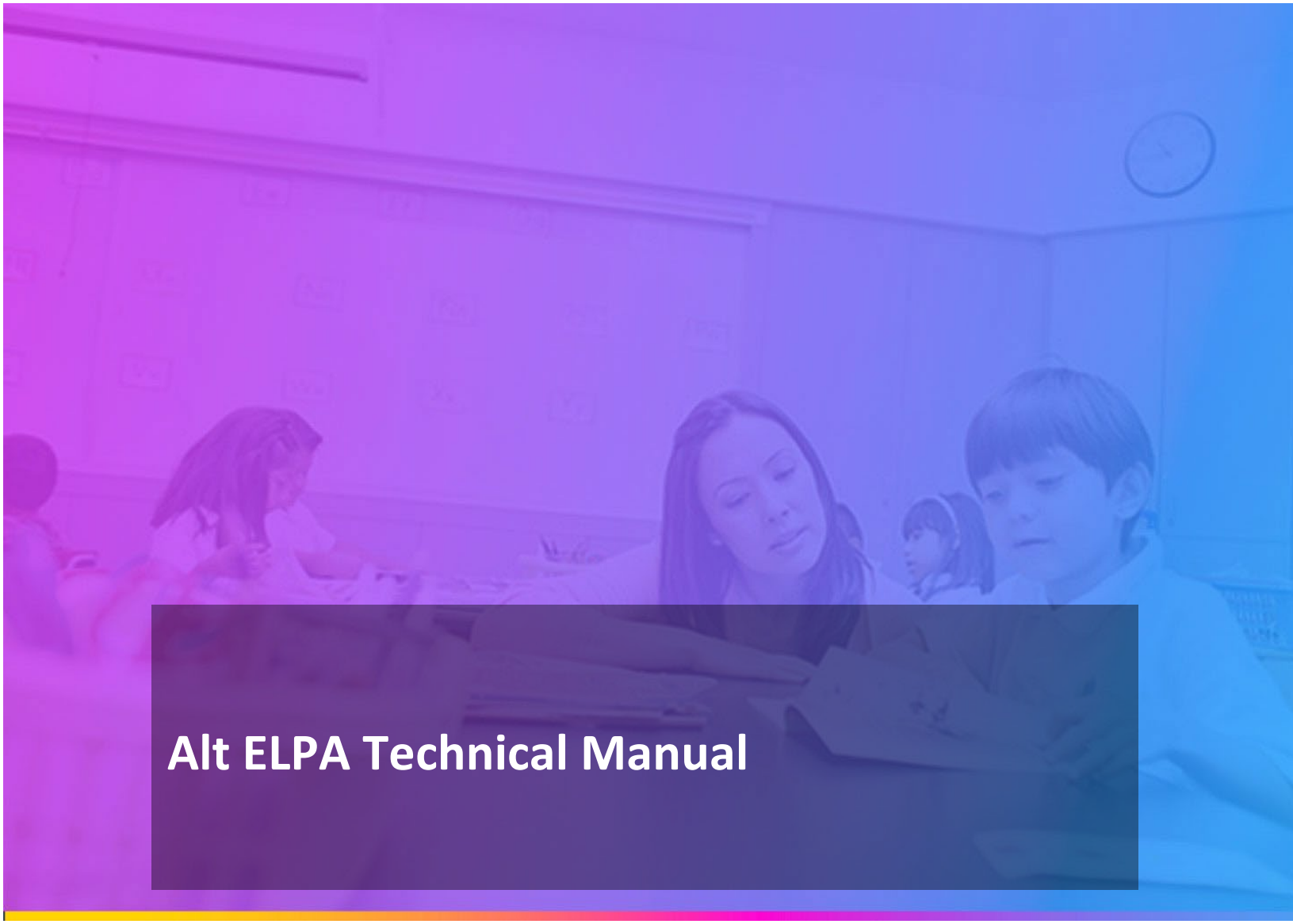
Alt ELPA
Alternate English Language Proficiency Assessment



Alt ELPA Technical Manual

Alternate English Language Proficiency Assessment

January 2025

To cite from this report, please use the following as your APA 7th edition reference: Alternate English Language Proficiency Assessment (Alt ELPA). (2025). *Alt ELPA technical manual*. Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

info@altelpa.org

Version History

Number	Changes	Date
1.0	Initial version	1/17/2025

Table of Contents

Alt ELPA Technical Manual[1]

Preface

Purpose of the Alt ELPA Technical Manual

The purpose is to provide technical information on the Alternate English Language Proficiency Assessment (hereinafter: Alt ELPA). Information will be updated as applicable on a periodic basis.

Report Structure

The report is divided into 11 chapters and also includes several appendices. Chapter 1 provides an overview of the Alt ELPA assessment and background. Chapters 2 and 3 cover test content, design, and development. Chapter 4 summarizes procedures related to test administration and training, while Chapter 5 summaries procedures related to scoring. Chapter 6 discusses classical item and test analyses, including differential item functioning. Chapter 7 provides an overview of the item calibrations and psychometric models used in the Alt ELPA. Chapter 8 summarizes the methodology used for standard setting. Chapters 9 and 10 cover reliability and validity, respectively, to date, including discussion of the Alt ELPA validity argument. Chapter 11 summarizes the types of reporting available.

Where applicable, relevant "*Critical Element*" numbers are referenced within parentheses in section headings and subheadings. These numbers correspond to the critical elements and sub-elements from *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process* (U.S. Department of Education, 2018), with an emphasis on coordinated critical elements.

Annual Technical Report

An annual technical report reflecting the year of test administration complements the contents of this manual, beginning with the *Alt ELPA 2022-23 Field Test Technical Report* (ELPA21, 2024). The annual technical report focuses more on student score distributions and information relevant to the test administration year.

[1] Report prepared by ELPA21, with guidance and support from Dr. Li Cai, members of the ELPA21 Technical Advisory Committee and the ELPA21 Research & Evaluation Committee. Special thanks to all of the members of the Collaborative for the Alternate Assessment for English Language Proficiency (CAAELP). ELPA21 is also grateful for the contributions from our partners and collaborators: Cambium Assessment, Inc., Center for Applied Special Technology, Creative Measurement Solutions, Cognia, Inc., HumRRO, and National Center on Educational Outcomes.

1. Assessment Overview and Background

1.1 Overview of the Assessment System *(Critical Element 2.1)*

Brief History and Background

The English Language Proficiency Assessment for the 21st Century (ELPA21) suite of English language proficiency (ELP) assessments for K-12 consists of a general summative, a general screener, an alternate summative, and an alternate screener. ELPA21 began in 2012 as an Enhanced Assessment Grant (EAG)-funded effort, culminating in the first general summative administration in spring 2016.

The Alternate English Language Proficiency Assessment (Alt ELPA) was designed by the Collaborative for the Alternate Assessment of English Language Proficiency (CAAELP), an Iowa-led, grant-funded project tasked with designing a fair and reliable assessment for English learners with the most significant cognitive disabilities. In addition to Iowa, the following nine states participated in the development of the Alt ELPA: Arizona, Arkansas, Connecticut, Louisiana, Nebraska, New York, Ohio, Oregon, and West Virginia.

To develop the assessment, CAAELP collaborated with researchers at UCLA CRESST, state departments of education, classroom educators who work with English learners with significant cognitive disabilities, and a technical advisory committee specializing in alternate assessment. Through their efforts, and those of many other specialized contributors, the Alt ELPA completed its operational field test administration in school year 2022–23. In the fall of 2024, a pilot of an alternate screener was launched.

The Alt ELPA Summative Assessment

The Alt ELPA summative assessment is a year-end assessment, which is generally administered in February and March of each school year to all eligible English learners who meet a state's criteria for being included in alternate assessments. The Alt ELPA consists of unique assessments for six grade levels or grade bands (K, 1, 2-3, 4-5, 6-8, and 9-12). Each grade level or grade band assessment consists of four short testlets—one per language domain—with a variety of test item types, including innovative selected-response, constructed-response, and technology-enhanced formats.

The Alt ELPA summative assessment is delivered online through a robust test delivery platform that allows for integration with students' assistive and augmentative communication devices. The administration of the Alt ELPA can be customized to the needs of each individual student, with test administration adaptations, accessibility features, and accommodations that can be personalized for each test taker. An assessment that is fair, accessible, inclusive, and engaging will provide a valid and reliable representation of a student's abilities.

The Alt ELPA has 3 overall proficiency determination categories, as described below:

- *Proficient* – Students show a level of English language proficiency reflected in the Alternate ELP standards that enables full participation or only slightly limits participation in the grade-appropriate classroom activities reflected in the Alternate Academic standards. This is indicated on the Alt ELPA by attaining Level 3 or higher in all modalities. Once Proficient on the Alt ELPA, students may be considered for reclassification.
- *Progressing* – Students show a level of English language proficiency reflected in the Alternate ELP standards that moderately limits participation in the grade-appropriate classroom activities reflected in the Alternate Academic standards. This is indicated on the Alt ELPA by attaining above Level 1 and below Level 3 in at least one modality. Students scoring Progressing on the Alt ELPA are eligible for ongoing program support.
- *Emerging* – Students show a level of English language proficiency reflected in the Alternate ELP standards that significantly limits participation in the grade-appropriate classroom activities reflected in the Alternate Academic standards. This is indicated on the Alt ELPA by attaining Level 1 in all modalities. Students scoring Emerging on the Alt ELPA are eligible for ongoing program support.

Student performance is placed into the proficiency categories above based on their scores in the Receptive and Productive modality. Performance in a modality is described by four performance levels:

- Level 1 – Beginning
- Level 2 – Intermediate
- Level 3 – Early Advanced
- Level 4 – Advanced

A modality performance level of 3 or 4 indicates the student is demonstrating they have the English language skills in that modality, as described in the Alternate English language proficiency standards, to participate in grade-appropriate academic content, as described in the state's alternate content standards. Students who achieve Level 3 or 4 in both modalities are considered Proficient, and are eligible to be exited from English language services.

Purposes of the Assessment (Critical Element 2.1.1)

The Alt ELPA is a standards-based ELP assessment for eligible English learners with the most significant cognitive disabilities in kindergarten through Grade 12. The purpose of this assessment is to measure students' progress toward the attainment of English language proficiency in the four recognized language domains of Listening, Reading, Speaking, and Writing, and includes the academic English language students need to access and achieve grade-appropriate content taught in English. This purpose is consistent with the requirements for assessing and reporting student achievement of English language proficiency under the Every Student Succeeds Act (ESSA). This assessment is designed for English learners with the

most significant cognitive disabilities, who are unable to participate in the general English language proficiency assessment even with accommodations.

The Alt ELPA offers eligible English learners with the most significant cognitive disabilities a way to demonstrate their English language proficiency on an assessment based on alternate performance expectations for English language development. The assessment's results are intended to provide important information about what students know and can do in English. The results are intended to help parents and educators establish appropriate English language proficiency expectations and inform decision making regarding appropriate English language services and targeted English language instruction.

Uses of the Assessment Information (Critical Element 2.1.1)

Alt ELPA assessment results provide valuable information that supports states in meeting federal requirements, inform instruction and accountability, and facilitate students' attainment of academic English proficiency, so that all English learners with the most significant cognitive disabilities have the same opportunities as their English learner and non-English learner peers to leave high school prepared for life, college, and career success.

The Alt ELPA assessment scores serve multiple uses. In addition to meeting Title I and Title III requirements, the scores inform teachers of English learners' instructional needs, identify school-based resource needs, provide a means to monitor students' progress towards English proficiency, determine proficiency for program exit decisions, and provide evidence of program effectiveness and accountability. Specifically, Alt ELPA scores are intended to meet three objectives:

- **Measuring Progress**: Alt ELPA scores can be used to monitor progress by English learners with the most significant cognitive disabilities towards English language proficiency. The scores describe individual and group strengths by domain and over time. Reliably measuring progress over time meets multiple state needs such as informing student placement and program reclassification, determining instructional needs of English Learners with the most significant cognitive disabilities and the support needs of teachers, evaluating program effectiveness for subgroups of students, and adjusting educational programming and resources as needed.
- **Reclassification**: Alt ELPA scores can be used to determine proficiency relative to grade-appropriate performance standards, allowing students to be reclassified (i.e., exited from English learner status). Once Proficient, English learners with the most significant cognitive disabilities will have acquired the content-specific English language practices that enable them to produce, interpret, collaborate on, and succeed in content-related and grade-appropriate academic tasks developed for their peers: non-English learners with the most significant cognitive disabilities.

- **Accountability**: Alt ELPA scores may be used for accountability purposes by identifying which institutions are meeting accountability targets and which may be in need of assistance.[2]

1.2 Who are the students?

The Alt ELPA serves K-12 English learners with the most significant cognitive disabilities.

English learners with the most significant cognitive disabilities are students:

- who are not proficient in the English language and have been identified as needing English language development services;
- who meet the Federal definition of an English learner (ESEA as amended by ESSA §8101(20) and 20 USC 20));
- who meet the state definition for having a most significant cognitive disability; and
- whose Individualized Education Program (IEP) teams have determined an alternate assessment is appropriate for the student (CCSSO, 2019; Christensen, Gholson, & Shyyan, 2018; Shyyan & Christensen, 2018; Thurlow, Liu, Goldstone, Albus, & Rogers, 2018)

For the operational field test administration during school year 2022-23, students from the following nine states participated: Arizona, Arkansas, Connecticut, Iowa, Louisiana, Nebraska, Ohio, Oregon, and West Virginia. Close to 4,000 students participated in the field test across all nine states, with approximately 400 to 800 students for each of the six grade levels/grade bands (K, 1, 2-3, 4-5, 6-8, 9-12).

The states adopted the alternate ELP standards in the fall of 2023, and the first operational test administration began in spring of 2024.

1.3 Theory of Action, Validity Framework, and Interpretive Argument

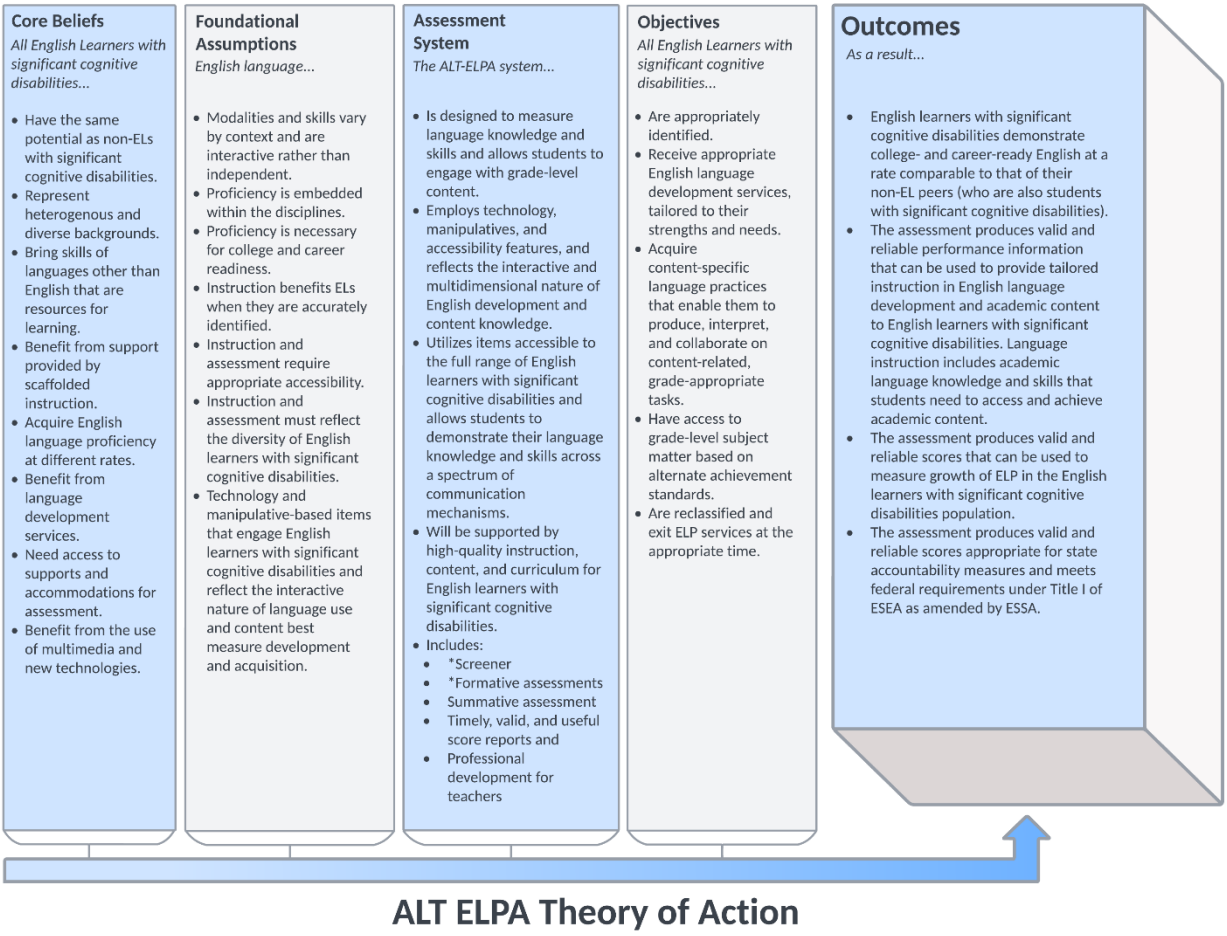
The Alt ELPA Theory of Action (see Figure 1) guided the development of the Alt ELPA summative assessment. The assessment is based on a set of core beliefs and foundational assumptions that distinguish Alt ELPA from other alternate ELP assessments. The assessment system reflects the synthesis and application of these core beliefs and assumptions to specific outcomes that address the needs and challenges of emerging English learners with the most significant cognitive disabilities, which are reflected in the Alt ELPA Theory of Action.

In meeting the objectives outlined in the Theory of Action, the Alt ELPA demonstrates that English learners with the most significant cognitive disabilities, when supported by appropriate high-quality instruction, content, and curriculum, are prepared to acquire content-specific language practices that enable them to produce, interpret, and collaborate on content-related, grade-appropriate tasks.

[2] As determined by the U.S. Department of Education's current accountability legislation.

Figure 1

Alt ELPA Theory of Action

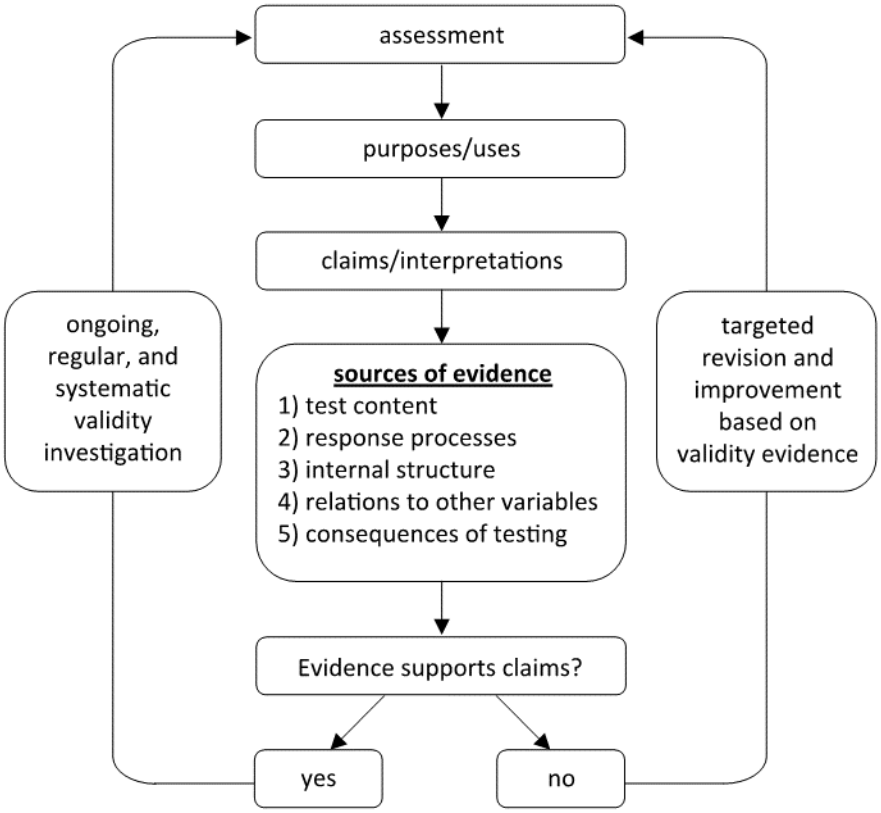


This technical manual will provide the ongoing collection of evidence evaluating the foundational assumptions that underlie the Alt ELPA assessment system. Validity "refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 11). The Alt ELPA validity framework (adapted from Wolf, Farnsworth, & Herman, 2008), as illustrated by Figure 2, articulates the ongoing and iterative processes for both maintaining and improving the assessment to ensure that assessment results reflect intended purposes. The framework begins with the assessment, followed by an articulation of purposes and uses, followed by an articulation of claims and interpretations, which are then evaluated through five primary sources of validity evidence as outlined by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014): evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence based on consequences of testing. Collected evidence is evaluated to

determine whether they support or do not support the assessment claims—results of which are used to either maintain to revise the assessment program.

Figure 2

The Alt ELPA Validity Framework



Not reflected in the figure, but fundamental to assessment validity is fairness. While a thorough discussion of fairness in testing is beyond the scope of this technical manual, it is important to acknowledge that fairness is critical at all stages of test development. While the concept of fairness in testing often has no clearly defined meaning or agreed upon operational definition, the Alt ELPA's design and development incorporated many standard practices of fairness to the extent possible. These practices include ensuring that the test is *accessible* and *appropriate* for the given population, as well as employing principles of universal design, minimizing construct-irrelevant variance, minimizing potential bias, and providing safeguards against inappropriate score interpretations for intended uses throughout the entire test development process.

2. Test Content

2.1 Standards and Performance Expectations *(Critical Element 1.2, 6.3)*

Alt ELP Standards

The Alt ELPA was developed using the *ELP Standards for English Learners with Significant Cognitive Disabilities* (Council of Chief State School Officers [CCSSO], 2019), hereinafter: Alt ELP Standards. The Alt ELP Standards are organized similarly to the CCSSO ELP Standards for a general population (2014). All of the ELP standards for a general population were retained as originally stated with the exception of Standard 4, which omitted the requirement to support oral and written claims with reasoning and evidence. Figure 3 shows the table of the 10 Alt ELP Standards from CCSSO (2019).

Figure 3

Alt ELP Standards

1	construct meaning from oral presentations and literary and informational text through grade-appropriate listening, reading, and viewing
2	participate in grade-appropriate oral and written exchanges of information, ideas, and analyses, responding to peer, audience, or reader comments and questions
3	speak and write about grade-appropriate complex literary and information texts and topics
4	construct grade appropriate oral and written claims
5	conduct research and evaluate and communicate findings to answer questions or solve problems
6	analyze and critique the arguments of others orally and in writing
7	adapt language choices to purpose, task, and audience when speaking and writing
8	determine the meaning of words and phrases in oral presentations and literary and informational text
9	create clear and coherent grade-appropriate speech and text
10	make accurate use of standard English to communicate in grade-appropriate speech and writing

(CCSSO, 2019)

Standards 1 through 7 involve the language necessary for English learners with significant cognitive disabilities to engage in the central content-specific practices associated with ELA & literacy, mathematics, and science. They begin with a focus on extracting meaning and then progress to engagement in these practices.

Standards 8 through 10 focus on some micro-level linguistic features and serve the other seven standards.

Performance Level Descriptors

Policy performance level descriptors (PLDs) which set performance expectations were developed and revised through several iterations with input from various stakeholders, including states, educators with relevant expertise, and the CAAELP Technical Advisory Committee (TAC). Note that skills in lower levels are subsumed in higher levels; an assumption is that students have the opportunity to learn and are provided appropriate supports and accommodations; and "grade-appropriate" refers to grade-appropriate expectations for students with the most significant cognitive disabilities. The three policy PLD levels are listed below:

- **Emerging:** English learners with the most significant cognitive disabilities show a level of English proficiency reflected in the Alternate ELP standards that significantly limits participation in the grade-appropriate classroom activities reflected in the Alternate Academic standards (e.g., convey information related to familiar texts, topics, or experiences, using short, simple sentences with a few frequently occurring words, phrases, and expressions).
- **Progressing:** English learners with the most significant cognitive disabilities show a level of English proficiency reflected in the Alternate ELP standards that moderately limits participation in the grade-appropriate classroom activities reflected in the Alternate Academic standards (e.g., convey information related to familiar texts, topics, experiences, or events using simple and compound sentences with frequently occurring general academic and content-specific words and phrases).
- **Proficient:** English learners with the most significant cognitive disabilities show a level of English proficiency reflected in the Alternate ELP standards that enables full participation or only slightly limits participation in the grade-appropriate classroom activities reflected in the Alternate Academic standards (e.g., convey information related to familiar texts, topics, experiences, or events using a variety of and increasingly complex sentences with an increasing number of general academic and content-specific words and phrases).

2.2 Definitions of Constructs Being Measured *(Critical Element 1.2)*

Language Domain Definitions for the Targeted Student Population

The four language domain definitions are general definitions of each domain assessed by the Alt ELPA for students who are English learners with the most significant cognitive disabilities. These general domain definitions reflect review and input from educators with experience and expertise in assessment, English language proficiency development, English learners, students with significant cognitive disabilities, English learners with significant cognitive disabilities, language pathology, and curriculum and instruction. The TAC, which is comprised of national experts in measurement, assessment validity, alternate assessment, English learners, students with disabilities, and special education, also reviewed and provided input that is reflected in these domain definitions.

- **Listening:** For English learners with the most significant cognitive disabilities, listening is the receiving of language with thoughtful attention and processing sounds to understand their meaning/intent. For the purpose of this assessment for students who are English learners with the most significant cognitive disabilities, listening is demonstrated via a purposeful and appropriate reaction (e.g., spoken response, eye gaze, gesture, movement) to a meaningful sound (e.g., word, sentence) to enable inferences about the degree to which the student, for example, comprehends, analyzes, and infers from spoken language.
- **Reading:** For English learners with the most significant cognitive disabilities, reading is the process of recognizing and understanding/making meaning from symbols, letters, words, etc. For the purpose of this assessment for students who are English learners with the most significant cognitive disabilities, reading is demonstrated through an intentional and appropriate response (e.g., eye gaze, gestures and sign language, movement) to written symbols (e.g., letters, words, sentences, pictures, tactile representations) to enable inferences about the degree to which the student, for example, comprehends, analyzes, and infers from written language.
- **Speaking:** For English learners with the most significant cognitive disabilities, speaking is the action of conveying information or expressing one's thoughts and feelings (e.g., mands, tacts) or a response in a conversation in a manner that is rule-based (e.g., not babble). And, for the purpose of this assessment for students who are English learners with the most significant cognitive disabilities, speaking may include articulation support3, gestures or sign language, picture information systems, and both high- and low-tech Augmented and Alternative Communication (AAC) devices, consistent with the student's IEP.
- **Writing:** For English learners with the most significant cognitive disabilities, writing is the process of using symbols (e.g., letters of the alphabet, punctuation and spaces, pictures, words, tactile representations) to convey or record information in a readable form to a particular audience and for a particular purpose or to communicate thoughts and ideas (e.g., mands, tacts) in a manner that is rule-based (e.g., not scribble). And, for the purpose of this assessment for students who are English learners with the most significant cognitive disabilities, writing includes selecting/pointing to pictures, letters, or words, selecting a tactile representation, and modified signing with a transcriber, consistent with the student's IEP.

Proficiency Definition for the Targeted Student Population

English language proficiency for English learners with the most significant cognitive disabilities is the ability to understand and make meaning, as well as to communicate in English, independently or with supports if/as needed. Proficiency in English requires both receptive and productive English language skills—across the four domains of speaking, listening, reading, and writing, with supports, if/as needed, for a given student—such that the student is able to

3 Articulation supports are supports that help the student physically produce a sound, syllable, or word. Such supports address a physical problem the student has in producing such sounds.

access, engage with, and learn grade-level academic content taught in English, as well as meaningfully participate in academic contexts comparable to their peers with the most significant cognitive disabilities who are not English learners.

2.3 Assessment Model and Structure

Assessment Model

The Alt ELPA is designed to be an end-of-the-year summative assessment. It uses a computer-based fixed form for all test-takers. There are 6 forms, one for each of the grade levels or grade bands: Kindergarten, 1, 2-3, 4-5, 6-8, and 9-12.

Assessment Structure

Each fixed form consists of 10 test items for each of the four language domains (Listening, Reading, Speaking, and Writing), which include a blend of selected-response and constructed-response item types. A practice item, which is not scored, is sequenced at the start of each language domain.

For the 2022-23 operational field test administration, two forms were administered, as illustrated by Table 1. For this test administration, within each domain, items were approximately sequenced in order of lowest to highest PLD.

Table 1

2022-23 Operational Field Test Administration Summary of Test Forms for All Grade Bands by Language Domain

Domain	Form A	Form B
Listening	1 practice item 10 field test items	1 practice item 10 field test items
Reading	1 practice item 10 field test items	1 practice item 10 field test items
Speaking	1 practice item 10 field test items	1 practice item 10 field test items
Writing	1 practice item 10 field test items	1 practice item 10 field test items

Additional items developed during the CAAELP phase continue to be field tested in subsequent test administration years. Table 2 illustrates the distribution of operational items and field test items across four available test forms for 2023-24 and 2024-25. Operational items

were sequenced roughly in order of lowest to highest "difficulty" based on item calibrations from the prior year (i.e., IRT b-parameters).

Table 2

2023-24 and 2024-25 Operational Administrations Summary of Test Forms for All Grade Bands by Language Domain

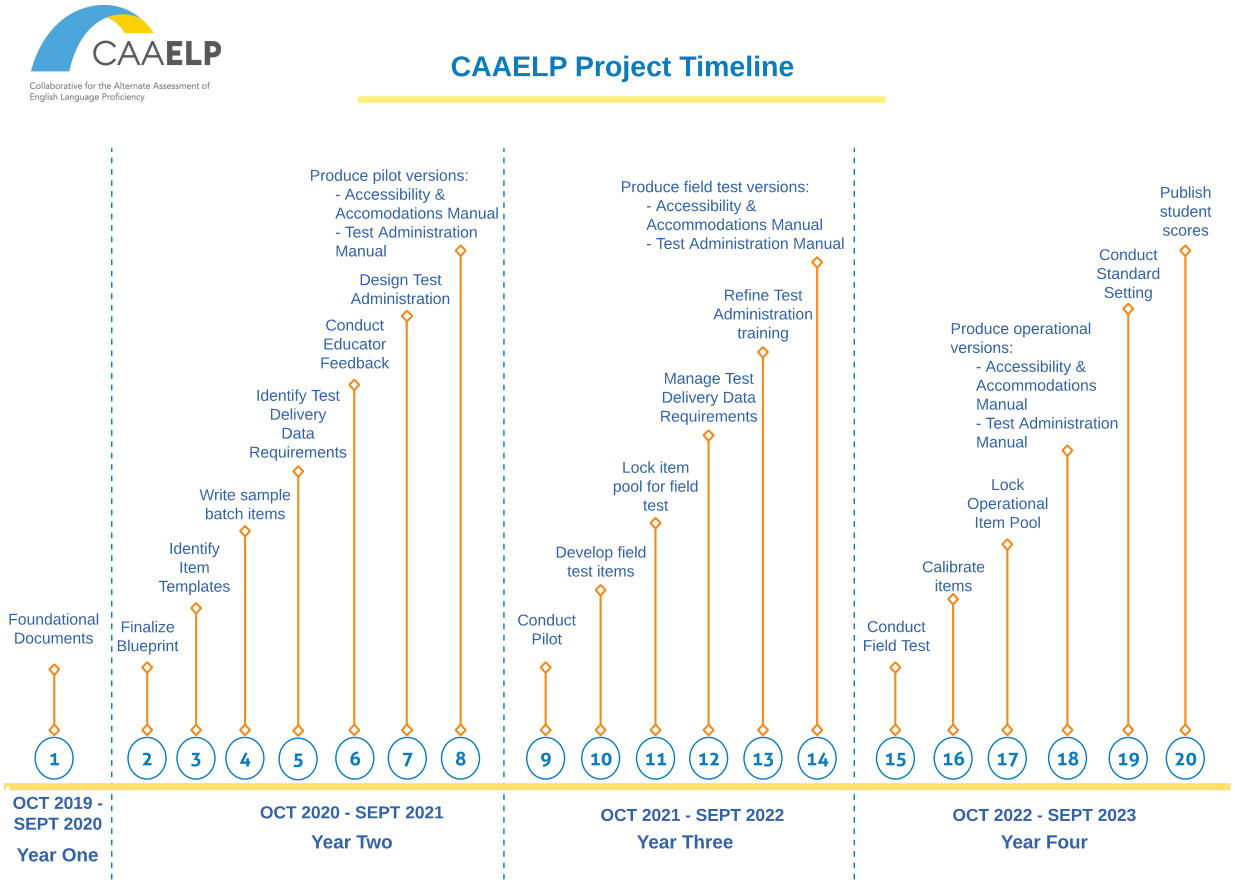
Domain	Form A	Form B	Form C	Form D
Listening	1 practice item 10 operational 6 field test items	1 practice item 10 operational 0 field test items	1 practice item 10 operational 0 field test items	1 practice item 10 operational 0 field test items
Reading	1 practice item 10 operational 0 field test items	1 practice item 10 operational 6 field test items	1 practice item 10 operational 0 field test items	1 practice item 10 operational 0 field test items
Speaking	1 practice item 10 operational 0 field test items	1 practice item 10 operational 0 field test items	1 practice item 10 operational 6 field test items	1 practice item 10 operational 0 field test items
Writing	1 practice item 10 operational 0 field test items	1 practice item 10 operational 0 field test items	1 practice item 10 operational 0 field test items	1 practice item 10 operational 6 field test items

3. Test Design and Development *(Critical Elements 2.1, 2.2)*

This chapter describes the test design and development of the Alt ELPA, including the principled and iterative approach to assessment design that guided the entire development process. Figure 4 shows the overall CAAELP Project Timeline, which began after establishing "foundational documents" described in the first subsection. This chapter details the milestones in the timeline related to item development, including a description of the assessment items and task types in the Alt ELPA and how they were developed and field tested. This chapter also discusses how fairness and inclusion were considered throughout the test design and item development processes.

