, noise buffers, white noise	• CCSSO • ELPA21 • PARCC • WIDA	• Embedded (CCSS0, ELPA21, PARCC) • Non-embedded (WIDA)	Universal Tool
Braille (refreshable and embossed)	• CCSSO • ELPA21 • PARCC • SBAC • WIDA (paper test only, Reading and Writing domains only)	• Embedded (SBAC) • Embedded and Non-embedded (PARCC), • Non-embedded (CCSSO, ELPA21, WIDA)	Accommodation
Color Contrast	• CCSSO • ELPA21 • PARCC • SBAC • WIDA	• Embedded (ELPA21, PARCC, SBAC, WIDA) • Non-embedded (CCSSO, SBAC)	• Universal Tool (WIDA) • Designated Support (ELPA21, CCSSO, PARCC, SBAC)
Color Overlays	• CCSSO • ELPA21	Non-embedded	• Universal Tool (WIDA)

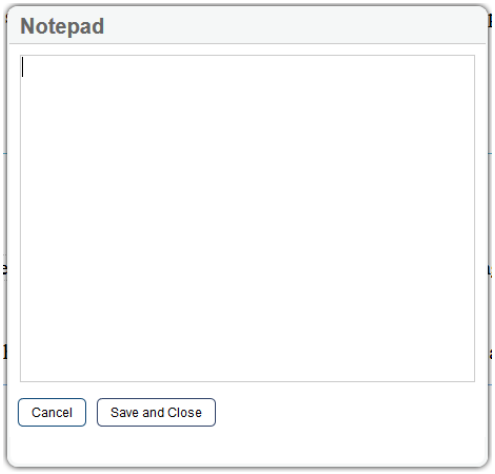
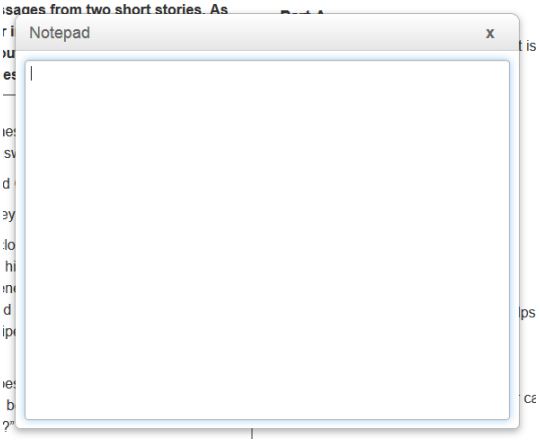
Support	Sources	Delivery system location (embedded/non-embedded)	Tier
	• SBAC • WIDA		• Designated Support (CCSSO, ELPA21, SBAC)
Digital Notepad; Also known as "Notepad" (PARCC) and "Sticky Notes" (WIDA)	• CCSSO • ELPA21 • PARCC • SBAC • WIDA (Writing domain only)	Embedded	Universal Tool
Highlighter	• CCSSO • ELPA21 • PARCC • SBAC • WIDA	Embedded	Universal Tool
Keyboard Navigation; also known as Keyboard shortcuts/equivalents (WIDA)	• CCSSO • ELPA21 • SBAC • WIDA	Embedded	Universal Tool
Large Print Edition	• CCSSO • ELPA21 • PARCC • WIDA	Non-embedded	Accommodation
Line Reader	• ELPA21 • WIDA	• Embedded (ELPA21, WIDA) • Non-embedded (WIDA)	• Universal Tool (WIDA) • Designated Support (ELPA21)
Magnification	• CCSSO • ELPA21 • SBAC • PARCC • WIDA	• Embedded (CCSSO, ELPA21, PARCC, SBAC, WIDA) • Non-embedded (WIDA)	• Universal Tool (PARCC, WIDA) • Designated Support (CCSSO, ELPA21, SBAC)
Noise Buffers; also part of "Audio Aids" (WIDA)	• ELPA21 • SBAC • WIDA	Non-embedded	• Universal Tool (WIDA) • Designated Support (ELPA21, SBAC)
Repeat Item Audio	• CCSSO (Listening domain only) • ELPA21 (Listening domain only)	Embedded	Accommodation

