TECHNICAL REPORT

PART I – ASSESSMENT OVERVIEW

(ARKANSAS, ARIZONA, CONNECTICUT, IOWA, LOUISIANA, NEBRASKA, OHIO, OREGON, AND WEST VIRGINIA)

Alternate English Language Proficiency Assessment for the 21st Century—

Listening, Reading, Speaking, and Writing

Grades K–12

2023–2024 Test Administration

Submitted to
ELPA21

Submitted by
Cambium Assessment, Inc.
4200 Wilson Blvd, Ste. 610
Arlington, VA 22203-1800

June 2025

Table of Contents

List of Tables

Chapter 1. Introduction

This technical report focuses on the 2023–2024 test administration of the Alternate English Language Proficiency Assessment (Alt ELPA) for Arizona, Arkansas, Connecticut, Iowa, Louisiana, Nebraska, Ohio, Oregon, and West Virginia. A companion report, the *Alt ELPA Technical Manual*, provides more technical details on the development of the Alt ELPA, including test design, item development, scoring processes, psychometric models, standards setting, reliability, validity, and reporting.

This technical report is divided into the following two parts with appendices:

1. Part I includes an introduction and general overviews of methods and reporting structure.
    - Chapter 1. Introduction
    - Chapter 2. Scoring
    - Chapter 3. Standard Setting
    - Chapter 4. Reliability
    - Chapter 5. Validity
    - Chapter 6. Reporting
    - Chapter 7. Quality Control
    - Chapter 8. Classical Item Analyses
2. Part II includes chapters that provide results specific to the 2023–2024 administration of the Alt ELPA summative assessment.
    - Chapter 1. Test Administration
    - Chapter 2. 2023–2024 Summary
    - Chapter 3. Reliability
    - Chapter 4. Validity
    - Chapter 5. Reporting
    - Chapter 6. Classical Item Analysis
3. Part III includes the appendices of the 2023–2024 summary for each of the nine states, as listed here, and the nine states combined. The pooled analyses are based on the data from all nine states.
    - Appendix for Arizona—2023–2024 Summary
    - Appendix for Arkansas—2023–2024 Summary
    - Appendix for Connecticut—2023–2024 Summary
    - Appendix for Iowa—2023–2024 Summary
    - Appendix for Louisiana—2023–2024 Summary
    - Appendix for Nebraska—2023–2024 Summary
    - Appendix for Ohio—2023–2024 Summary
    - Appendix for Oregon—2023–2024 Summary
    - Appendix for West Virginia—2023–2024 Summary
    - Appendix for Pooled Analysis—2023–2024 Summary

Each appendix contained the following sections:

- Student Participation
- Raw Score Summary
- Raw Score Distributions
- Scale Score Summary
- Percentage of Students by Domain Performance Level
- Percentage of Students by Modality Performance Level
- Percentage of Students by Overall Proficiency Category
- Testing Time
- Cronbach’s Alpha
- Marginal Reliability
- Classification Accuracy and Consistency (individual state appendices only; Part II contains these results for all states pooled together)
- Conditional Standard Error of Measurement (CSEM)
- Dimensionality
- Ability vs. Difficulty
- Mock-Ups for Reporting (except for pooled appendix)

1.1 Background

The Alt ELPA was designed by the Collaborative for the Alternate Assessment of English Language Proficiency (CAAELP), an Iowa-led, grant-funded project sponsored by the U.S. Department of Education (Competitive Grants for State Assessments) and 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[1], Ohio, Oregon, and West Virginia. The Iowa Department of Education received the award in 2019, with a subaward to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles, for a grant period of performance from October 2019 through September 2023.

From the outset, the CAAELP vision statement was to “[e]mbrace the language capabilities and full potential of English learners with the most significant cognitive disabilities—through a fair and accurate alternate assessment.”

The Alt ELPA completed its operational field test administration in school year 2022–2023, and its first operational test administration in school year 2023–2024.

[1] New York participated in the development of the Alt ELPA, but ultimately did not participate in the operational administration.

1.1.1 Targeted Population

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

- 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 that an alternate assessment is appropriate for the student.

1.1.2 The Alt ELPA Summative Assessment

The Alt ELPA summative assessment is a year-end assessment for 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 (i.e., Kindergarten, grades 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 the following three overall proficiency determination categories:

1. *Proficient.* Students show a level of English language proficiency reflected in the Alt 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 both modalities. Once Proficient on the Alt ELPA, students may be considered for reclassification.
2. *Progressing.* Students show a level of English language proficiency reflected in the Alt 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 scoring at or above Level 1 but below Level 3 in at least one modality. Students scoring Progressing on the Alt ELPA are eligible for ongoing program support.
3. *Emerging.* Students show a level of English language proficiency reflected in the Alt 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 both modalities. Students scoring Emerging on the Alt ELPA are eligible for ongoing program support.

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

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

A modality performance Level 3 or 4 indicates that the student is demonstrating that 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.

Additional detail on the purposes of the assessment and the test design and development can be found in the companion report, the *Alt ELPA Technical Manual*, Chapters 1–3.

1.2 General Overview of the Reporting Structure

The Alt ELPA assessment results are available in the Centralized Reporting System (CRS) and, for certain states, CRS-generated paper family reports to be sent home with the students. In addition to the individual student's score report, the CRS produces aggregate score reports for teachers, schools, districts, and states. Additionally, the CRS allows users to monitor the student participation rate. Furthermore, to facilitate comparisons, each aggregate report contains summary results for the selected aggregate unit, as well as all aggregate units above the selected aggregate.

