TECHNICAL REPORT

PART I – ASSESSMENT OVERVIEW

(ARKANSAS, IOWA, LOUISIANA, NEBRASKA, OHIO, AND WEST VIRGINIA)

English Language Proficiency Assessment for the 21st Century –

Listening, Reading, Speaking, and Writing

Grades K–12

2022–2023 Test Administration

Submitted to:
ELPA21

Submitted by:
Cambium Assessment, Inc.
1000 Thomas Jefferson Street, NW
Washington, DC 20007

April 2024

Table of Contents

List of Tables

Chapter 1. Introduction

This technical report focuses on the 2022–2023 test administration, scoring, standard setting, test form reliability and validity, scoring, reporting, and quality control applied for Arkansas, Iowa, Louisiana, Nebraska, Ohio, and West Virginia. This technical report has four parts:

1. Part I includes an introduction, a general overview of reporting structure, and material that is common to both summative and screener assessments:
    - 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 and Test Analyses
2. Part II includes chapters that delineate different aspects of the 2022–2023 administration of the summative assessment, including:
    - Chapter 1. Test Administration
    - Chapter 2. 2022–2023 Summary
    - Chapter 3. Reliability
    - Chapter 4. Validity
    - Chapter 5. Reporting
    - Chapter 6. Classical Item and Test Analyses
3. Part III includes chapters that delineate different aspects of the 2022–2023 administration of the screener assessment, including:
    - Chapter 1. Test Administration
    - Chapter 2. 2022–2023 Summary
    - Chapter 3. Reliability
    - Chapter 4. Validity
    - Chapter 5. Reporting

4. Part IV includes the appendices of the 2022–2023 summary for each of the six states, as listed here, and the six states combined. The pooled analyses are based on the data from all six states.
   - Appendix for Arkansas—2022–2023 Summary
   - Appendix for Iowa—2022–2023 Summary
   - Appendix for Louisiana—2022–2023 Summary
   - Appendix for Nebraska—2022–2023 Summary
   - Appendix for Ohio—2022–2023 Summary
   - Appendix for West Virginia—2022–2023 Summary
   - Appendix for Pooled Analysis—2022–2023 Summary

   Each Appendix contains the following sections:

   Section 1: Summative Assessment—Student Participation

   Section 2: Summative Assessment—Raw Score Summary

   Section 3: Summative Assessment—Raw Score Distribution

   Section 4: Summative Assessment—Scale Score Summary

   Section 5: Summative Assessment—Percentage of Students by Domain Performance Level

   Section 6: Summative Assessment—Percentage of Students by Overall Proficiency Level

   Section 7: Summative Assessment—Testing Time

   Section 8: Summative Assessment—Cronbach’s Alpha

   Section 9: Summative Assessment—Marginal Reliability

   Section 10: Summative Assessment—Conditional Standard Error of Measurement (CSEM)

   Section 11: Summative Assessment—Classification Accuracy and Consistency

   Section 12: Summative Assessment—Inter-Rater Analysis

   Section 13: Summative Assessment—Dimensionality

   Section 14: Summative Assessment—Ability vs. Difficulty

   Section 15: Summative Assessment—Mock-Ups for Reporting

   Section 16: Screener Assessment—Student Participation

   Section 17: Screener Assessment—Raw Score Statistics

Section 18: Screener Assessment—Raw Score Distribution

Section 19: Screener Assessment—Scale Score Summary

Section 20: Screener Assessment—Percentage of Students by Domain Performance Level

Section 21: Screener Assessment—Percentage of Students by Overall Proficiency Level

Section 22: Screener Assessment—Testing Time

Section 23: Screener Assessment—Marginal Reliability

Section 24: Screener Assessment—Conditional Standard Error of Measurement (CSEM)

Section 25: Screener Assessment—Classification Accuracy and Consistency

Section 26: Screener Assessment—Inter-Rater Analysis

Section 27: Correlations Between Summative and Screener Assessments

Section 28: Student Progress from Screener to Summative

Section 29: Screener Assessment—Mock-Ups for Reporting

1.1 Background

The English Language Proficiency Assessment for the 21st Century (ELPA21) is a testing program that supports educators as they implement the 2014 English Language Proficiency (ELP) standards (Council of Chief State School Officers [CCSSO], 2014) and college- and career-readiness standards. The ELPA21 Program, hereafter referred to as “the Program,” is an assessment system that measures students’ ELP and provides valuable information to inform instruction and facilitate the development of academic English proficiency so that all English learners (ELs) leave high school prepared for college and career success. The assessment system includes assessments on listening, reading, speaking, and writing for students in pre-K (except for the state of Ohio, which uses different rules to screen pre-K students), kindergarten, grade 1, grades 2–3, grades 4–5, grades 6–8, and grades 9–12.

Ohio, in contrast with the other states, administers two types of screeners, OELPS-BK and OELPS-K. The difference between the two screeners is in the proficiency determination rules, not in the content. The OELPS-BK is the Ohio English Language Proficiency Screener for the Beginning of Kindergarten. Students enrolling in kindergarten in the first half of the kindergarten year (on or before December 31) are administered the OELPS-BK. Kindergarteners taking the OELPS-BK will be proficient (not an English learner) if they earn domain levels of 3 or higher in all nonexempt domains of the screener. The OELPS-K is the Ohio English Language Proficiency Screener administered to kindergarteners enrolling in the latter half of the kindergarten year (after December 31). Kindergartners taking the OELPS-K will be proficient (not an English learner) if they earn domain levels of 4 or higher in all nonexempt domains of the screener.

Applying a different proficiency standard to students in [or before] kindergarten is not unique to Ohio. The OELPS-BK and OELPS-K are the same assessment that only differ in the definition of proficiency. In the OELPS-BK, *Proficient* is defined as achieving Level 3 or above in all non-exempted domains, while in the OELPS-K, *Proficient* is defined as achieving Level 4 or above in all non-exempted domains. Ohio screens Pre-K students after they have completed preschool instruction.

Starting in 2021–2022 for all states, pre-K (or BK for Ohio) students are considered overall proficient with all 3 or above in each domain rather than all 4 or above. For kindergarten and higher grades, students need to obtain 4 or above in each domain for proficiency.