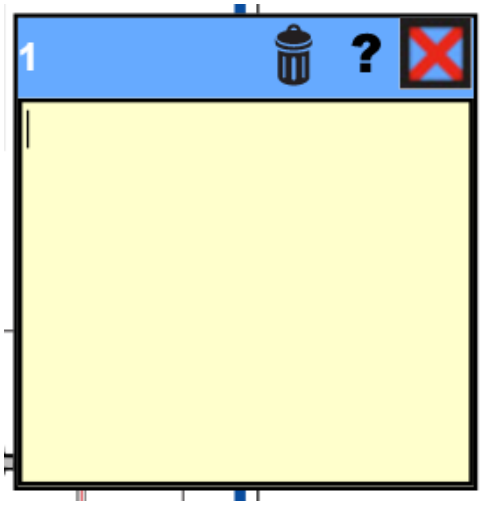
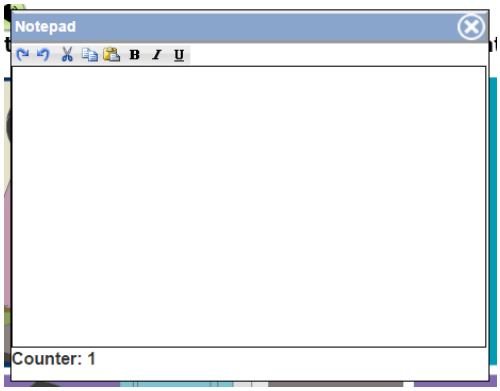
Support	Sources	Delivery system location (embedded/non-embedded)	Tier
	• WIDA (Listening, Speaking, Writing domains only)		
Scratch Paper	• CCSSO • ELPA21 • PARCC • SBAC • WIDA	Non-embedded	Universal Tool
Scribe	• CCSSO • ELPA21 • PARCC • SBAC • WIDA (Reading, Listening, Writing domains only)	Non-embedded	Accommodation
Speech to Text	• CCSSO • ELPA21 • PARCC • SBAC • WIDA (Reading, Listening, Writing domains only)	Non-embedded	Accommodation

Note. This table displays a list of common supports that are implemented across the nation's existing ELP consortia (ELPA21 and WIDA), content assessment consortia (PARCC and SBAC), as well as the CCSSO accessibility and accommodations manual. Sequential steps were taken to analyze the data. First, common supports in both ELPA21 and WIDA were identified and listed in the first table column. Next, multi-tiered accessibility and accommodations manuals from SBAC, PARCC, and the CCSSO were reviewed because of possible similarities in their ELA domain to the ELP domains of reading and writing. Only the ELA domain was reviewed in this step because of the nature of the construct and differences in the necessary supports for other domains (e.g., calculator in the mathematics domain). The location of each commonly identified support was recorded (embedded in the test delivery system, external to the test delivery system) and any variations were noted. Finally, the tier (universal tool, designated support, accommodation) was recorded for each commonly used support, and variations were noted.

Because of the focus of the analysis (commonalities *across* ELPA21 and WIDA), there may be features (embedded and non-embedded universal tools, designated supports, and accommodations) used by individual consortia (ELPA21, WIDA, SBAC, and PARCC) that were not included in this table. Each feature included and excluded from this list deserves careful review and deliberation to determine if it should be included in the ELPAC CBA.

Table C2: Example of Design Variations for the Commonly-used Embedded Digital Notepad Tool

Consortia and description	Digital notepad example
SBAC • Location: Embedded • Domain: Domain not specified. • Description: "This tool is used for making notes about an item. The digital notepad is item-specific and is available through the end of the test segment. Notes are not saved when the student moves on to the next segment or after a break of more than 20 minutes." (SBAC, 2016, p. 6)	 Source: Smarter Balanced Assessment Consortium (SBAC). (n.d.). Smarter Balanced Assessment Consortium: Practice and training tests. Retrieved from http://www.smarterbalanced.org/assessments/practice-and-training-tests/
PARCC ("Notepad") • Location: Embedded • Domain: ELA • Description: "The student selects the "Notepad" icon in the toolbar. The student writes notes using embedded Notepad tool on the ELA/literacy assessments. The student may disable this feature by selecting "Notepad" in the toolbar again." (PARCC, 2016, p. 15)	 Source: Partnership for the Assessment of Readiness for Careers and Colleges (PARCC). (n.d.). PARCC Practice Tests. Retrieved from https://parcc.pearson.com/practice-tests/

Consortia and description	Digital notepad example
WIDA ("Sticky Notes") • Location: Embedded • Domain: Writing • Description: "A tool which the student can use to make notes to assist in responding to Writing items. This tool is only available on the Writing test" (WIDA, 2015, p. 5)	 Source: WIDA. (n.d.). WIDA consortium test demo. Retrieved from https://wbte.drcedirect.com/WIDA/portals/wida
ELPA21 • Location: Embedded • Domain: Reading, Listening, Speaking, Writing • Description: "The student uses this feature as virtual scratch paper to make notes or record responses. The digital notepad is item-specific and is available through the end of each test domain. Notes are not saved when the student moves on to a different test domain or after a break of more than 20 minutes." (ELPA21, 2015, p. 10)	 Source: English Language Proficiency Assessment for the 21st Century (ELPA21). (n.d.). ELPA21 assessment program interactive demo. Retrieved from https://elpa-practice.nextera.questarai.com/student/WebClient/PracticeTest/ [Note: the preceding Web address is no longer valid.]

