TECHNICAL REPORT

PART III—SCREENER ASSESSMENT

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

English Language Proficiency Assessment for the 21st Century (ELPA21)— Listening, Reading, Speaking, and Writing

Grades Pre-K–12

2024–2025 Test Administration

Submitted to
ELPA21

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

October 2025

Table of Contents

Chapter 1. Test Administration 1
1.1 Testing Window 1
1.2 Test Design 1
1.3 Test Administration Manual 7
1.3.1 Directions for Test Administration 7
1.3.2 Training/Practice Tests 7
1.3.3 Business Scoring Rules for the Screener Assessment 8
Chapter 2. 2024–2025 Summary 9
2.1 2024–2025 Student Participation 10
2.2 2024–2025 Student Scale Score and Performance-Level Summary 15
2.3 2024–2025 Testing Time for Online Screener Tests 21
Chapter 3. Reliability 22
3.1 Marginal Standard Error of Measurement 22
3.2 Marginal Reliability 22
3.3 Classification Accuracy and Consistency 23
3.4 Inter-Rater Reliability Analysis 27
Chapter 4. Validity 34
4.1 Relationship Between Screener and Summative Tests 34
4.1.1 Correlation between Screener and Summative Tests 35
4.1.2 Student Progress from Screener to Summative Assessments 35
Chapter 5. Reporting 36
References 37

List of Tables

Table 1.1 2024–2025 ELPA21 Screener Testing Windows by State 1

Table 1.2 Threshold Raw Scores Proceeding From Step 2 to Step 3 4

Table 1.3 Number of Items and Score Points by Domain and Grade Band—Online Screener 6

Table 1.4 Number of Items and Score Points by Domain and Grade Band—Paper Screener 6

Table 1.5 Number of Items and Score Points by Domain and Grade Band—Braille Screener 6

Table 2.1 Student Participation in 2023–2024 and 2024–2025 by State and Grade 12

Table 2.2 Student Participation in 2024–2025 Summative and Screener Assessments 13

Table 2.3 Scale Score Summary by Grade—Listening and Reading* 16

Table 2.4 Scale Score Summary by Grade—Speaking and Writing* 17

Table 2.5 Scale Score Summary by Grade—Comprehension and Overall* 18

Table 2.6 Percentage of Students by Performance Level—Listening and Reading* 19

Table 2.7 Percentage of Students by Performance Level—Speaking and Writing* 20

Table 2.8 Percentage of Students by Overall Proficiency Category 21

Table 3.1 Marginal Reliability by Score and Grade* 23

Table 3.2 Overall Classification Accuracy and Consistency for Domain Performance Levels 24

Table 3.3 Classification Accuracy for Each Cut Score by Domain and Grade* 25

Table 3.4 Classification Consistency for Each Cut Score by Domain and Grade* 26

Table 3.5 Classification Accuracy and Consistency for Overall Proficiency Categories 27

Table 3.6 Summary of AS-H1 Agreement by Grade Band 31

Table 3.7 Summary of H1-H2 Agreement by Grade Band 31

Table 3.8 Autoscore Performance Compared to Human-Human Agreement on Exact Agreement and Quadratic Weighted Kappa on the Held-out Validation Sample (Condition Codes Removed) 32

Table 3.9 Means and Standard Deviations of Final Scores for the Past Four Years 33

List of Figures

Figure 1.1 2024–2025 ELPA21 Screener Online Test Design 3

Figure 1.2 2024–2025 ELPA21 Screener Paper Test Design 5

Chapter 1. Test Administration

The English Language Proficiency Assessment for the 21st Century (ELPA21) screener assessment was administered to students in the following grade bands: kindergarten (K) and grades 1, 2–3, 4–5, 6–8, and 9–12. All states except Mississippi had data for screener pre-K students. For the screener assessment, as with the summative assessment, each form involves four domains (listening, reading, speaking, and writing). Students can be exempted from as many as three domains. The assessments do not have a time limit.

1.1 Testing Window

The 2024–2025 screener testing windows for the eight states discussed in this report are shown in Table 1.1.

Table 1.1 2024–2025 ELPA21 Screener Testing Windows by State

State	ELPA21 Screener
Arkansas	8/5/2024–7/18/2025
Iowa	8/1/2024–7/11/2025
Louisiana	8/1/2024–7/18/2025
Mississippi*	2/3/2025–5/30/2025
Nebraska	8/1/2024–7/11/2025
Ohio	8/6/2024–6/30/2025
Tennessee*	7/1/2024–06/20/2025
West Virginia	8/6/2024–6/20/2025

* New states starting from 2024-25.