The Program conducted test development and item development for the summative ELP assessment as part of a U.S. Department of Education (USDOE) grant, which commenced in 2013 and ran through the first operational test administration in 2016. As part of the development process, Questar Assessment, Inc., built multiple fixed-length forms for each assessment. Items were field tested in spring 2015, and the first ELPA21 operational test administration took place in spring 2016. Following this test administration, the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) conducted item analyses, held data review meetings, and performed item calibration to obtain scoring parameters. Pacific Metrics, the organization contracted for standard setting, held a standard-setting workshop in July 2016. Based on recommendations from the workshop, the Program made decisions with respect to domain cut scores that further translated into performance levels for each grade. Cambium Assessment, Inc. (CAI) used the final item parameters, cut scores, and proficiency definitions to score and report the test results.

Details about test development, item development, field-test form building, item data review, item calibration, and standard setting can be found in the respective reports provided by the Program or obtained from the respective supporting vendors.

In 2017, the Program introduced the ELPA21 screener. The purpose of the screener was to determine students’ eligibility for English language development services. It assessed a student’s language proficiency in the required domains of listening, reading, speaking, and writing. The screener assessment items were drawn from the same item pools and were based on the same ELP standards as the summative assessment. The screener followed the same quality control procedures as the summative. Each state may have its own rules for deciding whether a student needs to take the screener assessment.

The 2022–2023 ELPA21 Program included summative and screener assessments. The summative and screener assessments were administered to students in six grade bands: kindergarten, grade 1, grades 2–3, grades 4–5, grades 6–8, and grades 9–12. Pre-K students can also participate in the screener assessment. The assessments do not have a time limit. Each assessment involves four domain (listening, reading, speaking, and writing) assessments. Students can be exempted from as many as three domains.

1.2 General Overview of the Reporting Structure

For both the summative and screener assessments, the ELPA21 results are available in the Centralized Reporting System (CRS[1]) and 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.

[1] The Centralized Reporting System (CRS) was used by all the states in the past school year. In the 2021–2022 test administration, Arkansas, Iowa, Louisiana, Nebraska, Ohio, and West Virginia adopted CRS Reporting for score reporting. Oregon is part of ELPA21; however, Oregon used computer-adaptive testing (CAT), so Oregon data and analyses were not included in this technical report. The term *CRS* throughout this report refers to Reporting for all six states: Arkansas, Iowa, Louisiana, Nebraska, Ohio, and West Virginia. All rules applicable to score reporting apply to both the CRS and Reporting.

Chapter 2. Scoring

For both summative and screener assessments, four domain scores and two composite scores were computed. The composite scores included a comprehension score for listening and reading and an overall score that comprised all four domains.

2.1 Estimating Student Ability

The ELPA21 team reported scale scores for each domain assessment, the overall scores 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) was used to estimate domain scores. Item bi-factor models were used to estimate the overall and comprehension scores. ELPA21 uses a 2PL based MIRT model, so one-to-one correspondence between raw and scale scores would not have been 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. This is important in interpreting the raw score statistics in the appendices. Details of score estimation can be found in the ELPA21 Scoring Specification: School Year 2022–2023 (National Center for Research on Evaluation, Standards, and Student Testing [CRESST], 2021). The business scoring rules for each of the summative and screener assessments are described in Part II and Part III of this technical report.

2.2 Theta to Scale Score Transformation

Student performance was summarized in an individual domain score for each domain, a comprehension score that included listening and reading, and an overall score that included all four domains. Each untransformed logit score (θ) obtained from the MIRT scoring model was linearly transformed to the reporting scale using the formula $SS = a * \theta + b$, where a is the slope and b is the intercept. There was one set of scaling constants for the domain scores and another set of constants for the composite scores, as shown in Table 2.1. Scale scores were rounded to an integer.

Table 2.1 Scaling Constants on the Reporting Metric

Subject	Grade	Slope (*a*)	Intercept (*b*)
Domain Scores (listening, reading, speaking, and writing)	K–12	80	550
Comprehension Scores	K–12	600	5500
Overall Scores	K–12	600	5500

2.3 Lowest/Highest Obtainable Scores

ELPA21 used expected a posteriori (EAP) scoring, which did not assign fixed minimum- or maximum-obtainable scale scores. The observed minimums, means, maximums, and standard deviations of scale scores by domain and by subgroup are presented in Sections 4 and 19 of the pooled and state-specific appendices.

2.4 Handscoring

For ELPA21 screener and summative assessments, all speaking items and some writing items were handscored. Measurement Incorporated (MI) provided all handscoring except for screeners administered in Ohio, which were scored locally. The procedure for handscoring items was provided by the ELPA21 Program. Scoring rubrics and item content were reviewed by content experts as a part of the item review meetings. Consistency in handscoring required that scoring rules be applied with fidelity during scoring sessions.

2.4.1 Rules for Handscoring

The ELPA21 assessments contained constructed-response items that required handscoring. In the speaking and writing domains, short-text items were scored on 1-, 2-, 3-, 4-, and 5-point rubrics. The following procedures were employed to handscore these items: All constructed-response items were assigned to a human rater for a first read (R1). The score assigned in this first read was the item score of record and was used to compute scale scores. Twenty percent of constructed-response items for the summative assessment were randomly selected for a second read (R2) (i.e., 20% of student responses to any constructed-response item had both a first read and a second read). Ten percent of the constructed-response items for the screener assessment were randomly selected for a second read.

The scores from these two reads were used to compute rater consistency statistics (% exact agreement, % adjacent agreement) included in Cambium Assessment, Inc.'s (CAI) annual technical reports. CAI and MI used second reads to monitor rater performance and provide ongoing feedback and training, as needed. Item scores from second reads were not used to compute scale scores.

First and second reads were performed by the same rater pool and occurred at approximately the same time. Raters did not know whether they were providing the first or second read.

If scores assigned in first and second reads differed by two or more score points (or if first and second raters differed in the selection of condition/scorability code), the student response was assigned to a supervisor for a third read (R3). The supervisor knew he or she was conducting a third read, had access to the results from the first and second reads, and would determine the score/code that should have been assigned. Third reads were performed only for the summative and not for the screener. CAI and MI used the results of the third read to provide ongoing feedback and training, as needed. Item scores from second reads were not used to compute scale scores.

Scores from all reads (first read, as well as second and third reads, if applicable) were included in the item's data file. CAI (presumably with MI's help) included detailed descriptions of scoring procedures in the annual technical report, including descriptions of ongoing feedback and training that was provided within a program year. Table 2.2 presents nonscorable codes for handscored items.