Note. Illustrations of the digital notepad tool are taken from each consortium's publically available online practice test. It is possible that state specific proprietary versions may render designs differently.

APPENDIX D: DRAFT OF A PROPOSED ELPAC CBA TEST BLUEPRINT

Grades K–2: **Listening**

PPT Listening task type	Aligned primary ELD standard(s)[3]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
Listen to a Short Exchange The student hears a two-turn exchange between two students (one female and one male) or one student and one teacher. The student then answers a question about the exchange.	PI.A.1 Exchanging information and ideas • Main Idea: What are the students talking about? • Key Detail: Who is the boy going to visit? PII.A.2 Understanding cohesion • Linking Detail: Who is the boy going to visit?	Discrete, 1 point	5	5	7	This task type should be redesigned so that the stimulus is delivered by recorded audio. Color graphics should be added.
Listen to a Story The student hears a short grade-appropriate fictional story that contains dialogue, then answers three questions about it.	PI.B.5 Listening actively • Main Idea: What are the students talking about? • Key Detail: Who is the boy going to visit? PII.A.1 Understanding text structure • Text Structure: What happens at the beginning of the story?	Set of 3 items, 3 points per set	9	9	9	This task type should be redesigned so that the stimulus is delivered by recorded audio. Color graphics should be added. Newly designed TEI types should be added to sets: **PI.B.5 Match items[4]**

[3] Multiple primary standards are listed for those task types with items that align to different standards.

[4] Definitions of new task types (match items, zone items, grid items, and select in passage items) are provided in Section 4.

PPT Listening task type	Aligned primary ELD standard(s)[3]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
Listen to an Oral Presentation The student hears an oral presentation on an academic topic, then answers three to four questions about it.	PI.B.5 Listening actively • Main Idea: What is the information about? • Key Detail: What did I tell you about [X]?	Set of 3–4 items, 3–4 points per set	6	6	6	This task type should be redesigned so that the stimulus is delivered by recorded audio. Color graphics should be added. Newly designed TEI types should be added to sets: **PI.B.5 Match items** **PI.B.5 Zone items**
		Total Number of Task Types	3	3	3	
		Total Number of Items	20	20	22	
		Total Number of Points	20	20	22	

Note. ELD = English language development.

Grades K–2: **Speaking**

PPT Speaking task type	Aligned primary ELD standard(s)	Aligned secondary ELD standard(s)[5]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
Talk About a Scene The student is presented with an illustration of a familiar scene (e.g., classroom, library) and answers six questions about it.	PI.A.1 Exchanging information and ideas	PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details	Set of 6 items, 9 points per set	9	9	9	Objects in the scene should be highlighted onscreen for emphasis. (In the PPT, the Test Examiner must point.) Color graphics should be added. An avatar should pose questions to the test taker.
Speech Functions The Test Examiner describes a situation and asks what the student would say in the situation.	PI.A.4 Adapting language choices	PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details	Discrete, 2 points	0	0	6	This task type should be redesigned so that an avatar (still image or animated) poses the situation directly to the student to ensure that the task is a more direct measure of the standard P1.A.4.
Support an Opinion The student listens to a presentation about two activities, events, materials,	PI.C.11 Supporting opinions	PII.B.3 Using verbs and verb phrases	Discrete, 2 points	4	4	2	This task type should be redesigned so that an avatar (still image or animated) poses the situation directly to the student to ensure that the task is a

[5] Secondary standards vary based on the stem types. These secondary standards are used in service of the PI Standards and are implicitly accounted for in the constructed-response rubrics.

PPT Speaking task type	Aligned primary ELD standard(s)	Aligned secondary ELD standard(s)[5]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
or objects, and is asked to give an opinion about why one is better than the other.		PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas					more direct measure of the standard P1.C.11.
Retell a Narrative [Speaking with Listening] The student listens to a story that follows a series of pictures, and then the student uses the pictures to retell the story.	PI.C.9 Presenting	PI.B.5 Listening actively PI.C.12 Selecting language resources PII.A.1 Understanding text structure PII.A.2 Understanding cohesion PII.B.3 Using verbs and verb phrases	Discrete, 4 points	4	4	4	The picture being narrated should be highlighted onscreen for emphasis. (In the PPT, the examiner must point.) Color graphics should be added. An avatar will pose the question to the test taker.