Figure 4

The CAAELP Project Timeline



3.1 Test Design *(Critical Element 2.1)*

Principled and Iterative Approach

The Alt ELPA was designed and developed using a principled approach, and the approach was iterative to ensure alignment and coherence across key elements of the assessment. The following provides background to the assessment's principled approach which is eclectic in that it draws from several principled approaches to design and reflects common elements across these approaches that include:

1. clearly defined assessment targets;
2. statement of intended score interpretations and uses;
3. model of cognition, learning, or performance;
4. aligned measurement models and reporting scales; and
5. assessment activities to align with assessment targets and intended score interpretations and uses.

See Ferrara, Lai, Reilly, and Nichols (2017) for more detail (Table 3.2 and pp. 53-65).

The CAAELP Assessment Design Team, which was comprised of state members and partners and led by ELPA21 staff, reviewed and discussed three approaches to principled assessment design: Evidence Centered Design (ECD), Assessment Engineering (AE), and Principled Design for Efficacy (PDE). Principled assessment design, development, and implementation offers a "logical, systematic approach to test creation" (Zieky, 2014, p. 79). A principled approach to assessment design provides a framework for decisions that includes intended test-score interpretations and uses, numbers and types of assessment activities, testing time, delivery mode, and various ways of specifying assessment activities (e.g., item specifications, item templates) (Ferrara, et al., 2017). While there is no fixed approach to principled assessment design, there are common elements and practices across approaches.

Generally, a principled approach to assessment design helps to (not comprehensive):

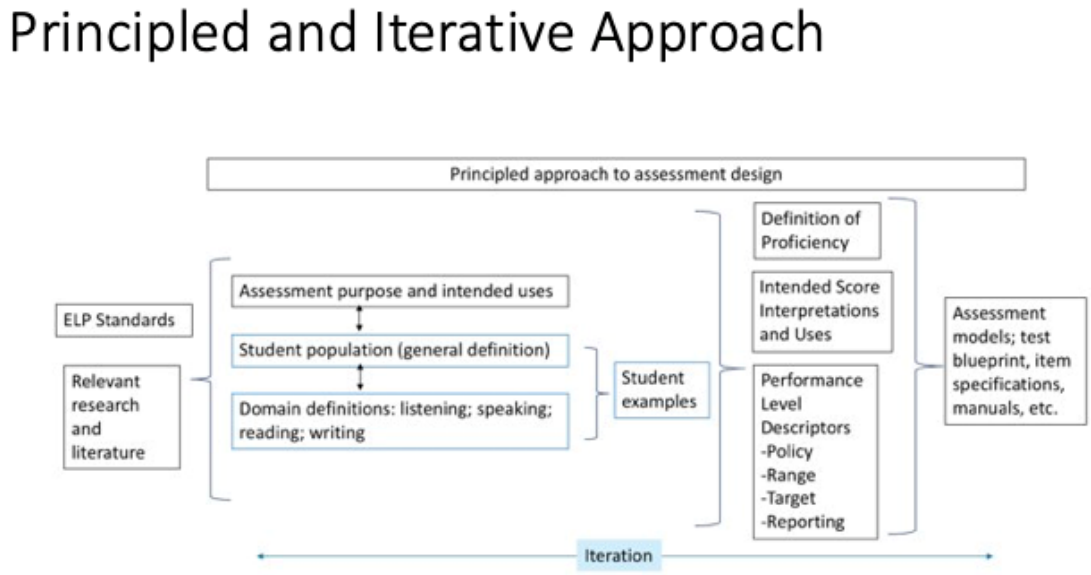
- Define the intended targets of measurement and the constructs to which inferences about students can be made.
- Articulate the relationships among performance level descriptors, task design, and properties of measurement scales, for example, to inform the body of evidence needed to support valid inferences about what students know and can do.
- Ensure meaningful outcomes (e.g., for educators, administrators, policy makers), including informing the interpretive argument.

An excerpt from Ferrara, Lai, Reilly, and Nichols (2017, pp. 45-51, see Appendix A) provides overviews of ECD, AE, and PDE, particularly in terms of how each addresses the six elements of principled approaches: clearly defined assessment targets; statement of intended score interpretations and uses; models of cognition, learning, or performance; alignment measurement models and reporting scales; manipulation of assessment activities, and ongoing accumulation of evidence to support validity arguments.

Figure 5 illustrates the principled and iterative approach guiding the assessment design of the Alt ELPA.

Figure 5

Illustration of the Alt ELPA's Principled and Iterative Approach to Assessment Design



As such, critical initial steps informed by the principled and iterative approach to assessment design led to a series of "foundational documents" (incorporated within the present Technical Manual) which guided the remainder of the CAAELP Project, and thus the Alt ELPA:

1. Description of principled approach to assessment design and development (*the present section)*
2. Statement of assessment purpose and intended uses (*see Chapter 1.1*)
3. Student population definition (*see Chapter 1.2*)
4. Language domain definitions for the targeted student population (*see Chapter 2.2*)
5. Proficiency definition for the targeted student population (*see Chapter 2.2*)
6. Performance level descriptors (PLDs) (*see Chapter 2.1*)
7. Assessment model (*see Chapter 2.3*)
8. Language Processes and Complexity Framework (see Appendix B)
9. Test blueprint (*see Chapter 3.3*)

Fairness and Inclusion (Critical Element 4.2)

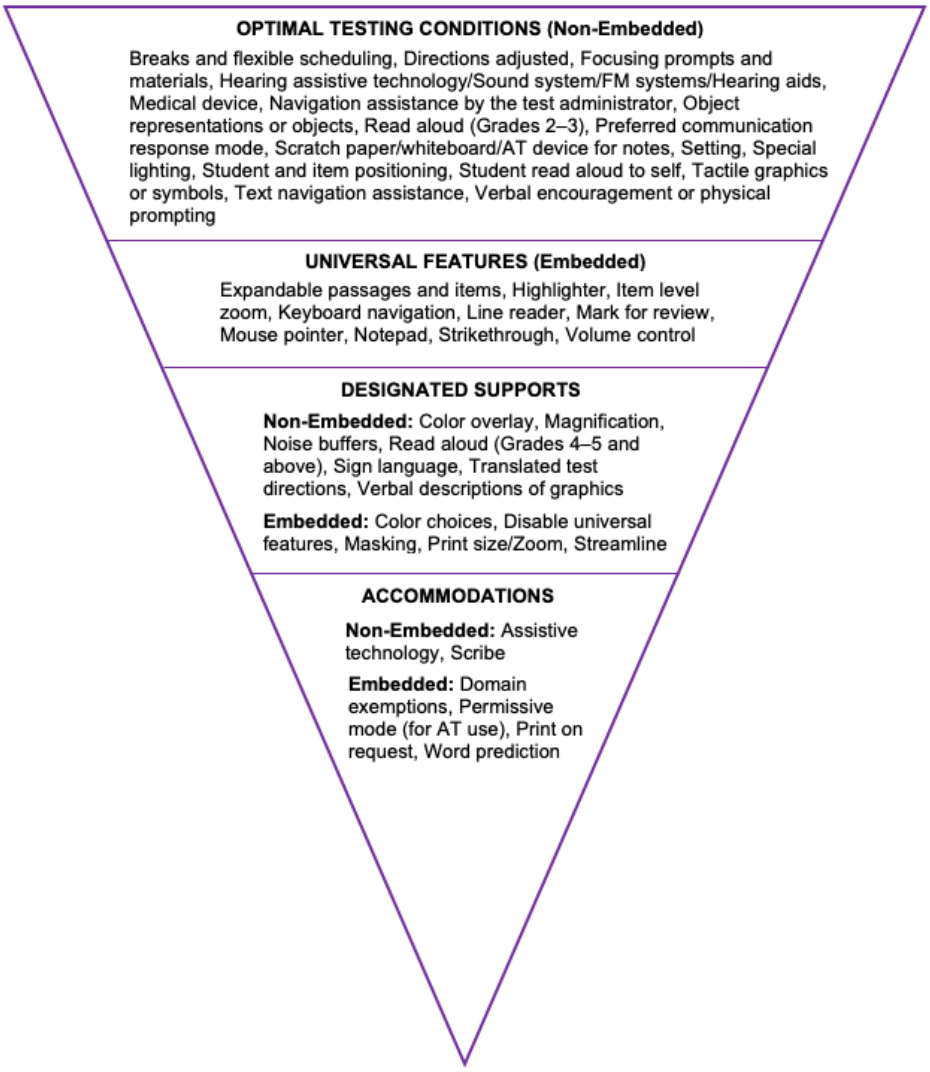
The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME), hereinafter, the *Standards*, noted that "fairness is a fundamental validity issue and requires attention throughout all stages of test development and use." As such, fairness and inclusion for the targeted population was considered at the outset of the Alt ELPA development and throughout the process. The CAAELP Assessment Design Team engaged with the National

Center on Educational Outcomes (NCEO), nationally-renowned experts on accessibility, accommodations, and universal design, throughout the test design and development cycle. NCEO lead a team of state participants to develop participation guidelines for the Alt ELPA through several iterations. The final product is a flow chart that exists as both a stand-alone document as well as incorporated into the *Alt ELPA Accessibility and Accommodations Manual* (which was also spearheaded by NCEO in collaboration with a team of state participants). State participants consisted of state department of education representatives who were EL specialists, Title III coordinators, IDEA specialists, and other related EL or special education assessment specialists.

Accessibility considerations were also developed at the outset of test design and development, and included in the above-referenced *Alt ELPA Accessibility and Accommodations Manual.* NCEO experts led the iterative process of collecting desired testing supports from the team members, developed accessibility tiers for the Alt ELPA, and consulted with all stakeholders including CAAELP leadership for a final version of the manual. The manual, in brief, describes the Alt ELPA's accessibility model, which reflects a tiered approach to accessibility tools that are embedded in the testing platform: (a) universal features available to all English learners with significant cognitive disabilities for the Alt ELPA; (b) designated supports available to all English learners based on need and identified by an adult; and (c) accommodations available only to certain students with significant cognitive disabilities based on their documented needs. Figure 6 shows each of these categories of accessibility tools. This model also reflects accessibility features that are not embedded in the testing platform: (a) optimal testing conditions; (b) designated supports; and (c) accommodations. Note that some accessibility features that are provided during alternate content assessments of reading, writing, mathematics, science, and other content areas may not be provided for the Alt ELPA because they would change the construct measured, making interpretations from test results invalid.

Figure 6

Alt ELPA's Tiered Accessibility Model



3.2 Item Development (*Critical Element 2.2*)

Item and Task Types (Critical Element 2.2)

The Alt ELPA items are either machine scored or scored locally on site by the test administrator using associated rubrics provided with the items. The Alt ELPA does not record the student's spoken responses. Students taking the assessment interact with the following item types (response formats):

- **Selected Response (SR)**: SR items appear in all domains (Listening, Reading, Speaking, and Writing). SR items include a stimulus, a stem, and two or three response options (depending on the PLD). They are either machine scored, or hand scored by the test

administrator using a rubric, worth 1 point at all complexity levels —low, mid, and high. In a standard SR item with three response options, low-complexity items contain two unrelated distractors. Mid-complexity items have one unrelated distractor and one plausible, but incorrect distractor. High-complexity items have two plausible, but incorrect distractors that are parallel in construction.

- **Constructed Response (CR)**: CR items appear in the Speaking and Writing domains only. CR items are typically short responses. They are hand scored using a 3-point rubric targeting high complexity Performance Level Descriptors (PLDs) 3 or 4.
- **Constructed Response-Guided Prompt (CR-GP)**: CR-GP items are structured writing prompts appearing in the Writing domain only. Students produce a piece of writing or compose a permanent product. CR-GP items are either machine scored or hand scored using a 3-point rubric targeting high complexity Performance Level Descriptors (PLDs) 3 or 4. High-PLD 3-students would construct responses using a set of provided answer choice options. High-PLD 4-students would construct their own responses.
- **Technology-Enhanced (TE)**: TE items appear in the Listening, Reading, and Writing domains. They are machine scored and worth 1 or 2 points and target all complexity levels — low, mid, and high, but typically more complex. TE items include a stimulus and a stem written as an imperative prompt, and an interactive component. TE items contain only one interaction type per item: hot spot, gap match, graphic gap match, or in-line choice.

Specific task types for each language domain are listed in the item specifications for each grade level or grade band (K, 1, 2-3, 4-5, 6-8, 9-12).

Item Development Timeline and Process (Critical Element 2.2)

Item development was led by Cognia Inc. from Winter 2020 through Fall 2022, following the articulation of PLDs (which was led by the CAAELP Assessment Design Team), and also included pilot studies led by ELPA21.

The item development process followed a structured and iterative approach as outlined below:

Item Specifications

- Winter 2020-Spring 2021: Cognia developed item specifications for each item type to guide the creation of items. Item specifications are grade-specific documents and contain information regarding terminology and definitions commonly used in item development, standards by modality and domain, descriptions of each item type, grade-level vocabulary, item-level approaches, use of graphics, alternative text, and PLDs for all language domains.

Sample Item Development

- June 2021: Cognia convened the initial Item Review Committee (IRC) to review the first batches of sample items. The IRC comprised of educators and practitioners from

CAAELP member states who work directly with the targeted student population. Feedback from this committee led to revisions of the items and refinement of the item specifications, ensuring they effectively guided further field test item development.

- July 2021: Workshops with another set of educators who work with the targeted population were conducted to train and write new items, incorporating revisions suggested by the initial IRC. This collaborative approach ensured that items were sound and aligned with PLDs.
- Late Fall 2021: Sample item development concluded, resulting in approximately 100 items ready for pilot testing.

Pilot Studies

- Summer 2021: Phase I of the pilot
- February 2022: Sample items were pilot tested through a series of cognitive laboratory interviews conducted with students from the targeted population. The pilot studies were conducted by ELPA21 staff with input from stakeholders. Results indicated that the pilot items were accessible, appropriate, and solicited intended language processes. See

- Appendix C for an Executive Summary of the pilot studies.
- Fall 2021: An initial test map blueprint was developed, providing a high-level plan that outlined the framework for field testing and operational testing. This blueprint informed the standards, PLDs, and levels for ongoing item development.

Field Test Item Development

- Summer 2021 – Summer 2022: 1,000 field test items were targeted for development.
- Fall 2021: Cognia continued with training item writers as needed for item development.
- Fall 2021: CAAELP Assessment Design Team members, including state representatives, reviewed passages.
- January - May 2022: Pre-IRC reviews of field test items were conducted by CAAELP Assessment Design Team state representatives. Feedback from these reviews were incorporated to improve and refine the items.
- June - July 2022: Item review committee meetings were held to evaluate and revise field test items based on feedback. This process led to substantial enhancements to the item bank. Additional details on the review committees are described in the next sub-section below.
- October 2022: The finalized field test item bank, containing over 1,000 items, was handed off within the test delivery vendor's online system in preparation for test delivery.

Cognia concluded their contributions to the Alt ELPA by delivering final item specifications in November 2022, and addressing final questions from the CAAELP team regarding the combined final item specifications in the spring of 2023. Throughout this timeline, Cognia's meticulous approach involved multiple review and revision cycles and collaboration with various stakeholders in order to meet the targeted goal of a comprehensive item bank ready for operational use.

Item Review Committees and Processes

As noted in the timeline above, item review committees were convened both for sample item development and for field test item development. Committee members were first recruited from states participating in CAAELP with efforts to ensure a representative sample to the extent possible across all of the states, locales within each state (i.e., urban, suburban, rural), and comprising of educators and practitioners who work closely with the targeted student population across all grade levels. Committee members held positions such as special education coordinator or coach, English as a Second Language (ESL) facilitator or coordinator or coach, intervention specialist, special education or ESL teacher, deaf educator, autism specialist, and specialists focusing on specific disabilities. Some committee members

participated in more than one committee or event throughout the test development process due to the limited number of specialists in this field.

The IRC meetings held in June and July 2022 by Cognia focused on content, bias, and sensitivity of field test items. Panelists were oriented with background information on the project, including an explanation of why the Alt ELPA is needed, descriptions of the item types, and an overview of accessibility and accommodations available for the assessment. As resources, panelists were provided the item specifications, the item style guide (which includes both editorial and graphic elements), item and passage sloping guidelines (i.e., guidelines used by item writers that include recommendations on areas where passages and items can be manipulated for either supports or ranging levels of complexity), and an alternative text guide.

Panelists were divided into one of three groups based on their background experience: (1) Kindergarten and Grade 1; (2) Grade Bands 2-3 and 4-5; or (3) Grade Bands 6-8 and 9-12. They worked on reviewing items both as a group and independently, and responded to each question (Yes or No) on a provided feedback form for each item. The feedback form was clustered into three main focal areas and consisted of the following directed questions:

1. Alignment and Content:
    a. Does the item align to the assigned grade level PLD for the standard?
    b. Is the content of the item clear and accurate (e.g., correct key, logical structure, and details)?
2. Accessibility
    a. Is the scenario accessible, realistic (e.g., age appropriate)?
    b. Are the terms used concrete, avoiding multiple meaning words?
    c. Are graphics, when included, clear and supportive?
3. Bias and Sensitivity: Does the item/item topic…
    a. … avoid presenting an advantage or disadvantage to any group of students?
    b. … avoid unintended impacts on the student who recently had a personal experience with the subject? (e.g., hurricanes, illness)?
    c. … avoid economic, regional, cultural, or gender bias?
    d. sensitively consider the student's physicality? (e.g., weight, visual/hearing/mobility impairment, etc.)

The feedback form also provided panelists the opportunity to vote to accept, revise, or reject the item, as well as the opportunity to make suggestions for changes. Feedback gathered from this event was used to make item revisions in order to finalize the field test item bank.

3.3 Test Form Construction *(Critical Element 2.1)*

Appendix D contains summary test blueprints for the Alt ELPA for each grade level or grade band, which shows the distribution of Alt ELP standards, PLDs, text complexity, and language processes across test forms. Also included in the appendix are tables of observable behaviors for each grade band by domain and by PLD. Initial iterations of the test blueprints were developed by the CAAELP Assessment Design Team, which included representatives from participating states as described earlier. Test forms for each grade level or grade band were constructed based on the test blueprints.

Multiple Assessment Forms (Critical Element 4.5)

For the first administration, operational field test year 2022-23, two forms (Form A and Form B) were generated based on the blueprints and designed to cover the same content across standards, text complexity and PLDs, but also to contain nearly identical item attributes (i.e., task types and score points). The processes of developing the assessment, the blueprints, and items using Principled Assessment Design and Embedded Standard Setting (described in Chapter 8) ensures alignment of the items to the Alt ELP standards from the outset, and thus, comparable forms. The multiple forms for each grade level or grade band are created using items from the same item bank and same test form planner to ensure the same depth and breadth of Alt ELP standards.

ELPA21 further examines psychometric properties of assessment items including descriptive analysis, concurrent item calibration, and evaluating score precision to ensure that scores from multiple assessment forms within a grade band are comparable across forms (see Chapter 7 for more information on item calibrations).

4. Administration and Training

4.1 Administration Procedures and Guidelines *(Critical Element 2.3)*

Alt ELPA Test Administration Manual (Critical Element 2.3.1)

The *Alt ELPA Test Administration Manual* (TAM) establishes clear, thorough, and consistent standardized procedures for administering the Alt ELPA. It contains rules for test administration, including technology preparation, test administrator directions, time and scheduling considerations, information on student preparation, accommodations and accessibility features, pausing a test, and rules for what is allowable for providing help to students during test administration. The TAM also includes information on how to access the practice version of the test, detailed login instructions, scripts for test directions from the beginning to the end of the test. Information on how to access Help Desk support is also included. Each TAM is customized for individual states, updated for each test administration, and contains links for resources throughout the manual, such as the *Alt ELPA Accessibility and Accommodations Manual*, so that test administrators can obtain more detailed information.

Alt ELPA Test Administrator Directions and Scoring Booklet (Critical Element 2.3.1)

The *Alt ELPA Test Administrator Directions and Scoring Booklet* provides specific, read-aloud directions for every constructed-response item appearing on the test form, and includes scoring rubrics, and local scoring rubrics. The booklet is customized every year for every grade level/grade band, and is meant to be used in conjunction with the *TAM*.

Accessibility and Accommodations (Critical Element 2.3.1)

The *Alt ELPA Accessibility and Accommodations Manual* is a manual intended primarily for district- and school-level educational and assessment staff. It provides information for selecting and using universal features, designated supports, and accommodations for students who need them. It clarifies which of these are embedded in the testing platform and which ones may be provided by the test administrator. It includes information on students' use of augmentative and alternative communication (AAC) devices. It also provides an overview of the Alt ELPA, including purposes of the assessment and participation criteria.

CAI, the test delivery vendor, also produces an *Assistive Technology Manual* available to all participating states through their web-based portals. This manual provides an overview of the embedded and non-embedded assistive technology tools that can be used to help students with accessibility needs complete the online tests. It includes lists of supported devices and applications for each type of assistive technology that students may need, as well as setup instructions for the assistive technologies that require additional configuration within CAI's test delivery system.

Training Courses (Critical Element 2.3.2)

ELPA21 provides self-paced, online training courses through a Learning Management System (LMS) for test administrators, test coordinators, and related personnel. Participating states can elect to have LMS courses available for their practitioners' Alt ELPA test administration training. The courses provide overall training for test administration, including accessibility and accommodations, score reporting, test security, and post-assessment tasks. Refresher courses, which are shorter versions of the initial training courses, are also available for test administrators with prior experience administering the Alt ELPA. An advantage of LMS is the ability to track attendance and usage through individual user logins.

Technology-Based Test Administration Considerations (Critical Element 2.3.3)

The Alt ELPA is entirely technology based. ELPA21 collaborates with the test delivery vendor, CAI, in ensuring that all technology-related system requirements are current and reflected in related resources. These resources are used by the test delivery vendor to ensure standard test delivery procedures for each test administration The *Alt ELPA Technology Requirements* document (available in Appendix E) describes the functions and formats required of the assessment production and delivery systems, including test-taking device requirements, item interaction types, stimuli types, as well as technology requirements for accessibility and accommodation features.

Contingency plans and systems for remediation are the same as those for the general ELPA delivered by CAI. The *CAI System Monitoring* plan lists a variety of potential scenarios and the established response procedures for remediation. The majority of these scenarios have automatic remediation procedures in place such that, at worst, a student or teacher testing in their classroom may experience nothing more than being logged out temporarily before being able to log back in and resume testing without loss of any previously submitted responses. CAI's scope of work for delivering all ELPA21 assessments also clearly outlines procedures for any unanticipated issues impacting the student testing system, including unique tracking IDs for each case and reporting procedures to ELPA21. Any technical malfunctions are reported to ELPA21 within 24 hours.

The TAM includes troubleshooting information for the test administrators and instructions on how to access the Help Desk staff, who are available during the testing window dates.

Other contingency plans (e.g., adverse weather conditions, health epidemic, etc.) should be developed at the local level based on local and state policies.

4.2 Monitoring Test Administration *(Critical Elements 2.4, 2.5.2)*

As described above, the Alt ELPA is entirely technology based and is currently delivered by a single test delivery vendor, CAI. CAI has systems in place to receive and process alerts of test irregularities to help ensure fidelity of test administration for all ELPA21 assessments. Any

irregularities are logged and tracked with unique tracking IDs. In addition, Help Desk reports provide detailed logs of inbound help requests, issue tracking by category, average speed to answer, and resolution times, and include any outages or escalation tickets. Help Desk reports are delivered weekly to ELPA21 during testing windows.

4.3 Test Security *(Critical Element 2.5)*

Test security protocols and procedures are the same for all ELPA21 assessments, and executed from the beginning to the end of a test development and administration cycle, including throughout all activities of item development and review, test administration, data delivery, and data destruction. Multiple procedures are in place to ensure test security for activities under the consortium. For test administration, the consortium provides guidance to states to ensure that assessments are supported by security protocols that establish both fairness for student engagement and validity in the interpretation of results. Each state administering the Alt ELPA must have its own procedures for setting up and monitoring security during and after test administration.

Item Development Security

During item development, ELPA21 follows routine protocols and procedures to ensure the security of testing materials to the extent possible. All item writers contracted by ELPA21 sign non-disclosure agreements (NDA) prior to beginning any work. New items are "written" directly into CAI's online authoring system, which requires secure login and cannot be screen captured. No items exist on paper. Login access is immediately terminated when employment is terminated. All item reviewers also sign NDAs and review items through virtual meetings which are not recorded. Any notes taken during item review meetings are maintained securely by ELPA21 staff or authorized contractors on a secure, cloud-based server.

Test Administration Test Security

The *Alt ELPA Test Administration Manual (TAM)* provides guidance on security and professional code of conduct for test administration (see p. 3), as well as a sample test security/confidentiality agreement for human readers, scribes, and translators (see Appendix E of the *Alt ELPA TAM*), as well a Test Security Chart (see Appendix F of the *Alt ELPA TAM*) listing potential test security issues by level of severity. A secure browser is required for all student assessments.

Prior to beginning any ELPA21 assessment, states must have processes in place to support test security, and may work with the platform vendor to meet these requirements. Each state is required to have in place:

- Comprehensive protocols to respond to possible security breaches (including test and/or item exposure). Minimum standards describing how to discriminate security breaches from other test incidents included in the Test Security Chart (as aforementioned).

- Plans supporting appropriate training on test security procedures for test administrators, test coordinators, principals, teachers, and test proctors. Such training should include, but not be limited to, training on item security and adherence to *TAM* policies.

With regard to overall test security, ELPA21 recommends that each state administering any ELPA21 assessment must have in place a process and associated timeline for ensuring that:

- Test administrators (and any other individuals who will be administering secure ELPA21 assessments) read and understand the *TAM*, the *Accessibility and Accommodations Manual*, and associated ELPA21 training modules before administering any ELPA21 assessments.
- There are clearly defined protocols that describe individuals (staff or otherwise) designated as test administrators or identified in other roles related to administering a secure ELPA21 assessment.
- There is a process in place for monitoring social media for the posting of any secure assessment materials.

States have the responsibility to:

4. Establish clear policies about monitoring during test administration.
5. Monitor, investigate, and report any test security incidents (including social media) throughout the test administration cycle and determine what, if any, action needs to be taken.
6. Ensure all breaches are reported to the state's ELPA21 representative immediately. The ELPA21 state representative must report breaches to ELPA21 as soon as possible, but no later than 24 hours after the incident occurred.
7. – Contact your state's test administration vendor at the help desk.
8. Review and revise monitoring procedures and/or implement corrective action as needed to avoid these situations in the future.
9. Develop a mechanism for tracking security incidents reported by schools/districts.
10. Develop a process for monitoring security incidents in social media.

Breaches must be reported to the ELPA21 state representative and the vendor help desk immediately. Irregularities must be reported within 24 hours of the incident to the ELPA21 state representative using an incident log or other mechanism for tracking security incidents. States should be prepared to provide the following information in the event of a breach:

- date
- state/district/school
- grade/domain (reading, writing, speaking, listening)
- testing mode (online, paper, or braille)

- number of students involved in incident
- description of incident
- district and/or school action taken
- state action taken

ELPA21 also maintains a breach log that is available to states and is a source for the consortium to help identify significant breaches, to advise states on possible actions, and help determine when a breach is resolved.

While ELPA21 provides test security guidelines for states administering its assessments within support documents, maintains the assessments, and keeps track of incidents across administrations, ELPA21 recognizes that any further remediation or penalties for behaviors that result in a test security breach lies with states and local districts.

4.4 Systems for Protecting Data *(Critical Element 2.6)*

Data Integrity (Critical Element 2.6.1)

ELPA21 provides states administering its assessments with resources and guidance to ensure its assessments are supported by security protocols and procedures necessary to ensure the security, confidentiality, and integrity of ELPA21 program data. The manner in which the information is implemented, however, may vary across states and local districts. Each state maintains a signed, legally binding Data Sharing Agreement (DSA) directly with UCLA CRESST (where ELPA21 is housed), which outlines how data will be stored, transferred, accessed, and maintained. DSAs all include language on allowable uses of the shared data, which are for studies to support the validity and the ongoing maintenance of the assessments through quantitative and qualitative analyses.

DSAs name specific individual(s) at UCLA CRESST who become the only named staff allowed to interact with data with personally identifiable information (PII), typically the CRESST Data Custodian and CRESST Information Technology Manager. A *Data Handling Manual*, used internally at UCLA CRESST, outlines in detail, a step-by-step protocol for the above-mentioned named staff to process the data and de-identify for ELPA21 researchers. All data is transferred through secure FTP sites and stored on local servers housed at the UCLA campus, which maintains rigorous security protocols. Individual DSAs specify expiration and termination dates, as well as data destruction dates with secure data destruction requirements. Data transfers that come directly from the test delivery vendor are also shared securely through encryption protocols outlined in the agreement with the test delivery vendor.

Securing Student-Level Assessment Data (Critical Element 2.6.2)

To support the security of student-level assessment data, the *Alt ELPA TAM* described earlier provides step-by-step procedures for online assessment log-in, assessment, and test material processing.

DSAs with each state administering ELPA21 assessments, as described above, outline how sensitive information is handled between states and ELPA21.

The *Data Handling Manual*, as described above, details how sensitive data is handled internally at ELPA21.

Protecting PII (Critical Element 2.6.3)

All student PII data is encrypted while in transit and at rest. Network file transfers containing program data are encrypted using TLS 1.2 encryption. Encryption is provided through commercial-grade, industry-standard cryptographic algorithms and protocols, using commercially reasonable key strengths. Refer to *FIPS 197* for a complete list of approved encryption algorithms and key lengths: http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf.

All data transferred between the operational platform vendor, scoring vendor, and ELPA21 is transmitted using secure communication protocols to protect data in transit. The diagram in Table 3 shows how data can be securely transferred between each component within a test delivery system.

Table 3

Process for Secure Data Transfer

ELPA21 Item Bank	• ELPA21 items must be hosted in a secure environment
Item Transfer	• Items must be transferred to the operational ELPA21 test vendor via secure protocols (ex. SFTP, FTPS, HTTPS)
Operational Vendor's Item Bank	• Store items in a secured operational database • A username and complex password must be utilized for access
Administrative Portal	• Required for assigning tests and particular grade-level ELPA21 forms to students • Usernames with complex passwords must be utilized for access to the portal
Secure Browser	• A "secure browser" is required to maintain test security • The browser must support multiple operating systems and devices
Secure Student Test Client	• The "secure browser" must connect to the student test client application via Secure Sockets Layer (SSL) encryption
Student Response Database	• Student data must be transmitted via encryption, connected to an encrypted database, and remain encrpyted while at rest
Transmit Student Data for Scoring	• Student data for scoring must be transferred to the scoring vendor via a secure protocol (ex. SFTP, FTPS, HTTPS)
Transmit Reports to Administrative Portal	• Reports must be generated in both student and aggregate formats • Data and reports must be accessible only to individuals with permissions to review the state client data
Administrative Portal	• State/District/School report distribution must be accessed via SSL encryption • Reports must be accessible only to approved personnel

All materials associated with the ELPA21 program will be securely disposed of upon completion of the contract or handled according to contract requirements. To satisfy this obligation, platform vendors will follow the guidelines provided by the National Institute of Standards and Technology (NIST), *Guidelines for Media Sanitization, Draft SP 800-88 Rev. 1* (2012): https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-88r1.pdf.

De-Identification Procedures. As aforementioned, only a select number of individuals named in states' DSAs (typically the CRESST Data Custodian and/or the CRESST Information Technology Manager) are allowed to directly interact with data containing PII in order to limit the chances of contact with secure information. The *Data Handling Manual* provides the step-by-step protocol for de-identifying data for ELPA21 researchers to conduct psychometric and operational analyses that support the ongoing maintenance and validity of all ELPA21 assessments, including the Alt ELPA.

Suppression Rules for Reporting. The Alt ELPA, based on a consensus from state representatives during the CAAELP development phase, employs the same suppression rule for reporting as the general ELPA assessments. That is, in all aggregated reporting of student data, the statistics are suppressed when the number of students (i.e., *N* count) is fewer than 10. The suppression rule applies to all situations where student data may be presented, including technical reports and technical presentations, in order to protect the integrity and confidentiality of personally identifiable information.

5. Scoring *(Critical Element 4.4)*

This chapter describes Alt ELPA scores and item scoring processes, including quality control for scoring. Technical details on psychometric models used to derive scores are found in Chapter 7.

5.1 Alt ELPA Scores

The Alt ELPA is comprised of four short testlets (one for each of the four language domains: Listening, Reading, Speaking, and Writing). Alt ELPA summative assessments provide eight scaled scores. The primary scores are the modality scores: receptive and productive. The receptive modality score reflects a student's performance on the listening and reading domains. The productive modality score reflects a student's performance on the speaking and writing domains. These scores are on a two-digit (0-99) scale. This scale is more intuitively meaningful to educators and families and distinguishes the Alt ELPA modality scale score from the general ELPA21 scale scores.

The Alt ELPA summative assessment also provides scaled scores on each of the four language domains. These scaled scores are reported on the same two-digit scale (0-99) as the modality scores.

Students' modality and domain scores are classified into four levels of performance: 1 – Beginning, 2 – Intermediate, 3 – Early Advanced, and 4 – Advanced. Performance levels represent an interpretation of the scaled score, and provide contextual information about a student's performance.