Table 2.2 Nonscorable Condition Codes for Handscored Items

Domain	Code	Description
Speaking	A	Blank
Speaking	B	Technological Issue
Writing	A	Blank

The following rules were adhered to when evaluating a potential nonscorable response in the speaking domain:

1. When a student responded with a word or phrase that could be tied to the stimulus, the response could receive a score point of “1.” The “0” score point responses followed the bulleted list contained in the rubric.
2. If no words were spoken by the student, it was considered a blank.
3. A teacher voice was not necessarily interpreted as interference; if the teacher was heard telling the student to speak but not telling them what to say, the scorer scored the student’s response.
4. A student response of “Yes”, “No”, or “I don’t know” was considered a refusal and should be scored a “0.”
5. A nonscorable code of “B” should be given for responses with a technical difficulty (e.g., speaking too close to the microphone causing unintelligible speech, broken recording with speech cut up, etc.).

Chapter 3. Standard Setting

For both the summative and screener assessments, the domain cut scores and the overall proficiency rules were set through a standard-setting meeting convened by the ELPA21 Program on July 19–22, 2016. Details about the standard-setting process can be found in the ELPA21 standard-setting technical report (CRESST & Pacific Metrics, 2016).

Five performance levels were established for each domain. The cut scores were set by grade, as listed in Table 3.1. The four cut scores set for each domain sorted students into Performance Levels 1–5. If a student scored below the first cut score (Cut 1), the student was classified as Performance Level 1. If a student scored at or above the first cut score but below the second cut score (Cut 2), the student was classified as Performance Level 2. This approach continued for Performance Levels 3 and 4. If a student scored at or above the fourth cut score, the student was classified as Performance Level 5.

Table 3.1 ELPA21 Domain Cut Scores by Grade

Grade	Domain	Cut 1	Cut 2	Cut 3	Cut 4	Grade	Domain	Cut 1	Cut 2	Cut 3	Cut 4
K	Listening	467	507	613	645	5	Listening	413	455	498	581
	Reading	473	514	592	627		Reading	468	511	588	627
	Speaking	487	535	598	625		Speaking	483	526	573	607
	Writing	497	562	651	673		Writing	438	486	598	628
1	Listening	435	467	549	594	6	Listening	410	440	498	565
	Reading	479	515	584	629		Reading	461	496	565	604
	Speaking	528	577	593	619		Speaking	465	511	562	595
	Writing	498	548	613	641		Writing	425	472	564	594
2	Listening	408	438	512	564	7	Listening	430	473	553	597
	Reading	457	489	555	595		Reading	486	534	609	642
	Speaking	490	529	555	588		Speaking	475	527	582	611
	Writing	452	493	555	591		Writing	474	520	597	625
3	Listening	409	448	536	598	8	Listening	432	478	565	613
	Reading	495	541	610	644		Reading	494	547	640	669
	Speaking	500	538	572	612		Speaking	476	528	590	619
	Writing	498	542	603	636		Writing	484	533	619	647
4	Listening	398	431	492	563	9–12	Listening	451	491	571	613
	Reading	453	488	550	594		Reading	488	539	631	662
	Speaking	462	506	544	584		Speaking	481	536	593	619
	Writing	437	481	568	600		Writing	485	533	615	641

Overall proficiency, defined as "proficiency determination," for a given student was established based on a profile of domain performance levels across all four tested domains. There were three proficiency determination categories: (1) Emerging, (2) Progressing, and (3) Proficient. The

following three rules determined a student's overall proficiency (note that for the purpose of assigning overall proficiency, nonexempt domains that were not attempted were treated as Performance Level 1. Therefore, students with one or more untested nonexempt domains cannot reach proficiency):

1. Students whose domain performance levels were 1 or 2 across all nonexempt domains were identified as Emerging.
2. Students whose domain performance levels were 4 or 5 across all nonexempt domains were identified as Proficient. In some cases, future kindergarten screeners used a different definition of proficiency. Starting in 2021–2022, all students taking the future kindergarten screener were proficient if all non-exempt domains were 3 or better. In Ohio, from the end of the before Kindergarten (BK) year and through December 31 of the kindergarten year, students taking the kindergarten screener before December 31 of the kindergarten year were proficient if all non-exempt domains were 3 or better.
3. Students with domain performance levels that did not fit with Emerging or Proficient (as defined previously) were identified as Progressing.

See details in Appendix B (Overall Proficiency Determination Look-up Tables) in the ELPA21 Scoring Specification: School Year 2022–2023 (CRESST, 2022).

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. The conventional approaches used are characterized as follows:

The *test-retest* method measures stability over time. With this method, the same test is administered twice to the same group at two different points in time. If test scores from the two administrations are highly correlated, then the test scores are deemed to have a high level of stability. For example, if the result is highly stable, those who scored high on the first test administration tend to obtain a high score on the second test administration. The critical factor, however, is the time interval. The time interval should not be too long, which could allow for changes in the test takers' true scores. Likewise, it should not be too short, in which case memory and practice may confound the results. The test-retest method is most effective for measuring constructs that are stable over time, such as intelligence or personality traits.

The *parallel-forms* method is used for measuring equivalence. With this design, two parallel forms of the test are administered to the same group. This method requires two similar forms of a test; however, it is very difficult to create two strictly parallel forms. When this method is applied, the effects of memory or practice can be eliminated or reduced, since the tests are not purely identical as they are with the test-retest method. The reliability coefficient from this method indicates the degree to which the two tests are measuring the same construct. While there are a wide variety of possible items to administer to measure any particular construct, it is only feasible to administer a sample of items on any given test. If there is a high correlation between the scores of the two tests, then inferences regarding high reliability of scores can be substantiated. This method is commonly used to estimate the reliability of achievement or aptitude tests.

The *split-half* method uses one test divided into two halves within a single test administration. It is crucial to make the two half-tests as parallel as possible, as the correlation between the two half-tests is used to estimate reliability of the whole test. In general, this method produces a coefficient that underestimates the reliability for the full test. To correct the estimate, the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910) can be applied. While this method is convenient, varying splits of the items may yield different reliability estimates.