PPT Speaking task type	Aligned primary ELD standard(s)	Aligned secondary ELD standard(s)[5]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
		PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas					
Summarize an Academic Presentation **[Speaking with Listening]** The student listens to an academic presentation while looking at a related picture or pictures. The student is prompted to retell the main points of the presentation using the illustration(s) and any key terms, if provided.	PI.C.9 Presenting	PI.B.5 Listening actively PII.A.2 Understanding cohesion PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details	Discrete, 4 points	4	4	4	A major redesign to this task type should be to include animation of the teacher talking and/or the academic content being presented in order to be a more engaging task type for the test taker. An avatar should pose the prompt to the student so that the task is presented in a more authentic context and more directly measures the standard PI.C.9.

PPT Speaking task type	Aligned primary ELD standard(s)	Aligned secondary ELD standard(s)[5]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
		PII.C.6 Connecting ideas					
Total Number of Task Types				4	4	5	
Total Number of Items				10	10	8	
Total Number of Points				21	21	25	

Grades K–2: Reading

PPT Reading task type	Aligned primary ELD standard(s)[6]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
Read-Along Word with Scaffolding With scaffolding from the examiner, the student provides the letter names and the initial letter sound for a word; reads the word; and chooses the picture that represents the word.	PIII Using Foundational Literacy Skills PI.B.6 Reading and viewing closely	Set of 2 items, 3 points per set	6	0	0	Read aloud should be delivered by recorded audio. Text highlighting should be synchronized with the audio so that the student can follow along.
Read-Along Story with Scaffolding The student listens and follows along as the examiner reads aloud a literary text and then answers print concepts and comprehension questions.	PIII Using Foundational Literacy Skills PI.B.6 Reading and viewing closely • Main Idea: What is the story about? • Key Detail: What does [X] do at the end of the story? • Text Elements: Where does the story take place?	Set of 4 items, 5 points per set	5	0	0	Read aloud should be delivered by recorded audio. Text highlighting should be synchronized with the audio so that the student can follow along.
Read-Along Information The student listens and follows along as the examiner reads aloud an informational text and then answers comprehension questions.	PI.B.6 Reading and viewing closely • Main Idea: What is the text about? • Key Detail: What does the text tell us about [X]?	Set of 3 items, 3 points per set	6	0	0	Read aloud should be delivered by recorded audio. Text highlighting should be synchronized with the audio so that the student can follow along.

[6] Multiple primary standards are listed for those task types with items that align to different standards.

PPT Reading task type	Aligned primary ELD standard(s)[6]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
	• Text Elements: What happens at the end of the text?					
Read and Choose a Word The student reads three grade-appropriate words and chooses the word that matches a picture.	PI.B.6 Reading and viewing closely	Discrete, 1 point	0	6	6	Graphics should be in color.
Read and Choose a Sentence The student reads three grade-appropriate sentences and chooses the one that describes a picture.	PI.B.6 Reading and viewing closely	Discrete, 1 point	0	5	6	Graphics should be in color.
Read a Short Informational Passage The student reads a short informational text and answers multiple-choice questions related to the text.	PI.B.6 Reading and viewing closely • Main Idea: What is the text about? Key Detail: What shape is [X]?	Set of 2–3 items, 1 point per item	0	3	5	Newly designed TEI types should be added to sets: **PI.B.6 Match items** **PI.B.6 Zone items**
Read a Literary Passage The student reads a literary text and answers multiple-choice questions related to the text.	PI.B.6 Reading and viewing closely • Main Idea: What is the text about?	Set of 3 items, 1 point per item	0	3	6	Newly designed TEI types should be added to sets: **PI.B.6 Match items** **PI.B. 6 Zone items**

PPT Reading task type	Aligned primary ELD standard(s)[6]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
Read an Informational Passage The student reads an informational passage and answers multiple-choice questions related to the text.	• Key Detail: What does the text tell us about [X]?	Set of 3 items, 1 point per item	0	3	3	Newly designed TEI types should be added to sets: **PI.B.6 Match items** **PI.B. 6 Zone items**
Total Points			17	20	26	

Grades K–2: Writing

As detailed in Section 4, ETS recommends that the K–2 Writing domain remain in PPT format.

PPT Writing task type	Aligned primary ELD standard(s)[7]	Aligned secondary ELD standard(s)[8]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
Label a Picture—Word, with Scaffolding The student is prompted by the examiner to write labels for objects in a picture.	PI.C.10 Composing/Writing	–	Set of 4 items, 6 points per set	6	6	6	NA
Write a Story Together with Scaffolding The student collaborates with the examiner to jointly compose a short literary text.	PI.A.2 Interacting via written English PI.C.10 Composing/Writing	–	Set of 4 items, 6 points per set	6	7	7	NA
Write an Informational Text Together The student listens to a short informational passage and then collaborates with the examiner to jointly compose a text about the passage.	PI.A.2 Interacting via written English PI.C.10 Writing	PI.C.12 Selecting language resources PII.A.1 Understanding text structure PII.A.2 Understanding cohesion	Set of 2 items, 5 points per set	0	5	5	NA

[7] Multiple primary standards are listed for those task types with items that align to different standards.