Proficiency is determined through the pattern and level of performance across the receptive (listening and reading) and productive (writing and speaking) modalities, which are calculated using a multidimensional item response theory (IRT) model (described in Chapter 7). Based on their performance level in the assessed (non-exempted) modalities, students are categorized into one of three proficiency categories: Emerging, Progressing, and Proficient.

The Alt ELPA also produces two composite scores (Comprehension and Overall). The composite scores represent a combination of a student's performance in multiple domains. The Comprehension scaled score is determined through the pattern and level of performance in the domains that represent language comprehension (listening and reading), while the Overall score is based on student performance in all four domains. Both scores are 3-digit scaled scores.

Demonstrating Proficiency for Students with Domain Exemptions

The Alt ELPA summative assessments complies with Federal requirements that stipulate that a student's domain exemptions should not preclude them from demonstrating proficiency. Since both listening and reading items provide information for the receptive modality, having responses in either domain is sufficient to draw inference on the student's proficiency in the receptive modality. Likewise, having responses in either the speaking or the writing domains is sufficient to draw inference on the student's proficiency in the productive modality. This approach allows students to demonstrate English language proficiency with one, two, or three domain exemptions.

5.2 Item Types and Scoring Processes

This section summarizes the item types along with the scoring process used for each type. All test items in the Alt ELPA are either machine scored or locally scored by the test administrator during the testing session.

For items in the receptive modality (listening and reading), students are presented with information in audio, written text, or both, and asked to respond. The listening domain subtest presents the students with oral conversations or presentations. Students are then directed to respond to what they heard by engaging with selected-response or technology-enhanced items, such as dragging and dropping a graphic or a selection of text. These items are machine scored. The reading domain subtest presents short correspondence, procedural, literary, or informational passages. Students respond to selected-response or technology-enhanced items, which are machine scored. An example of a technology-enhanced items is one where a student supplies a missing word in a sentence or passage by selecting from a drop-down menu.

To demonstrate their language skills and abilities in the productive modality (speaking and writing), students answer open-ended test items by speaking or writing constructed responses as well as responding to selected-response or technology-enhanced items in the writing domain. The speaking domain subtest has both selected-response and constructed-response items, but not technology-enhanced items. The selected-response items are machine scored, while speaking constructed responses are locally scored.

The writing domain subtest includes selected-response, constructed-response, constructed-response-guided prompts, and technology-enhanced items. The selected-response and technology-enhanced items are machine scored; the constructed-response and constructed-response-guided prompts are locally scored.

Table 4 summarizes the Alt ELPA's item and scoring process by domain.

Table 4

Item and Scoring Types by Domain

	Listening	Reading	Speaking	Writing
Item Type	• selected-response • technology-enhanced	• selected-response • technology-enhanced	• selected-response • constructed-response	• selected-response • constructed-response • constructed-response-guided prompt • technology-enhanced
Scoring Process	• machine	• machine	• machine • local scoring	• machine • local scoring

5.3 Scoring Quality Control

For items using local scoring (described above), general guidance for administering and scoring are found in the *Alt ELPA Test Administrator Manual (TAM)* (see pp. 6-9). Detailed directions and scoring rubrics for every constructed-response item are provided in the *Alt ELPA: Test Administrator Directions and Scoring Rubrics Booklet* for each grade band and applicable language domain.

During the CAAELP development phase, committees collectively decided that local scoring was the optimal process given the student population. English learners with the most significant cognitive disabilities may have personal communication systems that are best understood by a familiar listener, that is, a professional who frequently instructs or supports the student during instruction. Therefore, a student participating in the Alt ELPA summative assessment would

benefit from having an educator who is familiar with the student's personal communication system score the student's constructed responses.

Guidelines for ensuring valid scores through local scoring is also described in Appendix B of the *Alt ELPA TAM*, which summarizes the rationale, provides guidance on how to use scoring rubrics, and how to resolve potential disagreements with a second scorer (who are randomly assigned to observe a selection of lead scorers). Local scoring worksheets, with instructions, are provided for each grade band, for each test administration. Local scoring is an essential part of the test administration training, and is also available in the LMS modules described in Chapter 4. While the guidance for local scoring is provided by ELPA21, states must implement local scoring in accordance with their own local policies (which may or may not require two scorers for every student).

6. Classical Item and Test Analyses

Prior to item calibration using item response theory (IRT) models, item analyses based on classical test theory (CTT) were conducted to screen for items with inappropriate/undesirable features. Various item statistics were used to select items that differentiate students based on their English language proficiency in a valid, reliable, and fair manner. This chapter discusses the statistical indices used to judge the quality of items.

6.1 Item Analyses

Standardized mean scores

Standardized/relative mean score is defined as the average item score divided by the maximum possible score. The division by the maximum possible score "standardizes" the mean scores to lie between zero and one, so that the statistic is comparable across items with different maximum possible scores. For a dichotomously scored item, standardized mean score is equivalent to the proportion of students who answered the item correctly (also known as p-value). The standardized mean score can be thought of as a crude prediction of item difficulty. Based on the standardized mean scores, items that are extremely easy or extremely difficult (i.e., standardized mean score smaller than 0.1 or larger than 0.9) were flagged for review. Items that are too easy or too difficult do not provide much information on students' differential English language proficiency and thus are not desirable.

Relationship between Item score and total score

Items should also be able to differentiate between students who have higher proficiency and those who have lower proficiency in English language. In other words, in each item, high proficiency students are expected to score higher points than low proficiency students. The degree to which such expectation is met can be examined by observing whether the average total score monotonically increases with item score point. The degree of fulfillment can also be represented by a correlation coefficient between item score and total score. In both representations, the total score is the summed score for the modality test on which the item appeared, serving as a proxy of a student's proficiency in the modality. Put differently, it is assumed that items in the modality test "altogether/collectively" represent a student's proficiency in the modality, and it is expected that each item score covary with the total score. The total score is sometimes referred to as "rest score" in that its computation/summation omits the score of the item under scrutiny, to prevent obvious overlap between item score point and total score. When computing item-total correlation, item score is treated as an ordered categorical variable (with underlying latent normal variable), and total score is treated as a continuous variable, which leads to the choice of polyserial correlation coefficient (Olsson, Drasgow, & Dorans, 1982). When the item under scrutiny is dichotomously scored, a special case of polyserial correlation, biserial correlation coefficient is applied. The item-total

correlation is used to evaluate how well the item discriminates between high- and low-performing students. In general, the higher the correlation, the better the item is at distinguishing high- and low-performing students. Hence, items where the average total score did not monotonically increase with the item score point or the item-total polyserial correlation coefficient was less than 0.1 were flagged for review.

Proportion of responses in each item score category

The relationship between item score and total score needs to be interpreted in light of the proportion of responses in each item score category. When there are few endorsements in any score category, the relationship between item score and total score must be interpreted with extra caution. Especially given the small sample size of the target population, the lack of responses in any score category may cause unstable estimation in item calibration. Thus, items were flagged if less than 3% of the examinees occurred in any score point category. Table 5 shows the criteria for evaluating descriptive item statistics by flag number.

Table 5

Criteria for Evaluating Descriptive Item Statistics by Flag Number

Flag	Description
1	Relative mean > 0.9
2	Relative mean < 0.1
3	Proportion in each score category < 0.03
4	Item-total biserial/polyserial correlation < 0.1
5	Average total score not monotonically increasing with item score point

6.2 Differential Item Functioning (DIF) Analyses

Differential item functioning (DIF) exists when an item tends to favor one subgroup, while disadvantaging another, after students across two subgroups are matched on proficiency. DIF analysis refers to the procedure to detect DIF. In general, DIF analysis yields DIF statistics upon which cutoff values are imposed to derive conclusions on the degree of DIF underlying each item. Based on DIF analysis results, we flagged items that may potentially be unfair to students from diverse backgrounds. The flagged items underwent thorough review to determine whether the items may be retained, modified, or discarded/replaced.

For the Alt ELPA, DIF analysis was conducted to compare the functioning of items among students with different gender, economic status, and ethnicity. Among various demographic variables of interest/concern, those three[4] variables had sufficient sample size in both focal

[4] As Alt ELPA accrues more data, additional demographic variables may be used to evaluate DIF.

group and reference group, which is essential for prudent DIF analysis. Specifically, comparisons were made between female students (focal group) and male students (reference group), economically disadvantaged students (focal group) and students who are not economically disadvantaged (reference group), and Hispanic/Latino students (focal group) and non-Hispanic/Latino students (reference group).

A generalized Mantel-Haenszel (MH) procedure (Zwick, Donoghue, & Grima, 1993) was employed to evaluate DIF. The focal group and reference group were matched in proficiency by using the total score as a criterion. The total score is the summed score of the modality test in which the item was administered. We divided the total score into five intervals/bins of equal number of students, where the number of intervals/bins was selected to balance stability and sensitivity of DIF analysis. Presuming that students in the same interval/bin have matching proficiency, they are expected to have similar odds/chances of correctly/incorrectly responding to each item. To quantify the degree to which such expectation is met, different sets of DIF statistics were computed depending on the number of score categories. For dichotomously scored items, we computed Holland and Thayer's (1988) Mantel-Haenszel δ difference (MH D-DIF), its standard error (*SE* (MH D-DIF)), and effect size (ES) based on the standardized mean difference in item scores across groups (Zwick et al., 1993). For polytomously scored items, we computed Mantel and Haenszel's (1959) chi-square statistics (MH CHISQ) and ES.

Based on the statistics, items were classified into three categories (A, B, and C) according to the DIF classification convention illustrated in Table 6. The classification category of A indicates negligible DIF, B indicates moderate DIF, and C indicates large/severe DIF. An item is flagged if its classification category is C in any comparison. The DIF categorization takes into account both statistical and practical significance. Note that the hypothesis testing involved were conducted using a critical value approach (Zwick, 2012). For example, whether MH D-DIF is not significantly different from one at 5% level was tested by whether |MH D-DIF - 1| / *SE* (MH D-DIF) is larger than 1.96, the 97.5th percentile of the standard normal distribution. Likewise, whether MH D-DIF is significantly greater than one at 5% level was tested by whether (|MH D-DIF|-1) / *SE* (MH D-DIF) is larger than 1.65, the 95th percentile of the standard normal distribution.

Table 6

Criteria for DIF categorization

DIF Category	Dichotomous items	Polytomous items
A	MH D-DIF is not significantly different from zero at 5% level *OR* $\lvert\text{MH D-DIF}\rvert < 1$	MH CHISQ is not statistically significant at 5% level *OR* $\lvert\text{ES}\rvert \leq 0.17$
B	MH D-DIF is significantly different from zero at 5% level *AND EITHER* $1 \leq \lvert\text{MH D-DIF}\rvert < 1.5$ *OR* $1 \leq \lvert\text{MH D-DIF}\rvert$ AND MH D-DIF is not significantly different from one at 5% level	MH CHISQ is statistically significant at 5% level *AND* $0.17 < \lvert\text{ES}\rvert \leq 0.25$
C	$\lvert\text{MH D-DIF}\rvert$ is significantly greater than one at 5% level *AND* $1.5 \leq \lvert\text{MH D-DIF}\rvert$	MH CHISQ is statistically significant at 5% level *AND* $0.25 < \lvert\text{ES}\rvert$

Note. Source: Michaelides (2008).

7. Item Calibrations & Psychometric Models

This section provides a brief overview of item calibration, scoring, performance level assignment, and scale score reporting. Details can be found in Alt ELPA scoring specification (Alt ELPA, 2023), scoring business rule (Alt ELPA & CAI, 2023), and a report on calibration and scoring at different levels of aggregation (Kim & Cai, in press).

7.1 Psychometric Background

Item calibration and scoring procedure described in this section was used to produce scores at four different levels of aggregation: modality scores, domain scores, overall scores, and comprehension scores. All scores were obtained through the application of an item factor analysis model (Bock, Gibbon, & Muraki, 1988; Cai, 2010a, 2010b). Special cases of item factor analysis model include multidimensional item response theory (MIRT) model (Reckase, 2009) and item bifactor model (Cai, Yang, & Hansen, 2011). Item factor models used to produce modality scores, domain scores, overall scores, and comprehension scores will be referred to as modality model, domain model, overall model, and comprehension model, respectively, in the following sections.

Item factor analysis model is a confirmatory latent variable model that characterizes the relationship between observed responses and a set of continuous latent variables. The structure of latent variables was determined by theoretical studies (e.g., Thurlow, Christensen, & Shyyan, 2016) and a simulation study conducted prior to operational administration (Shin & Cai, 2023). Specifically, the simulation study tested the performance of hypothesized psychometric models under the conditions specific to the target population (i.e., small sample size and frequent domain exemptions). As a result, the hypothesized psychometric models were found to perform well and thus were operationalized with minor modifications.

For all models, item response probabilities were modeled using the logistic graded response model (Samejima, 1969). The graded response model describes the cumulative probability of achieving all possible item scores using logistic functions, and the probability of achieving each of the possible item scores is defined as the difference between adjacent cumulative probabilities. For each item, a slope and a set of ordered intercept parameters are estimated. Given that the primary unit of assessment is modality, item parameters were calibrated using the modality model that is most akin to the assessment design. In the remaining models, item parameters were fixed to the values calibrated in the modality model. By fixing the item parameters, all scores were linked to lie on the same scale.

7.2 Psychometric Model Specifications

Modality model, a primary calibration model, is a two-dimensional MIRT model with two correlated dimensions corresponding to the two modalities of language, Receptive and Productive modalities. The means, variances, and freely estimated covariances defined a

multivariate normal (MVN) population distribution. The means and variances were fixed to zeros and ones, respectively, to freely estimate all item parameters. In pursuit of model parsimony, item slope parameters were constrained to equality within each modality and gradeband. Put differently, items within the same modality in the same gradeband were assumed to have comparable levels of discriminability.

Domain model, also referred to as an augmented subscore model, is a four-dimensional MIRT model with four correlated dimensions corresponding to the four language domains, Listening, Reading, Speaking, and Writing. Having fixed the item parameters (i.e., slopes and intercepts) to the calibrated values from the modality model, we freely estimated the means, variances, and covariances that defined MVN population distribution. Domain scores obtained from the domain model are *augmented* subscores of modality scores, as the four domains "borrow strengths" from one another via correlations. Such augmentation is especially useful since there are a relatively small number of items in each domain (Haberman & Sinharay, 2010).

Overall model is a restricted hierarchical item factor analysis model, i.e., bifactor model, with a single primary dimension representing overall English language proficiency and two specific dimensions representing modality-specific variances unexplained by the primary dimension. The item parameters were again fixed to the calibrated values from the modality model. With equal slopes on the primary and specific dimensions, the overall model became a version of a testlet model (Wainer, Bradlow, & Wang, 2007) and was isomorphic to a second-order item factor analysis model (e.g., Rijmen, 2010). We freely estimated the variances that defined MVN population distribution, while constraining the modality-specific variances to equality. The constraint ensured that the primary dimension was not dominated by one of the modalities but instead represented an average or overall proficiency by equally weighing the two modalities. The covariances were fixed to zeros as the primary dimension and specific dimensions are assumed to be orthogonal. The means were fixed to zeros.

Comprehension model was a standard unidimensional IRT model fitted to items from the receptive modality. The population distribution of the single dimension was fixed to standard normal distribution (with mean zero and variance one), and the item parameters were again fixed to the calibrated values from the modality model. The comprehension scores from this model are different from the receptive modality scores produced by the modality model. The comprehension scores are based on the Listening and Reading items only. In the modality model, while receptive modality scores are largely determined by Listening and Reading items, they also "borrow strengths" from Speaking and Writing items through the strong correlation between receptive and productive modalities.

7.3 Psychometric Model Estimation

A total of 24 models, 4 models per gradeband, were estimated using flexMIRT® version 3.64 (Cai, 2021). Model parameters were estimated using full-information maximum marginal

likelihood estimation via Bock and Aitkin's (1981) expectation maximization (EM) algorithm. For modality model, we used 49 quadrature points equally spaced from -6 to +6 (on the logit scale) per dimension. For domain, overall, and comprehension models, we used 31 quadrature points per dimension, except for the domain model in gradeband HS that required 49 quadrature points per dimension along with an increased number of the maximum number of iterations in the E-step. In the domain models, the highest correlation was found in HS between Listening and Reading (i.e., 0.968), which could have necessitated more quadrature points for estimation. Standard errors were calculated via the supplemented EM algorithm (Cai, 2008). All calibrations met the termination criterion and were found to have converged to a stable solution in terms of first- and second-order tests (Houts & Cai, 2020).

7.4 Calibrated Parameters

Parameters calibrated from the models are item parameters (i.e., item-specific intercepts and gradeband- and modality-specific slopes) and group parameters that define univariate/multivariate normal distributions in each model. Appendix F includes estimated group parameters that are used as scoring priors.

7.5 Performance Levels

Performance level assignment

Performance levels (PLs) are assigned to modality and domain scores by applying the cut scores. In each grade and modality, three cut scores (L2, L3, and L4) mark the boundaries between four PLs (1, 2, 3, and 4). Since modality scores and domain scores lie on the same scale, the same set of cut scores can be applied to both modality scores and domain scores. That is, cut scores for the receptive modality are applied to Receptive modality scores, Listening domain scores, and Reading domain scores. Likewise, cut scores for the productive modality are applied to Productive modality scores, Speaking domain scores, and Writing domain scores. Specifically, cut scores in the Alt ELPA reporting scale (i.e., in scale scores; see Appendix G) are converted back to IRT logit scale to be applied to modality and domain scores in IRT logit scale prior to their transformation to scale scores. Once students are assigned modality- and domain-level PLs, their profile of modality-level PLs is used to assign their overall PL in three levels: Emerging, Progressing, or Proficient. Students whose estimated PL is 3 or 4 in both modalities are classified as 'Proficient;' Students whose estimated PL is 1 in both modalities are classified as 'Emerging;' All others are classified as 'Progressing.'

Posterior probabilities for performance levels

Importantly, along with the assigned PLs, posterior probabilities associated with each of the modality-level and overall PLs were computed. Students' posterior probabilities for each of the PLs serve two purposes. First, for each student, they convey the uncertainty or confidence in the student's PL assignment. Second, in aggregate, they are used to assess the reliability of the cut scores (see Chapter 9.2). The probability of being assigned to a certain PL is obtained by

integrating over the regions in each student's posterior distribution that are associated with the PL. The posterior distribution is approximated by a normal distribution with the vector of expected a posteriori (EAP) score estimates and associated error variance-covariance matrix as mean vector and covariance matrix, respectively.

For modality-level PLs, univariate normal distribution is used to approximate the posterior probabilities for each PL. Suppose student i has EAP score of $\hat{\theta}_i$ and associated standard error of $\sigma(\hat{\theta}_i)$. The student's posterior distribution of score is approximated by a normal distribution:

$$P(\theta_i|\boldsymbol{y}_i, \hat{\boldsymbol{\gamma}}) \sim N\left(\hat{\theta}_i, \sigma(\hat{\theta}_i)\right),$$

where $\boldsymbol{y}_i$ is the observed responses of student i and $\hat{\boldsymbol{\gamma}}$ is a vector of all model parameter estimates. Let p_{im} be the probability that the student's true score θ_i is at the PL of m defined by the lower bound of c_{m-1} and the upper bound of c_m. For $m = 1, \ldots, 4$, p_{im} is computed as

$$p_{im} = \int_{c_{m-1}}^{c_m} P(\theta_i|\boldsymbol{y}_i, \hat{\boldsymbol{\gamma}}) d\theta_i.$$

There are three unique cut scores (c_1, c_2, and c_3) in each modality, while the lower bound for the lowest PL is fixed to $c_0 = -\infty$, and the upper bound for the highest PL is fixed to $c_4 = \infty$.

Similarly, for the overall PL, a bivariate normal distribution is used to approximate the posterior probabilities for each PL (i.e., Emerging, Progressing, or Proficient), if a student participated in both modalities. Suppose student i has modality scores of $\hat{\boldsymbol{\theta}}_i$ and associated error variance-covariance matrix of $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_i)$. The student's posterior distribution of scores is approximated by a bivariate normal distribution:

$$P(\boldsymbol{\theta}_i|\boldsymbol{y}_i, \hat{\boldsymbol{\gamma}}) \sim MVN\left(\hat{\boldsymbol{\theta}}_i, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_i)\right).$$

The overall PL of Emerging is assigned when the modality-level PLs are both 1s. Hence, the posterior probability for the Emerging category is:

$$p_{i,\text{emerging}} = \int_{-\infty}^{c_{1,\text{receptive}}} \int_{-\infty}^{c_{1,\text{productive}}} P(\boldsymbol{\theta}_i)\, d\theta_{i,\text{receptive}} d\theta_{i,\text{productive}},$$

where $c_{1,\text{receptive}}$ and $c_{1,\text{productive}}$ are upper bounds for PL of 1 in the receptive and productive modalities, respectively. Similarly, the overall PL of Proficient is assigned when the modality-level PLs are both 3 or 4. Hence, the posterior probability for the Proficient category is:

$$p_{i,\text{proficient}} = \int_{c_{2,\text{receptive}}}^{\infty} \int_{c_{2,\text{productive}}}^{\infty} P(\boldsymbol{\theta}_i)\, d\theta_{i,\text{receptive}} d\theta_{i,\text{productive}},$$

where $c_{2,\text{receptive}}$ and $c_{2,\text{productive}}$ are lower bounds for PL of 3 in the receptive and productive modalities, respectively. Finally, the posterior probability for the Progressing category is

$$p_{i,\text{progressing}} = 1 - p_{i,\text{emerging}} - p_{i,\text{proficient}}.$$

When a student did not participate in either of the modalities (i.e., did not participate in any of the domains in one of the modalities), a univariate normal distribution is used instead.

7.6 Transformation to Scale Scores

An inverse logit transformation (i.e., 1 / (1+ exp ($-\theta$))) is used to map the estimated theta scores onto the scale of probability (ranging between 0 and 1). Then, for modality and domain scores, a constant of 100 was multiplied to the scale of probability, so that the scores ranged from 0 to 99. For overall and comprehension scores, a constant of 1,000 was multiplied to the scale of probability, so that the composite scores ranged from 0 to 999. The nonlinear transformation was implemented primarily to facilitate a more intuitive interpretation of the scores, and it also has the advantage of stabilizing the variance in a dataset of small sample size. The variance-stabilized scores can be used for simpler linear statistical procedures, such as cut score extrapolation used for standard setting, as will be illustrated in the next section.

8. Standards Setting *(Critical Element 6.2)*

8.1 Embedded Standard Setting (ESS) Executive Summary

This section is adapted from the executive summary of the *Alt ELPA Standard Setting Technical Report* (Creative Measurement Solutions, 2024; hereinafter: ESS technical report) which contains comprehensive details of the methodologies and multiple processes employed to set standards and establish cut scores that reflect the performance expectations in the Alt ELP standards. The full ESS technical report is organized into eight sections, each of which is summarized here and include key background information, methodologies, and findings.

Section 1. Introduction

A collaborative group of ten state departments of education and national assessment experts, collectively referred to as CAAELP, designed and developed an alternative English language proficiency assessment system for English language learners with the most significant cognitive disabilities. The consortium contracted with Creative Measurement Solutions LLC to design and implement a process to establish performance and proficiency levels for the Alt ELPA assessments. The standard setting design was supported by review and recommendations from various stakeholder groups including the CAAELP Assessment Design Team, the CAAELP Collaborative Council, the CAAELP Technical Advisory Committee (TAC), and educators who comprise the CAAELP Communities of Practice (CoP).

The Embedded Standard Setting (ESS) methodology was selected as the principal standard setting approach. ESS transforms standard setting from the traditional standalone workshop to a set of processes actively integrated throughout the principled assessment design (PAD) and development lifecycle. ESS processes directly contribute to the valid interpretation and use of test scores and improve test quality and the strength of validity arguments by maintaining a consistent focus on optimizing the evidentiary relationship between test items and the Alt ELPA Range performance level descriptors (PLDs).

ESS and supporting processes include:

- The development of Range PLDs, an articulation of the intended interpretations of the Alt ELPA across performance levels for each grade band. Educator review, discussion, and feedback supported the efficacy of the Range PLDs with respect to their description of the full breadth and depth of the Alt ELP Standards (CCSSO, 2019), as described in Section 2 of the ESS technical report.
- Item-PLD alignment, the association of each Alt ELPA item (and within-item score point) with an evidence statement in a Target performance level. Item-PLD alignment activities resulted in a pool of items aligned by design to the Alt ELPA Range PLDs and representing their full breadth and depth. A Summer 2021 Educator Content and Bias Review, ESS analyses, and results from the Inconsistent Item Review and Resolution

(R&R) workshop all provide evidence in support of the resulting alignments, as described in Section 3 of the ESS technical report.

- A Contrasting Groups Study, in which educators' classifications of students into performance and proficiency levels by survey provided an external source of validity evidence supporting the adoption of cut scores and the validity of the Alt ELPA Proficiency Determination rubric, as described in Section 4 of the ESS technical report.
- ESS analyses and the estimation of cut scores in ESS algorithms are employed to produce initial cut score estimates that optimize the coherence of the Target Levels and empirical data from the Spring 2023 Alt ELPA operational field test administration. The analyses (a) largely support the efficacy of the Target Levels and thus, the initial cut score estimates, as described in Section 5, and (b) identify items with Target Levels not supported by the data for subsequent review and resolution, as described in Section 7 of the ESS technical report.
- Vertical articulation, in which within-grade-band cut scores are smoothed to support an integrated and coherent cross-grade system of Alt ELPA cut scores. Vertical articulation is typically conducted following grade-specific standard setting activities to smooth the irregularities across grades that are commonly observed, as described in Section 6 of the ESS technical report.
- The Inconsistent Item Review and Resolution (R&R) Workshop, an activity in which educators iteratively review items identified as inconsistent, provides an opportunity to resolve Target Levels that are not supported by data. The goal of the R&R workshop is to strengthen the evidentiary chain of reasoning from the Alt ELPA Range PLDs to the Item-PLD alignments and ultimately, to score interpretation. The workshop resulted in an improvement in the consistency of the Item-PLD alignments of record, as described in Section 7 of the ESS technical report.
- In post-workshop activities, the full set of standard setting activities to date are considered together in a review by the various stakeholders and in consideration of CAAELP policy goals. The post-workshop activities produced a well-articulated system of cut scores that reflected stakeholders' expectations for impact data after modest smoothing, leading to the adoption of cut scores by the CAAELP Collaborative Council, as described in Section 8 of the ESS technical report.
- The standard setting validity evidence commonly cited in the measurement literature and federal peer review guidelines for states are described and supported via documentation. The summarized evidence demonstrates strong adherence to principles of test score validation and criteria articulated in the measurement literature and in the federal peer review guidelines to states, as described in Section 9 of the ESS technical report.

Findings from each of the above-referenced activities, briefly summarized below (and in detail in the full-length technical report), provide evidence that these ESS processes work together to promote the coherence of the Alt ELPA assessment system and support the validity of the adopted Alt ELPA cut scores.

Section 2. Proficiency and Performance Level Descriptor (PLD) Development

PLD development is the fundamental standard setting activity because the PLDs explicate the knowledge, skills, and abilities (KSAs) associated with each performance level as the interpretations required for valid decision-making. Four performance levels were established for each Alt ELPA modality:

- Level 1: Beginning
- Level 2: Intermediate
- Level 3: Early Advanced
- Level 4: Advanced

Three types of PLDs—Policy, Range, and Reporting PLDs—were articulated to establish and maintain the evidentiary chain from test score to intended interpretation:

> **Policy PLDs**: The Alt ELPA Policy PLDs communicate CAAELP's overarching vision for the Alt ELPA by articulating the intent of each proficiency level with respect to CAAELP policy goals.
>
> **Range PLDs**: The Alt ELPA Range PLDs operationalize and explicate the Alt ELPA Policy PLDs for each Alt ELPA domain. That is, the Range PLDs describe the English learner KSAs that students should have, for each domain, grade band, and performance level, to fulfill the intent of the Policy PLDs. The Range PLDs form a developmentally articulated progression across performance levels within grade bands and within levels across grade bands.
>
> **Reporting PLDs**: Reporting PLDs concisely summarize the Range PLDs. They are provided on individual student score reports for stakeholders to support score interpretation and decision-making.

ESS processes are pertinent to Policy and Range PLD development activities as they support the establishment of cut scores. Policy and Range PLDs were developed by CAAELP staff working with the Assessment Design Team. Range PLDs were developed to bring the Policy PLDs into alignment with the Alt ELP Standards (CCSSO, 2019; hereafter referred to as the Standards) through a process of domain definition in which the Standards are articulated across the range of performances. Range PLDs were developed for each domain and each of the six grade bands to support PAD domain definition, item development, Item-PLD alignment, and score interpretation. The Range PLDs are a structural variation of the Standards, which are articulated across only three levels.

The validity of the Range PLDs is supported by their reflection of the full breadth and depth of the Standards while providing interpretations supporting the Policy PLDs. Developed specifically to meet these goals, evidence that they did so was provided through educator

evaluations in two different workshops. First, in August 2021 an educator workshop was conducted to evaluate the blueprint and item pool, which were both developed in alignment to the Range PLDs. Seven of eight participating educators confirmed the sufficiency of the item pool blueprint and the Range PLDS, with respect to meeting the breadth and depth of the Standards.

Second, in the evaluation of the 2023 Educator Inconsistent Item Review and Resolution (R&R) Workshop, all twenty-one educators agreed with the following statement: "I was able to select the performance level that best targeted each item." Thus, each item could be aligned to an evidence statement in a specific performance level, further validating their link to the Standards and supporting the sufficient breadth and depth of the Range PLDs.

Thus, feedback from educators in the Summer 2021 and 2023 educator meetings provides evidence that the Alt ELPA Range PLDs developed by the CAAELP Assessment Design Team reflect their common foundation with, and the breadth and depth of, the Standards.

Section 3. Item-PLD Alignment

Item-PLD alignment was established in three phases reflecting a comprehensive effort to develop items strongly supporting test score interpretation and thus the goals of the principled approach to the Alt ELPA test design. These three phases were:

- Initial Item-PLD alignment by design. The initial Item-PLD alignment is referred to hereafter as the Target Level.
- Educator evaluation of the Target Levels in the Summer 2022 Item Content and Bias Review.
- Empirical evaluation of the Target Levels via ESS analyses with review and resolution of items identified as ESS-Inconsistent during the Summer 2023 Educator Inconsistent Item R&R workshop.

Each of these phases worked to support an item pool that reflects the Alt ELPA Range PLDs and thus score interpretation. First, alignment by design was conducted by CAAELP's item development partner, Cognia, who developed a pool of items intended to measure the breadth and depth of the Standards, with each item aligned to a specific Range PLD evidence statement. Cognia staff were trained by Creative Measurement Solutions in specific item development methodologies to achieve this goal (training materials available in the accompanying appendices of the ESS technical report). These alignments by design were considered hypothesized until confirmed in the next two phases. Second, as part of the Summer 2022 Educator Item Content and Bias Review, among the criteria evaluated and confirmed were the items' Target Levels. Third, when ESS analyses identified inconsistent items—items with Target Levels not supported by empirical data—educators in the Summer 2023 Inconsistent Item Review workshop reviewed the inconsistent items, and individually and independently aligned the items to a performance level based on the Range PLDs. When strong agreement was not

reached, they discussed differences in their individual Item-PLD alignments and made final alignment judgments.

The efficacy of the initial Item-PLD alignments in the first phase—alignment by design—is supported by the second and third phases. During the Summer 2022 Content and Bias Review, the vast majority of Item-PLD alignments were confirmed. Further, the ESS analyses indicated the proportion of Essentially Consistent items. An item was considered Essentially Consistent if the absolute value of the item's distance to the Target Level is less than or equal to 1 standard error of measurement (SEM) of the test. Approximately 60% of the items were Essentially Consistent with empirical data, rising to over 80% following the application of the recommendations of the R&R workshop in post-workshop ESS analyses. Together, these results largely support the Item-PLD alignments of record and reflect their iterative improvement based on ESS processes.

The three phases leading to empirically validated Item-PLD alignments reflect a comprehensive approach to support the goals of a principled approach to Alt ELPA test design—the development of test items that strongly support score interpretation and test validity via the evidentiary chain of reasoning from the Alt ELP Standards, to the Range PLDs, to items that provide evidence of student performance on the PLD evidence statements as intended, confirmed by subject matter experts (SMEs), and supported by empirical data.

Section 4. Contrasting Groups Study

A Contrasting Groups Study (CGS) was conducted to provide an additional source of validity evidence in support of the ESS cut scores. The CGS cut scores were estimated by comparing students' modality performance and overall proficiency classifications established by qualified educators with those students' Alt ELPA scores. Additionally, modality profiles associated with each overall proficiency level provided by survey respondents were compared to the Alt ELPA Proficiency Determination rubric adopted by CAAELP. Finally, correlations between students' CGS overall and modality levels and students' Alt ELPA scores were provided as evidence in support for the validity of the Alt ELPA.

The surveys were sent to all Alt ELPA test administrators. Respondents evaluated students' Receptive and Productive performances, and their overall English language proficiency based on their observations of students during regular classroom instruction. These data were analyzed for overall efficacy of cut scores, where the cut scores were estimated via logistic regression and the ESS algorithm. Efficacy analyses presented correlations ranging from .05 to .61, low in some cases but showing an overall tendency of linking higher ratings or judgements of teachers to higher scores on the Alt ELPA. The relatively modest numbers of survey responses were insufficient to support logistic regression analyses and thus, ESS analyses were conducted. The case counts (ranging from 14 to 28 per grade) were still low to support valid cut score estimation but provided some useful information in support of the adopted cut scores and CAAELP policy.