Chapter 2. Scoring

For the Alt ELPA assessment, the following scores are reported:

- Modality scores (receptive and productive):
  - A scale score for each modality
  - Standard error of measurement (SEM) of modality scale scores
  - A performance level associated with the modality scale score
- Domain scores (listening, reading, speaking, and writing):
  - A scale score for each domain
  - SEM for the domain scale score
  - A performance level associated with the domain scale score
- A comprehension score comprised of listening and reading domains:
  - A composite comprehension scale score
  - SEM for the compression composite scale score
- An overall score comprised of all four domains:
  - A composite overall scale score
  - SEM for the overall composite scale score
- An overall performance proficiency category based on the profile of modality performance levels.

2.1 Estimating Student Ability

The Alt ELPA assessment reports scale scores for each modality, sub-scores for each domain, the overall scale score for the whole assessment that includes four domains, and the comprehension scores for the partial assessment that includes the reading and listening domains. Multidimensional item response theory (MIRT) is used to estimate modality scores and to derive domain sub-scores. An item bi-factor model is used to estimate the overall scores and a unidimensional IRT model is used to estimate the comprehension score.

The Alt ELPA assessment uses a two-parameter logistic (2PL)-based MIRT model, so one-to-one correspondence between raw and scale scores is not possible. The MIRT model precludes one-to-one correspondence between domain raw and scale scores and allows the same domain raw score to fall into different performance levels depending on performance on the off-domain items. Similarly, the same modality raw score may fall into different performance levels depending on performance on the off-modality items. This is important in interpreting the raw score statistics in the appendices. Details of score estimation can be found in the Alt ELPA Scoring Specification: School Year 2023–2024 (National Center for Research on Evaluation, Standards, and Student Testing [CRESST], 2023). The business scoring rules for the Alt ELPA summative assessment are described in Part II of this technical report.

2.2 Theta-to-Scale-Score Transformation

Student performance is summarized in an individual modality score for each modality, an individual domain sub-score for each domain, a comprehension score that included listening and reading, and an overall score. Each untransformed logit score (θ) obtained from the IRT scoring model is transformed to the reporting scale using the following formula:

Modality scale score:

$$\widehat{SS}_{i,m} = 100 \times f\left(\hat{\theta}_{i,m}\right) = 100 \times \frac{1}{1+\exp\left(-\hat{\theta}_{i,m}\right)},$$

where $\hat{\theta}_{i,m}$ is the estimated theta score in modality *m* for student *i*.

Domain sub-score:

$$\widehat{SS}_{i,d} = 100 \times f\left(\hat{\theta}_{i,d}\right) = 100 \times \frac{1}{1+\exp\left(-\hat{\theta}_{i,d}\right)},$$

where $\hat{\theta}_{i,d}$ is the estimated theta score in domain *d* for student *i*.

Overall scale score:

$$\widehat{SS}_{i,g} = 1000 \times f\left(\hat{\theta}_{i,g}\right) = 1000 \times \frac{1}{1+\exp\left(-\hat{\theta}_{i,g}\right)},$$

where $\hat{\theta}_{i,g}$ is the estimated general dimension score for student *i*.

Comprehension scale score:

$$\widehat{SS}_{i} = 1000 \times f\left(\hat{\theta}_{i}\right) = 1000 \times \frac{1}{1+\exp\left(-\hat{\theta}_{i}\right)},$$

where $\hat{\theta}_{i}$ is the estimated comprehension theta score for student *i*.

Refer to Chapter 7 of the companion report, the *Alt ELPA Technical Manual*, for additional technical information on Alt ELPA scoring and psychometric models.

2.3 Lowest/Highest Obtainable Scale Scores

The 2023–2024 Alt ELPA summative assessment uses expected a posteriori (EAP) scoring, which does not assign fixed minimum or maximum theta scores. Through the non-linear transformation from logit theta scores to scale scores, the lowest and highest obtainable scale score (LOSS/HOSS) is 0 and 99, respectively, for modality scale scores and domain sub-scores; the LOSS/HOSS is 0 and 999, respectively, for the overall composite scale score and comprehension scale score.

Chapter 3. Standard Setting

The 2023–2024 Alt ELPA employed Embedded Standard Setting (ESS) as its principal standard-setting approach. ESS transformed the standard-setting process from the traditional stand-alone workshop to a set of processes actively integrated throughout the test development cycle. ESS processes consisted primarily of the following three phases:

1. *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).
2. *Educator evaluation of the target levels.* During Item Content and Bias and Sensitivity Committee meetings in summer 2022, educators evaluated items' target-level PLDs as one of the components of their review.
3. *Empirical evaluation of the target levels.* As part of the ESS analyses detailed in the *Alt ELPA Standard Setting Technical Report* (Creative Measurement Solutions, 2024), items (with empirical data inconsistent with Target Level PLDs) were reviewed by educators through the "Inconsistent Item Review and Resolution Workshop" in summer 2023. Over 80% of the alignment-by-design target levels were confirmed during this event, and there were no challenges to the Alt ELPA standards alignment.

A full-length report on ESS provides detailed information on the entire process. The Executive Summary of the ESS report appears in Chapter 8 of the *Alt ELPA Technical Manual,* and the table of cut scores appears in its appendices.