[8] Secondary standards vary based on the stem types. These secondary standards are used in service of the PI Standards and are implicitly accounted for in the constructed-response rubrics.

PPT Writing task type	Aligned primary ELD standard(s)[7]	Aligned secondary ELD standard(s)[8]	Discrete/set Point value	K	1	2	Recommended CBA enhancements
		PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas					
Describe a Picture The student looks at a picture and writes a brief description about what is happening in the picture.	PI.C.10 Writing	–	Discrete, 3 points	0	3	3	NA
			Total Number of Task Types	2	4	4	
			Total Number of Items	8	11	10	
			Total Number of Points	12	21	21	

Grades 3–12: Listening

PPT Listening task type	Aligned Primary ELD standard(s)[9]	Discrete/set Point value	3–5	6–12	Recommended CBA enhancements
Listen to a Short Exchange The student hears a two-turn exchange between two students (one female, one male) or one student and one teacher. The student then answers a question about the exchange.	PI.A.1 Exchanging information and ideas • Main Idea • Key Detail PII.A.2 Understanding cohesion • Linking Detail	Discrete, 1 point	6	3	A context-setting image should be added.
Listen to a Classroom Conversation Students hear a multiple-turn conversation between two students (female and male) or one student and one teacher, then answer three questions about it.	PI.A.1 Exchanging information and ideas • Main Idea • Key Detail PI.A.3 Supporting opinions and persuading others • Opinion	Set of 3 items, 3 points per set	6	3	A context-setting image should be added.
Listen to a Story The student hears a short grade-appropriate fictional story that contains dialogue, then answers three questions about it.	PI.B.5 Listening actively • Main Idea • Key Detail PII.A.1 Understanding text structure • Text Structure	Set of 3 items, 3 points per set	6	0	This task should be redesigned so that the stimulus is delivered by recorded audio. Color graphics should be added. Newly designed TEI types should be added to sets: **PI.B.5 Match items**

[9] Multiple primary standards are listed for those task types with items that align to different standards.

PPT Listening task type	Aligned Primary ELD standard(s)[9]	Discrete/set Point value	3–5	6–12	Recommended CBA enhancements
Listen to an Oral Presentation Students hear an oral presentation on an academic topic, then answer three to four questions about it.	PI.B.5 Listening actively • Main Idea • Key Detail PI.B.7 Evaluating language choices • Rhetorical Function PI.B.8 Analyzing language choices • Shades of Meaning PII.A.1 Understanding text structure • Text Structure	Set of 4 items, 4 points per set	4	8	Newly designed TEI types should be added to sets: **PI.B.5 Match items** **PI.B. 5 Grid items**
Listen to a Speaker Support an Opinion Students hear a discussion between two classmates, during which one classmate makes an argument in support of an opinion about an academic topic. Students then answer four questions about the discussion.	PI.A.3 Supporting opinions and persuading others • Main Idea • Key Detail PI.B.7 Evaluating language choices • Rhetorical Function PI.B.8 Analyzing language choices • Shades of Meaning PII.A.1 Understanding text structure • Text Structure	Set of 4 items, 4 points per set	0	8	Newly designed TEI types will be added to sets: **PI.B.3 Match items** **PI.B. 3 Grid items** **PI.B.7 Match items**
		Total Number of Task Types	4	4	
		Total Number of Items	22	22	
		Total Number of Points	22	22	

Grades 3–12: Speaking

PPT Speaking task type	Aligned primary ELD standard(s)[10]	Aligned secondary ELD standard(s)[11]	Discrete/set Point value	3–5	6–12	Recommended CBA enhancements
Talk about a Scene The student is presented with an illustration of a familiar scene (e.g., library, classroom) and answers six questions about it.	PI.A.1 Exchanging information and ideas	PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details	Set of 6 items, 9 points per set	9	9	Objects in the scene should be highlighted on-screen for emphasis. (In the PPT, the examiner must point.) Graphics should be in color. An avatar should pose questions to the test taker.
Speech Functions The examiner describes a situation and asks what the student would say in the situation.	PI.A.4 Adapting language choices	PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details	Discrete, 2 points	6	4	This task type should be revised so that an avatar (still image or animated) poses the situation directly to the student so that the task is a more direct measure of the standard P1.A.4
Speaking—Support an Opinion The student listens to a presentation about two activities, events, materials, or objects, and is asked to give an opinion about why one is better than the other.	PI.C.11 Supporting opinions	PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas	Discrete, 3 points	3	3	This task type should be revised so that an avatar (still image or animated) poses the situation directly to the student so that the task is a more direct measure of the standard P1.C.11.