First, the validity of the Alt ELPA Proficiency Determination rubric is supported by the CGS analyses. Specifically, 77.1% of the 293 profiles agreed with the Alt ELPA Proficiency Determination rubric when applied to the respondents' Productive and Receptive endorsements for the matched data cases. Second, the CGS results indicate that students tend to be lower performing in the Productive than the Receptive modality, which is paralleled by the ESS analyses reported in Section 5 of the ESS technical report. Third, the CGS analyses indicated greater proportions of students placing in higher performance levels in higher grades for both modalities, also paralleling the ESS analyses reported in Section 5. Finally, positive correlations between the respondents' endorsed modality levels and their Alt ELPA scores support the convergent validity of the Alt ELPA.

Section 5. ESS Analyses

ESS analyses use empirical data from the Spring 2023 Alt ELPA administration to provide four key outcomes. *First*, initial ESS cut scores emerge analytically and organically by optimizing the coherence of the Target Levels and empirical data. *Second*, the efficacy of each item's Target Level is evaluated. Evaluation criteria include:

a) the correlations of empirical item difficulty (IRT RP67 locations) and the ordinality of the Target Level (Level 1 = 1, etc.),

b) agreement rates between the Target and Empirical Levels, where an item's Empirical Level is determined by its IRT scale location relative to the initial ESS cut scores, and

c) weighted Kappa values that quantify the strength of agreement between the Target Levels and Empirical Levels.

Third, impact data—the proportion of students in each performance level—is estimated based on the initial ESS cut scores.

Fourth, lists of ESS-Inconsistent items are produced. These are items with Target Levels that do not agree with the Empirical Levels.

The efficacy of the initial Target Levels is largely supported by the data. Median correlations were moderate to good—adjusted median correlations of 0.74 and 0.68 were observed for the Productive and Receptive modalities, respectively. Median weighted Kappa values reflected a substantial strength of agreement—median Kappas of 0.72 and 0.71 were observed for the Productive and Receptive modalities, respectively. Thus, the results largely support the efficacy of the initial Target Levels, and therefore the initial ESS cut scores, which are used in subsequent ESS iterative activities including vertical articulation (see Section 6), the review and resolution of the items identified as inconsistent (see Section 7), and review by stakeholders leading to the adoption of final cut scores (see Section 8).

Section 6. Vertical Articulation

Under ideal circumstances, the estimation of initial ESS cut scores for each grade band and modality results in cross-grade impact data that is reasonable, meets stakeholder expectations, and supports CAAELP's policy goals. The appropriateness of impact data should be informed by theory and the expectations of SMEs who are knowledgeable about the population of interest—English learners with the most significant cognitive disabilities. When data are not as suggested by theory or as expected, then some statistical smoothing, referred to as vertical articulation, may be necessary. It is common to refine cut scores to support their vertical articulation by panelists during a standard setting workshop and by policymakers, with support by their technical advisors, following a standard setting.

The following guidelines supported Alt ELPA vertical articulation:

1. Relatively smooth transitions in the percentage of students at or above each cut score across grade bands should be observed. That is, there should be no "saw tooth" patterns in the percentage of students at or above each cut score unless supported by theory or expectations.
2. Differences in the impact data across modalities may be acceptable, but only to the degree that might be suggested by theory and/or expectations, given the SMEs' understanding of student proficiency and relative challenges across the modalities.
3. The Level 2, 3, and 4 cut scores should be sufficiently different within a grade band to have a reasonable proportion of students in each level. This is desirable because a

performance level that is obtained by a trivial proportion of students is likely to be unreliable. That is, it is likely that all students in such a level are within an SEM of either of the adjacent levels.

4. The percentage of students in the various levels should reasonably correspond to the expectations of SMEs who are knowledgeable about English learners with the most significant cognitive disabilities.
5. Shifts in cut scores to support smoothing should be as modest as possible.

For the Alt ELPA, determining the extent of necessary smoothing involved the evaluation of impact data based on the initial unsmoothed cut scores. Where deemed necessary, vertical articulation was applied according to the above guidelines. Findings and recommendations are as follows:

The need for smoothing for the Productive modality was modest. Transitions across grade bands were relatively smooth. However, the very small percentage of students (less than 1%) placing at or above Level 3, Early Advanced, especially in the lowest grade bands, was unexpected by the Assessment Design Team and suggested the need for some adjustment. For the Receptive modality, recommendations for smoothing were based on two observations. First, there was a disarticulation in the percentage of students in Level 1 between grade bands K, 1, and 2-3. Second, a relatively low percentage of students placed in Level 2, Intermediate, in grade bands 1 and 6-8, which would likely result in unreliable classifications.

The magnitude of the adjustments to resolve these articulation issues, reported in terms of the SEM of the assessments in Section 6, were generally within accepted industry standards and resulted in a well-articulated system of cross-grade cut scores.

Section 7. Educator Workshop: ESS Inconsistent Item Review and Resolution

ESS-Inconsistent items are defined as items with Target Levels that do not agree with their Empirical Levels associated with the initial ESS cut scores. The identification of inconsistent items is an important product of ESS analyses that provides an opportunity for their review and resolution. This was done through an online workshop in which educators reviewed the inconsistent items and independently evaluated the Item-PLD alignments with the goal of resolving the inconsistencies. The workshop consisted of three panels of educators—well qualified and representing important areas of practice and demographic distributions—evaluating the inconsistent items for two grade bands per panel.

Two rounds of activity were conducted. In Round 1, panelists worked individually to review each item and document their independent Item-PLD alignments. Items with 50% or less alignment agreement among panelists were reconsidered by panelists in Round 2. Round 2 gave panelists a chance to reach a better shared understanding of the item content characteristics relative to the PLDs prior to making their second and final independent judgments. Post-workshop analyses indicated that more than 60% of the inconsistent items had majority agreement after Round 1, and almost 90% had majority agreement after Round 2.

Panelists' consensus alignments confirmed the initial Target Levels for about one-third of the inconsistent items, supporting the initial Target Levels and indicating that some other factor (e.g., opportunity to learn or some construct irrelevant trait) resulted in the items' anomalous difficulties. Panelists' consensus alignments agreed with the Empirical Level for approximately one-third of the inconsistent items, resolving the inconsistencies for those items. The remaining items were either aligned to a new level (neither Target nor Empirical) or did not reach a majority level of agreement.

Recommendations for establishing these items alignments of record based on resolution type and status are as follows:

- When panelists' Consensus Level agrees with the initial Target Level, preserve the Target Level as the alignment of record.
- When panelists' Consensus Level agrees with the Empirical Level, change the alignment of record to the Empirical Level.
- When panelists' Consensus Level disagrees with both the Target and Empirical Levels, preserve the Target Level as the alignment of record, but flag the item for further review.
- When panelists do not reach a sufficient level of agreement, preserve the Target Level, but flag the item for further review.

Recommendations were noted by the consortium and discussed internally and with Technical Advisory Committee members. A decision was made to continue collecting more assessment data and review the additional empirical data.

The flagged items can be reviewed by different stakeholders and the item development team. These will be considered along with other item flags by classical item and test analysis (see Chapters 6) to inform possible flagged item interventions such as item revision, PLD refinement, or realignment of the item to a new level.

R&R Workshop evaluations, provided in Section 7 of the ESS technical report, resulted in median endorsements of 4—Strongly Agree—to all evaluation items. This provides evidence that panelists found the workshop to be productive, meaningful, and efficacious, and that they found that items could be aligned effectively to performance levels, supporting items' coverage of the breadth and depth of the PLDs. Item review and resolution has a potentially powerful and beneficial effect on score interpretation. As Item-PLD alignment inconsistencies are resolved, it brings intended and observed interpretations of the cut scores into coherence with the PLDs.

Section 8. Post-Workshop Analyses and the Adoption of Alt ELPA K-12 Cut Scores

Post-workshop activities include:

- Updated cut score estimates based on the post-workshop recommendations for items' alignments of record,

- A review of the cut score adjustments supporting vertical articulation,
- Updated Item-PLD Alignment efficacy metrics including correlations, agreement rates, and weighted Kappas, and comparison of these to pre-workshop values, and
- The adoption of the vertically articulated cut scores.

The post-workshop cut score estimates were identical to the initial cut scores, so the adjustments proposed in Section 6 of the ESS technical report for vertical articulation were maintained. Improvements in correlations, agreement rates, and weighed Kappa values were observed for both modalities and all grade bands. Though such improvements are expected, their magnitude speaks to the efficacy of the R&R workshops.

All metrics reflect an increase in the efficacy of the alignments of record. The median post-workshop adjusted correlations are 0.90 and 0.83 for the Productive and Receptive modalities, respectively. The median post-workshop weighted Kappas are 0.90 and 0.82 for the Productive and Receptive modalities, respectively.

With the exception of grades 9-12, which share a common set of cut scores, the vertically articulated cut scores reported in Section 6 of the ESS technical report were established for the highest grade in each grade band, by design, as the Alt ELP Standards were written to reflect the goals for the highest grade in each grade band. CRESST conducted statistical analyses, like those conducted for the General ELPA in 2016, to establish cut scores for the lower grades in grade bands 2-3, 4-5, and 6-8. In each of the grade bands, the cut scores established for the highest grade were extrapolated to lower grades via linear regression (see *Alt ELPA Scoring Specification* [Alt ELPA, 2023]). The proposed K-12 system of cut scores for each modality was adopted by the CAAELP Collaborative Council on Aug 17, 2023.

Section 9. Standard Setting Validity Evidence

The efficacy of standard setting processes, the resulting cut scores, and the validity of their interpretations are supported by evidence gathered specifically to inform two key perspectives: those articulated in the measurement literature and those required in support of federal peer review. Methods for gathering this evidence are embedded throughout principled assessment design test development processes. These methods are captured throughout this technical report, and Section 9 consolidates this validity evidence.

A preponderance of evidence supporting the ESS iterative processes and resulting cut scores is presented, including:

- Procedural, internal, and external validity evidence as outlined in the measurement literature; and
- Evidence meeting federal peer review Critical Element 6.2, which specifically relates to standard setting processes.

The evidence documented in Section 9 supports the validity of score interpretations associated with the adopted cut scores. Specifically, the documented evidence supports the

breadth and depth of the Range PLDs with respect to the Standards, the iterative establishment of the Item-PLD alignments of record, and the qualifications and training of panelists making judgments throughout the PAD process.

9. Reliability *(Critical Element 4.1)*

Reliability is defined as "the extent to which a test would give the same result on successive trials" (Wainer & Thissen, 1996). Reliability coefficients are indices that represent the degree to which a test yields consistent results with minimal errors. As such, they place a limit on the construct validity of a test (Peterson, 1994). In the context of Alt ELPA, the concept of reliability applies to the scores, as well as the cut scores used to discretize the scores into performance categories.

Reliability of scores and cut scores is computed and reported across all participants, and is not disaggregated by state. The test takers of Alt ELPA are assumed to be comparable across the states. The English learners with the most significant cognitive disabilities are themselves a narrowly defined student subgroup among all English learners. Moreover, the number of test takers in each state is too small to derive meaningfully interpretable reliability coefficients.

9.1 Reliability of Scores

We employed two approaches to estimate the reliability of scores. First, based on classical test theory (CTT), reliability of each domain and test form[5] (e.g., Listening A, Listening B, Reading A, Reading B, and so on) is computed in each grade band. Second, based on item response theory (IRT), reliability of modality-level, comprehension, and overall scores are computed in each grade band. This ensures that reliability is reported for scores at all levels of aggregation.

The first approach is based on CTT that assumes that the observed score (X) can be expressed as a sum of true score (T) and some error (E) that are orthogonal to (i.e., independent of) one another. The variance of the observed score is the sum of two orthogonal variance components,

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2.$$

Based on the assumption, reliability is quantified as the ratio of true score variance to observed score variance, σ_T^2/σ_X^2 or $1 - \sigma_E^2/\sigma_X^2$. Per the definition of reliability, the ratio can be obtained by *test-retest* method where the same test is administered twice to the same group of students. However, obtaining scores from two independent administrations is often not feasible due to issues such as logistics and cost constraints. Hence, the *internal consistency* method is used as an alternative. Internal consistency quantifies reliability as the degree to which items in a test yield consistent result. By treating each item as a one-item test, the correlation among the items represents the correlation among the one-item tests that are intended to measure the same construct. A representative index of internal consistency is Cronbach's alpha (Cronbach & Shavelson, 2004).

[5] This pertains to School Year 2022-23 due to the limited sample size for the field test administration. Beginning in School Year 2023-24, internal consistency will be evaluated for each domain/modality.

Another approach to assess the reliability of scores is based on IRT. Unlike in CTT where the error was assumed to be homoscedastic or uniform for all scores, IRT allows for heteroscedastic error across the ability continuum. The heteroscedastic error, denoted by σ^2_{error}, has an inverse relationship with test information. Test information indicates the amount of information provided by the test at different points along the ability scale, and its inverse indicates the uncertainty or lack of information at those points. The average of the heteroscedastic errors in IRT, denoted by $\bar{\sigma}^2_{\text{error}}$, is analogous to the error variance in CTT. Moreover, the variance of the expected a posteriori (EAP) scores, denoted by σ^2_{EAP}, is analogous to the true variance in CTT. That is, the variance of EAP scores is used as a proxy of the true score variance, since EAP scores are estimates of the true scores. Hence, the total score variance is decomposed into the EAP score variance and the average measurement error:

$$\sigma^2_{\text{total}} = \sigma^2_{\text{EAP}} + \bar{\sigma}^2_{\text{error}}.$$

The formula can also be understood as the total variance being decomposed into the between-student variance (the variability of scores between students) and the average within-student variance (the average uncertainty of each student's score). As in CTT, the marginal reliability in IRT, denoted as $\bar{\rho}$, is defined as the ratio of EAP score variance to total score variance, $\sigma^2_{\text{EAP}}/\sigma^2_{\text{total}}$ or $1 - \bar{\sigma}^2_{\text{error}}/\sigma^2_{\text{total}}$. $\bar{\rho}$ is a marginal value in that it is based on the average error variance $\bar{\sigma}^2_{\text{error}}$ instead of heteroscedastic error variance σ^2_{error}.

Internal Consistency

Cronbach's alpha (Cronbach & Shavelson, 2004) is used to evaluate the internal consistency of items in each test. A high Cronbach's alpha coefficient indicates that the items in the test are strongly related to each other, as expected for items that measure the same underlying construct (i.e., listening, reading, speaking, or writing). To take into account the sampling distribution and sample size, bootstrap method is employed to compute 95% confidence intervals of Cronbach's alpha (Yuan et al., 2003).

Conditional Standard Error of Measurement

The conditional standard error of measurement (CSEM) or σ_{error} is a standard error associated with the EAP scores. They are conditional values in that the size of σ_{error} depends on the EAP scores. The relationship between EAP scores and CSEM is reviewed to assess which values of EAP scores are associated with smaller or larger CSEM.

Marginal Standard Error of Measurement

The marginal standard error of measurement (MSEM) or $\bar{\sigma}_{\text{error}}$ is computed as the square root of $\bar{\sigma}^2_{\text{error}}$ which is the average of the conditional measurement error σ^2_{error}. Smaller value of MSEM indicates that the EAP scores, on average, have greater precision. To facilitate interpretation, the ratio of MSEM and the standard deviation of the EAP scores is computed: $\bar{\sigma}_{\text{error}}/\sigma_{\text{EAP}}$. The ratio characterizes the noise-to-signal ratio, or the ratio of within-person

variance to between-person variance. The lower value of the ratio indicates that the average error variance is smaller relative to the true variance.

Marginal Reliability

Marginal reliability (Sireci, Thissen, & Wainer, 1991) is defined as the proportion of true score variance among the observed score variance. Marginal reliability, denoted by $\bar{\rho}$, is computed as $\sigma^2_{\text{EAP}}/\sigma^2_{\text{total}}$ or $1 - \bar{\sigma}^2_{\text{error}}/\sigma^2_{\text{total}}$. Ranging from 0 to 1, a higher marginal reliability indicates greater scoring precision.

9.2 Classification Accuracy and Consistency

The reliability of achievement classification is defined as the proportion of students who would be classified in the same way on two replications of the procedure (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). The reliability of classification is assessed in two ways. First, classification accuracy (CA) signifies how precisely students are classified into each PL (Rudner, 2001; 2005). Second, classification consistency (CC) characterizes how consistently students are classified to each PL across two (hypothetical) independent administrations of equivalent forms. In Alt ELPA, since the overall PL is determined by a combination of modality-level PLs, the CA and CC are examined for modality-level cutscores, as well as the thresholds between the overall categories (Emerging, Progressing, and Proficient).

For a classification into L levels, the computation of CA and CC is based on two $L \times L$ matrices, denoted by $\mathbf{A}$ and $\mathbf{C}$, of which elements are the posterior probability of each student being classified to each PL (discussed in Chapter 7.5). Hence CA and CC indices are affected by factors such as the magnitude of measurement error (or the variance of posterior distribution), the distance between adjacent cut scores, the location of cut scores on the ability scale, and the proportion of students around the cut scores. The indices tend to be lower when the measurement error is larger, adjacent cut scores are closer to each other, and there are a greater portion of students distributed near the cut scores. Hence, they are interpreted with caution. Also, we focus our interpretation on the CA and CC of critical cuts, i.e., the cut score between levels 3 and 4 in modality-level classification, and the threshold between Progressing and Proficient in overall classification.

Classification Accuracy

For CA, we compute:

$$\mathbf{A} = \begin{pmatrix} n_{a11} & \cdots & n_{a1m} & \cdots & n_{a1L} \\ \vdots & & \vdots & & \vdots \\ n_{al1} & \cdots & n_{alm} & \cdots & n_{alL} \\ \vdots & & \vdots & & \vdots \\ n_{aL1} & \cdots & n_{aLm} & \cdots & n_{aLL} \end{pmatrix}$$

where

$$n_{alm} = \sum_{i \in \{j | \mathrm{PL}_j = l\}} p_{im}$$

is the probability of student i being classified to the PL of m (p_{im}) summed across all students who were actually assigned the PL of l. The row represents the *observed* PL, and the column represents the *expected* PL. Since the probabilities for each student i sum to 1 (i.e., $\sum_{m=1}^{L} p_{im} = 1$), the sum of all elements in the matrix becomes N, the total number of students. Based on $\mathbf{A}$, the CA for cut score c_l for $l = 1, \ldots, L-1$ is:

$$\mathrm{CA}_l = \frac{1}{N} \sum_{k=1}^{l} \sum_{m=1}^{l} n_{akm} + \frac{1}{N} \sum_{k=l+1}^{L} \sum_{m=l+1}^{L} n_{akm}.$$

The first term is the sum of posterior probabilities where the observed and expected PLs are level l or lower, and the second term is the sum of those where the observed and expected PLs are level $l+1$ or higher. That is, CA_l represents how well the cut score c_l *accurately* distinguishes between level l or lower and level $l+1$ and higher. Moreover, overall CA is the sum of posterior probabilities where the observed and expected PLs are equal:

$$\mathrm{CA} = \frac{1}{N} \sum_{j=1}^{L} n_{ajj}.$$

The overall CA conveys a more stringent criterion of accuracy than CA_l for each cut score l. Higher value of overall CA implies higher confidence and lower uncertainty in students' PL assignments. The overall CA of 0.8, for example, implies that the average posterior probability of students being classified to the PLs they were actually assigned to is 0.8. In other words, on average, we are 80% confident (and 20% uncertain) about the students' PL assignments.

Classification Consistency

For CC, we compute

$$\mathbf{C} = \begin{pmatrix} n_{c11} & \cdots & n_{c1m} & \cdots & n_{c1L} \\ \vdots & & \vdots & & \vdots \\ n_{cl1} & \cdots & n_{clm} & \cdots & n_{clL} \\ \vdots & & \vdots & & \vdots \\ n_{cL1} & \cdots & n_{cLm} & \cdots & n_{cLL} \end{pmatrix}$$

where

$$n_{clm} = \sum_{i} p_{il} p_{im}$$

is the probability of student i being in PL of l in the first administration (p_{il}) and in PL of m in the second administration (p_{im}) summed across all students. Since the execution of two independent administrations is not feasible, it is hypothesized that the posterior probabilities

are replicated exactly in the second hypothetical administration. The row represents the expected PL in the *first* administration, and the column represents the expected PL in the *second hypothetical* administration. Since the probabilities for each student i sum to 1 (i.e., $\sum_{m=1}^{L}\sum_{l=1}^{L} p_{il}p_{im} = 1$), the sum of all elements in the matrix again becomes N. Based on $\mathbf{C}$, the CC for cut score c_l for $l = 1, \ldots, L-1$ is:

$$\mathrm{CC}_l = \frac{1}{N}\sum_{k=1}^{l}\sum_{m=1}^{l} n_{ckm} + \frac{1}{N}\sum_{k=l+1}^{L}\sum_{m=l+1}^{L} n_{ckm}.$$

The first term is the sum of posterior probabilities where the expected PLs are level l or lower in both administrations, and the second term is the sum of those where the expected PLs are level $l+1$ or higher in both administrations. That is, CC_l represents how well the cut score c_l *consistently* distinguishes between level l or lower and level $l+1$ and higher. Moreover, overall CC is the sum of posterior probabilities where the expected PLs are equal between the two hypothetical administrations:

$$\mathrm{CC} = \frac{1}{N}\sum_{j=1}^{L} n_{cjj}.$$

The overall CC conveys a more stringent criterion of consistency than CC_l for each cut score l. Higher value of overall CC implies higher consistency in students' PL assignments. The overall CC of 0.64, for example, implies that the average posterior probability of students being classified as the same PLs in two administrations is 0.64. In other words, on average, there is a 64% chance that students will be assigned the *same* PLs across two independent administrations, and there is 36% chance that they will be assigned *different* PLs across the two administrations.

10. Validity

As discussed in Chapter 1.3, this technical manual will provide the ongoing collection of evidence evaluating the foundational assumptions that underlie the Alt ELPA assessment system. We evaluate the validity of the Alt ELPA assessment following guidelines outlined in the *Standards* (AERA, APA, & NCME, 2014), in which intended uses of the assessment are first articulated. Table 7 thus provides a crosswalk of Alt ELPA assessment purposes and uses of scores with the five core sources of validity evidence which are important for validating each statement. Also included as a source of validity evidence is fairness, which is fundamental to assessment validity. Fairness is interpreted as being responsive to individual differences, which is important for all populations but especially critical for English learners with the most significant cognitive disabilities. An assessment is not valid if it is not fair for the test takers. The crosswalk table is general, but provides a broad picture of how the ongoing collection of validity evidence supports the Alt ELPA's intended uses and interpretations.

Table 8 summarizes the validity evidence by source type, and further distinguishes the validity evidence by whether it is procedural or empirical. Given the very small and highly variable targeted population, the processes of developing and administering the assessment are fundamental to ensuring assessment validity, since it may not be feasible to collect representative empirical evidence for the target population.

Table 7

Alt ELPA Sources of Validity Evidence by Assessment Purposes and Uses Statements

	Sources of Validity Evidence					
Alt ELPA Assessment Purpose/Use	Test Content	Response Processes	Internal Structure	Relations to other variables	Conse-quences of testing	Fairness
To measure students'* progress toward the attainment of English language proficiency in the four recognized domains of Listening, Reading, Speaking, Writing, and includes the academic English language students need to access and achieve grade-appropriate content taught in English.	X	X	X	X		X
To offer a way for eligible students* to demonstrate their English language proficiency on an assessment based on alternate performance expectations for English language development.	X	X	X			X
To provide important information on what students* know and can do in English in order to help parents and educators establish appropriate English language proficiency expectations and inform decision making regarding appropriate English language services and targeted English language instruction.	X		X	X	X	X
To allow students* to be reclassified (i.e., exited from English learner status) based on English language proficiency relative to grade-appropriate performance standards.	X		X	X	X	X
To use scores for accountability purposes.	X		X	X	X	X

Note: * The Alt ELPA Summative assessment is for eligible English learners with the most significant cognitive disabilities.

Table 8

Summary of Alt ELPA Validity Evidence by Source Type

Sources of Validity Evidence	Procedural Evidence	Empirical Evidence	States' Evidence or Future Evidence
Test content	Process of test design, item writing and development (*TM Ch. 3.1-3.2*) Processes related to Embedded Standard Setting (*TM Ch. 8*)	Educator Item Review Committee (IRC) Ratings (*TM Ch. 3.2*) Embedded Standard Setting (Inconsistent Item Review & Resolution Workshop) (*TM Ch. 8*)	
Response processes		Student cognitive laboratory interviews (*TM Ch. 3.2*) Test taker response times (*TR Appendix Table S 7.1*)	Local scoring inter-rater reliability (by states)
Internal structure		Dimensionality analysis (*TR Ch. 10 & Appendix Sec. 12-13*) Scoring Priors (*TM Ch. 7*) DIF analyses (*TR Ch. 6*)	
Relations to other variables			Analysis of students' alternate content assessments
Consequences of testing		Classification accuracy (*TR Appendix Sec. 11*)	EL reclassification rates
Fairness	Process of test design, item writing and development (*TM Ch. 3.1-3.2*) Process of test administration (*TM Ch. 4.1*)	Student cognitive laboratory interviews (*TM Ch. 3.2*) DIF analyses and committee review (*TR Ch. 6 & 8*)	

Note. TM = Technical Manual (this manual); TR = Annual Technical Report.

The following sub-sections provide additional detail on validity evidence by source type, as outlined in Table 8 above, with a focus on those listed in Section 3 (Technical Quality—Validity) of the federal assessment peer review guidance.

10.1 Validity Based on Test Content *(Critical Element 3.1)*

The processes of employing Principled Assessment Design (see Chapter 3) and Embedded Standard Setting (ESS) (see Chapter 8) to design and develop the Alt ELPA summative assessment provide support for validity based on test content. A discussion of how these processes support validity based on test content can be found on pages 96-100 of the *Alt ELPA Standard Setting Technical Report* (Creative Measurement Solutions, 2024).

To summarize, both Alt ELPA PLDs and the Alt ELPA item pool reflect the depth and breadth of the Alt ELP standards. Through educator review workshops (reported in the *Final Report for the Alt ELPA Pilot Study* [Sato, Kao, & Lin, 2022]), educators agreed with the sufficiency of the item pool blueprint and the range PLDs with respect to meeting the breadth and depth of the Alt ELP standards. The Alt ELPA item pool was aligned with Alt ELP standards by measuring the depth and breadth of the range PLDs through a three-phased approach:

- Initial item-PLD alignment by design: Items were written to specific PLDs from the outset (referred to as the Target Level in the ESS report).
- Educator evaluation of the Target Levels: During Item Content, Bias, and Sensitivity Review Committees in the Summer of 2022, educators evaluated items' Target Level PLDs as one of the components of their review.
- Empirical evaluation of the Target levels: As part of the ESS analyses detailed in the *Alt ELPA Standard Setting Technical Report*, items (with empirical data inconsistent with Target Level PLDs) were reviewed through the "Inconsistent Item Review and Resolution Workshop" in the Summer of 2023 by educators. Over 80% of the alignment-by-design Target Levels were confirmed during this event, and there were no challenges to the Alt ELP Standards alignment.

Evidence of alignment is demonstrated through the ESS process. In the *Alt ELPA Standard Setting Technical Report*, Creative Measurement Solutions (2024) argued that

> Embedded Standard Setting takes a stronger view of alignment than traditionally adopted for alignment studies. That is, traditionally, item alignment is viewed in terms of alignment to the standard and the associated claims and measurement targets. Embedded Standard Setting requires a stronger form of alignment—to the standard, claim, measurement target, and a specific performance level. That is, ESS requires the articulation of the standards, claims, and measurement targets across each level of performance—requiring the specific articulation of the learning progression associated with the standard across the performance levels. ESS processes consider the initial Item-PLD alignment—the item's Target Level—a hypothesized alignment

> needing verification by empirical data. This is a stronger verification of alignment than engaged in traditional alignment studies, which only require the consensus of expert opinion. ESS requires the consensus of expert opinion rendered for the Target Level alignment plus verification by empirical data. And when empirical data does not support the Target Level alignment, additional item review is needed.
>
> Thus, evidence of alignment is supported by two separate independent alignment activities. First is the initial Target Level alignment, supported by empirical data. Second, is the additional item review for items whose Target Levels are not supported by empirical data (p. 98).

Test blueprints (see Chapter 3) also show how the item pool represents coverage of Alt ELP standards, PLDs, text complexity, and language processes.

10.1.2 Validity Based on Language Processes (Critical Element 3.2)

The process of item development and writing described in Chapter 3 also intentionally included targeted "language processes" by applying a Language Processes and Language Complexity Framework (available in Appendix B). The framework was developed in the absence of a common, agreed-upon way to identify and evaluate academic English language and the language demands that are reflected in academic state standards. The framework was first successfully applied retroactively to existing ELPA21 assessment items for a general population, which confirmed the presence of language processes in existing items. The framework was thus applied to the development of Alt ELPA assessment items and are reflected in the test item specifications and test blueprints. The presence of language processes in the Alt ELPA assessment items was corroborated during the pilot study through student cognitive laboratory interviews (also described in Chapter 3; see also, *Final Report for the Alt ELPA Pilot Study*, p. 37 [Sato et al., 2022]). Findings from the student cognitive laboratory interviews indicated that Alt ELPA items solicit language processes as intended.

10.2 Validity Based on Internal Structure *(Critical Element 3.3)*

Validity evidence based on internal structure can be found in the Technical Report, Chapter 10 on Validity (dimensionality analysis), and in this Technical Manual, Chapter 7 (and Appendix F: Scoring priors).

DIF analyses are reported in the Technical Report, Chapter 6, and committee reviews are reported in the Technical Report, Chapter 8.

10.3 Validity Based on Relations to Other Variables *(Critical Element 3.4)*

Validity evidence based on relations to other variables is typically collected by analyzing the association between test scores on ELP assessments (e.g., scale scores or performance levels) with other measures such as scores on content assessments. With participation in the

Alt ELPA every year numbering at only a few thousand across all states, and single digit participation for some grade levels in some participating states, collecting this type of validity evidence is a unique challenge.

After consulting with our Technical Advisory Committee, it was suggested that combining two consecutive years' of test administration data (i.e., 2022-23 and 2023-24), though not ideal, could help mitigate the effects of low sample size. As of the writing of this Technical Manual, several states participating in the Alt ELPA have either conducted such analyses with their alternate content assessment data, or plan to do so in 2025, or plan to share their alternate content assessment data with the consortium (in which case the consortium will conduct the data analyses in 2025 on their behalf). The consortium provided a suggested "research template" for states able to conduct their own analyses. It included the following sample tables and figures, based on combining two consecutive years' of test administration data:

- Sample table: Content assessment achievement levels by ELPA21 overall proficiency classification
- Sample table: Content assessment score distribution (*M*, *SD*, and quartiles) by ELPA21 overall proficiency classification
- Sample table: Correlations between ELP and content assessments (overall, modality, domains/subscores)
- Sample table: Summary of results from logistic regression of content assessment (met standard) on ELPA21 overall scale score
- Sample figure: Boxplot showing content assessment score distribution by ELPA21 overall proficiency classification

Findings from these studies will be summarized in future versions of this Technical Manual.

10.4 Fairness and Accessibility *(Critical Element 4.2)*

Chapter 3.1 describes how fairness and inclusion for the targeted population was considered right at the outset of test development as well as throughout the entire process of test development.

Chapter 3.2 describes on item review committees comprised of educators who work with the targeted population were convened to review items for content, bias, and sensitivity using clear guidelines, including guidelines for reviewing an item's accessibility.

The National Center on Educational Outcomes (NCEO), nationally-renowned experts on accessibility, accommodations, and universal design, was contracted throughout the test design and development cycle to develop, in collaboration with state partners, participation guidelines and spearhead the *Alt ELPA Accessibility and Accommodations Manual* (see Chapter 3.1) which is used during test administration (see Chapter 4.1).

The Technical Report Chapters 6 and 8 report on DIF analyses findings and results of committee reviews.

10.5 Ongoing Maintenance of the Assessment *(Critical Element 4.7)*

The Alt ELPA summative assessment follows ELPA21's comprehensive system for monitoring, maintaining, and improving the quality of its assessments. The conceptual framework for this system begins with identifying assessment purposes and uses (reflected in Table 7 above), identifying sources of evidence, planning data collection and analyses strategies, implementing those data collection and analyses plans, and evaluating the evidence against the purpose and use statements.

The review and maintenance cycles are built-in to the annual activities of the assessment program, with governance and committee structures that are set up to review technical information on a recurring basis. As a consortium, ELPA21 is structured with a Governing Board consisting of representatives from each member state. The Governing Board is the primary-decision making body for the ELPA21 program. The Board provides feedback and guidance on all aspects of the program and is informed by various ELPA21 Committees. Decisions related to technical analysis are informed by research and expert advice from two primary committees: (1) the ELPA21 Research & Evaluation (R&E) Advisory Committee, which consists of representatives of ELPA21 member states with backgrounds in psychometrics, assessment validity, and measurement; and (2) the ELPA21 Technical Advisory Committee (TAC), which consists of national experts in the field of assessment, measurement, psychometrics, educational linguistics, bilingual and multilingual education, students with disabilities, and special education. The R&E Advisory Committee meets monthly, while the TAC meets twice annually (every spring and fall). In addition to reviewing studies that support the ongoing maintenance of the assessments, the committees also advise on studies to be conducted before, during, and after any activity that has implications for outcomes and interpretations of the assessment program. Thus, these active, standing committee meetings provide an ongoing feedback loop to ensure the technical quality of the assessment program.

Technical reports on each assessment are produced annually following the year of test administration, which reports on the reliability, validity, scoring, and quality control of the assessment of each operational administration. These reports provide the opportunity to examine the quality of each assessment and plan for ongoing psychometric and research studies.