Chapter 4. Reliability

Reliability can be defined as the degree to which individuals' deviation scores remain relatively consistent over repeated administrations of the same test or alternate test forms (Crocker & Algina, 1986). For example, if a person takes the same or parallel tests repeatedly, he or she should receive consistent results for the test to be considered reliable. The reliability coefficient is one way to assess this consistency; it refers to the ratio of true score variance to observed score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}.$$

It is also conceptually defined as "the degree to which measures are free from error and therefore yield consistent results" (Peter, 1979, p. 6). As such, the reliability coefficient places a limit on the construct validity of a test (Peterson, 1994). There are various approaches for estimating the reliability of scores. Conventional approaches have included the test-retest method, the parallel-forms method, and the split-half method, which are no longer common for high-stakes testing.

The internal consistency method can be employed when it is not possible to conduct repeated test administrations. Whereas other methods often compute the correlation between two separate tests, this method considers each item within a test to be a one-item test. There are several other statistical methods based on this idea: Coefficient alpha (Cronbach & Shavelson, 2004), Kuder-Richardson Formula 20 (Kuder & Richardson, 1937), Kuder-Richardson Formula 21 (Kuder & Richardson, 1937), stratified coefficient alpha (Qualls, 1995), and Feldt-Raju coefficient (Feldt & Qualls, 1996; Feldt & Brennan, 1989).

Inter-rater reliability is the extent to which two or more individuals (coders or raters) agree. Inter-rater reliability addresses the consistency of the implementation of a rating system.

Another way to view reliability is to consider its relationship with the standard errors of measurement (SEMs)—the smaller the standard error, the higher the precision of the test scores. For example, classical test theory (CTT) assumes that an observed score (X) of each individual can be expressed as a true score (T) plus some error (E), $X = T + E$. The variance of X can be shown to be the sum of two orthogonal variance components:

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

Returning to the definition of reliability as the ratio of true score variance to observed score variance, the following formula can be determined:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_x^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}.$$

As the fraction of error variance to observed score variance approaches 0, the reliability then approaches 1.

In contrast to the homoscedastic (uniform) errors assumed in CTT, the SEMs in item response theory (IRT) vary over the ability continuum. These heteroscedastic errors are a function of a test information function (TIF) that provides different information about test takers depending on their estimated abilities. Often, the TIF is maximized over an important performance cut score, such as the proficiency cut score.

Because the TIF indicates the amount of information provided by the test at different points along the ability scale, its inverse indicates the lack of information at different points along the ability scale. This lack of information is the uncertainty, or the SEM, of the score at various score points. Conventionally, fixed-form tests are maximized near the middle of the score distribution, or near an important classification cut score, and have less information at the tails of the score distribution.

The reliability results are presented in Part II, Chapter 3 of this technical report.

4.1 Internal Consistency

Cronbach's alpha (Cronbach & Shavelson, 2004) is used to access the internal consistency of items in each domain and each modality of the Alt ELPA assessment. A high Cronbach's alpha coefficient indicates that the items in the domain or in the modality are related to each other, as expected for items intending to measure the same underlying concept (i.e., listening, reading, writing, and speaking).

4.2 Marginal Standard Error of Measurement

Another way to examine score reliability is with the marginal standard error of measurement (MSEM) (or $\bar{\sigma}_{error}$). MSEM is computed as the square root of $\bar{\sigma}^2_{error}$, which is the average of the squared SEMs of the IRT-based scale scores obtained by applying the Alt ELPA scoring procedures. Smaller values of MSEM indicate that the estimated test scores have greater precision, on average. The marginal reliability $\bar{\rho} = 1 - \frac{\bar{\sigma}^2_{error}}{\sigma^2_{total}}$ (refer to Section 4.3, Marginal Reliability and Conditional Standard Error of Measurement) and the test MSEM are inversely related. The ratio of MSEM and the standard deviation of expected a posteriori (EAP) scores ($\frac{\bar{\sigma}_{error}}{\sigma_{EAP}}$) can also indicate the measurement errors. The ratio characterizes the noise-to-signal ratio, or the ratio of within-person variance to between-person variance.

4.3 Marginal Reliability and Conditional Standard Error of Measurement

Marginal reliability (Sireci, Thissen, & Wain er, 1991) assesses the precision of scoring. It is based on the average of the conditional standard error of measurement (CSEM: σ_{error}) for the estimated theta scores. By definition, marginal reliability is the proportion of true score variance among the observed score variance. While Cronbach's alpha is computed using item-level scores, marginal reliability is estimated by using expected a posteriori (EAP) estimates, which are used to estimate the modality or domain scores. EAP is an estimate of the true score, but its variance underestimates the true score variance, so the marginal reliability within domain or modality can be estimated by

$$\bar{\rho} = \left(\frac{\sigma^2_{EAP}}{\sigma^2_{total}}\right) = 1 - \frac{\bar{\sigma}^2_{error}}{\sigma^2_{total}}$$

where $\bar{\sigma}^2_{error}$ is the average error variance (variance of the measurement error), $\sigma^2_{total} = \sigma^2_{EAP} + \bar{\sigma}^2_{error}$, and σ^2_{EAP} is the variance of the EAP estimate. The maximum value for the marginal reliability is 1. A higher reliability coefficient indicates greater scoring precision.