1.2 Test Design

Each 2024–2025 screener assessment had one online form, one paper-pencil form, and one braille form. Pre-K students were permitted to take the kindergarten tests. The pre-K and kindergarten tests have the same items but different rules in determining a student's overall proficiency. Pre-K students are considered Proficient at the overall level with a performance level at 3 or above in each domain, while students in kindergarten and higher grades are considered Proficient with a performance level at 4 or above in each domain.

In Ohio, the ELPA21 pre-K test is called the Ohio English Language Proficiency Screener for the Beginning of Kindergarten (OELPS-BK), and the ELPA21 kindergarten test is called the Ohio English Language Proficiency Screener for Kindergarten (OELPS-K). The two testing windows for the OELPS-BK are August 6 through December 31, 2024 (for kindergarten students screened on or before Dec. 31, 2024) and between May 27 and June 30, 2025. The testing window for the OELPS-K is from January 1 to June 30, 2025 (for kindergarten students screened on or after Jan. 1, 2025). It should also be noted that students who are enrolled in kindergarten take the OELPS-K

and students who have completed preschool but will not start kindergarten until fall take the OELPS-BK.

Online Form

The online form has three steps. Step 1 consists of practice items, while Steps 2 and 3 consists of operational items. To allow for domain exemptions and because test administrator (TA) input is required (at the end of Step 1 and for the scoring of speaking items in Step 2), the three steps are administered as nine segments, with various possible routes through a subset of those segments, as shown in Figure 1.1. The content of the segments includes the following:

- Segment 1 (Step 1) includes nonscored practice items. At the end of Segment 1, the TA indicates whether the student should proceed to the operational items. If the TA doesn't answer the end-of-Step 1 survey question (which could be due to the fact that the student doesn't finish all Step 1 practice questions), each non-exempt domain will be assigned 'N', and the overall proficiency will be assigned 'N' with the label of 'Not Attempted'. If the TA determines that the student should not proceed, the student is directed to Segment 9, and the test ends. Additional details follow in Section 1.3. In this case, the student is assigned an overall classification of "Proficiency Not Demonstrated," and domain performance levels are assigned as "Performance Not Determined." If the TA indicates the tester should proceed, then the student is routed to Segment 2 (Step 2A) unless the student is exempted from the speaking domain, in which case the student is routed to Segment 7 (modified version of Step 2).

- Segment 2 (Step 2A) consists of speaking items scored on-the-fly. After the student responds to these items, the TA assigns a score to each item. After Segment 2, most students are routed to Segment 3 (Step 2B). However, students who are exempted from the listening, reading, and/or writing domains proceed to Segment 5 (modified version of Step 2B).

- Segment 3 (Step 2B) consists of machine-scored operational items from the listening, reading, and writing domains. After the student completes Segment 3, a summed score is computed from all the item scores in Step 2 (Segments 2 and 3). If this summed score is below a threshold score, the test ends. If the summed score meets or exceeds the threshold score, the test is routed to Segment 4 (Step 3) (refer to Table 1.2 for threshold information).

- Segment 4 (Step 3) includes operational items from all four domains.

- Segment 5 (Step 2B for students who are exempted from the listening, reading, and/or writing domains) consists of machine-scored operational items from all non-exempted domains. Upon completion of Segment 5, students proceed to Segment 6 (modified version of Step 3), regardless of score.

- Segment 6 (Step 3 for students who are exempted from the listening, reading, and/or writing domains) consists of items from all non-exempted domains.

- Segment 7 (Step 2 for students who are exempted from the speaking domain) consists of machine-scored operational items from the listening, reading, and writing domains. Students are administered the form for which their exempted domains are suppressed.

Upon completion of Segment 7, students proceed to Segment 8 (modified version of Step 3), regardless of score.

- Segment 8 (Step 3 for students who are exempted from the speaking domain) consists of items from non-exempted listening, reading, and/or writing domains.
- Segment 9 (Step 1) contains a survey item that allows TAs to describe why the student did not engage with the screener assessment.