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:

$$\sigma_{\mathrm{X}}^2 = \sigma_{\mathrm{T}}^2 + \sigma_{\mathrm{E}}^2.$$

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

$$\rho_{\mathrm{XX}'} = \frac{\sigma_{\mathrm{T}}^2}{\sigma_{\mathrm{T}}^2 + \sigma_{\mathrm{E}}^2} = \frac{\sigma_{\mathrm{T}}^2}{\sigma_{\mathrm{X}}^2} = \frac{\sigma_{\mathrm{x}}^2 - \sigma_{\mathrm{E}}^2}{\sigma_{\mathrm{X}}^2} = 1 - \frac{\sigma_{\mathrm{E}}^2}{\sigma_{\mathrm{X}}^2}.$$

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 proficient 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 Chapter 3 of technical reports Part II and Part III.

4.1 Internal Consistency

Cronbach's alpha (Cronbach & Shavelson, 2004) is used to access the internal consistency of items in each assessment for each domain for the summative assessment. A high Cronbach's alpha coefficient indicates that the items in the domain 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 standard errors measurement of the IRT-based scale scores obtained by applying the ELPA21 scoring procedures. Smaller values of MSEM indicate that the estimated test scores have greater precision, on average. The marginal reliability $\bar{\rho} = 1 - \frac{\bar{\sigma}^2_{error}}{\sigma^2_{total}}$ (see Section 4.3 in the following paragraph) and the test MSEM are inversely related. The ratio of MSEM and the standard deviation of scale scores (i.e., signal-noise ratio) can also indicate the measurement errors. In other words, it shows the ratio of the error and total score ($\frac{\bar{\sigma}_{error}}{\sigma_{total}}$).

4.3 Marginal Reliability and Conditional Standard Error of Measurement

Marginal reliability (Sireci, Thissen, & Wainer, 1991) assesses the precision of scoring. It is based on the average of the conditional standard error of measurement (CSEM: σ_{EAP}) 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 was computed using item-level scores, marginal reliability was estimated by using expected a posteriori (EAP) estimates, which are used to estimate the domain scores. EAP is an estimate of the true score, but its variance underestimates the true score variance, so the marginal reliability within domain 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 ELPA21, since the overall proficiency is based on domain performance level, the CA and CC are examined at each cut score in each domain test. Five performance levels divided by four cut scores, cut scores 1–4, are established for each domain test.

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

At domain 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 *i*th student, and p_{il} can be estimated as follows:

$$\begin{aligned} p_{il} = p(c_{l-1} \leq \theta_i < c_l) &= p\left(\frac{c_{l-1} - \hat{\theta}_i}{se(\hat{\theta}_i)} \leq \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} < \frac{c_l - \hat{\theta}_i}{se(\hat{\theta}_i)}\right) \\ &= p\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)} < \frac{\hat{\theta}_i - \theta_i}{se(\hat{\theta}_i)} \leq \frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) = \Phi\left(\frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) - \Phi\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)}\right). \end{aligned}$$

For Level 1, $c_0 = -\infty$, and for Level L, $c_L = \infty$. If scaled score is to be used, the formula shown previously 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 domains. Similar to the case shown previously for individual domains, 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 domains. The 4 × 1 vector of score estimates $\widehat{\boldsymbol{\theta}}_{\iota}$ and the 4×4 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 in Table 2.1 on page 6. 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 domain. For most students (those without exemptions), the computation is as follows:

$$\begin{aligned} &\hat{p}_{i,(e,f,g,h)} \\ &= \int_{\text{cut}_{e,\text{listening}}}^{\text{cut}_{(e+1),\text{listening}}} \int_{\text{cut}_{f,\text{listening}}}^{\text{cut}_{(f+1),\text{listening}}} \int_{\text{cut}_{g,\text{listening}}}^{\text{cut}_{(g+1),\text{listening}}} \int_{\text{cut}_{h,\text{listening}}}^{\text{cut}_{(h+1),\text{listening}}} P(\boldsymbol{SS}|\boldsymbol{y}_i) dSS_{\text{listening}}\, dSS_{\text{reading}} dSS_{\text{speaking}} dSS_{\text{writing}}, \end{aligned}$$

where e, f, g, and h are the performance levels for listening, reading, speaking, and writing, respectively. Additionally, $\text{cut}_{1,d} = -\infty$ and $\text{cut}_{6,d} = \infty$.

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:

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

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 domain 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 the 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}, similar to CA, a similar $L \times L$ table is constructed by assuming the test is administered independently 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 computation of CA and CC for overall proficiency categories follows the same procedure as that for domain performance levels, as described previously.

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.

4.5 Inter-Rater Analysis

The fidelity of handscoring was monitored by having a subset of student responses (20% of responses for each item in the summative and 10% in the screener) independently scored by two raters. Each student response was scored holistically by a trained and qualified rater using the scoring criteria developed and approved by ELPA21, with a second read conducted on 20% of responses for the summative and 10% of responses for the screener for each task type. Responses were randomly selected for second readings and scored by raters who were not aware of the score assigned by the first rater, or even that the response had been scored previously. The rater pool consisted of teachers, test administrators (TAs), school administrators, or other qualified school staff. The detailed information of handscoring quality assurance (QA), including scorer qualifications, is described in 7.2.2, Quality Assurance in Handscoring.

For both the summative and screener assessments, handscorer reliability was examined using Cohen’s quadratic weighted Kappa coefficient (Cohen, 1968). The coefficient is a measure of agreement corrected for chance and allows differential weighting of disagreement. Cohen’s kappa is calculated for items with a maximum score of 1 point; quadratic weighted kappa (QWK) is calculated for items with a maximum score of 2 or more points.

Cohen's kappa is computed as

$$\frac{P_o - P_e}{1 - P_e}$$

where *P*o is the proportion of observed agreement computed as the sum of the diagonal proportions; *P*e is the proportion of chance agreement computed as the sum of the product of the row and column marginal proportions on the diagonal.

QWK is computed as

$$1 - \frac{\sum_{ij} w_{ij} o_{ij}}{\sum_{ij} w_{ij} e_{ij}}$$

where $w_{ij} = (i - j)^2/(k - 1)^2$ and k is the number of score points, o_{ij} is the observed proportion, and e_{ij} is the expected proportion computed as the product of the row and marginal proportions.

In addition, the frequencies and percentages of the exact match between first rater and second rater, the exact match plus +1/-1 score differences, and +2/-2 and above differences were computed.

Chapter 5. Validity

Validity refers to the degree to which "evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education (NCME), 2014). Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment." Both definitions emphasize evidence and theory to support inferences and interpretations of test scores. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) suggested five sources of validity evidence that can be used in evaluating a proposed interpretation of test scores. When validating test scores, these sources of evidence should be carefully considered.