4.4 Classification Accuracy and Consistency

When student performance is reported in terms of achievement levels, a reliability of achievement classification is computed in terms of the probabilities of consistent classification of students as specified in Standard 2.16 in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

Classification accuracy (CA) analysis investigates how precisely students are classified into each performance level. By definition, *classification consistency* (CC) analysis investigates how consistently students are classified into each performance level across two independent administrations of equivalent forms. Since obtaining test scores from two independent administrations is not feasible due to issues such as logistics and cost constraints, the CC index is computed with the assumption that the same test is independently administered twice to the same group of students.

For the Alt ELPA, since the overall proficiency is based on modality performance level, the CA and CC are examined at each cut score in each modality test. Four performance levels divided by three cut scores (i.e., cut scores 1–3) are established for each modality test.

In general, the CA and CC can be estimated using the following approach.

At modality Level l, the marginal posterior distribution of student i can be approximated as a normal distribution with mean equal to the estimated $\hat{\theta}_i$ and standard deviation of SEM $se(\hat{\theta}_i)$. That is, $\hat{\theta}_i \sim N\left(\theta_i, se(\hat{\theta}_i)\right)$. Let p_{il} be the probability of the true score at Performance Level l for the ith student, and p_{il} can be estimated as follows:

$$p_{il} = p(c_{l-1} \le \theta_i < c_l) = p\left(\frac{c_{l-1} - \hat{\theta}_i}{se(\hat{\theta}_i)} \le \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} < \frac{c_l - \hat{\theta}_i}{se(\hat{\theta}_i)}\right)$$

$$= p\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)} < \frac{\hat{\theta}_i - \theta_i}{se(\hat{\theta}_i)} \le \frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) = \Phi\left(\frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) - \Phi\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)}\right).$$

For the lowest level (Level 1, or Level one), $c_0 = -\infty$, and for the highest level (Level L), $c_L = \infty$. If scale score is to be used, the formula previously shown can be used based on the scale score distribution.

For proficiency categories, the probability of a particular profile is obtained by integrating over the posterior distribution of the assessed modalities. Similar to the case previously shown for individual modalities, this posterior can be approximated as a multivariate normal distribution with means equal to the vector of score estimates $\widehat{\boldsymbol{SS}}_\iota$ and covariance equal to the error variance-covariance matrix $\Sigma(\widehat{\boldsymbol{SS}}_\iota)$, the diagonal of which provides the squared SEMs for the estimated scores):

$$P(\boldsymbol{SS}|\boldsymbol{y}_i) \sim MVN\left(\widehat{\boldsymbol{SS}}_\iota, \Sigma(\widehat{\boldsymbol{SS}}_\iota)\right),$$

where $\boldsymbol{y}_i$ is the pattern of item responses across all modalities. The 2 × 1 vector of score estimates $\widehat{\boldsymbol{\theta}}_\iota$ and the 2×2 error covariance matrix $\Sigma(\widehat{\boldsymbol{\theta}}_\iota)$ may be obtained from the scoring output from software capable of performing multidimensional IRT scoring; $\widehat{\boldsymbol{SS}}_\iota$ and $\Sigma(\widehat{\boldsymbol{SS}}_\iota)$ may, in turn, be obtained by applying the transformations described earlier. The probability of a specific performance profile is obtained by integrating over the multivariate posterior distribution over the ranges of scores defining the performance level in each modality. For most students (those without exemptions), the computation is as follows:

$$\hat{p}_{i,(e,f)} = \int_{\text{cut}_{e,\text{receptive}}}^{\text{cut}_{(e+1),\text{receptive}}} \int_{\text{cut}_{f,\text{productive}}}^{\text{cut}_{(f+1),\text{productive}}} P(\boldsymbol{SS}|\boldsymbol{y}_i)\, dSS_{\text{productive}} dSS_{\text{receptive}},$$

where e and f are the performance levels for receptive and productive, respectively. Additionally, $\text{cut}_{1,d} = -\infty$ and $\text{cut}_{4,d} = \infty$.

The probability of a particular overall determination, given the response pattern $\boldsymbol{y}_i$ can be estimated by adding up the probabilities associated with each profile receiving that determination:

$$\hat{p}_i = \Sigma_{L_i \in \Im_D} p_{i,(e,f)},$$

where $\Im_D$ is the set of performance-level profiles that are assigned the overall determination D, as described in Chapter 3.

Different matrices are defined for CA and CC, respectively.

To compute CA and CC for modality performance levels, define the following matrix based on L performance levels ($L \times L$ matrix)

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

where $n_{alm} = \sum_{pl_i = l} p_{im}$ is the sum of the probabilities for expected performance level m at each observed performance level l (the level actually assigned). In the matrix, the row represents the observed level and the column represents the expected level.

Based on the previous matrix, the CA for the cut score c_l $(l = 1, \cdots, L-1)$ is:

$$CA_{c_l} = \frac{\sum_{k,m=1}^{l} n_{akm} + \sum_{k,m=l+1}^{L} n_{akm}}{N},$$

where N is the total number of students.