11. Reporting *(Critical Element 6.4)*

11.1 Types of Scores Reported

Each student receives an Individual Score Report (ISR) which reports several types of scores for the Alt ELPA summative assessment (as described in Chapter 5):

- Overall Scale Score (3 digits)
- Comprehension Scale Score (3 digits)
- Overall Proficiency Determination (Emerging, Progressing, Proficient)
    - Proficiency level determination descriptions
- Modality Scale Scores (2 digits)
    - Productive: speaking and writing
    - Receptive: listening and reading
    - Modality performance level descriptions (Levels 1-4)
- Domain Scale Scores (2 digits)
    - Speaking, writing, listening, and reading
    - Domain performance level descriptions (Levels 1-4)

Note that the overall and comprehension scale scores are provided for program evaluation purposes and are intended for use by policymakers and administrators.

Additional detail on score reporting can be found in the Alt ELPA Score Reporting Manual which is a guidance document for vendors, member states, and school districts, and is updated annually. The manual contains information on policy descriptors, Alt ELP standards, Alt ELPA range PLDs (which identify the performance levels for each domain and each grade/grade band based on the Alt ELP standards), and Alt ELPA reporting PLDs.

A sample ISR is shown in Figure 1 of the *Alt ELPA Quick Guide to Understanding Reports for Educators*, a publicly available document. The Quick Guide is designed to explain score reports in educator-friendly terms, and includes general information on the Alt ELPA, descriptions of the types of scores reported, definitions for each of the overall proficiency determinations, and other pertinent information related to interpreting results and appropriate uses of test scores.

The *Alt ELPA Parent Guide to Student Reports*, also a publicly available document, aims to explain score reports for parents and families of children participating in the Alt ELPA using language that is accessible for parents and families. Other than English, the guide is available in the following home languages of students currently participating in the Alt ELPA (based on the top three home languages from each participating state): Arabic, Chinese (simplified), Karen, Marshallese, Portuguese, Russian, Somali, Spanish, Swahili, and Vietnamese. Participating states may request additional languages for a nominal fee.

Score reports and accompanying user-friendly guides adhere to the guidelines for reporting and interpretation outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, p. 119).

11.2 Report Design Process

The ISR and accompanying resources are the outcome of extensive collaboration among state representatives, the consortium, and CAI (the test delivery vendor). Namely, CAAELP Team 3 (the Sustainability Team) spearheaded the report design process, engaged with state representatives on gathering iterative feedback on mock ISRs (which are generated within CAI's reporting system). State representatives were comprised of technical experts with background experience working with the targeted population. This iterative process ensured that score reporting met initial design requirements, reported on the progress toward English language proficiency of English learners with the most significant cognitive disabilities based on alternate ELP standards, and were accompanied by user-friendly guides to aid score interpretation and uses, to facilitate participating states in releasing results to all stakeholders in a timely manner.

Alt ELPA scores reflect students' progress toward attaining Alt ELP standards. The Alt ELPA reporting structure is consistent with the structure of the Alt ELP standards on which the intended interpretation and use of results are based, with additional granularity. There are three levels of Alt ELP standards, which reflect a continuum of the language skills expressed by the target population (which was determined by experts who work closely with the target population). As item development began, technical experts advised reporting four levels of PLDs to support more precise measurement, but cautioned against five levels (which would be too granular and thus not meaningful for the target population). CAAELP Team 3 collectively decided to transform the three levels of Alt ELP standards into four PLDs for measurement and reporting. Each item was thus written to a key standard and a target PLD (Levels 1-4), along with intended language processes and observable behaviors associated with the specified standard and PLD. Skills, processes, and behaviors for each PLD are specified in the appendices of the item specifications for each grade level/grade band. Students demonstrate the skills, processes, and behaviors associated with attainment of grade-level appropriate language needed to access and participate in content instruction (as expressed in states' alternate content standards), as measured by the range PLDs that guided item writing. The range PLDs were transformed into performance level descriptors for the purposes of score reporting.

11.3 Reports for Schools and Districts

In addition to the ISR, school, district, and state-level reports can be generated within each state's online dashboard through CAI's reporting system. Each state administering the Alt ELPA receives a customized dashboard and *Reporting System User Guide* published by CAI, which provides detailed instructions for specific state-, district-, school-, teacher-level views on how to navigate reports, generate aggregated results, define filters at the state, district, and

school level, and by specific subgroups, as well as view results for individual students in their classroom. The guide also includes instructions on how to export and print results.

Within CAI's reporting system, ISRs can be generated into any language that the state has made available. CAI's reporting portal meets WCAG standards for web accessibility.

References

Alternate English Language Proficiency Assessment [Alt ELPA] (2023). *Alt ELPA scoring specification: School year 2023-2024*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Alternate English Language Proficiency Assessment [Alt ELPA] & Cambium Assessment, Inc. [CAI] (2023). *Business rules for school year 2023-2024 Alt ELPA test scoring*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

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

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy R. (2004). Introduction to evidence centered design and lessons learned from its application in a global E-learning program. *International Journal of Testing*, 4(*4*), 295-301. https://doi.org/10.1207/s15327574ijt0404_1

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261-280.

Cai, L. (2021). *flexMIRT® version 3.64: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Yang, J. S. & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221-248.

Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581-612.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*(3), 307-335.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*(2), 309-329.

Christensen, L. L., Gholson, M. L., & Shyyan, V. V. (2018, April). *Establishing a definition of English learners with significant cognitive disabilities* (ALTELLA Brief No. 1). Retrieved from University of Wisconsin-Madison, Wisconsin Center for Education Research, Alternate English Language Learning Assessment project: http://altella.wceruw.org/resources.html

Council of Chief State School Officers (2019). *English language proficiency standards for English learners with significant cognitive disabilities*. Washington, DC, Council of Chief State School Officers.

Creative Measurement Solutions, LLC. (2024). *Alt ELPA standard setting technical report*. Author.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418.

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2017). Principled approaches to assessment design, development, and implementation. In A.A. Rupp & J.P. Leighton (Eds.) *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (1st Ed.). John Wiley & Sons, Inc.

Gierl, M.J. & Leighton, J. P. (2010, June). *Cognitive models for reading, math, and science: Implications for formative and summative assessment* [Paper Presentation]. Annual meeting of the Canadian Society for the Study of Education (CSSE), Montreal, Quebec, CANADA.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*(2), 209-227.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Brown (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Houts, C. R., & Cai, L. (2020). flexMIRT® user's manual version 3.6: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

Huff, K., & Plake, B. S. (2010). Innovations in setting performance standards for K–12 test-based accountability. *Measurement: Interdisciplinary Research and Perspectives*. *8*(2–3), 130–144. https://doi.org/10.1080/15366367.2010.508690

Kim, Y., & Cai, L. (in press). *Calibration, scoring, and reporting at multiple levels of granularity: Application to Alternate English language proficiency assessment*. (CRESST Report). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lai, J., Gierl., M. J., & Alves, C. (2010, April). *Using automated item generation to promote principled test design and development* [Paper Presentation]. Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, *14*(1).

Luecht, R., Dallas, A., & Steed, T. (2010, May). *Developing assessment engineering task models: A new way to develop test specifications* [Paper Presentation]. Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Michaelides, M. P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment Research & Evaluation, 13*, 7.

Mislevy, R. J., & Haertel, G. D. (2007). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, (*25*)4. https://doi.org/10.1111/j.1745-3992.2006.00075.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-Centered Assessment Design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*, 61–90. Lawrence Erlbaum Associates Publishers.

Nichols, P., Ferrara, S., & Lai, E. (2016). Design and development for next generation tests: Principled Design for Efficacy (PDE). In H. Jiao & R. W. Lissitz (Eds.), *The next generation of testing: Common Core standards, Smarter-Balanced, PARCC, and the nationwide testing movement* (pp. 49–82). Information Age Publishing.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika, 47*(3), 337-347.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*(2), 381-391.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Rijmen, F. (2010). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361-372.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. Practical Assessment, Research & Evaluation, 7(14), 1-5.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monographs No. 17). Richmond, VA: Psychometric Society.

Sato, E., Kao, J. C., & Lin, A. (2022). *Final report for the Alt ELPA pilot study*. Collaborative for the Alternate Assessment of English Language Proficiency (CAAELP).

Shin, N., & Cai, L. (2023). *Alt ELPA field test: A simulation study of psychometric models*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Shyyan, V. V., & Christensen L. L. (2018). *A Framework for Understanding English Learners with Disabilities: Triple the work* (ALTELLA Brief No. 5). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research, Alternate English Language Learning Assessment project: altella.wceruw.org/resources.html

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237–247.

Thurlow, M. L., Christensen, L. L., & Shyyan, V. V. (2016). *White paper on English language learners with significant cognitive disabilities*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes, English Language Proficiency Assessment for the 21st Century.

Thurlow, M. L., Liu, K. K., Goldstone, L., Albus, D., & Rogers, C. (2018). *Alt-ELPA21 participation guidelines.* Los Angeles, CA: Regents of the University of California.

U.S. Department of Education. (2018). *A State's guide to the U.S. Department of Education's assessment peer review process*. Washington, D.C.: Author.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29.

Wolf, M. K., Farnsworth, T., & Herman, J. (2008). Validity issues in assessing English Language Learners' language proficiency. *Educational Assessment, 13,* 80-107.

Yuan, K. H., Guarnaccia, C. A., & Hayslip Jr., B. (2003). A study of the distribution of sample coefficient alpha with the Hopkins symptom checklist: Bootstrap versus asymptotics. *Educational and Psychological Measurement, 63*(1), 5-23.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, 20, 79–87.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

Zwick, R. (2012). *A Review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report RR-12-08). Princeton, NJ: Educational Testing Service.

Appendix A:
Principled Approaches to Test Development

The following is excerpted from Ferrara, Lai, Reilly, and Nichols (2017, pp. 45-51) and provides overviews of ECD, AE, and PDE, particularly in terms of how each addresses the six elements of principled approaches: clearly defined assessment targets; statement of intended score interpretations and uses; models of cognition, learning, or performance; alignment measurement models and reporting scales; manipulation of assessment activities, and ongoing accumulation of evidence to support validity arguments. These principled approaches collectively influenced the development of the Alt ELPA.

Evidence-centered Design (ECD)

ECD is a framework for identifying, developing, and operationalizing theories and models of learning and cognition in assessment design and development. It makes explicit the assessment argument (e.g., Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006, Table 4.1) in the form of claims about what examinees know and can do based on evidence generated in the assessment process. The ECD process is organized in five layers. During the design, development, and implementation planning process, assessment designers cycle through these layers rather than move through the layers sequentially (Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003). ECD may be the most widely implemented of the principled frameworks and most widely recognized.

In the first layer, domain analysis, assessment designers gather information about the domain of interest that might be useful for assessment design and development, including models or theories of learning, models of performance, specialized vocabulary, and the kinds of technology and tools used in the domain. In the second layer, domain modeling, assessment designers organize the information gathered during the process of domain analysis in a design document to support later assessment design and development decisions. Design pattern tools are used in ECD to help document and organize this information (e.g., Mislevy & Haertel, 2006, Table 2). A design pattern is a table with fields that prompt the assessment designer to record the knowledge, skills, and abilities (KSAs), the important content, and the important performances, among other things, that assessment development should include to support the development of a family of assessment activities. This information is made more specific in the next layer, the conceptual assessment framework.

In the third layer, the conceptual assessment framework, assessment designers create three model architectures: student model(s), task model(s), and evidence model(s). These components further refine the information gathered and organized in the design document. The student model delineates aspects of the targets of inference that the assessment designer intends to make inferences about, given the purpose of assessment. Task models represent the content to be used to elicit student performance that will be used as evidence about the targets

of inference in the student model. The content is described in terms of features that may be classified as characteristic, variable, or irrelevant. These content features would have been identified earlier in the process during domain analysis and may be based on research findings, expert judgment, or may be untested assumptions.

Each task model is used to generate multiple assessment activities that are explicitly related via their content features and, potentially, with similar psychometric features. The evidence model represents instructions for interpreting students' performance and consists of three parts: work product specifications, evidence rules, and the statistical model.

Work product specifications describe the structure and format of the performance that will be captured, evidence rules describe how to code work products (e.g., using a rubric for students' use of argument in science) to capture aspects of the construct, and the statistical model describes how the coding of the responses will be aggregated to make inferences about what students know and can do.

Layers four and five in ECD are assessment implementation and assessment administration, respectively. During the assessment implementation process, the tools created in the conceptual assessment framework are used to write items and tasks, construct rubrics or other evaluation rules, and scale the assessment. During assessment administration, the assessment is administered and results are analyzed and reported; these practical implementation aspects are described in what is known as the four-process model.

ECD was implemented for the PARCC (see http://parcconline.org/), Smarter Balanced (see http://www.smarterbalanced.org/smarter-balanced-assessments/), NCSC (see http://ncscpartners.org/Media/Default/PDFs/NCSC-Policymaker-Handout-2-20-14. pdf), and Dynamic Learning Maps (see http://dynamiclearningmaps.org/content/test- development) statewide assessment programs required under Race to the Top. In addition, SRI International supports other organizations' use of ECD (see http://www. sri.com/work/projects/padi-applying-evidence-centered-design-large-scale-science- assessment) while Cisco Systems (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004), the Educational Testing Service, the College Board (Huff & Plake, 2010), and other groups or organizations have implemented ECD for their own assessment initiatives.

Assessment Engineering (AE)

The AE approach is a "highly structured and formalized manufacturing-engineering process" (Luecht, 2013, p. 3) with four stages: (1) construct mapping and evidence modeling, (2) task modeling, (3) designing item templates and writing items, and (4) calibrating items and quality control (Luecht, 2013). The stages are designed and implemented to achieve "three fundamental assertions" (Luecht, 2013, p. 6), which are that (a) the content requirements and complexity of items differ across the examinee proficiency and test score scale, (b) a "family" of items can be designed from a model of task complexity that specifies declarative and procedural knowledge and other requirements for responding to items in the family, and (c)

large numbers of items can be engineered within the same family with the same task complexity and psychometric (e.g., item difficulty) properties.

During the processes of construct mapping and evidence modeling, the assessment designer develops a construct map, which is a set of claims about examinees that are ordered along a complexity scale that coincides with the intended proficiency continuum and score reporting scale, similar to achievement level descriptors. During this stage, designers also create evidence models, or descriptions of what performance at each level of this ordered scale looks like. The second stage, task modeling, focuses on creating a set of specifications for a family of related task templates, which are themselves more detailed specifications for families of related items or assessment tasks. These specifications include detailed descriptions of the assessment targets, response demands of items in the task family, as well as other item or task features that may impact cognitive complexity, and so relate back to both the construct map and the evidence models.

In AE, the specifications are written using a highly controlled language called a task model grammar. These grammars are potentially programmable specifications for generating items in the same family so that they are isomorphic in terms of cognitive complexity (i.e., declarative, procedural, and other response demands) and in location on the proficiency scale. Once the task models are created, they can be arrayed along the complexity scale to create a task model map that portrays which locations along the proficiency scale will be given the greatest emphasis during task development.

Each task model is then implemented during the processes of designing item templates and writing items to develop item templates. The templates provide even more detailed specifications, including item format and scoring rules, manipulable features, and evaluation criteria. By systematically varying parameters within the manipulable features, item writers or programmed task model grammars can create multiple items from the same template. These items are expected to be co-located on the complexity/proficiency scale through their connection to the item templates and task models. During the final stage, calibrating items and quality control, items are field tested and calibrated using modern measurement models to confirm that the hypothesized complexity/proficiency ordering of items, templates, and task models actually holds, and to make adjustments where it does not.

The AE approach was used to demonstrate how to develop construct map versions of cognitive models (Gierl & Leighton, 2010), task models, and an associated task model map (Luecht, Dallas, & Steed, 2010), as well as to generate and evaluate 10,301 items based on 15 item templates (Lai, Gierl, & Alves, 2010) for the Critical Reading and Mathematics section of the College Board's Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT).

Principled Design for Efficacy (PDE)

The PDE approach builds on ECD (Nichols, Ferrara, & Lai, 2016). As a result, it shares several concepts and tools such as domain analysis and domain modeling but emphasizes

concepts and practices in unique ways. Specifically, the central role of KSA research from the learning sciences in construct definition and assessment activity design, as well as the emphasis on communication among stakeholders throughout the design and development process, stand out. The PDE approach is implemented as a principled enhancement to conventional, recognizable practices as illustrated in Figure 3.3 (discussed later) rather than a new, seemingly unfamiliar approach that can be off-putting to test developers and testing program managers.

The PDE approach to the design and development process consists of four stages and a framework with six design concepts. The four stages are named, designed, and carried out in ways that should be familiar and easily comprehensible to test developers and managers. During the first stage, construct definition, the assessment designer explores research literature from the learning sciences to define academic content standards or other assessment targets in terms of cognitive processes, knowledge structures, strategies, and mental models that are more fine-grained than educational content standards. The assessment designer uses research literature findings to describe features of stimuli and items that most effectively elicit the cognitive assessment targets, which are described as characteristic and variable content around which stimuli and items are developed.

During the second stage, content creation, assessment designers take advantage of the characteristic and variable content features to create stimuli and items that assess the full range of test-taker performance in relation to the assessment targets, as well as rubrics for evaluating examinee test performance. The third stage, generalization, focuses on using the stimuli and items written during content creation activities to create reusable guidelines and specifications. Finally, during the fourth stage, content re-creation, content developers use the guidelines and specifications to generate additional numbers of stimuli, items, and rubrics.

The six design concepts for the work in the four stages are “intended to facilitate reasoning and communication in assessment design and development” (Nichols et al., 2016, p. 56). The construct design concept represents the assessment targets. The evidence design concept articulates features of test-taker responses that will be collected, as well as how they will be evaluated and aggregated. The content design concept specifies the features of stimuli and items that are needed to elicit those responses. The other design concepts include communication with stakeholders (e.g., examinees, item developers), assessment implementation consistent with practical constraints, and the consequences or theory of action for the assessment, which captures the intended outcomes of the assessment as well as the mechanisms for achieving them.

PDE has been used to develop a theory of action, task models, and performance assessment tasks for a system-wide elementary and middle school formative assessment program for the Baltimore County (Maryland) Public Schools, NGSS assessments for the Maryland statewide assessment program, and the Insight Science and Dialogue for Language Learners systems, two digital-device-based learning and formative feedback systems, now in development at Pearson.

Appendix B:
Language Processes and Complexity Framework

The list of language processes from the framework is excerpted below. For the full framework with background information, see Appendix C of the *Final Report for the Alt ELPA Pilot Study* (Sato, Kao, & Lin, 2022).

Language Processes		*Operational Definition—The English **language needed to engage with and achieve** in the content (standard or item) consists of the **use of**:*
General Category	Specific Process	
Identifying	Identification	A word or phrase to name an object, action, event, idea, fact, problem, need, or process.
	Labeling	A word or phrase to name an object, action, event, or idea.
	Enumeration	Words or phrases to name distinct objects, actions, events, or ideas in a series, set, or in steps.
Classifying	Classification	Words, phrases, or sentences to assign/associate an object, action, event, or idea to the category or type to which it belongs.
	Organization	Words, phrases, or sentences to express relationships between/among objects, actions, events, or ideas, or the structure or arrangement of information. Discourse markers include coordinating conjunctions such as *and*, *but*, *yet*, *or*.
	Sequence	Words, phrases, or sentences to express the order of information (e.g., a series of objects, actions, events, ideas). Discourse markers include adverbials such as *first*, *next*, *then*, *finally*.
Comparing	Comparison/ Contrast	Words, phrases, or sentences to express similarities and/or differences, or to distinguish between two or more objects, actions, events, or ideas. Discourse markers include coordinating conjunctions *and*, *but*, *yet*, *or*, and adverbials such as *similarly*, *likewise*, *in contrast*, *instead*, *despite this*.
Inquiring	Inquiry	Words, phrases, or sentences to solicit information (e.g., *yes-no* questions, *wh-* questions, statements used as questions).
Describing	Description	Word, phrase, or sentence to express or observe the attributes or properties of an object, action, event, idea, or solution.
Defining	Definition	Word, phrase, or sentence to express the meaning of a given word, phrase, or expression.

Language Processes General Category	Specific Process	*Operational Definition—The English* ***language needed to engage with and achieve*** *in the content (standard or item) consists of the* ***use of****:*
Explaining	Causality	Phrases or sentences to express causal relationships, causes and effects related to one or more actions or events. Discourse markers include coordinating conjunctions *so* and *because*, and adverbials such as *therefore*, *as a result*, *thus.*
	Explanation	Phrases or sentences to express the rationale, reasons, or relationships related to one or more actions, events, ideas, or processes that are non-causal. Discourse markers include coordinating conjunctions *for*, and adverbials such as *for that reason*.
Summarizing	Retelling	Phrases or sentences to relate or repeat information. Discourse markers include coordinating conjunctions such as *and*, *but*, and adverbials such as *first*, *next*, *then*, *finally*.
	Summarization /Synthesis	Phrases or sentences to express important facts or ideas and relevant details about one or more objects, actions, events, ideas, or processes. Discourse structures include: beginning with an introductory sentence that specifies purpose or topic.
Interpreting	Interpretation	Phrases, sentences, or symbols to express understanding of the intended or alternate meaning of information.
Analyzing	Analysis/ Evaluation	Phrases or sentences to indicate parts of a whole and/or the relationship between/among parts of an action, event, idea, or process. Relationship verbs such as *contain*, *entail*, *consist of*, partitives such as *a part of*, *a segment of*, and quantifiers such as *some*, *a good number of*, *almost all*, *a few*, *hardly any* often are used. Phrases or sentences to express a judgment about the meaning, importance, or significance of an action, event, idea, or text.
Extended Thinking	Generalization	Phrases or sentences to express an opinion, principle, trend, or conclusion that is based on facts, statistics, or other information, and/or to extend that opinion/principle/etc. to other relevant situations/contexts/etc.
	Inference	Words, phrases, or sentences to express understanding of implied/implicit based on available information. Discourse markers include inferential logical connectors such as *although*, *while*, *thus*, *therefore*.
	Prediction	Words, phrases, or sentences to express an idea or notion about a future action or event based on available information. Discourse markers include adverbials such as *maybe*, *perhaps*, *obviously*, *evidently*.
	Hypothesis	Phrases or sentences to express an idea/expectation or possible outcome based on available information. Discourse markers include adverbials such as *generally*, *typically*, *obviously*, *evidently*.
Persuading	Argumentation	Phrases or sentences to present a point of view with the intent of communicating or supporting a particular position or conviction. Discourse structures include expressions such as *in my opinion*, *it seems to me*, and adverbials such as *since*, *because*, *although*, *however*.

Language Processes General Category	Specific Process	*Operational Definition—The English* ***language needed to engage with and achieve*** *in the content (standard or item) consists of the* ***use of****:*
	Persuasion	Phrases or sentences to present ideas, opinions, and/or principles with the intent of creating agreement around or convincing others of a position or conviction. Discourse markers include expressions such as *in my opinion*, *it seems to me*, and adverbials such as *since*, *because*, *although*, *however*.
	Negotiation	Phrases or sentences to engage in a discussion with the purpose of creating mutual agreement from two or more different points of view.
Critiquing	Critique	Phrases or sentences to express a focused review or analysis of an object, action, event, idea, or text.
Representing	Symbolization & Representation	Symbols, numerals, and letters, to represent meaning within a conventional context (e.g., +, -, CO_2, >, Δ, π, cos, $y=3x+4$, $c^2= a^2+ b^2$, $h/2(b_1+b_2)$, *cat* vs. cat).
None	No Academic Language Function	Item or standard does not contain *any* academic language functions; may contain linguistic skills (e.g., phonemic awareness, syllabication).

Appendix C:
Alt ELPA Pilot Study: Executive Summary

The following Executive Summary is an excerpt from the *Final Report for the Alt ELPA Pilot Study* (Sato, Kao, & Lin, 2022). For more information, see the full report.

The Collaborative for the Alternate Assessment of English Language Proficiency (CAAELP), a four-year (2019-2023) Competitive Grant for State Assessments Programs funded by the U.S. Department of Education's Office of Elementary and Secondary Education, conducted a pilot study to gather information that could inform its development of an alternate English language proficiency (ELP) assessment for English learners with the most significant cognitive disabilities, the Alt ELPA. The Alt ELPA will align to the *English Language Proficiency Standards for English Learners with Significant Cognitive Disabilities* (CCSSO, 2019) and will be administered annually to eligible English learners with the most significant cognitive disabilities in kindergarten through Grade 12. The purpose of this assessment is to measure students' progress toward the attainment of ELP in the four recognized language domains of listening, speaking, reading, and writing, and includes the academic English language students need to access and achieve grade-appropriate content taught in English. The Alt ELPA will satisfy requirements of the U.S. Elementary and Secondary Education Act (ESEA), as amended by the Every Student Succeeds Act (ESSA, 2015; Sections 3111(b)(2)(G), 1111(b)(1)(F), 1111(b)(2)(G), 34 CFR §§ 200.2(b)(2), (b)(4), (b)(5), 200.6(h)(2)).

The primary purpose of the pilot study was formative; that is, its purpose was to gather information that could be used to develop and refine test items and other assessment documents such as the Accessibility and Accommodations Manual and the Test Administration Manual. In particular, information relevant to accessibility and the appropriateness of the test, items, and conditions for administration was gathered.

As a means for gathering this information, CAAELP researchers conducted a two-phased pilot study, presenting educators and students with test items and engaging them in discussion relevant to the test. Phase 1 involved educator focus groups, and Phase 2 involved student cognitive interviews. The specific purposes of each phase of the pilot study were as follows:

Phase 1 Pilot: Educator Feedback on Assessment

1. To gather information about K-12 English learners with the most significant cognitive disabilities (e.g., characteristics, instructional experiences, accessibility supports and accommodations);
2. To solicit feedback from teachers about the accessibility and appropriateness of Alt ELPA items, as well as information about the language processes that the items are designed to measure; and

3. To solicit feedback from teachers about the planned assessment model and related test blueprints.

Phase 2 Pilot: Cognitive Interviews with Students

1. To gather information about the accessibility and appropriateness Alt ELPA items for English learners with the most significant cognitive disabilities; and
2. To gather information about the language processes that the items are designed to measure.

Phase 1 took place July through September 2021, and Phase 2 took place February through April 2022. A summary of findings from Phases 1 and 2 of the pilot study follows:

1. Alt ELPA items are accessible.

Phase 1 educators were presented a sample of items that represented a range of item types that could be included on the Alt ELPA in Grades K-12. Educators made recommendations that included: making illustrations clearer; limiting the number of details presented in text (e.g., two rather than three); breaking longer tasks into steps or parts; and considering that autistic students may struggle with passages and items that involve and ask about emotions or feelings of others. Educators also identified some words and concepts that might be unfamiliar to and challenging for some students. Despite this feedback, educators generally believed that the items would be accessible to a range of students eligible for the Alt ELPA, assuming a student was not exempt from the domain to which an item was associated (e.g., listening, speaking, reading, writing) and students were able to use appropriate accommodations.

For the Grade 4-12 items included in Phase 2 of the pilot study, all students except for one Grade band 4-5 student completed the items in their assigned cognitive interview protocol. For the student who did not complete the items in their protocol, the TA stopped the session, consistent with stopping rule guidance provided, because the student did not respond to the items or prompting. Although the student interacted with a few of the items, the student had difficulty paying attention to the items and the TA's prompting, and the student disengaged. According to the TA, the student's behavior during the cognitive interview was typical in terms of the student's level of attention and level of communication.

Some students were distracted during their cognitive interview session (e.g., by the protocol that the TA was reading, by objects in the room), but generally students attended to each item's directions, stimulus, and prompt with guidance from their TA, and they were able to engage with the items. That all but one student was able to engage with and complete items in their protocols suggests that students were able to access them. This applies to all item types included in the protocols.

Based on the expert judgment from Phase 1 and students' engagement with items in Phase 2, the item types developed for the Alt ELPA appear to be accessible to a range of students who could be eligible for this assessment.

2. Alt ELPA items are appropriate.

Phase 1 educators reviewed a sample of items that represented a range of item types that could be included on the Alt ELPA in Grades K-12. Educators identified some words and content that might be challenging for some students. However, educators believed that the items generally would be appropriate for a range of students eligible for the Alt ELPA; that is, most students likely would encounter questions in their instruction similar to those represented in the Alt ELPA items.

For the Grade 4-12 items included in Phase 2 of the pilot study, students either explicitly stated that they understood an item's directions, stimulus, prompt, and response options, or they indicated understanding with nods or gestures when asked by their TA. Some students asked for help understanding (e.g., asked TA to repeat information), but students generally were able to understand the items and the information presented.

For each item, as part of the retrospective protocol, students were asked the following: (1) How many times have you seen a question like this before? (2) Did you understand how to answer this question? and (3) How did you feel about this question? In response to Question 1, most students indicated that they saw questions like the items presented to them "a few times," "several times," and "a lot." Regardless of the frequency with which students encountered questions like the items in their protocols, students generally showed that they were able to interact with and provide a response to the items in their protocols. Regarding Question 2, most students indicated that they understood most or all of the items in their protocol. When students indicated that they did not understand how to answer a question, observations showed that the students appeared to engage with the item; therefore, students' indication of a lack of understanding may be reflective of their level of ELP and not necessarily the appropriateness of the item in terms of grade-level content. In response to Question 3, most students indicated that the items were "easy" or "about right." Some students indicated that some of the items were "hard." Student responses to Question 3 were as expected, given that the Alt ELPA items are written for a range of levels of ELP.

The focus of the analysis of the cognitive interview data was not whether the students answered each question correctly. Rather, the analysis focused on whether students understood the directions, stimuli, prompts, and response options in order to interact with the item and provide a response. Based on the expert judgment from Phase 1 and students' indications of understanding and interactions with items in Phase 2, the item types developed for the Alt ELPA appear to be appropriate for a range of students who could be eligible for this assessment.

3. Alt ELPA items solicit language processes as intended.

For their evaluation of language processes, Phase 1 educators typically identified more than one language process for an item because the language process used by a student likely is dependent on or reflective of their level of English language proficiency. For example, a student

with a lower ELP level may use the language of *identifying*, whereas a student with a higher ELP level may be able to use the language of *describing* for a given item. If the targeted language process for an item was identified by educators among a few identified, then it was determined that the item solicited the language process as intended. Based on educators' evaluations of items, the K-12 items were likely to solicit language processes as intended.

Data from Phase 2 observations (TA, researchers, observers) were examined along with available transcriptions of student responses to items. In most cases, items solicited more language from students during the cognitive interview because students were prompted to share their thinking about an item in terms of, for example, describing what they were thinking when they looked at an item's stimulus or chose a particular response. Therefore, similar to the analysis of Phase 1 educator evaluations of language processes, if the targeted language process for an item was among those identified in the observations and/or reflected in related student transcripts, then it was determined that the item solicited the language process as intended. Based on these data, all Grade 4-12 items included in the cognitive interview protocols appeared to solicit language processes as intended.

Based on the expert judgment from Phase 1 and observations and examinations of the language students used to respond to items in Phase 2, the items appear to solicit language processes as intended.

4. Additional findings

Test Blueprints: Breadth and Depth of Standards Coverage

Phase 1 educators were asked the following: Does our approach to developing the Alt ELPA seem reasonable to ensure a breadth and depth of standards coverage? The educators considered the standards for English language proficiency for English learners with the most significant cognitive disabilities (CCSSO, 2019), to which the Alt ELPA items are aligned and that the Alt ELPA will measure, the draft test blueprints, and Policy Performance Level Descriptors. The educators generally agreed that the Alt ELPA would likely cover the breadth and depth of the standards.

Additionally, Phase 1 educators reviewed the planned assessment model for the summative Alt ELPA, which is a stage-adaptive model. Educators were asked: Given the different possible test experiences we discussed, would a student likely experience a breadth (i.e., all ten standards) and a depth (i.e., a range of complexity/difficulty) of the ELP standards? The educators generally agreed that a student would likely experience a breadth and depth of the ELP standards if they were to complete the Alt ELPA assessment.

Findings from the pilot study were shared with item developers and members of the CAAELP project, and these findings served their formative purpose—they were considered or used, as appropriate, as items and other assessment materials (e.g., Accessibility and Accommodations Manual, Test Administration Manual) were developed and refined. This

report presents a description of the pilot study's methods and discussion of findings, limitations, and recommendations.