The first source of evidence for validity is the relationship between the test content and the intended test construct. For test score inferences to support a validity claim, the items should be representative of the content domain, and the content domain should be relevant to the proposed interpretation of test scores. To determine content representativeness, diverse panels of content experts conduct alignment studies in which experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a particular construct (discussions about test development, form construction, scaling, equating, and standard setting can be found in related ELPA21 documents). Test scores can be used to support an intended validity claim when they contain minimal construct-irrelevant variance.

For example, scores on a mathematics item targeting a specific mathematics skill that requires advanced reading proficiency and non-content-related vocabulary may display substantial construct-irrelevant variance. Thus, the intended construct of measurement is confounded, which impedes the validity of the test scores. Statistical analyses, such as factor analysis or multi-dimensional scaling of relevance, are also used to evaluate content relevance. Evidence based on test content is a crucial component of validity because construct underrepresentation or irrelevancy could result in unfair advantages or disadvantages to one or more groups of test takers.

The second source of evidence for validity is based on "the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA, APA, & NCME, 2014). This evidence is collected by surveying test takers about their performance strategies or responses to particular items. Because items are developed to measure particular constructs and intellectual processes, evidence that test takers have engaged in relevant performance strategies to correctly answer the items supports the validity of the test scores.

The third source of evidence for validity is based on internal structure: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. Differential item functioning (DIF), which determines whether particular items may function differently for subgroups of test takers, is one method for analyzing the internal structure of tests. Other possible analyses to examine internal structure are confirmatory factor analysis, goodness-of-model-fit to data, and reliability analysis.

A fourth source of evidence for validity is the relationship of test scores to external variables. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) divided this source of evidence into three parts: (1) convergent and discriminant evidence, (2) test-criterion relationships, and (3) validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs. Conversely, discriminant evidence delineates the test from other measures intended to assess different constructs. To analyze both convergent and discriminant evidence, a multi-trait, multi-method matrix can be used. Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy mainly depends on the purpose of the test, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another. Validity generalization is related to whether the evidence is situation-specific or can be generalized across different settings and times. For example, sampling error or range restriction may need to be considered to determine whether the conclusions of a test can be assumed for the larger population.

The fifth source of evidence for validity is based on whether the intended and unintended consequences of the test use should be included in the test validation process. Determining the validity of the test should depend on evidence directly related to the test; this process should not be influenced by external factors. For example, if an employer administers a test to determine hiring rates for different groups of people, an unequal distribution of skills related to the measurement construct does not necessarily imply a lack of validity for the test; however, if the unequal distribution of scores is in fact due to an unintended, confounding aspect of the test, this would interfere with the test’s validity. As described in this document, test use should align with the intended purpose of the test.

Supporting a validity argument requires multiple sources of validity evidence. This allows for one to evaluate if sufficient evidence has been presented to support the intended uses and interpretations of the test scores. Thus, determining the validity of a test first requires an explicit statement regarding the intended uses of the test scores and, subsequently, evidence that the scores can be used to support these inferences. The validity results are shown in Chapter 4 of Part II and Part III of this technical report.

Chapter 6. Reporting

For both the summative and screener assessments, the ELPA21 results were available in the Centralized Reporting System (CRS) for schools and districts to print out, and CRS-generated paper family reports to be sent home with students. The screener results were reported online only. Arkansas and Ohio ordered summative paper score reports that were shipped to districts.

6.1 Centralized Reporting System

The CRS generated a set of online score reports describing student performance for students, parents, educators, and other stakeholders for both the summative and screener assessments. 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 can 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 domain performance level (overall and by subgroup) • Average domain scale 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 domain performance levels • Domain scale scores with domain performance levels and level descriptor

6.1.1 Types of Online Score Reports

The CRS was designed to help educators, students, and parents answer questions regarding how well students have performed in the assessment for each domain. The CRS was designed with great consideration for stakeholders who are not technical measurement experts (e.g., teachers, parents, students). It ensures that test results are easy to interpret and accessible. Simple language is used so that users can quickly understand assessment results and make valid inferences about student achievement. In addition, the CRS was designed to present student performance in a uniform format. For example, similar color is used for groups of 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, a teacher, and a 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 S15.1 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 and 29 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 has been integrated into every aspect of the ELPA21 summative and screener assessments. ELPA21, the states, Questar, Cambium Assessment, Inc. (CAI), and Measurement Incorporated (MI) have built in multiple layers of reviews and verifications to ensure that outputs are 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 ELPA21 and associated vendors. This chapter describes CAI and MI quality control procedures related to test administration, scoring, and reporting.

7.1 Quality Control in Test Configuration

For online summative and screener testing, the test configuration files contained 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 was checked and confirmed independently numerous times by multiple teams prior to the testing window. Scoring was also verified before the testing windows opened.

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 assessment 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 was checked to ensure that it was displayed 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. 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.

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 testing window opened. Detailed protocols were developed for the review process of

[2] ITS is CAI’s item bank for ELPA21. 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.

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 ELPA21 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 are presented here:

- Item layout issues: Some items were not rendering as anticipated in the TDS, and the test was not moving. The item layouts were updated for these items to render correctly.
- Item drop-down zoom issue: A zoom issue with the Editing Task Choice (ETC) (i.e., student identifies an incorrect word or phrase and chooses the replacement from several options) items where the drop-down content was not enlarged was identified. 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 ELPA21 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 QA of scoring includes the assurance of the online data, the precision of handscoring, the correctness of machine scoring, and the strictness when applying the business rules in scoring. This section describes the details of QA in scoring.

MI handscored the writing constructed-response items and speaking items. For online tests, the responses for the handscored items were transferred between CAI and MI on a rolling basis via Ledger.[3] Therefore, as soon as a student submitted a test to the TDS, the responses to handscored items were transformed into XML format, and were then sent to Ledger, from which MI retrieved responses for handscoring. When scoring was complete, the record was sent to Ledger, from which CAI downloaded the record for final scoring. The data transmission process was automatic.

[3] Ledger is an electronic system that CAI and MI use to transmit data from one vendor to the other for purposes of transmitting and reporting handscored item scores. Individual responses can be tracked at all times through Ledger before a record is reported.