The overall CA is computed as

$$\mathrm{CA} = \frac{\sum_{i=1}^{L} n_{aii}}{N}.$$

For example, the CA at cut score 2 is the sum of the n_{alm} values $(\sum_{k,m=1}^{l} n_{akm})$ assigned in the levels equal to or below cut score 2 at both expected and observed levels and $(\sum_{k,m=l+1}^{L} n_{akm})$ assigned in the levels above cut score 2 at both expected and observed levels divided by the total number of students.

$$\begin{pmatrix} n_{a11} & n_{a12} & n_{a13} & \cdots & n_{a1\mathrm{L}} \\ n_{a21} & n_{a22} & n_{a23} & \cdots & n_{a2\mathrm{L}} \\ n_{a31} & n_{a32} & n_{a33} & \cdots & n_{a3\mathrm{L}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{a51} & n_{a52} & n_{a53} & \cdots & n_{a5\mathrm{L}} \end{pmatrix}$$

For CC using p_{il}, a similar $L \times L$ table is constructed by assuming the test is independently administered twice to the same student group,

$$\begin{pmatrix} n_{c11} & \cdots & n_{c1L} \\ \vdots & \vdots & \vdots \\ n_{cL1} & \cdots & n_{cLL} \end{pmatrix},$$

where $n_{clm} = \sum_{i=1}^{N} p_{il} p_{im}$ is the sum of the probabilities multiplied by each paired combination of performance levels. p_{im} can be computed based on the same equation for p_{il}, as described previously.

The CC for the cut score c_l $(l = 1, \cdots, L-1)$ is

$$CC_{c_l} = \frac{\sum_{k,m=1}^{l} n_{ckm} + \sum_{k,m=l+1}^{L} n_{ckm}}{N}.$$

The overall CC is computed as

$$\mathrm{CC} = \frac{\sum_{i=1}^{L} n_{cii}}{N}.$$

The CA and CC indexes are affected by the interaction of the magnitude of se(θ), the distance between adjacent cut scores, the location of the cut scores on the ability scale, and the proportion of students around a cut point. The larger the se(θ), the closer the two adjacent cut scores, and the greater the proportion of students around a cut point, the lower the indexes.

Chapter 5. Validity

Validity “refers to the degree to which evidence and theory support the interpretation of test scores for the proposed uses of tests” (AERA, APA, & NCME, 2014, p. 11). The Alt ELPA assessment system follows guidelines for evaluating validity as outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014), in which five core sources of validity evidence are examined with each statement of intended use of the assessment. Sources of validity evidence include evidence related to test content, response processes, internal structure, relations to other variables, and the consequences of testing. Fairness is also fundamental to assessment validity, especially for the targeted population. Discussion of the Alt ELPA validity framework, assessment uses, and description of the ongoing collection of validity evidence for the Alt ELPA is provided in the *Alt ELPA Technical Manual*, a perennial report accompanying the annual technical reports that provides comprehensive technical information on the Alt ELPA.

The present technical report focuses on reporting validity evidence from the previously referenced test administration year, specifically on validity evidence for internal structure and fairness. Domain test internal structure was measured using domain dimensionality and reported in Part II of this technical report. The assumption for test internal structure was that each domain measured an essentially unidimensional construct. A principal component analysis with an orthogonal rotation (Cook, Kallen, & Amtmann, 2009; Jolliffe, 2002) was used to investigate the dimensionality for each domain test and the overall test. Results are shown in the scree plots in Section 12 of the Pooled Appendix for the Alt ELPA summative assessment.

Differential item functioning (DIF) analysis was conducted on field-test items and is also reported in Part II of this technical report.

Chapter 6. Reporting

The Alt ELPA assessment results were available in the Centralized Reporting System (CRS) for schools and districts to print out and, for certain states, CRS generated paper family reports to be sent home with students.

6.1 Centralized Reporting System

For the 2023–2024 Alt ELPA, the CRS generated a set of online score reports describing student performance for students, parents, educators, and other stakeholders. Because the score reports on student performance were updated each time students completed tests, authorized users (e.g., school principals, teachers) could view student performance on the tests and use the results to improve student learning. In addition to the individual student's score report, the CRS produced aggregate score reports for teachers, schools, districts, and states. Additionally, the CRS allowed users to monitor the student participation rate.

Furthermore, to facilitate comparisons, each aggregate report contained summary results for the selected aggregate unit, as well as all aggregate units above the selected aggregate. For example, if a school was selected, the summary results of the district to which the school belonged and the summary results of the state were also provided so that the school performance could be compared with district and state performance. If a teacher was selected, the summary results for the school, the district, and the state were also provided for comparison purposes. Table 6.1 lists the typical types of online reports and the levels at which they can be viewed (i.e., state, district, school, teacher, roster, and student) across the six states.

Table 6.1 Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
State District School Teacher Roster	• Number of students tested and percentage of students determined proficient (overall and by subgroup) • Average composite scale scores (overall and comprehension) and standard error of the averages (overall and by subgroup) • Percentage of students at each modality and domain performance level (overall and by subgroup) • Average modality (receptive and productive) scale scores and standard error of the averages (overall and by subgroup) • Average domain sub-scores (listening, reading, speaking, and writing) and standard error of the averages (overall and by subgroup) • On-demand student roster report
Student	• Overall and comprehension scale scores and standard error of the scale scores • Proficiency status based on the modality performance levels

Level of Aggregation	Types of Online Score Reports
	• Modality scale scores with modality performance levels and level descriptor • Domain sub-scores with domain performance levels and level descriptor

6.1.1 Types of Online Score Reports

The CRS is designed to help educators, students, and parents answer questions regarding how well students have performed in the assessment for each modality and each domain; CRS is also designed with great consideration for stakeholders who are not technical measurement experts (e.g., teachers, parents, students). The CRS ensures that test results are accessible and easy to interpret. Simple language is used so that users can quickly understand assessment results and make valid inferences about student achievement. The CRS presents student performance in a uniform format. For example, color is used to group similar elements, such as achievement levels, throughout the design. This design strategy allows state-, district-, and school-level users to compare similar elements and to avoid comparing dissimilar elements.