[10] Multiple primary standards are listed for those task types with items that align to different standards.

[11] Secondary standards vary based on the stem types. These secondary standards are used in service of the PI Standards and are implicitly accounted for in the constructed-response rubrics.

PPT Speaking task type	Aligned primary ELD standard(s)[10]	Aligned secondary ELD standard(s)[11]	Discrete/set Point value	3–5	6–12	Recommended CBA enhancements
Retell a Narrative [Speaking with Listening] The student listens to a story that follows a series of pictures, and then the student uses the pictures to retell the story.	PI.C.9 Presenting	PI.B.5 Listening actively PI.C.12 Selecting language resources PII.A.1 Understanding text structure PII.A.2 Understanding cohesion PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas	Discrete, 4 points	4	0	The picture being narrated should be highlighted on-screen for emphasis. (In the PPT, the examiner must point.) Color graphics should be added. An avatar should pose the prompt to the test taker.
Present and Discuss Information [Speaking with Reading] The student views a graph, a chart, or an image that provides information (e.g., a bar graph showing different ways that people exercise each day). The student is	PI.C.9 Presenting PI.A.3 Supporting opinions and persuading others	PI.B.6 Reading and viewing closely PII.A.2 Understanding cohesion PII.B.3 Using verbs and verb phrases	Set of 2 items, 6 points per set	0	6	The two items in the set should be presented while the graphic set leader remains on the screen. (This is a more fluid presentation of the task than on paper.) Text should be highlighted for emphasis instead of the examiner pointing.

PPT Speaking task type	Aligned primary ELD standard(s)[10]	Aligned secondary ELD standard(s)[11]	Discrete/set Point value	3–5	6–12	Recommended CBA enhancements
prompted to read the information and then asked to respond to two prompts.		PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas				An avatar should pose the claim for the student to respond to so that this task is a more direct measure of the standard P1.A. 3.
Summarize an Academic Presentation [Speaking with Listening] The student listens to an academic presentation while looking at a related picture or pictures. The student is prompted to retell the main points of the presentation using the illustration(s) and any key terms, if provided.	PI.C.9 Presenting	PI.B.5 Listening actively PII.A.2 Understanding cohesion PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas PII.C.7 Condensing ideas	Discrete, 4 points	4	4	A major redesign to this task type should be in include animation of the teacher talking and/or the academic content being presented in order to be a more engaging task type for the test taker. An avatar should pose the prompt for the student to respond to so that the task is presented in a more authentic context and more directly measuring the standards PI.C.9.
Total Number of Task Types				5	5	
Total Number of Items				12	12	
Total Number of Points				26	26	

Grades 3–12: **Reading**

PPT Reading task type	Aligned primary ELD standard(s)[12]	Discrete/set Point value	3–5	6–12	Recommended CBA enhancements
Read and Choose a Sentence The student reads three grade-appropriate sentences and chooses the sentence that describes a picture.	PI.B.6 Reading and viewing closely	Discrete, 1 point	2	0	Graphics should be in color.
Read a Short Informational Passage The test taker reads a short informational text and answers multiple-choice questions related to the text.	PI.B.6 Reading and viewing closely PI.B.7 Evaluating language choices PI.B.8 Analyzing language choices PII.A.1 Understanding text structure PII.A.2 Understanding cohesion	Set of 2–3 items, 1 point per item	6	6	Newly designed TEI types should be added to sets: **PI.B.6/7/8 Match items** **PI.B.6/7/8 Select in passage items** **PI.B.6 Grid items**
Read a Student Essay The student reads an informational essay to provide feedback before it is submitted to the teacher. The student answers a set of multiple-choice comprehension questions.	PI.B.6 Reading and viewing closely PI.B.7 Evaluating language choices PI.B.8 Analyzing language choices PII.A.1 Understanding text structure	Set of 6 or 8 items, 1 point per item	6	8	Newly designed TEI types should be added to sets (see above).

[12] Multiple primary standards are listed for those task types with items that align to different standards.