After the administration of paper-pencil tests, student responses were entered into the CAI Data Entry Interface (DEI) on the state testing portal for all ELPA21 domain tests, except for the writing constructed-response items. The responses of the writing constructed-response items were mailed to MI for scoring via secure shipping. After scoring, MI transmitted the scores to the Ledger system, from which CAI retrieved the item scores for final scoring. To answer speaking items, students who took paper-pencil tests spoke into the DEI directly, and the item responses followed the online procedure for scoring.

For braille assessments, test administrators (TAs) entered item responses into the braille DEI. The data were processed following the online data processing procedure, and the secure testing materials were returned to MI.

7.3.1 Quality Assurance in Online Data

The TDS has a real-time, built-in quality monitoring component. After a test was administered to a student, the TDS passed the resulting data to CAI's Quality Monitor (QM[4]) System. CAI's QM System conducted 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 contained no data from items that had been invalidated.

Data passed directly from the QM System 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 ensured that data in the extracted files matched the DOR prior to delivery to the state.

7.3.2 Quality Assurance in Handscoring

MI's scoring process was designed to employ a high level of quality control. The quality control procedures were implemented at each stage of the scoring process, which includes scorer recruitment, leader recruitment, training, and various reports that helped to ensure scoring quality. CAI does not do plagiarism detection, but MI flags test irregularities in the hand scoring process when there is teacher interference.

Scorer Recruitment/Qualifications

MI retains scorers who have years of experience in handscoring, and those scorers make up approximately 65% of the scorer pool. To complete the scorer staffing for this project, MI placed advertisements on various job boards, in local papers, in publications, and at regional colleges and universities. Recruiting events were held, and applications for scorer positions were screened by MI recruiting staff. Candidates were personally interviewed, and references and proof of a four-year college degree were collected. Candidates completed placement tests for English language arts (ELA) (reading and writing) and mathematics. In this screening process, preference was given to candidates with previous experience scoring large-scale assessments. The scorer pool

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

consisted of educators, writers, editors, and other professionals who were valued for their experience, but who were also required to set aside their own biases about student performance and accept the scoring standards.

Leadership Recruitment/Qualifications

Scoring directors and team leaders had experience as successful scorers and leaders on previous MI projects and had strong backgrounds in scoring content-specific projects. These individuals demonstrated strong organizational, leadership, and management skills. All scoring directors, team leaders, and scorers were required to sign confidentiality agreements prior to training with ELPA21 materials and/or handling secure materials.

Each room of scorers was assigned a scoring director or assistant scoring director. This individual led the handscoring for the duration of the project and was monitored by the scoring project manager. The scoring director conducted the team leader training and was responsible for training the scorers.

Team leaders assisted the scoring directors and assistant scoring directors with scorer training and monitoring by working with their teams in small group discussions and answering individual questions that scorers may not have felt comfortable asking in a large group. Once scorers were qualified, the team leaders were responsible for maintaining the accuracy and workload of team members. The ongoing monitoring identified scorers who were having difficulty scoring based on accuracy and resulted in individual scorers receiving one-on-one retraining. If this process did not correct inaccuracies in scoring, individual scorers were released from the project.

Training

In rangefinding meetings, the full range of responses that represent each score point and produce scoring training materials, including qualification, anchor, practice, and validity sets, were identified. The rangefinding process first involved MI review and selection of responses for rangefinding. During rangefinding, participants reviewed items and rubrics, iteratively scored, discussed, and reached consensus on responses, and identified which ones to use as anchor and training responses.

To train ELPA21 scorers, MI scoring staff used approved rubrics and training materials taken from the rangefinding meetings. The training materials comprised anchor, qualifying, and training responses provided by the ELPA21 Program. Training materials included a comprehensive annotated scoring guide for each item. The guide contained the anchor set that scorers referenced while evaluating live student responses. The scoring guides also contained several typical student responses presented in score point order.

Guides included detailed annotations explaining how the scoring criteria applied to each response’s specific features and why the response merited a particular score. Guides included responses that were the most useful in making scoring decisions, including some that fell within the upper and lower ranges of the score point to help scorers define the lines between score points.

Anchor and qualifying sets were designed to help the scorers learn to apply the criteria illustrated in the scoring guide, ensure that they become familiar with the process of scoring student responses, and assess the scorers' understanding of the ELPA21 scoring criteria before they could begin live scoring.

The item-specific rubrics served as the scorers' constant reference. Scorers were instructed on how to apply the rubrics and were required to demonstrate a clear comprehension of each anchor set by performing well on the training materials that were presented for each grade and item.

Team leaders assisted the scoring directors with the training and monitoring of scorers. The scoring director conducted the team leader training before the scorer training. This training followed much of the same process as the scorer training, but additional time was allotted for review, discussion, and addressing anticipated scorer questions and concerns. To facilitate scoring consistency, it was imperative that each team leader imparted the same rationale for each response that other team leaders used. Once team leaders qualified, leadership responsibilities were reviewed and team assignments were given. A ratio of one team leader per 8–10 scorers ensured adequate monitoring of the scorers.

Scorer training involved an intensive review of the rubric and anchor responses that were provided by the scoring director to help the scorers internalize the scoring criteria. The scoring director and team leaders led a thorough discussion of the training materials with the entire group. All responses were discussed using the annotations from the rangefinding meetings. A similar process was followed in training for writing and speaking items.

Once the scoring guidelines were discussed, scorers were required to apply the scoring criteria by qualifying (i.e., scoring with acceptable agreement to the "true" scores decided at rangefinding) in at least one of the qualifying sets. Scorers who failed to achieve the qualifying criteria were given additional training to improve their accuracy. Scorers who did not perform at the required level of agreement for a given item or related group of items by the end of the qualifying process were not permitted to score live student work. The required level at the end of the qualifying process are the qualifying sets in which the reader must score a 70% or higher with no nonadjacent scores.

Training was an ongoing process that did not end after the qualifying rounds. Feedback was an integral part of several reliability checks that were performed throughout the project. Primarily, team leaders monitored scorers' reliability by conducting read-behinds/listen-behinds on an as-needed basis. This is a process whereby team leaders re-read and check scores of each scorer on their team. This is to catch potential scorer drift (i.e., shifts in scoring over time) so that the scorer can have immediate feedback and be retrained in a timely fashion, if needed. The percentage of read-behinds conducted for an individual scorer is not fixed but varied based on current levels of performance. Scorers receive one-on-one retraining based on monitoring results. Scorers are removed from scoring an item or related group of items if they cannot score consistently with the rubric and the anchor responses after retraining. When live scoring began, one of the team leader's primary jobs was to conduct read-behinds for their team members to ensure that they were scoring accurately. As this process continued, the team leader could start to recognize if the individual readers had a firm grasp of the criteria for the particular task type that was being scored or who may have needed some additional coaching. Once this was established using the read-behinds, the reader's notes were sent for score clarifications and reader reliability reports. The team leader could then determine who needed fewer read-behinds or who needed more monitoring.