Once authorized users log in to the CRS and select "Score Reports," the online score reports are presented hierarchically. The CRS starts by presenting summaries on student performance by grade at a selected aggregate level. To view student performance for a specific aggregate unit, users can select the specific aggregate unit from a drop-down menu with a list of aggregate units (e.g., schools within a district, teachers within a school) to choose from. For more detailed student assessment results for a school, teacher, and roster, users can select the grade on the online score reports.

Generally, the CRS provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Table 6.1 summarizes the typical types of online score reports available at the aggregate level and the individual student level. Detailed information about the online score reports and instructions on how to navigate the online score reporting system can be found in the *Centralized Reporting System User Guide* for each state, accessible by the Help button in the CRS, as shown in Figures S15.1 and S29.1 in each state's Appendix.

6.1.2 Subgroup Reports

The aggregate score reports at a selected aggregate level are provided for students overall and by subgroups. Users can see student assessment results by any subgroup. Table S12.1 in each state's Appendix presents the subgroup data and subgroup categories for each state. It is noted that the subgroup data and subgroup categories are not included in the pooled Appendix for pooled analysis.

6.1.3 Paper Reports

The CRS enables users to print reports, as described earlier. The CRS also allows users to print a family report for each student. A mock-up of score reports can be found in Sections 15 of the

Appendix for each state. It is noted that the mock-up for score reports is not included in the pooled Appendix for pooled analysis.

Chapter 7. Quality Control

Thorough quality control was integrated into every aspect of the 2023–2024 Alt ELPA assessment. The National Center for Research on Evaluation, Standards, and Student Testing (CRESST), the states, and CAI built in multiple layers of reviews and verifications to ensure that outputs were of the highest quality in areas such as materials prepared for item-writing workshops, test form constructions, test booklet development and printing, post-test score quality control processes, and reporting. Quality control for item-writing workshops, test form construction, and test booklet development and printing can be found in the related documents prepared by CRESST and associated vendors. Constructed-response items were scored locally by test administrators (TAs) who were familiar with the test takers. Guidance on local scoring is provided in the *Alt ELPA Test Administrator Manual* and described in Chapter 5 of the *Alt ELPA Technical Manual*. The present chapter describes CAI's quality control procedures related to test administration, scoring, and reporting.

7.1 Quality Control in Test Configuration

For online testing, the test configuration files contain the complete information required for test administration and scoring, such as the test blueprint specifications, slopes, and intercepts for theta-to-scale score transformation, cut scores, and item information (e.g., answer keys, item attributes, item parameters, passage information). The accuracy of the configuration file is checked and confirmed independently numerous times by multiple teams prior to the testing window. Scoring is also verified before the testing windows open.

7.2 Platform Review

CAI's online Test Delivery System (TDS) supports a variety of item layouts for online test administration to many populations of students, including students who need designated supports and accommodations to test online. Each item on the 2023–2024 Alt ELPA went through an extensive platform device review on different operating systems, including Windows, Linux, and iOS, to ensure that the item displayed consistently across all platforms.

Platform Review is a process in which each item is checked to ensure that it is displaying appropriately (i.e., rendered) on each tested platform. A *platform* is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and Platform Review now takes place on various platforms that are significantly different from one another.

Platform Review was conducted by CAI's quality assurance (QA) team for the 2023–2024 Alt ELPA. The team leader projected every item from CAI's Item Tracking System (ITS[2]), and team members, each behind a different platform, viewed the same item to ensure that it rendered as expected.

[2] ITS is CAI's item bank for the Alt ELPA. It contains all information related to each item, such as item content categories at all levels, item type, maximum score points, item statistics from each test administration, etc.

7.2.1 User Acceptance Testing and Final Review

Both internal and external user acceptance testing (UAT), usually the state's, were conducted before the 2023–2024 testing window opened. Detailed protocols were developed for the review process of the TDS, and reviewers were given thorough instructions to note or report issues related to system functionality, item display, and scoring.

During the internal UAT, CAI staff took all Alt ELPA online tests that covered the entire range of possibilities of item responses and the complete set of scoring rules in the TDS. When issues were found, CAI took immediate actions to address them. The examples of issues identified and the actions taken during the internal UAT were as follows:

- Item layout issues: Some items were not rendering as anticipated in the TDS and the test was not moving. The item layouts were updated to ensure that items rendered correctly.
- Item drop-down zoom issue: A zoom issue was identified with Editing Task Choice (ETC) items (items where students identify an incorrect word or phrase and choose the replacement from several options) where the drop-down content was not enlarged. The items were updated to support different zoom levels in the drop-down menus.
- Student eligibility issues: Braille eligibilities were not working as expected. The test IDs needed to be updated in the TDS to resolve the issue.
- User eligibility issues: The user eligibilities were not working as expected. They were updated based on the state rules.
- Tool configuration issues: Some tools were not consistent across the tests. The tools were updated based on the state and Alt ELPA guidelines.