Appendix D:
Alt ELPA Summary Blueprints

Proportion of Standards Coverage by Domain

Table D1

Proportion of Standards Coverage by Domain Across All Grade Bands

Domain	Standard	Proportion of Test
Listening	1	22.5–25.0%
	2	12.5–22.5%
	5	10.0–22.5%
	6	0.0–20.0%
	8	12.5–20.0%
Reading	1	17.5–25.0%
	2	5.0–25.0%
	5	5.0–15.0%
	6	0.0–20.0%
	8	10.0–20.0%
Speaking	1	0.0–2.5%
	2	5.0–22.5%
	3	17.5–25.0%
	4	5.0–25.0%
	5	5.0–22.5%
	6	0.0–17.5%
	7	7.5–25.0%
	8	0.0–2.5%
Writing	2	2.5–22.5%
	3	20.0–25.0%
	4	15.0–25.0%
	5	7.5–20.0%
	6	0.0–15.0%
	7	15.0–25.0%

Distribution of Standards Per Test Form

Figure D1

Distribution of Standards Per Test Form—Kindergarten

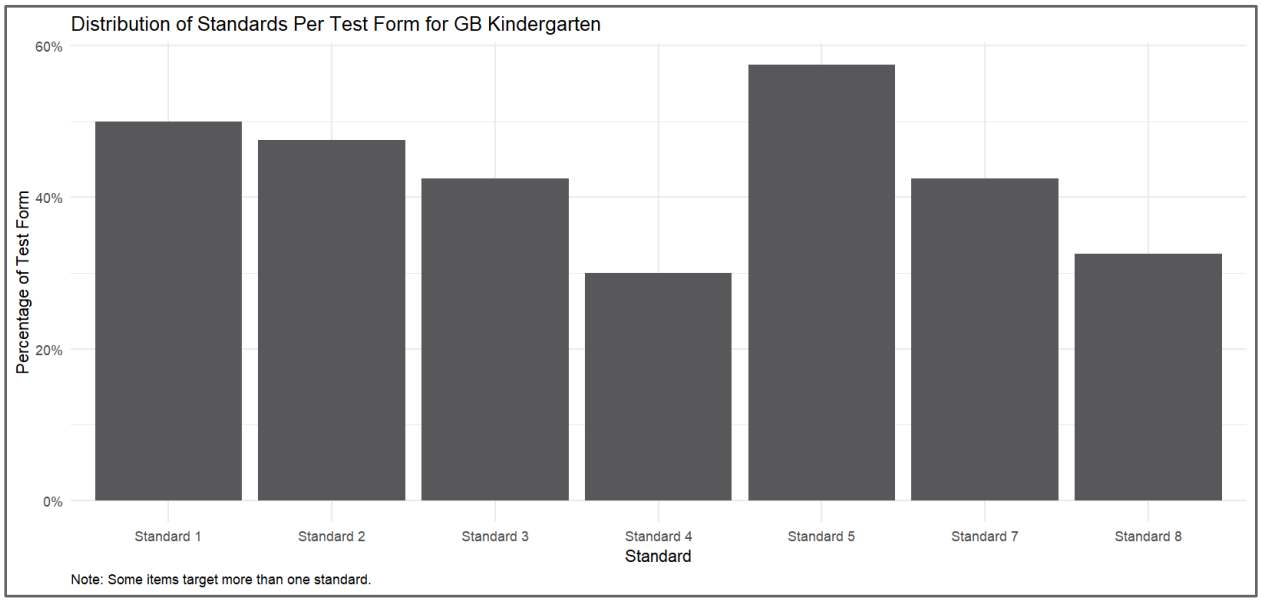


Figure D2

Distribution of Standards Per Test Form—Grade 1

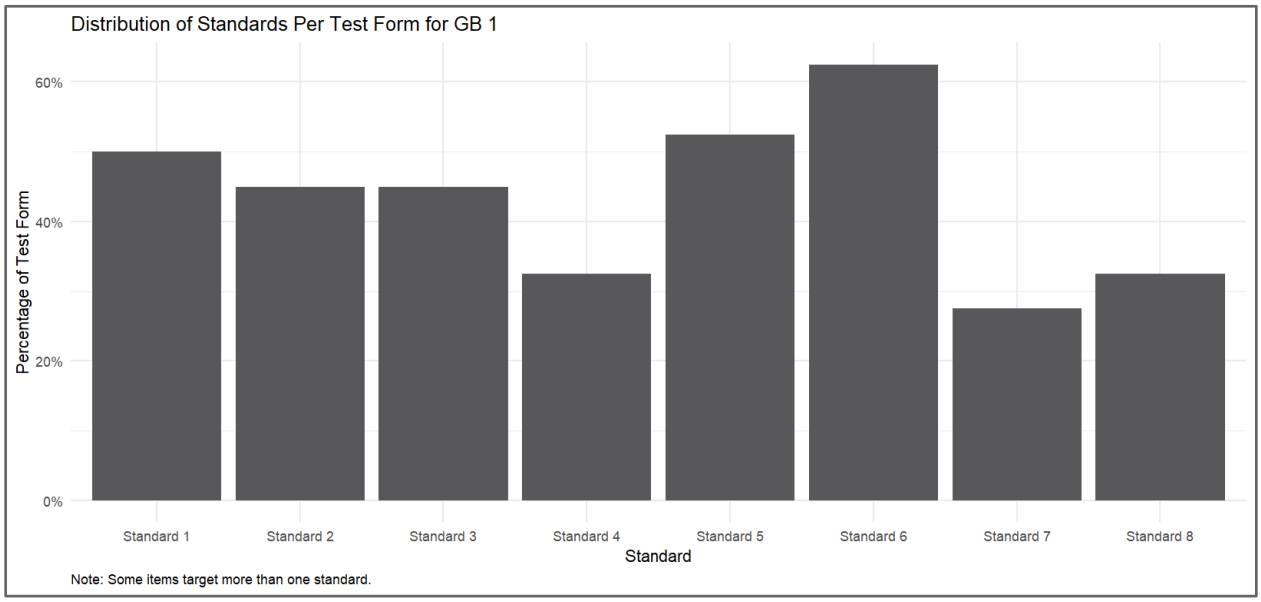


Figure D3

Distribution of Standards Per Test Form—Grade Band 2-3

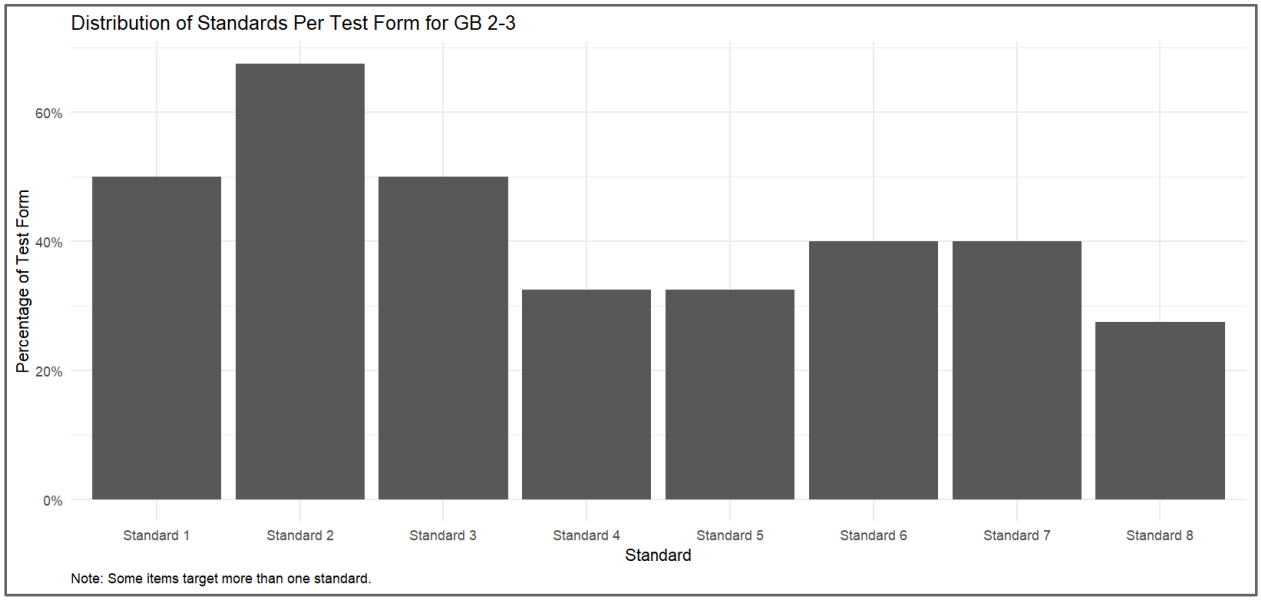


Figure D4

Distribution of Standards Per Test Form—Grade Band 4-5

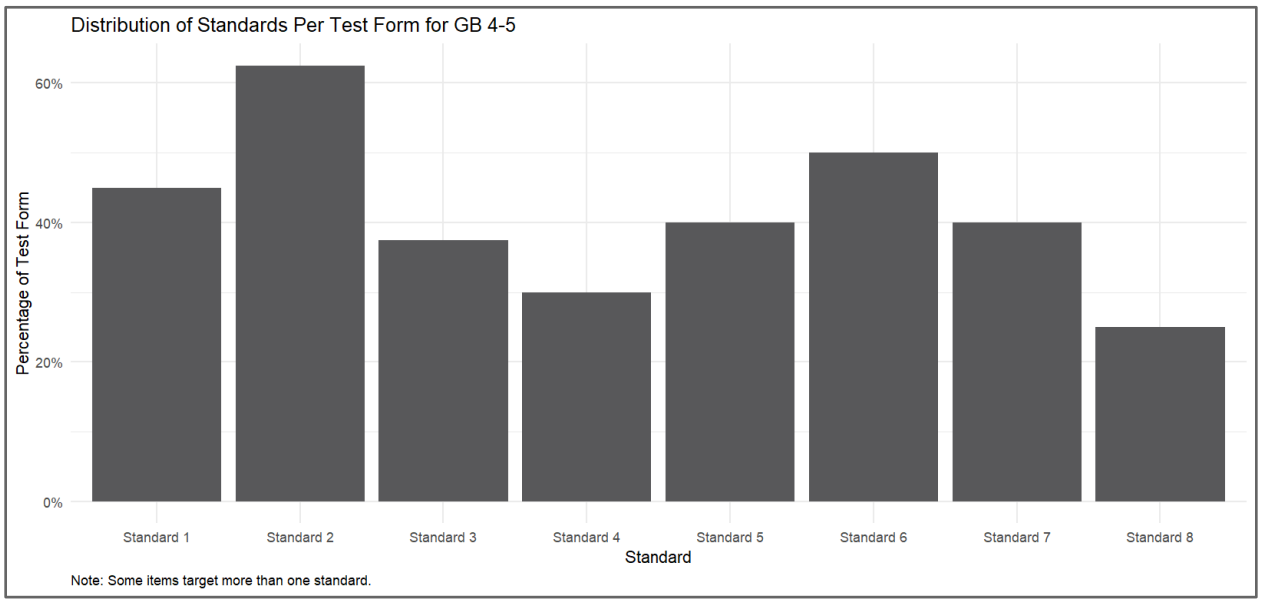


Figure D5

Distribution of Standards Per Test Form—Grade Band 6-8

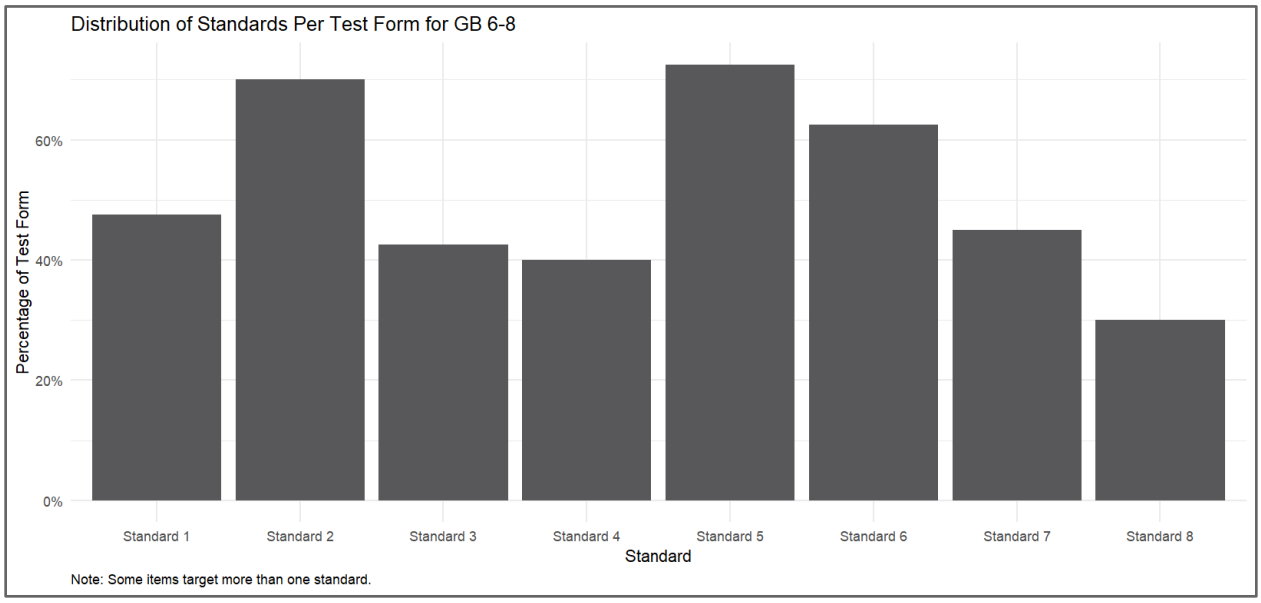
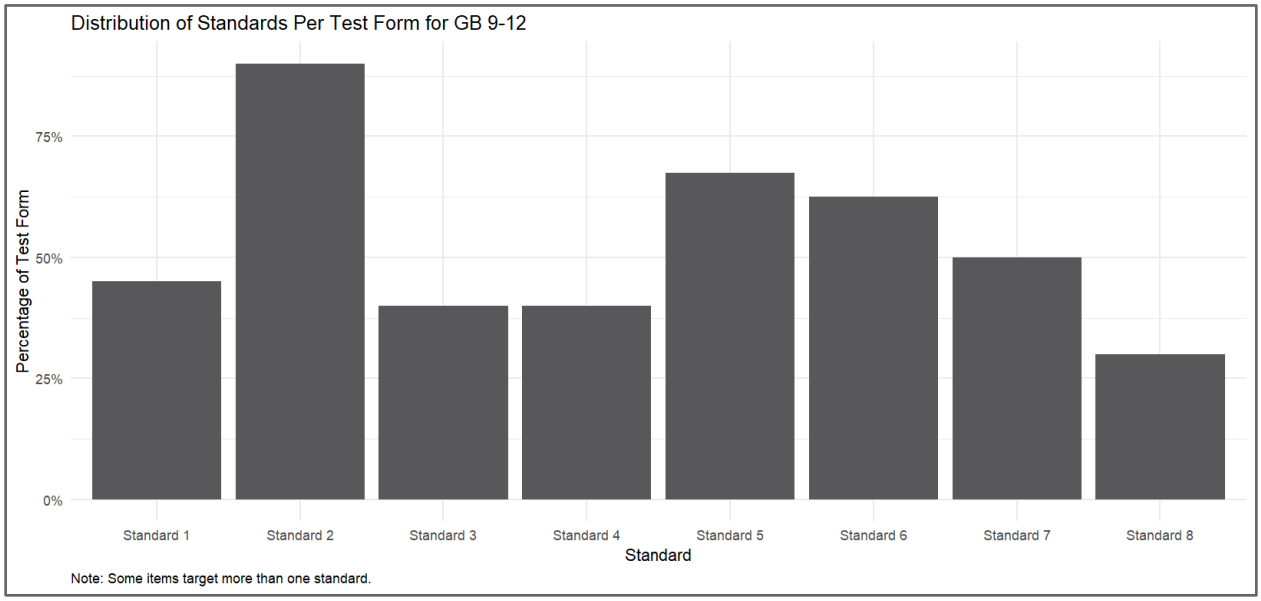


Figure D6

Distribution of Standards Per Test Form—Grade Band 9-12



Distribution of PLDs Per Test Form

Table D2

Distribution of PLDs for the Kindergarten Form

Range PLD	PLD Label	Number of Items	Percentage
PLD 1	Beginning	10	25.0%
PLD 2	Intermediate	8	20.0%
PLD 3	Early Advanced	15	37.5%
PLD 4	Advanced	7	17.5%
Totals		**40**	**100%**

Table D3

Distribution of PLDs for the Grade 1 Form

PLD	PLD Label	Number of Items	Percentage
PLD 1	Beginning	5	12.5%
PLD 2	Intermediate	9	22.5%
PLD 3	Early Advanced	15	37.5%
PLD 4	Advanced	11	27.5%
Totals		**40**	**100%**

Table D4

Distribution of PLDs for the Grade Band 2-3 Form

PLD	PLD Label	Number of Items	Percentage
PLD 1	Beginning	7	17.5%
PLD 2	Intermediate	11	27.5%
PLD 3	Early Advanced	17	42.5%
PLD 4	Advanced	5	12.5%
Totals		**40**	**100%**

Table D5

Distribution of PLDs for the Grade Band 4-5 Form

PLD	PLD Label	Number of Items	Percentage
PLD 1	Beginning	7	17.5%
PLD 2	Intermediate	10	25.0%
PLD 3	Early Advanced	15	37.5%
PLD 4	Advanced	8	20.0%
Totals		**40**	**100%**

Table D6

Distribution of PLDs for the Grade Band 6-8 Form

PLD	PLD Label	Number of Items	Percentage
PLD 1	Beginning	3	7.5%
PLD 2	Intermediate	15	37.5%
PLD 3	Early Advanced	15	37.5%
PLD 4	Advanced	7	17.5%
Totals		**40**	**100%**

Table D7

Distribution of PLDs for the High School (Grade Band 9-12) Form

PLD	PLD Label	Number of Items	Percentage
PLD 1	Beginning	5	12.5%
PLD 2	Intermediate	10	25.0%
PLD 3	Early Advanced	13	32.5%
PLD 4	Advanced	12	30.0%
Totals		**40**	**100%**

Distribution of Text Complexity Per Test Form

Table D8

Distribution of Text Complexity for the Kindergarten Form

Text Complexity	Number of Items	Percentage
Text: Low	12	30.0%
Text: Mid	15	37.5%
Text: High	13	32.5%
Totals	**40**	**100%**

Table D9

Distribution of Text Complexity for the Grade 1 Form

Text Complexity	Number of Items	Percentage
Text: Low	8	20.0%
Text: Mid	16	40.0%
Text: High	16	40.0%
Totals	**40**	**100%**

Table D10

Distribution of Text Complexity for the Grade Band 2-3 Form

Text Complexity	Number of Items	Percentage
Text: Low	9	22.5%
Text: Mid	19	47.5%
Text: High	12	30.0%
Totals	**40**	**100%**

Table D11

Distribution of Text Complexity for the Grade Band 4-5 Form

Text Complexity	Number of Items	Percentage
Text: Low	9	22.5%
Text: Mid	18	45.0%
Text: High	13	32.5%
Totals	**40**	**100%**

Table D12

Distribution of Text Complexity for the Grade Band 6-8 Form

Text Complexity	Number of Items	Percentage
Text: Low	8	20.0%
Text: Mid	19	47.5%
Text: High	13	32.5%
Totals	**40**	**100%**

Table D13

Distribution of Text Complexity for the High School (Grade Band 9-12) Form

Text Complexity	Number of Items	Percentage
Text: Low	11	27.5%
Text: Mid	14	35.0%
Text: High	15	37.5%
Totals	**40**	**100%**

Language Processes Per Test Form

Table D14

Language Processes for the Kindergarten Form

Language Process	Domain(s)	PLD(s)
Classifying	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Comparing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Defining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Describing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Explaining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Identifying	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Inquiring	Writing	PLD 1
Retelling	Listening	PLD 4
Sequencing	Listening, Speaking, Writing	PLD 3, PLD 4
Summarizing	Listening, Reading	PLD 1, PLD 2, PLD 3

Table D15

Language Processes for the Grade 1 Form

Language Process	Domain(s)	PLD(s)
Classifying	Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Comparing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Defining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Describing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Explaining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Identifying	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Inquiring	Listening, Reading, Speaking, Writing	PLD 2, PLD 3, PLD 4
Retelling	Listening, Reading	PLD 2, PLD 3
Summarizing	Listening, Reading, Writing	PLD 1, PLD 4

Table D16

Language Processes for the Grade Band 2-3 Form

Language Process	Domain(s)	PLD(s)
Classifying	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3
Comparing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3
Defining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Describing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Explaining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Identifying	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Inquiring	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Retelling	Listening, Reading	PLD 4
Sequencing	Listening, Reading	PLD 2, PLD 3
Summarizing	Reading	PLD 1

Table D17

Language Processes for the Grade Band 4-5 Form

Language Process	Domain(s)	PLD(s)
Classifying	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Comparing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Defining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Describing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Explaining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Identifying	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Inquiring	Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Retelling	Listening, Reading	PLD 4
Summarizing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4

Table D18

Language Processes for the Grade Band 6-8 Form

Language Process	Domain(s)	PLD(s)
Classifying	Listening, Reading, Speaking, Writing	PLD 2, PLD 3, PLD 4
Comparing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Defining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Describing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Explaining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Identifying	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Inquiring	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Retelling	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Summarizing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4

Table D19

Language Processes for the Grade Band 9-12 Form

Language Process	Domain(s)	PLD(s)
Classifying	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Comparing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Defining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Describing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Explaining	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Identifying	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Inquiring	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4
Retelling	Listening, Reading, Speaking	PLD 2, PLD 3, PLD 4
Summarizing	Listening, Reading, Speaking, Writing	PLD 1, PLD 2, PLD 3, PLD 4

Observable Behaviors by Domain and by PLD

Table D20

Observable Behaviors by Domain and by PLD—Kindergarten

Domain	PLD	Observable Behaviors
Listening	PLD 1	- Identify information that is relevant to the topic of shared research by selecting from two choices. “Which one is a dog?” (e.g., when provided with a picture of a loaf of bread and dog). - Match identical pictures or choose a picture of a familiar object, based on the read-aloud, oral presentation, or picture book, from a field of two pictures. - Attend to, point to, and/or identify realia, people, labeled pictures, items, shapes, etc., from the read-aloud, oral presentation, or picture book.
Listening	PLD 2	- Match pictures that are related by a concept (e.g., opposites), based on a read-aloud, oral presentation, or book/picture book. - Label/identify key vocabulary words/items/details, from a read-aloud, oral presentation, or book/picture book, using realia, visuals, or a word bank of key vocabulary to give a key detail. - Answer questions from a read-aloud, oral presentation or book/picture book with the use of picture cards (from a field of 3). - Answer a cloze/fill in the blank question or use a sentence frame to give key information from a read-aloud, oral presentation, or book/picture book.
Listening	PLD 3	- Match antonyms or synonyms (e.g., “The dog is big, show me an animal that is not big.”) from a read-aloud, oral presentation, or book/picture book. - Put items into the correct sequence of events, based on a read-aloud, oral presentation, or book/picture book, when provided with a field of 3 items. - Orally complete a sentence starter based on a read-aloud, oral presentation, or book/picture book (e.g., “The girl is happy but the boy is _____,” or “Plants need sun, _____ and ____ to grow.”). - Provide a simple retell in the appropriate sequence using key vocabulary (more elaborate details may be absent) from a read-aloud, oral presentation, or book/picture book, when provided with visual support as well as additional cues. - Repeat one word/word approximations or 1-3 words in response to simple questions (e.g., wh- questions, “Show me...”). - Match or select pictures related to key words or phrases from the story. - Select a response from options (e.g., for prepositional phrases — a picture of a ball in different positions).

Domain	PLD	Observable Behaviors
Listening	PLD 4	- Match antonyms or synonyms (e.g., “The dog is big, show me an animal that is not big.”) from a read-aloud, oral presentation, or book/picture book. - Put items into the correct sequence of events, based on a read-aloud, oral presentation, or book/picture book, when provided with a field of 3 items. - Orally complete a sentence starter based on a read-aloud, oral presentation, or book/picture book (e.g., “The girl is happy but the boy is _____,” or “Plants need sun, _____ and ____ to grow.”). - Provide a simple retell in the appropriate sequence using key vocabulary (more elaborate details may be absent) from a read-aloud, oral presentation, or book/picture book, when provided with visual support as well as additional cues.
Reading	PLD 1	- Touch words or items in book as they are read. - Point to pictures that correspond with frequently used vocabulary in the text (e.g. teacher says truck and student points to picture of a truck). - Match identical pictures or choose a picture of a familiar object, based on the read-aloud, oral presentation, or picture book, from a field of two pictures. - Attend to, point to, and/or identify realia, people, labeled pictures, items, shapes, etc., from the read-aloud, oral presentation, or picture book.
Reading	PLD 2	- Match pictures that are related by a concept (e.g., opposites), based on a read-aloud, oral presentation, or book/picture book. - Label/identify key vocabulary words/items/details, from a read-aloud, oral presentation, or book/picture book, using realia, visuals, or a word bank of key vocabulary to give a key detail. - Answer questions from a read-aloud, oral presentation or book/picture book with the use of picture cards (from a field of 3). - Answer a cloze/fill in the blank question or use a sentence frame to give key information from a read-aloud, oral presentation, or book/picture book. - Match or point to a picture after teacher gives examples. - Select a response from choices for definitions. (e.g., “Is the apple smooth or rough?”) - Identify what happy, sad, or angry looks like by pointing to pictures. - Repeat key words from the story. - With modeling, select a response (e.g., teacher says, “The dog is furry. Show me a picture of something furry in the book.”).
Reading	PLD 3	- Match antonyms or synonyms (e.g., “The dog is big, show me an animal that is not big.”) from a read-aloud, oral presentation, or book/picture book.

Domain	PLD	Observable Behaviors
		- Put items into the correct sequence of events, based on a read-aloud, oral presentation, or book/picture book, when provided with a field of 3 items. - Orally complete a sentence starter based on a read-aloud, oral presentation, or book/picture book (e.g., "The girl is happy but the boy is _____," or "Plants need sun, _____ and ____ to grow."). - Provide a simple retell in the appropriate sequence using key vocabulary (more elaborate details may be absent) from a read-aloud, oral presentation, or book/picture book, when provided with visual support as well as additional cues.
Reading	PLD 4	N/A
Speaking	PLD 1	- Indicate a feeling (e.g., if they are happy or sad in response to a question such as, "Are you happy or sad?"). - Choose between two options about familiar topics or experiences (e.g., "Do you want to go outside?", "Do you want water or milk?").
Speaking	PLD 2	- Identify and provide basic information about a topic, experience, or event (e.g., student identifies the happy person in the story or the sad person, and the student explains how he/she can tell that the person is happy or sad). - Respond to a simple question that asks about familiar experiences (e.g., "What do you like to do on the playground?" "What do you like to do at school?"). - Choose an answer appropriate for the task when presented with choices (2 or 3 choices) (e.g., Time for lunch. Student chooses picture of food, not picture of swing). - Indicate awareness of appropriate language use based on context (e.g., school versus playground). - Identify how to appropriately address a teacher versus a friend, etc.
Speaking	PLD 3	- Select from options or respond in writing or orally to questions such as, "Tell me what you want to do now," and "Tell me what the boy in the story did." - Respond to a question that asks about familiar topics, experiences, or events (e.g., "What did you do last summer?" "Tell me about a game you like to play?"). - Choose from multiple pictures of animals those that are dogs, or choose multiple examples of dogs from pictures (pictures of a German Shephard, Chihuahua, etc.). - Name/choose multiple examples of animals (e.g., animals with feathers, animals on a farm), classifying or sorting activities. - May use more complex sentences and compound sentences. - Use words and phrases such as: please, I like, Can I, It is... - Indicate awareness of appropriate language use based on context (school versus playground).

Domain	PLD	Observable Behaviors
		- Use learned words appropriate for social and academic contexts (e.g., playground and classroom language). - Use one or more words to express an opinion or preference. For example, student says “hot dog” with words/word approximations, AAC output, etc. when asked “Are hot dogs or hamburgers better?” - Answer questions about an opinion or preference.
Speaking	PLD 4	- Select from options or respond in writing or orally to questions such as, “Tell me what you want to do now,” and “Tell me what the boy in the story did.” - Respond to a question that asks about familiar topics, experiences, or events (e.g., “What did you do last summer?” “Tell me about a game you like to play?”). - Use words and phrases such as: please, I like, Can I, It is... - Indicate awareness of appropriate language use based on context (school versus playground). - Use learned words appropriate for social and academic contexts (e.g., playground and classroom language).
Writing	PLD 1	- Respond yes/no to a question. - Select from two choices (to communicate a feeling or opinion) when asked a question (e.g., “Which do you like, frogs or butterflies?”). - Select from the two picture options to respond to questions (“Joe is crying. Is he happy or sad?”). - Match similar pictures as directed (orally during listening domain) (e.g., of animals, people). - Choose between two options about familiar topics, experiences, or events. - Respond to a simple, familiar yes/no question (e.g., “Are you ready for lunch?”). - Indicate a feeling (e.g., if they are happy or sad in response to a question such as, “Are you happy or sad?”). - Choose between two options about familiar topics or experiences (e.g., “Do you want to go outside?”, “Do you want water or milk?”).
Writing	PLD 2	- Choose an answer appropriate for the task when presented with choices (2 or 3 choices) (e.g., Time for lunch. Student chooses picture of food, not picture of swing). - Indicate awareness of appropriate language use based on context (e.g., school versus playground). - Identify how to appropriately address a teacher versus a friend, etc.
Writing	PLD 3	- Select from options or respond in writing or orally to questions such as, “Tell me what you want to do now,” and “Tell me what the boy in the story did.”

Domain	PLD	Observable Behaviors
		- Respond to a question that asks about familiar topics, experiences, or events (e.g., “What did you do last summer?” “Tell me about a game you like to play?”). - Choose from multiple pictures of animals those that are dogs, or choose multiple examples of dogs from pictures (pictures of a German Shephard, Chihuahua, etc.). - Name/choose multiple examples of animals (e.g., animals with feathers, animals on a farm), classifying or sorting activities. - May use more complex sentences and compound sentences.
Writing	PLD 4	- Use words and phrases such as: please, I like, Can I, It is... - Indicate awareness of appropriate language use based on context (school versus playground). - Use learned words appropriate for social and academic contexts (e.g., playground and classroom language). - Select from options or respond in writing or orally to questions such as, “Tell me what you want to do now,” and “Tell me what the boy in the story did.” - Respond to a question that asks about familiar topics, experiences, or events (e.g., “What did you do last summer?” “Tell me about a game you like to play?”). - Choose from multiple pictures of animals those that are dogs, or choose multiple examples of dogs from pictures (pictures of a German Shephard, Chihuahua, etc.). - Name/choose multiple examples of animals (e.g., animals with feathers, animals on a farm), classifying or sorting activities. - May use more complex sentences and compound sentences.

Table D21

Observable Behaviors by Domain and by PLD—Grade 1

Domain	PLD	Observable Behaviors
Listening	PLD 1	- Choose a picture of a familiar object (or familiar word phrase) from a field of two or three. - Identify up to three key words related to the content of information presented orally.
Listening	PLD 2	- Choose the three pictures from a field of five that show key details. - Fill in the blanks about key details using a sentence frame. - Respond to simple wh- questions. - Match pictures to words/text and identify pictures. - Select pictures related to key details, setting, and characters. - Choose from a field of pictures (or word phrases) the key details of the presentation/text.
Listening	PLD 3	- Take turns in conversation. - Raise hand to indicate a desire to speak.

Domain	PLD	Observable Behaviors
		- Participate in a conversation about a chosen topic (e.g., staying on topic during a discussion). - Answer simple wh- questions about text/information presented. - Choose from a set of word phrases or sentences to answer questions about key details or retell a key detail. - Communicate at least two key details that support the main idea from a short text on a familiar topic that is read to the student.
Listening	PLD 4	- Answer questions using a preferred expressive modality with prompts and supports. - Provide input by using a preferred expressive modality to indicate animals discussed in the shared research topic. - Participate in a shared research project, which uses visuals and grade-appropriate books to provide a solution to a given problem using a preferred way to communicate. - Tell about an animal (e.g., its name, that it has stripes, spots, that it swims, etc.). - Respond to, “Show me an animal you like and why?” Student describes this animal, the sound it makes, how it moves, and why he/she likes this animal.
Reading	PLD 1	- Choose one of two pictures that reflect the main idea from a short text on a familiar topic that is read to the student.
Reading	PLD 2	- Choose the three pictures from a field of five that show key details. - Fill in the blanks about key details using a sentence frame. - Respond to simple wh- questions. - Match pictures to words/text and identify pictures. - Select pictures related to key details, setting, and characters. - Choose from a field of pictures (or word phrases) the key details of the presentation/text. - Name/point to/identify verbs or pictures of verbs and nouns or pictures of nouns in a text/information presented. - Choose the picture that matches the word. - Match the noun to a picture.
Reading	PLD 3	- Answer simple wh- questions about text/information presented. - Choose from a set of word phrases or sentences to answer questions about key details or retell a key detail. - Name/point to/identify nouns, verbs, or prepositional phrases. - Use some prepositional phrases such as “beside” or “around.” - Choose the picture that matches the word.
Reading	PLD 4	- Answer simple wh- questions about text/information presented.

Domain	PLD	Observable Behaviors
		- Choose from a set of word phrases or sentences to answer questions about key details or retell a key detail.
Speaking	PLD 1	N/A
Speaking	PLD 2	- Indicate preference; indicate what he/she likes best. - Select/label a feeling or opinion about a familiar topic by selecting from two choices. - Say “orange pumpkin” or “heavy” when describing a pumpkin. - Write “orange” or “big” when describing a pumpkin. - Tell/give two facts about their class, school, state, or nation. - State/name/list/label parts of a tree. - Respond to, “Who blew down the house?”, “What was the house the wolf blew down made of?” when shown a picture of the wolf and the straw house. - Use word phrases. - List up to two facts or ideas on a familiar topic. - Communicate what a character did in a story. - Respond to simple wh- questions with two or three choices given (e.g., “Which one is the brick house?” “What color is the sky?”). - Choose the main idea from information presented. - Choose a picture/symbol that represents presented information. - Respond to yes/no and simple wh- questions. - Participate in short conversations using short phrases.
Speaking	PLD 3	- Answer questions using a preferred expressive modality with prompts and supports. - Provide input by indicating animals discussed in a shared research topic. - Participate in a shared research project, which uses visuals and grade-appropriate books to provide solutions to a given problem using a preferred way to communicate. - Tell about an animal – its name, that is has stripes, spots, that it swims, etc. - Respond to, “Show me an animal you like and why?” Student describes this animal, the sound it makes, how it moves, and why he/she likes this animal.
Speaking	PLD 4	- Use a word or short phrase to express an opinion or a feeling. For example, student says, “like hot dog” with words/word approximations, AAC output, etc. when asked “Are hot dogs or hamburgers better?” - Use learned words and expressions, appropriate for social and academic contexts (e.g., playground, classroom) with prompts and supports. - Use words learned through conversations, reading, and being read to with prompts and supports.