Development and rangefinding of the materials used with the 2017 test administration were completed by a previous vendor. For 2021–2022, MI conducted a field-test score validation of the new short-response speaking items. This information is available from the Program.

7.3.3 Handscoring Quality Assurance Monitoring Reports

MI scorer accuracy was monitored throughout the scoring sessions by producing real-time, on-demand reports to ensure that an acceptable level of scoring accuracy was maintained. Interscorer reliability was tracked and monitored with multiple quality control reports that were reviewed by MI scoring staff. These reports were reviewed by the program manager, scoring project director, scoring directors, and team leaders. The following reports, available in daily, cumulative, and summary formats, were used during handscoring:

- *Interscorer Reliability Reports* displayed how often scorers were in exact agreement and supported maintaining an acceptable agreement rate. These reports provided rates of exact, adjacent (raters do not match within one point), and nonadjacent (raters more than one point apart) interscorer agreement, as well as mismatches between scores and nonscorable codes, and within nonscorable codes. They also indicated the number of responses read by each scorer.
- *Score Point Distribution Reports* displayed the percentage of responses that had been assigned each of the score points and nonscorable codes.
- *Validity Reports* tracked how the scorers performed by comparing predetermined scored responses to scores assigned by the selected scorer on the same set of responses. If the assigned score of the selected scorer fell outside of a determined percentage of agreement, remediation occurred and additional responses were reviewed by the team leader of the individual(s) who needed to be monitored more closely.
- *Item Status Reports* tracked each item and indicated the status (e.g., “first read complete,” “tabled”). This report was used to monitor the overall status and progress of handscoring.

Maintaining Consistency

MI used numerous processes to ensure scorer accuracy and to detect drift. The objective of the scoring process is to ensure that scorers rate student responses in a manner consistent with ELPA21 standards within a single ELPA21 test administration, as well as across multiple test administrations.

The validity selection process involved MI scoring staff selecting 30–75 responses per item from live responses from the current test administration to serve as validity responses. Validity responses were selected to illustrate trends identified by leadership in live responses, but not strongly reflected in the anchor sets, represent particular types of responses identified as challenging to score during training, and assess transfer of scorers’ knowledge of the anchor responses. Vetting of new validity responses involved identification and recommendation by team leaders while conducting read-behinds/listen-behinds, review and approval by scoring directors, and review and approval by the scoring project director.

The validity responses were used during handscoring to verify scorer accuracy. Validity responses were dispersed intermittently to the scorers throughout scoring at a rate of at least 10% of the total responses. These validity responses were blind reads, meaning that scorers saw these responses the same way they saw the actual live student responses; there was no distinguishable difference. This helped ensure the internal validity of the process. All scorers who received validity responses had already successfully completed the training and qualifying process.

Next, the scores that the scorers assigned to the validity responses were compared to the predetermined scores in order to determine the validity of the scorers' scores. For each item, the percentage of exact agreement and the percentage of high and low scores were computed. The same data were also computed for each specific scorer. Using these pieces of data, various validity reports could be produced in real time and used to monitor for potential drift.

If results indicated that there was drift for a particular response, item, or scorer, immediate action was taken to correct it. This action could include individual scorer retraining, room-wide retraining/recalibration, and/or rescoring responses where it was determined a scorer had been errantly assigning scores. Sometimes, when a particular validity response generated low agreement, an example of a similar response could be found in the existing training materials. If this was the case, a review of that particular training response was pursued in order to realign the scorer.

In most cases, including the 2022–2023 test administration, there was no room drift. Leadership can review particular types of responses and determine if there is a possible or potential shift in the scoring of those responses by using the questions provided by notes, reader reliability reports, and read-behinds. The scoring directors create recalibration sets that consist of commonly seen types of responses. These recalibration sets are given to the teams at the beginning of every week to help deter any negative trends or drifts. Additional recalibration sets are created if the scoring director starts to see a trend of a drift and can be given at any time it is determined warranted. All recalibration sets are approved by scoring management before being given to the scoring teams.

Recalibration sets consisting of a validation set representing a variety of score points in random score point order were also used to maintain consistency. Sets varied in size from three to five responses based on particular issues observed during scoring. The recalibration sets were distributed at the beginning of the morning on a weekly basis. MI also recalibrated approximately once a week with scorers who had missed a required day's scoring session and were required to recalibrate. Those scorers achieving a less-than-acceptable percentage of correct scores on these responses were monitored closely throughout that day. Scorers who did not demonstrate improvement received personal and extensive retraining. These scorers continued to be monitored on an individual basis until the next recalibration round took place.

By implementing these scoring procedures—using the same training materials whenever possible, using a suite of real-time reports, and making training decisions based on report data—MI maximized scoring reliability and validity.

7.3.4 Quality Control on Final Scores

CAI's scoring engine was used to produce final scores upon receiving handscores. 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 the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), if needed. Discrepancies were again investigated and resolved until a 100% match was reached.

7.4 Quality Assurance in Reporting

In 2021–2022, two types of score reports were produced for both the summative and screener assessments: online reports and printed reports (family reports only).

7.4.1 Online Report Quality Assurance

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

7.4.2 Paper Report Quality Assurance

Statistical Programming

The 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 implemented independently 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 reside 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 the director of psychometrics, as well as 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 verify 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 can 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 ran 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. In addition, CAI compared data independently calculated by CAI psychometricians with data on the reports. 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 and Test Analyses

The purpose of this chapter is the item analysis summary of all the operational and field-test items that were embedded in the 2022–2023 operational test administration. Domain correlations were calculated for operational items only. Differential item functioning (DIF) analyses were conducted for field-tested machine-scored items only. Typically, for handscored items, a sample of responses for each item is drawn and sent to Measurement Incorporated (MI) for handscoring. After the sampled responses are scored, Cambium Assessment, Inc. (CAI) conducts item analyses for those handscored items. However, for the 2022–2023 test administration, there are no field-tested handscored items being scored. For machine-scored items, all item responses were used for item analyses without sampling.