When the TDS was updated, the tests were taken again to ensure that the issues were fixed. The process was repeated until all issues were resolved during the UAT period prior to operational testing.

State staff also conducted a hands-on review of the system prior to the testing window opening. The states approved the TDS before the system was opened for testing.

Before the Centralized Reporting System (CRS) opened, CAI and the state staff conducted internal and external UAT of the system similar with that of the TDS to ensure that the CRS would function as intended when opened to the public for score reporting.

7.3 Quality Assurance in Scoring

The quality assurance (QA) of scoring included the assurance of the online data, the correctness of machine scoring, and the strictness when applying the business rules in scoring. This section describes the details of QA in scoring.

7.3.1 Quality Assurance in Online Data

The TDS has a real-time, built-in quality monitoring component. After a test is administered to a student, the TDS passes the resulting data to CAI's Quality Monitor (QM) subsystem[3]. CAI's QM subsystem conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item, keys for multiple-choice items, score points in each item, and total number of items, and that the test record contains no data from items that have been invalidated.

Data pass directly from the QM subsystem to the Database of Record (DOR), which serves as the repository for all test information and from which all test information for reporting is retrieved. The Data Extract Generator (DEG) is the tool that is used to retrieve data from the DOR for delivery to each state. CAI staff ensure that data in the extracted files match the DOR prior to delivery to the state.

7.3.2 Quality Control on Final Scores

CAI's scoring engine was used to produce final scores for the 2023–2024 Alt ELPA. Before operational scoring, CAI created mock-ups of student records to verify the accuracy of the scoring engine. Both CAI's Analysis Team (responsible for the scoring engine) and psychometricians independently computed scores on the mock-ups of student records. The Psychometrics and Statistics Team performed score verification using a different software and compared the scoring results with those from CAI's scoring engine. Specifically, if the Psychometrics and Statistics Team found score discrepancies from the scoring engine, they discussed them with the Analysis Team to find out the causes of discrepancies. After the Analysis Team updated the scores in the scoring engine, the Psychometrics and Statistics Team compared the scores again. The process was performed iteratively until a 100% match was reached.

During operational scoring, CAI's psychometricians independently scored students and compared the scores with the results from the scoring engine. Discrepancies were iteratively resolved until a 100% match was reached.

Before final scores were delivered to the state, they were also compared with the unofficial scores from CRESST, if needed. Discrepancies were again investigated and resolved until a 100% match was reached.

7.4 Quality Assurance in Reporting

In 2023–2024, two types of score reports were produced: (1) online reports and (2) printed reports (family reports only) for some states.

7.4.1 Online Report Quality Assurance

Every assessment underwent a series of validation checks. Once the QM subsystem signed off, data were passed to the DOR, which served as the centralized location for all student scores and

[3] The QM subsystem is CAI's quality monitoring system. It ensures that the information in a student record, such as item key or score point, is correct.

responses, ensuring that there was only one place where the official record was stored. Only after scores passed the QA checks and were uploaded to the DOR were they passed to the CRS, which was responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score was reported in the CRS until it passed all of the QM subsystem's validation checks.

7.4.2 Paper Report Quality Assurance

Statistical Programming

The 2023–2024 family reports contained custom programming and required rigorous QA processes to ensure their accuracy. All custom programming was guided by detailed and precise specifications in CAI's reporting specifications document. Upon approval of the specifications, analytic rules were programmed and each program was extensively tested on test decks and real data from other programs. Two senior statisticians and one senior programmer reviewed the final programs to ensure that they implemented agreed-on procedures. Custom programming was independently implemented by two statistical programming teams working from the specifications. The scripts were released for production only when the output from both teams matched exactly. Quality control, however, did not stop there.

Much of the statistical processing was repeated, and CAI implemented a structured software development process to ensure that the repeated tasks were implemented correctly and identically each time. CAI's software developers wrote small programs called *macros* that took specified data as input and produced data sets containing derived variables as output. Approximately 30 such macros resided in CAI's library. Each macro was extensively tested and stored in a central development server. Once a macro was tested and stored, changes to the macro were required to be approved by the director of score reporting and by the project directors for affected projects.

Each change was followed by a complete retesting with the entire collection of scenarios on which the macro was originally tested. The main statistical program was made up mostly of calls to various macros, including macros that read-in and verified the data and conversion tables and macros that did the many complex calculations. This program was developed and tested using artificial data generated to test both typical and extreme cases. In addition, the program went through a rigorous code review by a senior statistician.

Display Programming

The paper report development process used graphical programming, which took place in a Xerox-developed programming language called Variable Data Intelligent PostScript Printware (VIPP) and allowed virtually infinite control of the visual appearance of the reports. After designers at CAI created backgrounds, VIPP programmers wrote code that indicated where to place all variable information (i.e., data, graphics, and text) on the reports. The VIPP code was tested using both artificial and real data. CAI's data generation utilities could read the output layout specifications and generate artificial data for direct input into the VIPP programs. This allowed the testing of these programs to begin before the statistical programming was complete.

In later stages, artificial data were generated according to the input layout and run through the score reporting statistical programs, and the output was formatted as VIPP input; this enabled CAI to test the entire system. Programmed output went through multiple stages of review and revision by graphics editors and the Communications and Reporting Team to ensure that design elements were accurately reproduced and data were correctly displayed.