Figure 1.1 2024–2025 ELPA21 Screener Online Test Design

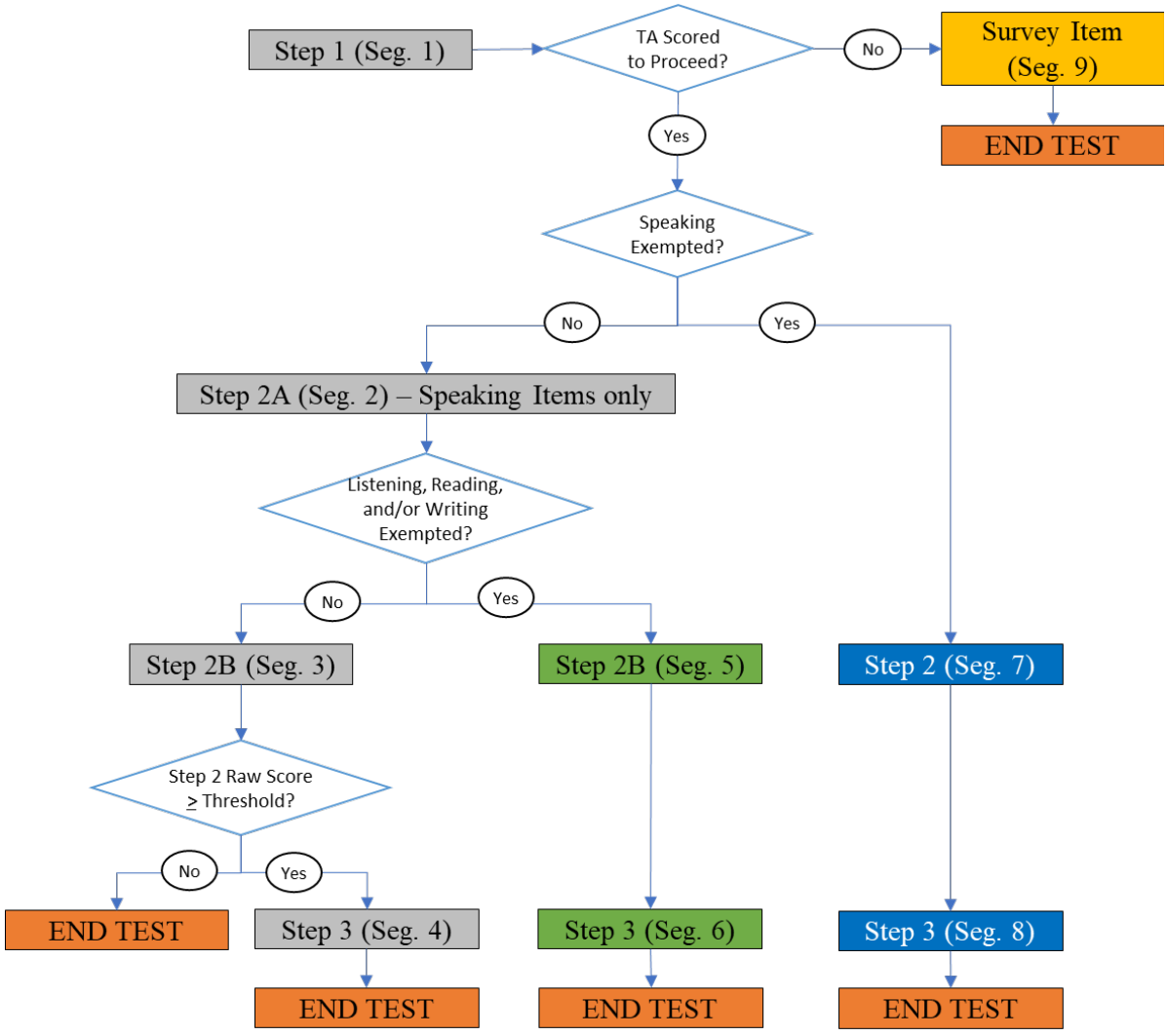


Table 1.2 Threshold Raw Scores Proceeding From Step 2 to Step 3

Grade/Grade Band	Threshold Score	Step 2 Max Score
Pre-K	20	26
K	23	26
1	24	27
2–3	25	28
4–5	26	31
6–8	28	33
9–12	26	29

Paper-Pencil Form

The paper-pencil form has five segments, as displayed in Figure 1.2.

- Segment 1 (Step 1) includes nonscored practice items. At the end of Segment 1, the TA indicates whether the student should proceed to the operational items. If the TA determines that the test should not proceed, the test ends. If the TA doesn't answer the end-of-Step 1 survey question (which could be due to the fact that the student doesn't finish all Step 1 practice questions), each non-exempt domain will be assigned 'N', and the overall proficiency will be assigned 'N' with the label of 'Not Attempted'.