CAI psychometricians used CAI's Database of Record (DOR) to generate data files for the classical item analysis. The data files will include all students in the system. The DOR produces a fixed-width text file as well as a layout file with which to read the text file. The cleaned data files are comma delimited files and contain only the students with raw and scale scores for the current analysis.

In the data file, values for multiple-choice items range from 1–5 to indicate which option a student selected. For all non-multiple-choice items, the data file indicates the student score on an item. If students did not respond to that item, a missing response would be used for that item.

8.1 Item Analysis

For all the operational and field-test items, including machine-scored field-test items only for the 2022–2023 test administration, CAI conducted classical item analysis and analysis of DIF. The machine-scored items were analyzed in order to provide item statistics for future rubric validation.

CAI employs classical item analysis procedures to ensure that items function as intended with respect to the underlying scales. CAI's analysis program (Workspace), a statistical software, computes the required statistics for each item to check the item's integrity and to verify the appropriateness of its difficulty level. The key statistics that we compute and examine are as follows, and Table 8.1 outlines the flagging criteria of the key statistics when evaluating field-test items.

- **Item Difficulty.** Items that are either extremely difficult or extremely easy will be flagged for review. For multiple-choice items, we compute the proportion of examinees in the sample selecting the correct answer (*p*-value), as well as those selecting each of the incorrect responses. For constructed-response items, item difficulty will be calculated both as the item's mean score and 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). The percent of students in each score point category will also be calculated. Items are flagged for reviews if the *p*-value for multiple-choice items is less than 0.30 or greater than 0.95, and the *p*-value for other types of items is greater than 0.95. Items will also be flagged if less than 5% of the population occurs in any score point category.

- **Distractor Analysis.** Distractor analysis for the multiple-choice items is used to identify items that may have marginal distractors or ambiguous correct responses. In the distractor analysis, the correct response should be the most frequently selected option among high-scoring examinees. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. The point biserial correlation for distractors is the correlation between the item score, treating the target distractor as the correct response, and the scale score derived based on all items, restricting the analysis to those students selecting either the target distractor or the keyed response. Items are flagged for subsequent reviews if the point biserial correlation for the distractor response is greater than 0.05.
- **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 for multiple-choice items is calculated as the correlation between the item score and the domain scale score. For polytomous items, we compute and report the mean scale score value for students scoring within each of the possible score categories. Items are flagged for subsequent reviews if the point biserial/polyserial correlation for the keyed (correct) response is less than 0.25.

Table 8.1 Flagging Criteria

Rule	**Flagging Criteria**	**Rationale**
p-values	For 1-point items, flag if *p* < 0.30 or *p* > 0.95	Items are too difficult and *p*-value is less than expected from random chance or item is too easy for population
Average Proportion Correct (Relative Mean)	For polytomous items, flag if relative mean is > 0.95	Item difficulty is too difficult or too easy
Point-biserial/polyserial	Flag if < 0.25	Non-discriminating item
Distractor *p*-value	Flag if *p*-value for distractor is larger than *p*-value for key	Potentially problematic item
Distractor point-biserial/polyserial	Flag if distractor point biserial/polyserial is > 0.05	Potentially mis-keyed item
DIF	Flag if DIF classification category is +C or -C in any of the analysis groups (see the following section)	Potentially biased item

8.2 Domain Intercorrelations

This section explores the internal structure of the assessment using the scores provided at the domain level. the relationship of the domain scores is just one indicator of the test dimensionality.

Scale scores based on each domain were computed for this analysis. It is not reasonable to expect that the domain scores are completely orthogonal. This would suggest that there are no relationships among domain scores and would make justifying a unidimensional item response theory (IRT) model difficult. However, if the domains were not perfectly correlated, one could justify a multidimensional model.

8.3 Differential Item Functioning (DIF) Procedure

Analysis of the content alone is not sufficient to determine the fairness of a test. Rather, it must be accompanied by statistical processes. While a variety of item statistics were reviewed during form building to evaluate the quality of items, one notable statistic that was used was DIF (Dorans & Holland, 1993). Items were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF, according to the DIF classification convention illustrated in the analysis plan in the following paragraph.

CAI typically uses a generalized Mantel-Haenszel (MH) procedure to evaluate DIF. The generalizations include (1) adaptation to polytomous items and (2) improved variance estimators to render the test statistics valid under complex sample designs. Scale score (domain) is divided into 10 intervals based on the percentile of the scale scores to compute the MH chi-square DIF statistics for balancing the stability and sensitivity of the DIF scoring category selection. The analysis program computes the MH chi-square value, the log-odds ratio, the standard error of the log-odds ratio, and the MH-delta for the multiple-choice items, the MH chi-square, the standardized mean difference (SMD), and the standard error of the SMD for the constructed-response items. Table 8.2 and Table 8.3 present the DIF analysis groups and Table 8.3, respectively. The classification category of A indicates DIF is negligible, B indicates slight to moderate, and C indicates moderate to large. An item is flagged if its classification category is C in any of the DIF analysis groups.

Table 8.2 DIF Groups

Focal Group	Reference Group
Female	Male
African American	Not African American
Asian	Not Asian
Hispanic	Not Hispanic
White	Not White
IEP or 504 Plan	Not IEP or 504 Plan

Table 8.3 DIF Classification Rules for Items

Category	Rule
Dichotomous Items	
C	MH_{X^2} is significant and $\left|\hat{\Delta}_{MH}\right| \geq 1.5$
B	MH_{X^2} is significant and $1 \leq \left|\hat{\Delta}_{MH}\right| < 1.5$
A	MH_{X^2} is not significant or $\left|\hat{\Delta}_{MH}\right| < 1$
Polytomous Items	
Category	Rule
C	MH_{X^2} is significant and $|SMD|/\,|SD| \; > .25$
B	MH_{X^2} is significant and $0.17 < \; |SMD|/\,|SD| \; \leq .25$
A	MH_{X^2} is not significant or $|SMD|/\,|SD| \; \leq 0.17$

References

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

Center for Research on Evaluation, Standards, and Student Testing (CRESST) & Pacific Metrics. (2016). *ELPA21 standard setting technical report*.

Cohen, Jacob (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70* (4), 213–220.

Council of Chief State School Officers (CCSSO). (2014). *English Language Proficiency (ELP) Standards with correspondences to K–12 practices and Common Core state standards.* https://www.k12.wa.us/sites/default/files/public/migrantbilingual/pubdocs/elp/wa-elp-standards-k12.pdf

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

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