PPT Reading task type	Aligned primary ELD standard(s)[12]	Discrete/set Point value	3–5	6–12	Recommended CBA enhancements
	PII.A.2 Understanding cohesion PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas PII.C.7 Condensing ideas				
Read a Literary Passage The test taker reads a literary text and answers multiple-choice questions related to the text.	PI.B.6 Reading and viewing closely PI.B.7 Evaluating language choices PI.B.8 Analyzing language choices PII.A.1 Understanding text structure PII.A.2 Understanding cohesion	Set of 6 items, 1 point per item	6	6	Newly designed TEI types should be added to sets (see above).
Read an Informational Passage The test taker reads an informational passage and answers multiple-choice questions related to the text.		Set of 5–6 items, 1 point per item	6	6	Newly designed TEI types should be added to sets (see above).
		Total Number of Task Types	5	4	
		Total Number of Items	26	26	
		Total Number of Points	26	26	

Grades 3–12: Writing

PPT Writing task type	Aligned primary ELD standard(s)[13]	Aligned secondary ELD standard(s)[14]	Discrete/set Point value	3-5	6-12	Recommended CBA enhancements
Describe a Picture In addition to looking at a picture, the student is prompted to examine a paragraph written by a classmate. The student is asked to expand, combine, and correct different sentences written by a classmate. The student is then asked to add a sentence to the paragraph.	PI.A.2 Interacting via written English PII.B.3 Using verbs and verb phrases PII.B.5 Modifying to add details PII.C.7 Condensing ideas	–	Set of 4 items, 8 points per set	8	8	Graphics should be in color. The student should proceed through the Writing domain independently, as directions should be delivered via recorded audio.
Write about an Experience The student is provided with a common topic, such as a memorable classroom activity or event. The student is prompted to write about the topic from his or her own personal experience.	PI.C.10 Writing	PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas	Discrete, 2 points	4	4	The student should proceed through the Writing domain independently as directions should be delivered via recorded audio.
Write about Academic Information [Writing with Reading] The student interprets academic information from a	PI.C.10 Writing PI.C.11 Justifying/arguing	PI.B.6 Reading and viewing closely PI.C.12 Selecting language resources	Discrete, 3 points	5	5	The two items in the set should be presented while the graphic set leader remains on the screen. (This is a more

[13] Multiple primary standards are listed for those task types with items that align to different standards.

[14] Secondary standards vary based on the stem types. These secondary standards are used in service of the PI Standards and are implicitly accounted for in the constructed-response rubrics.

PPT Writing task type	Aligned primary ELD standard(s)[13]	Aligned secondary ELD standard(s)[14]	Discrete/set Point value	3-5	6-12	Recommended CBA enhancements
graphic organizer created for a group project and answers two questions about it.		PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas PII.C.7 Condensing ideas				fluid presentation of the task than on paper.) An avatar should pose the questions for the student to respond to so making the task more authentic. The student should proceed through the Writing domain independently as directions should be delivered via recorded audio.
Justify an Opinion The student is writing a letter or response to a school newspaper, principal, or other staff member. A school-related topic (e.g., wearing school uniforms, best type of exercise) is introduced. The student is asked to provide his/her opinion along with appropriate support.	PI.C.11 Supporting opinions	PI.C.12 Selecting language resources PII.A.1 Understanding text structure PII.B.3 Using verbs and verb phrases PII.B.4 Using nouns and noun phrases PII.B.5 Modifying to add details PII.C.6 Connecting ideas	Discrete, 4 points	4	4	Additional context-setting material should be presented to the student. An avatar should pose the prompt to the student making the task more authentic. The student should proceed through the Writing domain independently as directions will be delivered via recorded audio.

PPT Writing task type	Aligned primary ELD standard(s)[13]	Aligned secondary ELD standard(s)[14]	Discrete/set Point value	3-5	6-12	Recommended CBA enhancements
Additional new CBA task type	Potential Aligned Standard(s)					
Write about Academic Information: Listening and Writing The student listens to/watches a clip of teacher talking about an academic subject. The student is asked to write a summary. If added, this task type would replace a task type aligned to P1.C. 10.	PI.C.10 Writing		Discrete, 4 points	4	4	New task: This task should take advantage of the computer-based platform to allow for easy administration of an integrated task. (This type task was very successful in the ELPA PPA pilot, but was dropped due to logistical challenges for administration.)
			Total Number of Task Types	4	4	
			Total Number of Items	8	8	
			Total Number of Points	21	21	

APPENDIX E: PROPOSED HIGH-LEVEL TIMELINE FOR DESIGN, DEVELOPMENT, AND IMPLEMENTATION OF THE ELPAC SUMMATIVE ASSESSMENT CBA