- Segment 2 (Step 2) includes operational items from all four domains. After data entry is completed for Segment 2, a summed score is computed from all the item scores in this segment. If this summed score is below a threshold score, the test ends. If the raw score meets or exceeds the threshold score, the test is routed to Segment 3 (Step 3) (refer to Table 1.2 for threshold information).

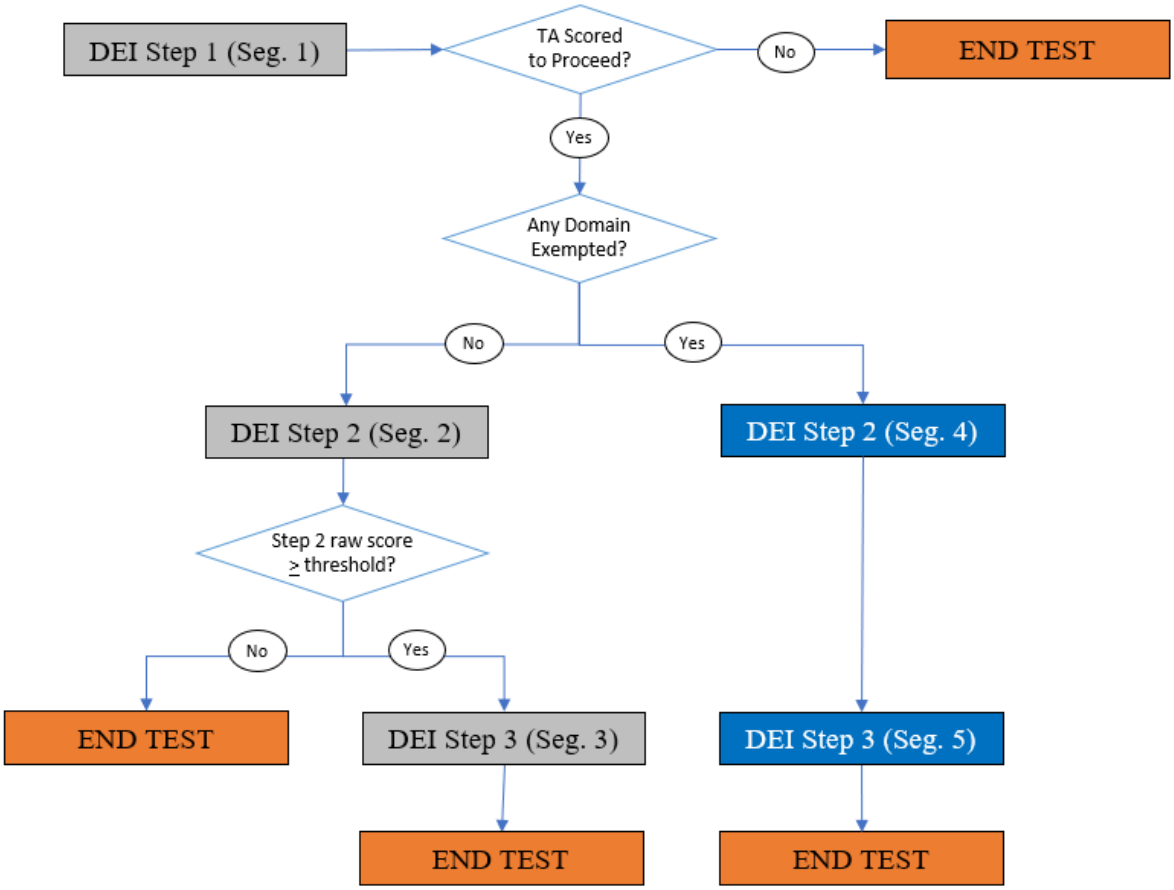
- Segment 3 (Step 3) includes operational items from all four domains.

- Segment 4 (Step 2 for students with any domain exemption) includes operational items from all non-exempted domains.

- Segment 5 (Step 3 for students with any domain exemption) include operational items from all non-exempted domains. Tests proceed from Segment 4 to Segment 5, regardless of score.

After test administration of the paper-pencil form, student responses are entered into the Data Entry Interface (DEI) from Cambium Assessment, Inc. (CAI) on the state testing portal for all ELPA21 domain tests. Practice test items are not entered into the DEI and are not scored.