Domain	PLD	Observable Behaviors
		- Use language appropriately (e.g., language used in anger versus learned language to express emotions; communicate intent). - Indicate awareness of appropriate language use based on context (e.g., school vs. playground). - Demonstrate volume control based on location. - Sort a set of six or more cards into playground language and classroom language.
Writing	PLD 1	- Provide input when given two choices, with prompts and supports. For example, student selects from choices to respond to the following. “We are talking about the zoo. Here is a lion. Show me another lion.” - Name animals with four legs, animals that like to swim, animals that have stripes. - Answer yes/no questions. - - Answer simple questions on familiar topics such as: In a cafeteria: “Do you want chocolate ice cream or vanilla ice cream?” - Identify feelings of persons/characters in texts (e.g., “Bruce won the race. Does he look happy or sad?”). - Express a single word idea (e.g., "happy," "sad") about a grade-appropriate text or topic.
Writing	PLD 2	- Choose one of three pictures that show a key detail supporting the main idea from a short text on a familiar topic that is read to the student.
Writing	PLD 3	- Use a word or short phrase to express an opinion or a feeling. For example, student says, “like hot dog” with words/word approximations, AAC output, etc. when asked “Are hot dogs or hamburgers better?” - Use learned words and expressions appropriate for social and academic contexts (e.g., playground, classroom) with prompts and supports. - Use words learned through conversations, reading, and being read to with prompts and supports. - Use language appropriately (e.g., language used in anger versus learned language to express emotions; communicate intent). - Indicate awareness of appropriate language use based on context (e.g., school vs. playground). - Demonstrate volume control based on location. - Sort a set of six or more cards into playground language and classroom language. - Take turns in conversation. - Raise hand to indicate a desire to speak. - Participate in a conversation about a chosen topic. For example, the student is not talking about elephants when the discussion is about fire drill safety.

Domain	PLD	Observable Behaviors
		- Answer questions using a preferred expressive modality with prompts and supports. - Provide input by using a preferred expressive modality to indicate animals discussed in the shared research topic. - Participate in a shared research project, which uses visuals and grade appropriate books to provide a solution to a given problem using a preferred way to communicate. - Tell about an animal – its name, that it has stripes, spots, that it swims, etc. - Respond to "Show me an animal you like and why?" by describing the animal, the sound it makes, how it moves, and why he/she liked this animal.
Writing	PLD 4	- Use a word or short phrase to express an opinion or a feeling. For example, student says, "like hot dog" with words/word approximations, AAC output, etc. when asked "Are hot dogs or hamburgers better?"

Table D22

Observable Behaviors by Domain and by PLD—Grade Band 2-3

Domain	PLD	Observable Behaviors
Listening	PLD 1	N/A
Listening	PLD 2	- Participate in a research project by answering simple questions about a topic. - Agree/disagree with a stated opinion. - Identify the subject of research by answering questions about the topic. - Respond to simple questions with 2-3 word phrases. - Answer basic questions (e.g., "What is the girl doing?"). - Answer simple questions using sentence frames and starters. - Answer questions with words or phrases. - Answer questions around the main idea or characters. For example, What is the main idea? Who is the main character? Choose one of the 2-3 options. - Put three events in order of beginning, middle, and end (options could be images or simple sentences). - Answer yes/no and simple "wh" questions about the main topic or specific words.
Listening	PLD 3	- Describe what happened in simple sentences. - Provide sentences about a topic. - Discuss ideas from the story. - Ask and answer questions about the story. - Ask for clarification if he or she does not understand a word heard orally.
Listening	PLD 4	- Describe what happened in simple sentences.

Domain	PLD	Observable Behaviors
		- Provide sentences about a topic.
Reading	PLD 1	- Identify common sight words, items, or phrases (e.g., "Which word is...?"). - Demonstrate understanding by giving objects or pictures according an to attribute (the blue paper, the yellow flower, the long string, etc.) upon request. - Say a line that is often repeated in the text.
Reading	PLD 2	- Participate in a research project by answering simple questions about a topic. - Agree/disagree with a stated opinion. - Identify the subject of research by answering questions about the topic. - Answer questions with words or phrases. - Answer questions around the main idea or characters. For example, What is the main idea? Who is the main character? Choose one of the 2-3 options. - Put three events in order of beginning, middle, and end (options could be images or simple sentences). - Answer yes/no and simple “wh” questions about the main topic or about specific words. - Identify the word/word combination and picture associated with the picture. - Name a word to match environmental print and may inquire about environmental print that is new to them. - Respond to questions such as, "When the sign says stop, do I keep going?" - Look at/read/point to words/phrases and expressions that are related to common events, topics and ideas in their daily life.
Reading	PLD 3	- Describe what happened in simple sentences. - Provide sentences about a topic. - Discuss ideas from the story. - Ask and answer questions about the story.
Reading	PLD 4	- Discuss ideas from the story. - Ask and answer questions about the story. - Describe what happened in simple sentences. - Provide sentences about a topic.
Speaking	PLD 1	- Choose appropriate language from two-word phrase cards (e.g., social, self-help language) with prompts and supports. - Select the word phrase that is “playground language” when presented with a set of word phrases (“Hey there.” or “Hello Mr. Graham.”). - Use sounds, gestures or expressions appropriate for social and self-help contexts (e.g., greetings, needs).
Speaking	PLD 2	- Respond to simple questions with 2-3 word phrases. - Answer basic questions (e.g., "What color is the car?" "What is the girl doing?").

Domain	PLD	Observable Behaviors
		- Answer simple questions using sentence frames and starters. - Provide a preference, when asked to choose between a field of two or more. - Respond with a yes/no response and give one reason why when asked an open-ended question. - Relate what he or she has read, using sentence starters or pictures, and adding short phrases to show understanding of the text. (e.g., "In the story, Tom feels sad because___", "The ball is _____ ") - Tell about the main events or important topics from the text, with prompting and support. - Write about the main events or important topics from the text, with prompting and support. - Dictate information to a scribe to produce written text (e.g., Do you like/not like the story/text?). - Produce written text (e.g., Do you like/not like the story/text?).
Speaking	PLD 3	- Discuss ideas from a story. - Ask and answer questions about the story. - Participate in a research project by answering questions about a topic or asking questions about a topic provided (or pictures of topics). - Identify information provided as true or false (real or not). For example, "Zebras have spots" is not a true statement. - Sort objects or pictures according to common characteristics. - Identify word/noun by pointing to a picture/object. - Provide an opinion on a provided topic and be able to tell why he/she has that opinion. Sentence frames can be used to help with responses (e.g., I like_____ because_____.) - Use language appropriately (e.g., language used in anger versus learned language to express emotions; communicative intent). - Indicate awareness of appropriate language use based on context (school versus playground). - Demonstrate volume control based on location (e.g., use a quieter "inside voice" when in the classroom but yelling to friends is acceptable on the playground). - Write about topic given by teacher. - Talk about topic given by teacher.
Speaking	PLD 4	- Provide an opinion on a provided topic and be able to tell why he/she has that opinion. Sentence frames can be used to help with responses (e.g., I like_____ because_____.)
Writing	PLD 1	- Choose appropriate language from two-word phrase cards (e.g., social/self-help language) with prompts/supports. - Select the word phrase that is "playground language" when presented with a set of word phrases ("Hey there." or "Hello Mr. Graham.").

Domain	PLD	Observable Behaviors
		- Use sounds, gestures or expressions appropriate for social and self-help contexts (e.g., greetings, needs). - Provide one word to complete a sentence, using a sentence frame (e.g., "How does the ___ in the story feel?", "What word tells ___?", "Which picture shows_____?"). - Provide a preference, when asked to choose between a field of two or more. - Share a "like" or "dislike" when asked about a familiar topic. - Label or match objects or pictures. - Produce non-verbal responses and/or vocalization interactions.
Writing	PLD 2	N/A
Writing	PLD 3	- Respond to questions like "Do you like/not like the story?" - Provide reasons for his/her opinions. - Provide details to support the main idea/main topic, for example, what is this story about? - Discuss ideas from a story. - Ask and answer questions about a story. -)rovide an opinion on a provided topic and be able to tell why he/she has that opinion. Sentence frames can be used to help with responses (e.g., I like_____ because_____.)
Writing	PLD 4	- Respond to questions such as: "Do you like/not like the story?" - Provide reasons for his/her opinions. - Provide details to support the main idea/main topic, for example, what is this story about?

Table D23

Observable Behaviors by Domain and by PLD—Grade Band 4-5

Domain	PLD	Observable Behaviors
Listening	PLD 1	- Identify key vocabulary. Example: "This is a lion. Can you point to another lion?" - With support and guidance, identify their name, environmental print, etc. - Choose from two pictures, for example, "Which boy is on a bike?" - Identify common sight words, colors, items, or phrases (e.g. which word is ...?)
Listening	PLD 2	- Identify common words, colors, items, or phrases with limited prompting or among several options. - Complete sentences. For example, "In this story, the lion lived at the ____ (zoo)." - Identify the main topic. - Use pictures to assist in a retell.

Domain	PLD	Observable Behaviors
		- Answer wh- questions about an oral presentation. - Answer wh- questions about text. - Respond appropriately to beginning formulaic expressions such as “time for lunch,” “go to the bathroom,” and “time to go.” - Match a word to a picture. - When prompted with questions such as, “What does the author feel about X/Y? How do you know?” provide a reasonable response. - Given two or more pictures, identify which one represents the author’s topic. - Point to a picture showing the author’s feelings about the topic, for example, “Can you find a picture or word in the text to show why the author is sad?” - Given categories (informative, persuasive, and entertain), match texts to their categories. May use pictures of covers of familiar books.
Listening	PLD 3	- Use pictures to assist in a retell. - Answer wh- questions about text. - Answer wh- questions about an oral presentation. - Locate answers to wh- questions in written text. - Respond to wh- questions (e.g., what does the speaker/author want to happen, why does the speaker/author want...) - Respond using sentence frames such as: “The author or speaker believes....and I agree with him/her because...”. - Organize the author/speaker’s points by completing a graphic organizer (e.g., web page, book, magazine.)
Listening	PLD 4	- Use pictures to assist in a retell. - Answer wh- questions about text. - Answer wh- questions about an oral presentation. - Locate answers to wh- questions in written text.
Reading	PLD 1	- Identify key vocabulary. Example: “This is a lion. Can you point to another lion?” - With support and guidance, identify their name, environmental print, etc. - Choose from two pictures, for example, “Which boy is on a bike?” - Identify common sight words, colors, items, or phrases (e.g. which word is ...?)
Reading	PLD 2	- Identify common words, colors, items, or phrases with limited prompting or among several options. - Complete sentences. For example, “In this story, the lion lived at the ____ (zoo).” - Identify the main topic. - Use pictures to assist in a retell.

Domain	PLD	Observable Behaviors
		- Answer wh- questions about an oral presentation. - Answer wh- questions about text. - Respond appropriately to beginning formulaic expressions such as “time for lunch,” “go to the bathroom,” and “time to go.” - Match a word to a picture. - Participate in short written exchange by providing multiple responses using task specific word banks. - Use appropriate social skills in short conversations such as turn taking in a discussion. - Respond to wh- questions about a text (e.g., “Have you. . .”). Sentence starters can be used. - Provide specific details about events/topics. - Write or dictate a short sentence about a topic (e.g., a short letter to someone). - Using conversation frames, participate in a conversation about familiar topics, such as, “Should students wear uniforms?” - When prompted with questions such as, “What does the author feel about X/Y? How do you know?” provide a reasonable response. - Given two or more pictures, identify which one represents the author’s topic. - Point to a picture showing the author’s feelings about the topic, for example, “Can you find a picture or word in the text to show why the author is sad?” - Given categories (informative, persuasive, and entertain), match texts to their categories. May use pictures of covers of familiar books.
Reading	PLD 3	- Provide specific details about events/topics. - Answer wh- questions about a text. - Write or dictate a short sentence about a topic (e.g., a short letter to someone). - Use appropriate social skills in short conversations such as turn taking in a discussion. - Use pictures to assist in a retell. - Answer wh- questions about text. - Answer wh- questions about an oral presentation. - Locate answers to wh- questions in written text. - Sort several picture cards into categories using established criteria. - Given a model to follow, provide some citations, like a title or author name of a more than one type of source (e.g., web page, book, magazine.)
Reading	PLD 4	- Use pictures to assist in a retell. - Answer wh- questions about text.

Domain	PLD	Observable Behaviors
		- Answer wh- questions about an oral presentation. - Locate answers to wh- questions in written text.
Speaking	PLD 1	- Match appropriate greetings, vocabulary, tone, or mechanics (upper case/lower case) to situations and people. - Use appropriate words and timing to respond to an adult (e.g., greeting, farewell). - Indicate a like or dislike of a text or topic. - Identify the correct feeling or emotion of a character. - Identify the topic of presented information (Was this about bicycles or trains?). - Indicate information the author shared in the text.
Speaking	PLD 2	- Respond appropriately to, for example, “How old are you?” - Indicate a choice/preference (e.g., food preference) or feeling (e.g., cold, tired). - When the student feels cold, indicate cold, not “sleepy.” - Given a picture of a teacher, choose the best title: Mrs. Smith, Mr. Smith, Dr. Smith. - Given choices, write a closing and signature for a friendly letter. - Given two word or phrase choices, select the better greeting, title, vocabulary. - Given a situation, choose the better vocabulary, tone or gesture. - Use the correct word(s) in the correct context. - When prompted with questions such as, “What does the author feel about X/Y? How do you know?” provide a reasonable response. - Given two or more pictures, identify which one represents the author’s topic. - Point to a picture showing the author’s feelings about the topic, for example, “Can you find a picture or word in the text to show why the author is sad?” - Given categories (informative, persuasive, and entertain), match texts to their categories. May use pictures of covers of familiar books.
Speaking	PLD 3	- Sort several picture cards into categories using established criteria. - Given a model to follow, provide some citations, like a title or author name of a more than one type of source (e.g., web page, book, magazine). - Respond to wh- questions (e.g., what does the speaker/author want to happen, why does the speaker/author want...). - Respond using sentence frames such as: “The author or speaker believes....and I agree with him/her because...”. - Organize the author/speaker’s points by completing a graphic organizer (e.g., web page, book, magazine.)

Domain	PLD	Observable Behaviors
		- Provide specific details about events/topics. - Answer wh- questions about a text. - Write or dictate a short sentence about a topic (e.g., a short letter to someone). - Use appropriate social skills in short conversations such as turn taking in a discussion.
Speaking	PLD 4	- Write about topic. - Talk about a topic. - Initiate a greeting or farewell using appropriate words and timing to respond to a peer. - Compose written texts about a text or topic using multiple simple sentences. - Communicate information about texts with prompting or with visual aids.
Writing	PLD 1	- When given descriptions of a preference of the speaker, respond to the following, “Does the speaker feel X/Y?” or “Is the speaker happy/sad?” etc. - When given three possible points, indicate which point the speaker made. - Provide a preference when asked to choose between two or three objects.
Writing	PLD 2	- Identify the topic of the text. For example, ‘This story was about ___.’ - With prompting, share responses about written or oral text. This could be answering simple questions about the text or topic. - Label, dictate, or compose a narrative or expository text. Narrative should include clear beginning, middle, and end. Expository should include topic and 1-2 supporting details. - Communicate using their preferred communication mode to share details from the story (i.e., “What is the character doing on this page?” “The girl is...”).
Writing	PLD 3	- Write about topic. - Talk about a topic. - Initiate a greeting or farewell using appropriate words and timing to respond to a peer. - Provide an opinion and tell why he/she has that opinion. - Sort several picture cards into categories using established criteria. - Given a model to follow, provide some citations, like a title or author name of more than one type of source (e.g., web page, book, magazine). - Compose written texts about a text or topic using multiple simple sentences. - Communicate information about texts with prompting or with visual aids.

Domain	PLD	Observable Behaviors
Writing	PLD 4	- Compose written texts about a text or topic using multiple simple sentences. - Communicate information about texts with prompting or with visual aids. - Provide an opinion and tell why he/she has that opinion. - Respond to wh- questions (e.g., what does the speaker/author want to happen, why does the speaker/author want...). - Respond using sentence frames such as: "The author or speaker believes....and I agree with him/her because...". - Organize the author/speaker's points by completing a graphic organizer (e.g., web page, book, magazine.)

Table D24

Observable Behaviors by Domain and by PLD—Grade Band 6-8

Domain	PLD	Observable Behaviors
Listening	PLD 1	N/A
Listening	PLD 2	- Organize a provided sentence card in an appropriate sequence for retelling information in a literary text. - Produce 1-2 words to explain what a story is about. - Make a list or tell steps in a process described in a presentation. - Present a summary of information. - Read a claim and supporting evidence and identify which evidence supports the claim. - Listen to and record information. - Respond to questions about information gathered through a questionnaire. - Listen to information and identify one key fact. - Listen to information and identify a conclusion. - Read the information gathered from a questionnaire and summarize the information. - Read a list of results of a questionnaire and choose the best summary. - Read a short text and identify the best summary. - Present a summary of information and respond to questions. - Summarize information from a survey. - Present a summary of the information gathered and respond to questions about the information. - Use a questionnaire to gather information and summarize the information. - Respond to a request for an opinion when provided key words. - Construct a short sentence.

Domain	PLD	Observable Behaviors
		- Choose when provided two options. - Respond to questions about an informational text. - Demonstrate an understanding of the words after listening to a speaker's presentation. - Determine the meaning of words in a text.
Listening	PLD 3	- Create a short report. - Identify a book for the class to read and say why. - Identify similarities and differences in the characteristics of literary and informational texts. - Express an opinion. - Share out information gathered about preferences and the reasons for the preferences. - Gather information about preferences and the reasons for the preferences. - Retell information shared in a discussion. - Respond to, “What is this story about?” after reading or listening to a grade-appropriate literary text. - Provide supporting ideas for a teacher-provided summary of an informational text. - Identify the main idea and supporting details in an oral presentation. - List the sequence of steps. - Identify an important point made by a speaker or an author. - Identify reasons supporting a point made by a speaker or an author.
Listening	PLD 4	N/A
Reading	PLD 1	N/A
Reading	PLD 2	- Organize a provided sentence card in an appropriate sequence for retelling information in a literary text. - Produce 1-2 words to explain what a story is about. - Make a list or tell steps in a process described in a presentation. - Demonstrate an understanding of the words after listening to a speaker's presentation. - Determine the meaning of words in a text. - Identify information from a speaker’s presentation. - Identify information presented by an author. - Respond with agreement or disagreement and reasons to a main point in an informational text.
Reading	PLD 3	- Respond to, “What is this story about?” after reading or listening to a grade-appropriate literary text. - Provide supporting ideas for a teacher-provided summary of an informational text. - Identify the main idea and supporting details in an oral presentation. - List the sequence of steps.

Domain	PLD	Observable Behaviors
		- Create a short report. - Identify a book for the class to read and say why.
Reading	PLD 4	- Identify vocabulary related to text about a topic. - Identify and explain words in a text. - Identify and explain words in a speaker's presentation. - Respond to task requirements including summarizing, explaining, comparing, sequencing, or identifying cause-effect. - Use an adapted dictionary with pictures to find the meaning of a word.
Speaking	PLD 1	- Identify classroom, community, family, and other familiar words presented on word cards. - Respond yes or no to questions about preferred subjects for stories. - Fill in a blank such as “My favorite story is___.”
Speaking	PLD 2	- Compose 1-2 sentences. - Summarize in a few sentences basic facts from an informational text. - Write or tell in phrases what happened first in a story.
Speaking	PLD 3	- Select appropriate vocabulary when writing a letter to the principal versus writing a text to a friend. - Use sentences in letters but not when making a word list of supplies needed for a science project. - Provide a list of words that belong to a specific word category (e.g., school supplies, transportation). - Write (summarize or sequence) about an informational text. - Identify the main characters in a familiar story. - State the steps in a familiar sequence. - Identify an important point made by a speaker or an author. - Identify reasons supporting a point made by a speaker or an author.
Speaking	PLD 4	- Give an expanded response to support information presented in a presentation. - Provide several statements to support an author’s opinion about an informational topic. - Select appropriate vocabulary when writing a letter to the principal versus writing a text to a friend. - Use sentences in letters but not when making a word list of supplies needed for a science project. - Provide a list of words that belong to a specific word category (e.g., school supplies, transportation). - Write (summarize or sequence) about an informational text. - Identify the main characters in a familiar story. - State the steps in a familiar sequence.
Writing	PLD 1	- Identify the most frequent of three options. - Compare through short phrases. - Compare facts from a short text.

Domain	PLD	Observable Behaviors
Writing	PLD 2	- Present a summary of information. - Read a claim and supporting evidence and identify which evidence supports the claim. - Listen to and record information. - Respond to questions about information gathered through a questionnaire. - Listen to information and identify one key fact. - Listen to information and identify a conclusion. - Read the information gathered from a questionnaire and summarize the information. - Read a list of results of a questionnaire and choose the best summary. - Read a short text and identify the best summary. - Present a summary of information and respond to questions. - Summarize information from a survey. - Present a summary of the information gathered and respond to questions about the information. - Use a questionnaire to gather information and summarize the information. - Give a simple response to support information presented in a presentation. - Select words to enter into sentence frames about reasons for liking a favorite book.
Writing	PLD 3	- Create a short report. - Identify a book for the class to read and say why. - Identify similarities and differences in the characteristics of literary and informational texts. - Express an opinion. - Share out information gathered about preferences and the reasons for the preferences. - Gather information about preferences and the reasons for the preferences. - Retell information shared in a discussion. - Write (summarize or sequence) about an informational text. - Identify the main characters in a familiar story. - State the steps in a familiar sequence. - Select appropriate vocabulary when writing a letter to the principal versus writing a text to a friend. - Use sentences in letters but not when making a word list of supplies needed for a science project. - Provide a list of words that belong to a specific word category (e.g., school supplies, transportation).
Writing	PLD 4	- Select appropriate vocabulary when writing a letter to the principal versus writing a text to a friend. - Use sentences in letters but not when making a word list of supplies needed for a science project.

Domain	PLD	Observable Behaviors
		- Provide a list of words that belong to a specific word category (e.g., school supplies, transportation). - Give an expanded response to support information presented in a presentation. - Provide several statements to support an author's opinion about an informational topic.

Table D25

Observable Behaviors by Domain and by PLD—Grade Band 9-12

Domain	PLD	Observable Behaviors
Listening	PLD 1	- Match an argument to a source after shown two points from a speaker presentation. - Match an argument to a source after shown two points from a speaker presentation.
Listening	PLD 2	- Answer the question "Who was the story about?" - Match a word, phrase, or description to the main character. - Match a visual to the main character. - Identify attributes (gender, clothing, etc.) of the main character in a story. - Follow picture directions. - Follow directions. - Fill in a sentence frame to state what a story is about. - Match an object (e.g., picture representing a school topic) to information provided. - Communicate in writing or orally with one- or two-word statements important similarities and differences between two objects. - Follow instructions in a simple manual (e.g., instructions for setting an alarm clock). - Show understanding of key terms choosing the answer from three options sentences. - Participate in a two-turn conversation. - Use simple sentences. - Select and display a photo or picture that reflects a short statement. - Match a topic sentence to an image. - Tell a peer through simple phrases about a favorite subject in school. - Choose captions for pictures. - Write modified stories or essays using sentence starters or writing frames, pictures, or word banks. - Conduct, complete, and/or evaluate a survey. - Write a list.

Domain	PLD	Observable Behaviors
Listening	PLD 3	- Identify important similarities and differences when presented an information card and objects (e.g., rock attributes). - Express an opinion in writing or orally by responding to a presented argument. - Write instructions for a peer through a simple manual or set of sentences. - Explain the reason an author or speaker gives to support a claim. - Identify the main idea of a story and two supporting details. - Identify the theme of simple news article or text and provide details to support it. - Read and interpret a report. - Identify two central ideas in an informational text. - Identify and explain related vocabulary (ecosystem, photosynthesis, species) in a science unit. - Respond to task requirements including synthesize, summarize, explain, compare, sequence, cause-effect.
Listening	PLD 4	- Participate in a multi-turn conversation. - Use simple and compound sentences. - Write a story using print materials to illustrate. - Create informational text using a graphic organizer or chart. - Read poetry. - Complete an application for a job. - Create advertisements for an in-school business (e.g., coffee shop; supply store; office support service). - Appropriately respond to in-school business or work-experience interactions (see list above), greetings, taking orders, filling orders, supporting colleagues, communicating job completion.
Reading	PLD 1	- Match an argument to a source after shown two points from a speaker presentation. - Identify the most frequent of three options (e.g., favorite subject in school after asking peers and recording responses on a bar graph). - Indicate whether a speaker agrees with a claim and chart agreements and disagreements.
Reading	PLD 2	- Answer the question “Who was the story about?” - Match a word, phrase, or description to the main character. - Match a visual to the main character. - Identify attributes (gender, clothing, etc.) of the main character in a story. - Follow picture directions. - Follow directions. - Fill in a sentence frame to state what a story is about. - Participate in a two-turn conversation. - Use simple sentences.

Domain	PLD	Observable Behaviors
		- Select and display a photo or picture that reflects a short statement. - Match a topic sentence to an image. - Tell a peer through simple phrases about a favorite subject in school. - Choose captions for pictures. - Write modified stories or essays using sentence starters or writing frames, pictures, or word banks. - Conduct, complete, and/or evaluate a survey. - Write a list.
Reading	PLD 3	- Participate in a multi-turn conversation. - Use simple and compound sentences. - Write a story using print materials to illustrate. - Create informational text using a graphic organizer or chart. - Read poetry. - Complete an application for a job. - Create advertisements for an in-school business (e.g., coffee shop; supply store; office support service). - Appropriately respond to in-school business or work-experience interactions (see list above), greetings, taking orders, filling orders, supporting colleagues, communicating job completion. - Identify the main idea of a story and two supporting details. - Identify the theme of simple news article or text and provide details to support it. - Read and interpret a report. - Identify two central ideas in an informational text. - Identify and explain related vocabulary (ecosystem, photosynthesis, species) in a science unit. - Respond to task requirements including synthesize, summarize, explain, compare, sequence, cause-effect.
Reading	PLD 4	- Participate in a multi-turn conversation. - Use simple and compound sentences. - Write a story using print materials to illustrate. - Create informational text using a graphic organizer or chart. - Read poetry. - Complete an application for a job. - Create advertisements for an in-school business (e.g., coffee shop; supply store; office support service). - Appropriately respond to in-school business or work-experience interactions (see list above), greetings, taking orders, filling orders, supporting colleagues, communicating job completion. - Identify the main idea of a story and two supporting details. - Identify the theme of simple news article or text and provide details to support it. - Read and interpret a report.

Domain	PLD	Observable Behaviors
		- Identify two central ideas in an informational text.
Speaking	PLD 1	N/A
Speaking	PLD 2	- Speak or write one to two sentences using sentence starters. - Write a summary about a graph. - Find three to five words (using word cards, online tool, etc.) that represent key points from a text.
Speaking	PLD 3	- Identify important similarities and differences when presented an information card and objects (e.g., rock attributes). - Express an opinion in writing or orally by responding to a presented argument. - Write instructions for a peer through a simple manual or set of sentences. - Explain the reason an author or speaker gives to support a claim. - Write (summarize or sequence) about an informational text. - Speak or write one or two sentences to summarize a literary or informational text. - Present information about a schedule (e.g., school day) to peers. - Describe characters in a familiar story. - Identify a favorite literary text and state several reasons for that choice. - Describe several facts that support a claim (recycling is important). - Respond to questions about an author's opinion (e.g., did the author think water pollution is a big or small problem?) and support those answers with facts. - Respond to a specific question. - Respond to a friend with a greeting different from the response to the classroom teacher. - Complete forms such as job applications and school registrations, providing personal information requested. - Select appropriate vocabulary when writing a letter to the principal versus writing a text to a friend. - Make a word list of supplies needed for a project.
Speaking	PLD 4	- Write (summarize or sequence) about an informational text. - Speak or write one or two sentences to summarize a literary or informational text. - Present information about a schedule (e.g., school day) to peers. - Describe characters in a familiar story. - Identify a favorite literary text and state several reasons for that choice. - Describe several facts that support a claim (recycling is important).

Domain	PLD	Observable Behaviors
		- Respond to questions about an author's opinion (e.g., did the author think water pollution is a big or small problem?) and support those answers with facts.
Writing	PLD 1	- Identify classroom, community, family, and other familiar words presented on word cards.
Writing	PLD 2	- Identify the correct words in a response to a specific question. - Indicate which of two responses is appropriate for a friend versus for a classroom teacher. - Select basic personal information (e.g., name, address, phone, etc.). - Make requests about an academic task (help with getting book). - Identify symbols to add to a card for family member or peer. - Fill in sentence starters using appropriate vocabulary when writing a letter to the principal vs. writing a text to a friend. - Match jobs or roles to workplaces. - Participate in a two-turn conversation. - Use simple sentences. - Select and display a photo or picture that reflects a short statement. - Match a topic sentence to an image. - Tell a peer through simple phrases about a favorite subject in school. - Choose captions for pictures. - Write modified stories or essays using sentence starters or writing frames, pictures, or word banks. - Conduct, complete, and/or evaluate a survey. - Write a list.
Writing	PLD 3	- Write (summarize or sequence) about an informational text. - Speak or write one or two sentences to summarize a literary or informational text. - Present information about a schedule (e.g., school day) to peers. - Describe characters in a familiar story.
Writing	PLD 4	- Identify important similarities and differences when presented an information card and objects (e.g., rock attributes). - Express an opinion in writing or orally by responding to a presented argument. - Write instructions for a peer through a simple manual or set of sentences. - Explain the reason an author or speaker gives to support a claim. - Respond to a specific question. - Respond to a friend with a greeting different from the response to the classroom teacher.

Domain	PLD	Observable Behaviors
		- Complete forms such as job applications and school registrations, providing personal information requested. - Select appropriate vocabulary when writing a letter to the principal versus writing a text to a friend. - Make a word list of supplies needed for a project. - Write (summarize or sequence) about an informational text. - Speak or write one or two sentences to summarize a literary or informational text. - Present information about a schedule (e.g., school day) to peers. - Describe characters in a familiar story.

Appendix E:
Alt ELPA Technology Requirements

The technology requirements in this document describe the functions and formats required of the assessment production and delivery systems that will serve the Alt ELPA and its members. A standards-based approach to technology will ensure compatibility between production and delivery systems, ensure longevity of the assessment system, and protect the investment made by Alt ELPA in its solutions.

Test-Taking Device Requirements

The following are the hardware and software requirements for the devices students use to take the tests.

Screen: So that the student can see the screen clearly, it must be a minimum of 9.5 inches diagonal ("10 inch class"). Minimum resolution is 1024x768.

Audio: The device must have audio output capabilities. Typically, this will consist of speakers. The student may use headphones or a headset, especially if they are accustomed to that mode. When headphones are used and a teacher is helping administer the test, a headphone splitter should be used so that both the teacher and student can hear audio content.

Keyboard: A physical keyboard, on-screen touch keyboard, or assistive device may be used. Regardless of the device type, it must be an input device with which the student is familiar. When an on-screen touch keyboard is used, it must leave enough remaining space on the screen for the student to see and interact with the assessment item. Thus, the remaining screen space, not consumed by the keyboard, should be at least 9.5 inches diagonal with a minimum resolution of 1024x768. If a tablet is used, an external keyboard is strongly recommended.

Pointing Device: A pointing device such as mouse, pen, touch screen, or assistive device must be provided. The device must be one with which the student is familiar.

Internet Connectivity: There should be a minimum of 500 Kbps reserved for each concurrent test. For example, if 10 students will be taking a test at the same time, the facility should have 5000 Kbps (5 Mbps) of bandwidth above and beyond that being used by other activities that share the internet connection (e.g. other classes, administration, testing, etc.).

Item Types and Stimuli

Alt ELPA assessment items consist of three parts.

- **Stimulus**: Each item is associated with a stimulus. Item stimuli include images, audio, short texts or extended passages.
- **Prompt:** This is the text and/or media that introduces the item and prompts the student to perform an action. Also known as a "stem."

- **Interaction:** The part of the item that records the student's response. This includes two to three response options, technology enhanced interaction, or a rubric which the test administrator (TA) uses in scoring a student's response.

The following is a summary of the types of prompt, interaction, and stimulus that may be presented *online* in the Alt ELPA assessments. Certain item types, such as speaking or writing constructed response are completed by the student without computer interaction. Instead, the student responds to a stimulus and prompt, either through speech or writing, using their preferred mode of communication. The student's response is scored off-line by a TA using a rubric specific to the item. The TA then enters the score for that item into the online system.

The following images are intended to help communicate how items may appear. The test authoring and delivery systems are not expected to exactly mimic this presentation.

Prompts

A prompt may also be called a "stem." Examples of prompts are included with the interaction types that follow.

Text Prompt

A textual prompt is provided in HTML format which may include embedded images and media.

Audio Prompt

Audio prompts are embedded in the HTML text prompt. At the point of embedding, an audio control is displayed with a "play" button and a progress bar.

Items should never play automatically. Playback starts when the play button is touched or clicked. Following the first play, clicking a "play" button allows the student to have the prompt repeat.

When audio is included as a supplementary feature, or as an accessibility feature, it does not play until the student selects the play feature.

Item Interaction Types

Multiple Choice

A multiple-choice item (also known as 'selected response') requires that the student select one answer from a field of two to three answer choices. Answers may be oriented vertically or horizontally. Answers may be textual, image, or text with embedded images. A circular "radio button" appears to the left of each possible answer. Selecting an answer causes the button to be filled in with a dot. Selecting a different answer causes a previously selected answer to be cleared.

Image answer with horizontal orientation

Listen to Sam talk about coloring a picture. Then answer the question.

I have a coloring book. I want to color a picture.

What can Sam use to color the picture?

Ⓐ  Ⓑ  Ⓒ 

Textual answer with vertical orientation

Read part of the text again.

Parents can send a text to their child during the school day. This is the quickest way for parents to tell their children something important.

What does the word quickest mean in the sentence?