ELPAC year (est.)	Contract year	Quarter	Task/event
July 2018– June 2019	Year One	July – Sept.	• Contract in place: July 1 • **Project Kickoff:** July 1 • **Test Design Documents Drafted** o Test blueprint (draft) o Assessment claims (draft) o Item writing guidelines (sufficient to support pilot/prototype) o Metadata definitions (draft) o Accessibility and accommodations table (draft) • **Initial work on Test Delivery System** o Requirements gathering session o Update IT solution design o Create schedule for systems deployment o Create conceptual plan for score reporting • Milestone: System specifications for pilot delivery system ready to support pilot/prototype item development • **IBIS set up to support item writing for pilot/prototype** o Perform all necessary IBIS set up for CBA task types o Enhance CAASPP IBIS item previewer for use with ELPAC items o Create multi-tiered accessibility tables
		Oct. – Dec.	• **Develop and test the Test Delivery System** • **Item development for prototype/pilot** • Milestone: Pilot/prototype items finalized and handed off to test delivery system • **Accessibility Framework completed** o Consulting with California School for the Blind • **Recruit schools/students for Pilot/Prototype** • **Detailed schedule through CBA launch completed**
		Jan. – Mar.	• **Pilot/Prototype delivery system ready for UAT** • **Conduct UAT for Pilot/Prototype delivery system** • **Milestone: Administration of small-scale pilot/prototype**

ELPAC year (est.)	Contract year	Quarter	Task/event
		April – June	• **Conduct Cog Labs Study** • **Evaluation of pilot/prototype responses** • **Create sampling plan for CBA Field Test** • **Define process for inclusion of students with visual impairments in CBA field test** • **Test Design Documents updated** based on results of pilot/prototype and results of cog labs study (test blueprint, assessment claims, item writing guidelines, metadata definitions, and accessibility framework). o Outcome: Documents are suitable for use in development of item pool for field test. • **Item authoring guidelines finalized (item style guide, graphics style guide, display guide)** • **Test delivery engine configuration finalized (including all specifications of user interface)** • **Revise IBIS set up to reflect task type revisions based on pilot/prototype results** • **Conduct inventory of existing PPT item pool** and document path for possible inclusion of PPT items in CBA field test • **Item Development Plan** • Milestone: Ready to start item development for CBA field test

July 2019– June 2020	Year Two	July – Sept.	• **Item Development for CBA field test**
			  o Item Writing Workshop with California educators
			  o Testing vendor internal item review and development
			  o CDE reviews
			  o Content/Bias and Sensitivity panel meetings
			  o CDE resolution of Content/Bias and Sensitivity panel recommendations
			  o Recording of audio materials
			• Milestone: Item pool for CBA field test complete/handed off to Test Delivery System

		Oct. – Dec.	• **Assembly of forms for CBA field test**
			• **Systems development and testing needed to support CBA field test**
			• **Develop interactive Demos and Tutorials** (to be available in January), including sample items
			  o Version for general population
			  o Version for students with visual impairments
			• **Develop trainings for CBA field test administration**
			• **Deliver trainings for CBA field test administration**

ELPAC year (est.)	Contract year	Quarter	Task/event
		Jan. – Mar.	• **Administration of large scale CBA field test: March** o Includes accessible versions for students with visual and hearing impairments • **Range finding/scoring of Speaking items** in CBA field test • **Range finding/scoring of Writing items** in CBA field test
		April – June	• **Analysis of CBA field test results**: classical item analyses; dimensionality analyses; preliminary IRT analyses; DIF analyses • **AI Scoring Analysis** • **Conduct survey/interview study for Test Examiners/educators on the effectiveness of interactive demos/tutorials** • Milestone: **Confirm or adjust blueprint for CBA operational forms** based on field test results • Milestone: Policy decision regarding plans for use of AI scoring on ELPAC CBA • **Test Design Documents Finalized** based on FT results. Outcome: Documents are suitable to support operational use: CBA test blueprint; CBA assessment claims; item writing guidelines; accessibility metadata; accessibility framework

July 2020– June 2021	Year Three	July – Sept.	• **Complete AI Scoring Analysis**
			  o Item-level analysis
			  o Test-level analysis
			• **Review AI Scoring Recommendations and Establish Policy for Operational Implementation of AI Scoring**
			  o AI scoring most likely to be implemented for 2nd operational administration
			• **Develop Operational CBA forms**
			  o Includes accessible forms such as Braille-ready file/printer output file, tactile graphics, etc.
			• **Review PLDs to determine if they need to be updated based on CBA blueprint.**
			• **Create operational versions of all demos and tutorials**
			  o Version for general population
			  o Version for students with disabilities

		Oct. – Dec.	• Milestone: Hand off operational forms to Test Delivery System • **Systems set up and systems testing for Operational Administration**
		Jan. – Mar.	• Milestone: CBA operationally ready Jan. 1, 2021 • **First CBA Operational Administration**

ELPAC year (est.)	Contract year	Quarter	Task/event
			• **Range finding/scoring of Speaking items** in Operational Administration • **Range finding/scoring of Writing items** in Operational Administration
		April – June	• **Statistical analysis of first operational administration** o Vertical Scaling o Operational Scaling • **Standard setting meeting** • Milestone: Review and approval of cut scores • **Reporting of operational scores**