Once CAI received the final data and VIPP programs, the CAI Communications and Reporting Team reviewed proofs that contained actual data based on CAI's standard QA documentation. A large sample of reports was reviewed by several CAI staff members to ensure that all data were correctly placed on reports. This rigorous review was typically conducted over several days and took place in a secure location at CAI. All reports containing actual data were stored in a locked storage area. Prior to printing the reports, CAI provided a live data file and individual student reports (ISRs) with sample districts for the state staff to review. CAI worked closely with each state to resolve questions and correct any problems. The reports were not delivered until the state approved the sample reports and data file.

Chapter 8. Classical Item Analysis

The purpose of this chapter is to present the classical item analysis for both operational and field-test items included in the 2023–2024 operational test administration. It also includes differential item functioning (DIF) analyses for field-test items.

CAI conducted classical item analysis for operational items and ELPA21 conducted classical item analysis and DIF analysis for field-test items.

8.1 Item Analysis for Operational Items

CAI employs classical item analysis procedures to monitor and ensure that operational items function as intended with respect to the underlying scales. Key statistics computed are as follows:

Item Difficulty. For dichotomous items, item difficulty is computed as the proportion of examinees in the sample selecting the correct answer or earning the full point (*p*-value); for polytomous items, item difficulty is computed as the average proportion correct (analogous to *p*-value and indicating the ratio of item's mean score divided by the number of points possible, referred to as "relative mean" in the previous year's technical report).

Item discrimination. The item discrimination index indicates the extent to which each item differentiates between those examinees who possess the skills being measured and those who do not. In general, the higher the value, the better the item is able to differentiate between high- and low-achieving students. The discrimination index is calculated as the correlation between the item score and the domain scale score.

8.2 Item Analysis for Field-Test Items

For the 2023–2024 Alt ELPA, ELPA21 conducted classical test analyses and DIF analysis for the field-test items embedded in each test form. Various descriptive statistics based on classical test theory (CTT) were computed. Among them were average proportion correct (i.e., mean of item score divided by the maximum possible score), proportion of responses at each score category, biserial/polyserial correlation between item score and total score, and average total score by item score. The descriptive statistics were compared with the suggested thresholds to flag items for review. Refer to Chapter 6.1 of the *Alt ELPA Technical Manual* for details of these descriptive statistics.

Table 8.1 shows the criteria for evaluating descriptive item statistics by flag number.

Table 8.1 Criteria for Evaluating Descriptive Item Statistics by Flag Number

Flag	Description
1	Average proportion correct > 0.9
2	Average proportion correct < 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

8.3 DIF Analysis for Field-Test Items

DIF analysis was also conducted on all field-test items to examine whether items may potentially advantage or disadvantage a subgroup of students based on demographic variables of interest or concern. Items flagged for DIF were to be further reviewed by content experts to determine whether the DIF flag was due to item bias or due to statistically spurious findings.

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 or concern, those three variables had relatively sufficient sample size in both the focal group and reference group, which was 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 (refer to Chapter 6.2 of the *Alt ELPA Technical Manual* for additional detail on the procedures). Based on the statistics, items were classified into three categories (A, B, and C) (Michaelides, 2007). The “A” classification category indicated negligible DIF, “B” indicated slight-to-moderate DIF, and “C” indicated a moderate-to-large level of DIF. An item was flagged if the classification category was +C or -C in any comparison group.

Table 8.2 shows the criteria used for DIF categorization for both dichotomous and polytomous items.

Table 8.2 Criteria for DIF Categorization

DIF Category	**Dichotomous items**	**Polytomous items**
A	MH D-DIF is not significantly different from zero at 5% level *OR* |MH D-DIF| < 1	MH CHISQ is not statistically significant at 5% level *OR* |ES| ≤ 0.17
B	MH D-DIF is significantly different from zero at 5% level *AND EITHER* 1 ≤ |MH D-DIF| < 1.5 *OR* 1 ≤ |MH D-DIF| AND MH D-DIF is not significantly different from one at 5% level	MH CHISQ is statistically significant at 5% level *AND* 0.17 < |ES| ≤ 0.25
C	|MH D-DIF| is significantly greater than one at 5% level *AND* 1.5 ≤ |MH D-DIF|	MH CHISQ is statistically significant at 5% level *AND* 0.25 < |ES|

Note. Source: Michaelides (2008).

References

Alternate English Language Proficiency Assessment (Alt ELPA). (2025). *Alt ELPA technical manual*. Los Angeles, CA: University of California, National Center for 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*.

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. https://files.eric.ed.gov/fulltext/ED483410.pdf

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland, and H. Wainer (Eds.), *Differential Item Functioning* (pp. 35–66). Hillsdale, NJ; Lawrence Erlbaum Associates, Publishers.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). Macmillan.

Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education*, *9*(3), 277–286.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391

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.

National Center for Research on Evaluation, Standards, and Student Testing (CRESST). (2023). *Alternate ELPA scoring specification: School year 2023–24*. Author.

Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, *16*(1), 6–17. https://doi.org/10.2307/3150868

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, *21*(2), 381–391. https://www.jstor.org/stable/2489828

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education*, *8*(2), 111–120. https://doi.org/10.1207/s15324818ame0802_1

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 234–247. https://www.academia.edu/17360936/On_the_Reliability_of_Testlet_Based_Tests

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