Ⓐ biggest

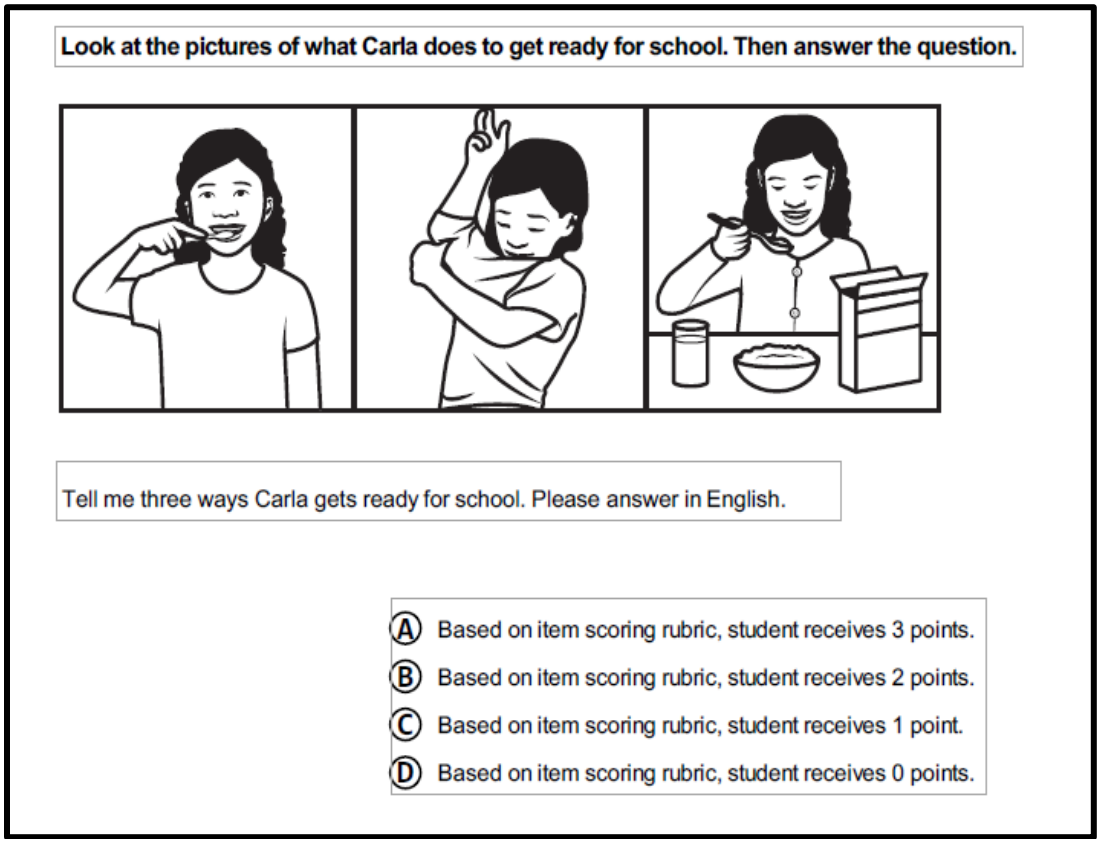
Ⓑ fastest

Ⓒ loudest

Constructed Response

Constructed response items require a student to provide a response to an open-ended question. The student's response is scored off-line by a TA using a rubric specific to the item. The TA then enters the score for that item into the online system.

Score recording using vertical orientation



QTI

Multiple choice items are encoded as a QTI *choiceInteraction*.

Inline Choice

The student is presented with text with one or more inline drop-down selection boxes. Images may also be embedded in the text. Students are prompted to select one of the responses in each of the selection boxes.

Listen as I read about a girl named Rosa who wants to grow tomatoes. After I read, you will choose words to finish some sentences about Rosa and her tomatoes.

Rosa wants to grow tomatoes in her yard. Tomatoes need a lot of sunlight to grow. Rosa recorded the number of hours of sunlight that parts of her yard received in a day. Then, she put that information in a table. The information showed that the area by the fence received the most sunlight. So, Rosa decided to plant the tomatoes there.

Hours of Sunlight in the Yard

Area	Hours of Sun
Under the tree	2 hours
Next to the house	4 hours
By the fence	6 hours

Here are words from Rosa's table.

Tomatoes need a lot of sunlight to grow. If Rosa grows the tomatoes [Ⓐ under the tree Ⓑ next to the house Ⓒ by the fence], they will only get two hours of sunlight each day. The best place for Rosa to grow her tomatoes is [Ⓐ under the tree Ⓑ next to the house Ⓒ by the fence] because they will get six hours of sunlight each day.

Choose the words that will best complete the sentences.

QTI

Encoded as a QTI *inlineChoice* interaction.

Hot Text

The student is presented with a prompt and then asked to move one or more images or words to the right place. The student moves the images or words by clicking and dragging with a mouse or by tapping and dragging with a touch screen.

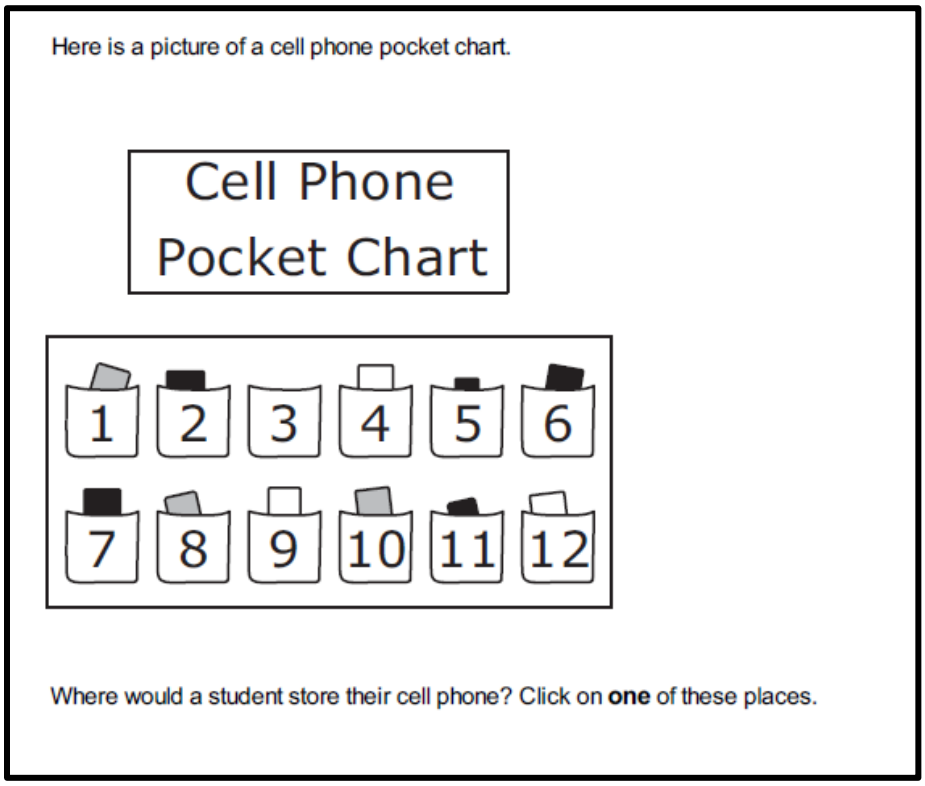
<IMAGE TBD>

QTI

Encoded as a QTI *hottextInteraction*

Image Hot Spot

In an image hot spot interaction, the student is prompted to click somewhere in a graphical image. Pre-defined selections are highlighted when the mouse cursor is above them. In some cases, the student will receive multiple prompts. The student is expected to respond to the prompt by clicking or tapping on the appropriate location on the image.



QTI

Encoded as a QTI *hotspotInteraction*

Graphic Drag and Drop

The student is presented with a prompt and then asked to move one or more images or words to the right place. The student moves the images or words by clicking and dragging with a mouse or by tapping and dragging with a touch screen. In some cases, the movable graphics will snap to designed spots on the image when they are reasonably close. Graphics to be dragged may be words. It must be possible for graphics to have transparent backgrounds so that they look correct when dragged over the background image, when necessary.

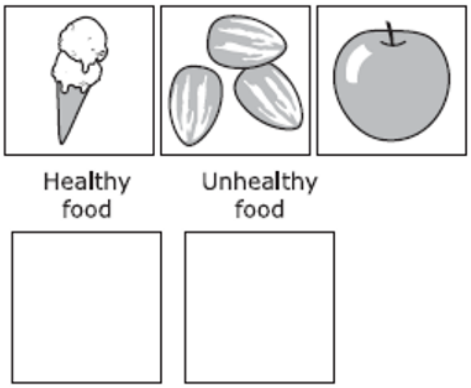
Listen and follow the directions.

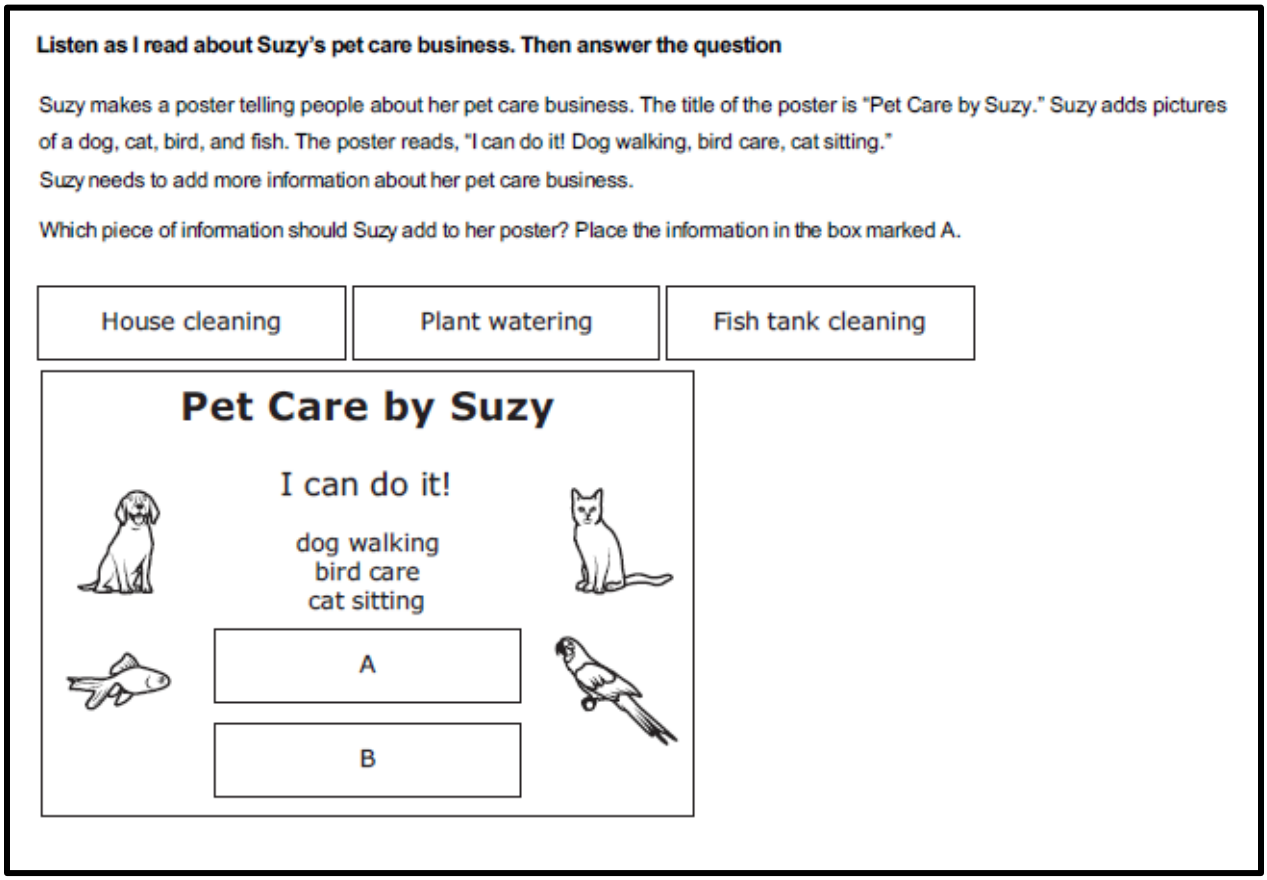
(Indicate the graphics and read the text.)

Here are three pictures of healthy and unhealthy foods: ice cream cone, apple, almonds.
Here are two blank boxes. The first blank box is labeled "Healthy food." The second blank box is labeled "Unhealthy food."
Use only **two** pictures to sort the foods into **two** groups: healthy and unhealthy.

Which food was described as healthy? Place it in the box under the heading Healthy Food.
(Allow student time to respond.)

Which food was described as unhealthy? Place it in the box under the heading Unhealthy Food.
(Allow student time to respond.)



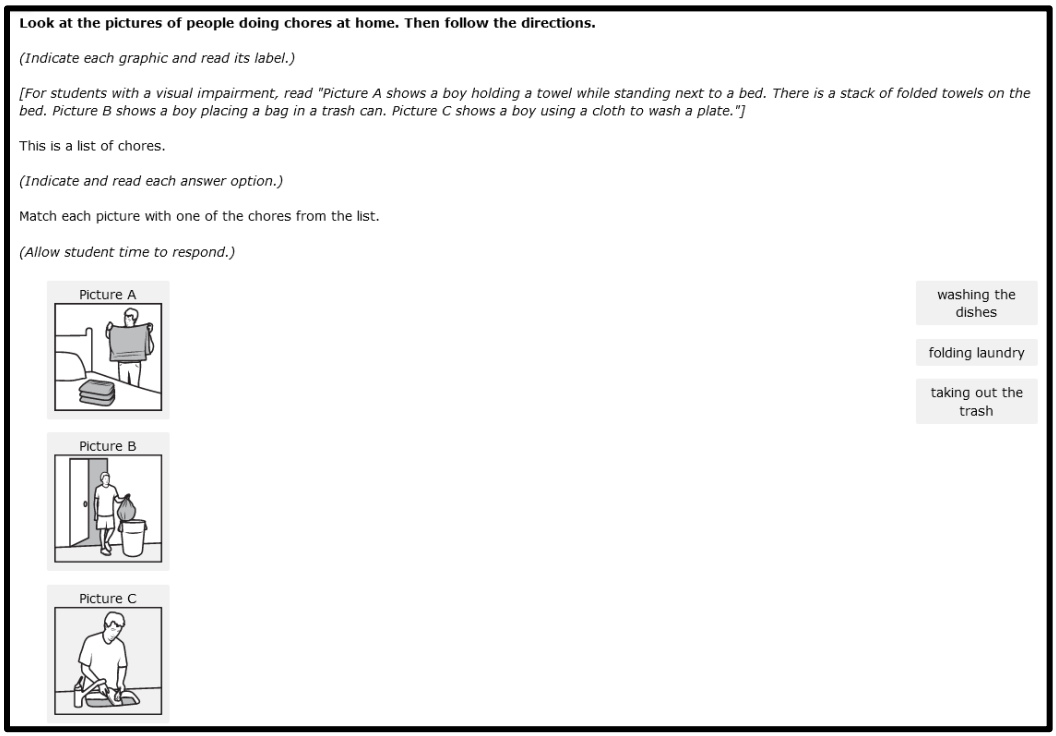


QTI

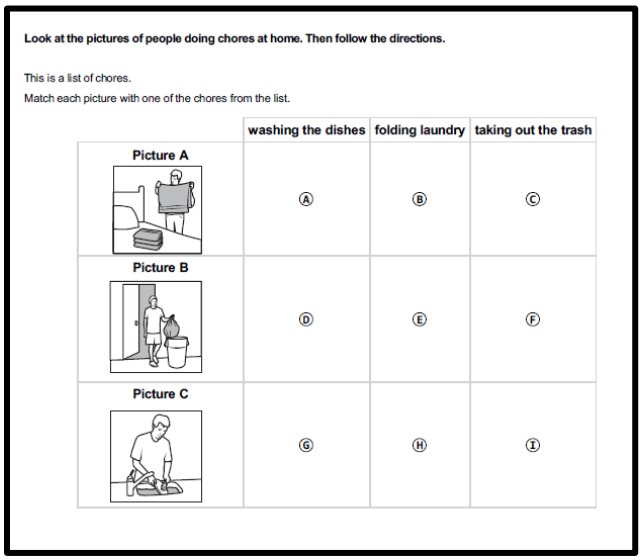
Encoded as a QTI *gapMatchInteraction* or a *graphicGapInteraction*

Match Interaction

The student is presented with a prompt and then asked to match one side of responses to the other side of responses by clicking response option on left and right or by tapping response options on left and right with a touch screen.



(Note: below is paper version of same item as generated by ITS)

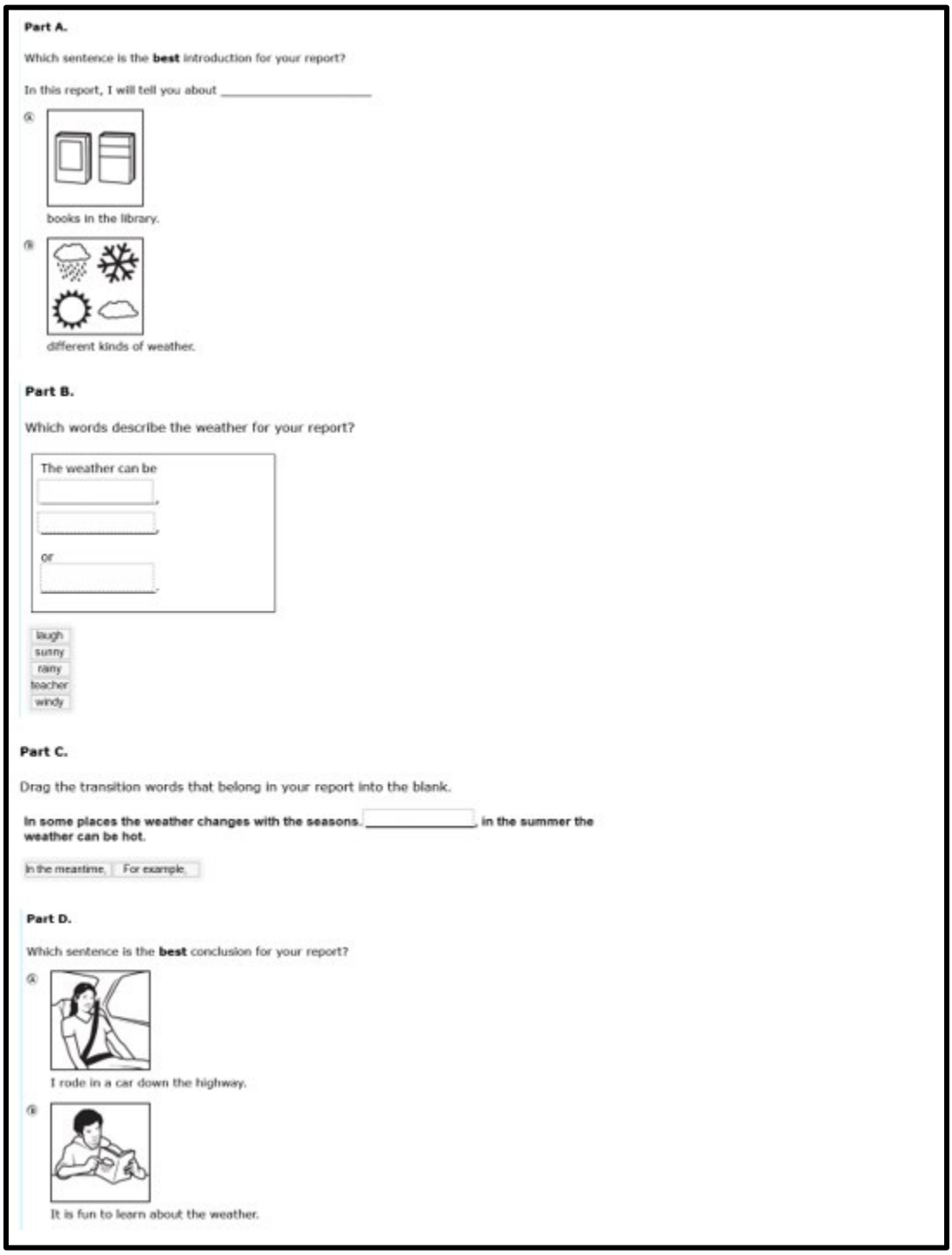


QTI

Encoded as a QTI *MatchInteraction*

Composite Interaction

Some items combine a prompt with multiple interactions. The interactions may be of the same type or of different types.



QTI

Encoded as a QTI composite item exclusively using the QTI interaction types listed earlier in this document.

Stimulus Types

Text Stimulus

A textual stimulus is provided in HTML format which may include embedded images and media.

The stimulus may be divided into pages that are presented one at a time and the pages advance when the student presses the navigation button (e.g., next).

Audio Stimulus

Audio stimuli are embedded in the HTML text stimulus. At the point of embedding, an audio control is displayed with a “play” button and a progress bar. While the audio is playing, the progress bar moves.

Items should never play automatically. Playback starts when the play button is touched or clicked.

When audio is a primary part of the stimulus, and there is also audio that is a primary part of the prompt, each part should have its own play button.

When audio is included as a supplementary feature, or as an accessibility feature, it does not play unless the student selects the play button.

A stimulus may be composed of multiple images each with associated audio. In such cases, the images and audio are presented in a coordinated fashion.

Media Types

Stimuli, Prompts, and Answers may have embedded media. The following formats are supported:

- Image: PNG
- Image: JPEG
- Image: SVG
- Audio: MP3 and OGG (Both formats must be supplied to ensure compatibility with all browsers.)
- Video: MP4 and OGV (Both formats must be supplied to ensure compatibility with all browsers.)

Accessibility

The accessibility technology requirements are guided by the Alt ELPA Accessibility and Accommodations Manual, by Universal Design for Learning, and the recommendations of the ELPA21 White Paper on English Language Learners with Significant Cognitive Difficulties (2016).

The scope of this document is strictly limited to technology requirements for assessment delivery. The full scope of assessment accessibility includes policies, procedures, content design and much more. Accessibility of the assessment production system is also out of scope.

Design for Accessibility

Accessibility should be considered from the outset of production whether the thing being produced is a document, a learning resource, an assessment, or a software application.

For example, a matching item that requires students to draw lines between items might be more challenging to students with vision or motor disabilities. A logically equivalent item can be presented with one set of items in rows and the other in columns with students selecting the cells where matching items intersect. The latter item type is more accessible to students with all abilities while measuring the same student skill.

With certain disabilities, an equivalent skill must be measured. For example, students with visual disability may use text-to-speech or braille in the place of reading. Likewise, students with hearing disability may use sign language video in the place of listening.

Embedded Accessibility Feature Requirements

Accessibility features are divided into three categories:

- **Universal Features** are available to all students.
- **Designated Features** must be identified for students in advance
- **Accommodations** require that the student have an IEP or Section 504 plan indicating that the feature should be made available to the student.

Within these categories, features are defined as Embedded and Non-Embedded. Embedded Features are those that are delivered to the student by the online test delivery system. Non-embedded features are external to the delivery system.

Non-embedded features may still be technological in nature. For example, assistive technology software and hardware may make a web-based application accessibility to students with visual disability. The next section, Web Accessibility Requirements and Compatibility with Assistive Technology, will address compatibility with Non-Embedded Technological features.

This section is dedicated to Embedded Features. An additional subset are Item-Embedded features. These are embedded features that require special augmentation of the items themselves. For example, Audio Support requires pre-recorded audio of printed text to be prepared and embedded in an assessment item. Item-embedded features are indicated in the listing below.

Required Embedded Accessibility Features

These are the requirements for any Alt ELPA online test-delivery platform. They are not a substitute for the Alt ELPA Accessibility and Accommodations Manual as they do not give guidelines on how accessible

content should be prepared or context about delivery.
Please consult the Alt ELPA Accessibility and Accommodations Manual for current information. If the guidance is in conflict, the Accessibility and Accommodations Manual takes precedence.

Embedded Universal Features

Embedded Universal Features for Reading (R), Writing (W), Listening (L), and Speaking (S)

Universal Feature	R	W	L	S	Description
Amplification (e.g., audio aids, volume control)	X	X	X	X	Volume may be raised or lowered, as needed. Student may use headphones for amplification.
Color Adjustments (e.g., contrast, overlay, choice)	X	X	X	X	The text color and screen background color may be adjusted to meet the student's needs.
Disable Universal Features	X	X	X	X	This feature allows disabling of any universal feature that might interfere with student performance or be distracting to the student.
Keyboard Navigation	X	X	X	X	Navigating through test content may be made by using a keyboard (e.g., arrow keys).
Online Tools (e.g., highlighter, mark items, masking, strikethrough, zoom)	X	X	X	X	These include a variety of tools within the platform, including: Highlighter: This digital feature may be used for marking desired text, items, or response options, with a choice of four colors. Highlighted text remains available throughout the test. Mark Items: Items may be flagged for future review during the assessment. Markings are not saved when moving to another test domain or after pausing the test for more than 20 minutes. Masking: This feature allows blocking off answer choices. Strikethrough: This feature may be used to eliminate those answer choices that do not appear correct to the student. The student

					must clearly indicate the choice is not correct.
Replay Audio	X	X	X	X	All tasks can be replayed as needed. Student may use headphones for this feature.
Re-record				X	Answers in the speaking domain may be recorded an unlimited number of times. Student may use headphones for this feature.
Writing Tools		X			Use of writing tools to format and edit written responses, including cut and paste, copy, underline, italicize, bold, and undo/redo. Spell check is allowed unless the Test Administration Manual indicates it is not allowed for a specific item.
Zoom	X	X	X	X	Embedded zoom magnification allows for up to a 400% increase. Magnifying features will work in conjunction with other allowed accessibility features and accommodations. Zoom may be set for either item level (available on demand) or test level (test platform is pre-set to be enlarged before test begins).

Accommodations

Embedded Accommodations for Reading (R), Writing (W), Listening (L), and Speaking (S)

Accommodations	R	W	L	S	Description
Word Prediction		X		X	Word prediction prompts the user with a list of likely word choices from which to select. The choices are based on words previously typed. Word prediction is allowed unless the Test Administrator Manual indicates it is not allowed for a specific item.

Web Accessibility Requirements and Compatibility with Assistive Technology

Assistive technology includes hardware and software that make computer applications available to students with particular disabilities. Here are some examples:

-

- Refreshable braille displays present text that would normally be shown on a screen in braille form.
- Alternative input devices such as eye trackers allow students with motor impairments to control and input into computer applications.

The Alt ELPA does not advise or require that software be designed for compatibility with any specific assistive technology. Rather, applications should be developed according to applicable standards that are recognized by assistive technology.

Applicable Standards

The following standards address different aspects to ensure compatibility with built-in browser features and assistive technology.

WCAG:

The Web Content Accessibility Guidelines (WCAG) family of standards are published by the W3C and ISO. They indicate how world-wide web content should be designed to be accessible. For example, all images should have an "alt" text tag with a description of the image and all navigation should be possible through the keyboard interface without use of a pointing device.

WCAG 2.0 is the international standard for accessibility. WCAG 2.1 augments version 2.0 according to advancements that have occurred since the 2.0 standard was written. The standards have three compliance levels, A, AA, and AAA.

Section 508

Section 508 is the U.S. Federal Government standard for accessibility required of federal agencies and their contracts. It has reach beyond web applications. Generally speaking, a web application meeting WCAG 2.0 AA compliance should also meet Section 508 requirements.

WAI-ARIA

Web Accessibility Initiative – Accessibility Rich Internet Applications (WAI-ARIA) is a W3C standard for adding semantic information to web content. The focus is on dynamic applications using browser-side JavaScript though other applications are supported. For example, ARIA tags may be used to indicate the position of a slider control for presentation by assistive technology.

Requirements for Compatibility with Assistive Technology

Alt ELPA Test Delivery Solutions must do the following for compatibility with assistive technology.

- Solutions must meet WCAG 2.0 AA Conformance for both the test delivery engine and the test content.
- Applications must use HTML 5.0 semantic tagging. For example, the menu or navigation section of a document should be designated with a <nav> tag rather than regardless of the visual style.
- Where semantic tags don't give sufficient information, WAI-ARIA markup should be used to provide additional indicators of the purpose or meaning of HTML content and controls.
- SSML markup should be used to guide speech synthesis. However, SSML tags should be limited to those commonly supported by web browsers and speech synthesis engines.

Item and Test Package Formats

Assessment production and delivery systems may be managed by different service providers. To ensure compatibility now and in the future, items and test packages must be in industry standard formats.

Item Format Requirement

Items should be encoded in IMS QTI 2.2 format. Vendors of the assessment production and delivery tools should anticipate upgrading to IMS QTI 3.0 before the 2022 school year.

The following table indicates the QTI interaction type that should be used for each assessment item type.

Interaction	**QTI Type**
Multiple Choice	choiceInteraction
Inline Choice	inlineChoiceInteraction
Hot Text	hottextInteraction
Image Hot Spot	hotspotInteraction
Text/Image Drag and Drop	gapMatchInteraction OR graphicGapMatchInteraction
Match Interaction	matchInteraction

Composite Items

Some assessment items will have multiple interactions. In those cases, a QTI composite item is appropriate.

Stimulus

In some interactions, a single stimulus (activity, reading passage, audio passage, video, etc.) may be associated with multiple assessment items. In those cases, the stimulus should be encoded as a QTI AssessmentStimulus and the association of the items with the stimulus should be recorded in the item metadata according to QTI standards.

Test Package Format Requirement

Test packages should be built according to IMS Question and Test Interoperability (QTI): Assessment, Section and Item (ASI) XSD Binding. As of this writing, Version 2.2.2 is current. When the upgrade to QTI 3.0 is made then the corresponding packaging standards should be followed.

Test packages and the items they contain should pass the IMS conformance tests.

Appendix F:
Scoring Priors

Table F1

Group Parameter Estimates in Modality Model

Grade Band	Domain	Mean	Variance	Covariances	
				Receptive	Productive
KG	Receptive	0.000	1.000	1.000	
	Productive	0.000	1.000	0.881	1.000
1	Receptive	0.000	1.000	1.000	
	Productive	0.000	1.000	0.893	1.000
2-3	Receptive	0.000	1.000	1.000	
	Productive	0.000	1.000	0.858	1.000
4-5	Receptive	0.000	1.000	1.000	
	Productive	0.000	1.000	0.888	1.000
6-8	Receptive	0.000	1.000	1.000	
	Productive	0.000	1.000	0.916	1.000
HS	Receptive	0.000	1.000	1.000	
	Productive	0.000	1.000	0.925	1.000

Table F2

Group Parameter Estimates in Domain Model

Grade Band	Domain	Mean	Variance	Covariances			
				Listening	Reading	Speaking	Writing
KG	Listening	0.025	0.854	1.000			
	Reading	-0.020	1.259	0.962	1.000		
	Speaking	-0.162	1.627	0.801	0.797	1.000	
	Writing	0.048	0.784	0.920	0.877	0.899	1.000
1	Listening	0.012	0.936	1.000			
	Reading	-0.008	1.146	0.954	1.000		
	Speaking	-0.162	1.974	0.811	0.857	1.000	
	Writing	0.073	0.620	0.890	0.912	0.901	1.000
2-3	Listening	0.003	0.993	1.000			
	Reading	0.000	1.100	0.943	1.000		
	Speaking	-0.116	1.739	0.788	0.811	1.000	
	Writing	0.034	0.686	0.847	0.877	0.925	1.000
4-5	Listening	-0.001	0.884	1.000			
	Reading	0.008	1.306	0.924	1.000		
	Speaking	-0.051	1.545	0.862	0.824	1.000	
	Writing	0.012	0.761	0.880	0.852	0.919	1.000
6-8	Listening	-0.012	0.768	1.000			
	Reading	0.040	1.555	0.942	1.000		
	Speaking	-0.002	1.519	0.878	0.860	1.000	
	Writing	0.004	0.837	0.904	0.907	0.928	1.000
HS	Listening	0.003	1.013	1.000			
	Reading	0.006	1.047	0.968	1.000		
	Speaking	-0.010	1.374	0.883	0.872	1.000	
	Writing	0.001	0.849	0.936	0.935	0.945	1.000

Table F3

Group Parameter Estimates in Overall Model

Grade Band	Domain	Mean	Variance	Covariances		
				Overall	Receptive	Productive
KG	Overall	0.000	0.880	1.000		
	Receptive	0.000	0.120	0.000	1.000	
	Productive	0.000	0.120	0.000	0.000	1.000
1	Overall	0.000	0.892	1.000		
	Receptive	0.000	0.109	0.000	1.000	
	Productive	0.000	0.109	0.000	0.000	1.000
2-3	Overall	0.000	0.857	1.000		
	Receptive	0.000	0.143	0.000	1.000	
	Productive	0.000	0.143	0.000	0.000	1.000
4-5	Overall	0.000	0.887	1.000		
	Receptive	0.000	0.113	0.000	1.000	
	Productive	0.000	0.113	0.000	0.000	1.000
6-8	Overall	0.000	0.916	1.000		
	Receptive	0.000	0.084	0.000	1.000	
	Productive	0.000	0.084	0.000	0.000	1.000
HS	Overall	0.000	0.926	1.000		
	Receptive	0.000	0.072	0.000	1.000	
	Productive	0.000	0.072	0.000	0.000	1.000

Appendix G:
Cut Scores

Table G1

Cut scores by Modality and Grade in Alt ELPA Reporting Scale

Grade	Receptive modality			Productive modality		
	L2	L3	L4	L2	L3	L4
KG	62	71	83	74	84	92
1	57	65	84	68	83	95
2	51	61	80	62	81	88
3	56	66	83	67	84	90
4	39	52	84	42	73	81
5	45	58	87	49	78	85
6	34	43	80	42	65	84
7	35	45	81	43	67	85
8	36	46	81	45	68	86
9	36	47	83	50	68	77
10	36	47	83	50	68	77
11	36	47	83	50	68	77
12	36	47	83	50	68	77