Figure 1.2 2024–2025 ELPA21 Screener Paper Test Design



Braille Form

The braille form includes two segments. In Segment 1, the TA indicates whether the student should proceed to the operational items. If so, the student is routed to Segment 2, which contains operational items for all domains. If the TA indicates the student should not proceed, then the test ends. If the TA doesn't answer the end-of-Step 1 survey question (which could be due to the fact that the student doesn't finish all Step 1 practice questions), each non-exempt domain will be assigned 'N', and the overall proficiency will be assigned 'N' with the label of 'Not Attempted'.

Summary of Forms

The non-domain-exempted form summaries of the screener tests are listed in Table 1.3–Table 1.5. for the online, paper-pencil, and braille forms, respectively. They present the number and score points of operational items by domain and grade band.

Table 1.3 Number of Items and Score Points by Domain and Grade Band—Online Screener

	Grade/Grade Band											
	Pre-K/K		1		2–3		4–5		6–8		9–12	
Domain	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*
Listening	13	13	11	11	11	11	10	10	17	18	14	17
Reading	9	9	13	13	11	13	21	23	13	13	16	17
Speaking	6	14	6	15	6	14	7	21	9	27	9	27
Writing	10	10	11	11	14	17	9	21	7	23	6	20
Total	38	46	41	50	42	55	47	75	46	81	45	81

Table 1.4 Number of Items and Score Points by Domain and Grade Band—Paper Screener

	Grade/Grade Band											
	Pre-K/K		1		2–3		4–5		6–8		9–12	
Domain	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*
Listening	13	13	11	11	11	11	10	10	17	18	14	17
Reading	9	9	13	13	11	13	21	23	13	13	16	17
Speaking	6	14	6	15	6	14	7	21	9	27	9	27
Writing	10	10	11	11	14	17	9	21	7	23	6	20
Total	38	46	41	50	42	55	47	75	46	81	45	81

Table 1.5 Number of Items and Score Points by Domain and Grade Band—Braille Screener

	Grade/Grade Band											
	Pre-K/K		1		2–3		4–5		6–8		9–12	
Domain	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*
Listening	9	9	9	9	10	10	11	11	11	12	10	13
Reading	11	11	9	9	8	10	13	15	11	11	12	13
Speaking	6	14	6	16	6	16	8	29	8	25	8	25
Writing	8	8	8	8	10	13	8	16	7	23	8	26
Total	34	42	32	42	34	49	40	71	37	71	38	77

1.3 Test Administration Manual

1.3.1 Directions for Test Administration

For the 2024–2025 administration, a test administration manual (TAM) was developed for each state. The TAM guides Test Administrators (TAs) in test administration. The TAM for the screener assessment included the following key topics:

- Overview of the ELPA21 screener assessment
- TA qualifications
- Preliminary planning
- Required materials
- Administrative considerations
- Student preparation/guidance in Step 1
- Administrative guidance in Step 2 and Step 3
- Test security instructions in each of the three steps
- Contact information for user support

1.3.2 Training/Practice Tests

To help TAs and students familiarize themselves with the online registration and Test Delivery System (TDS), training/practice tests (Step 1 in screener tests) were provided before and during the testing windows. Training/practice tests can be accessed through a nonsecure browser or a secure browser. For screener assessments, the tests become secure automatically when students proceed to Step 2.

The screener training/practice tests have two components: one for TAs to create and manage the training/practice test sessions and a second for students to take an actual training/practice test.

The *Practice Test Administration* site introduces TAs to

- logging in;
- starting a test session;
- providing the session ID to the students signing in to the test session;
- monitoring students' progress throughout their test sessions; and
- stopping the test session.

The *Practice Tests* site introduces students to

- signing in;
- verifying student information;
- selecting a test;
- waiting for the TA to check the test settings and approve participation;
- preparing to begin the test (adjusting the audio level, checking the microphone for recording speaking responses, and reviewing test instructions);
- taking the test; and
- submitting the test.

1.3.3 Business Scoring Rules for the Screener Assessment

Business rules and instructions applied to the 2024–2025 screener assessment include the following:

- All pending and expired test records in Step 2 should be scored.
- If a single item in Step 2 is attempted, all domains without domain exemptions are considered attempted, and all non-attempted items in Step 2 should be given a score of zero.
- If a student's test is stopped by the automatic stopping rule after Step 2, items in Step 3 should be treated as "not presented." If the student's test continues to Step 3, all items in Step 3 that the student does not respond to should be scored as zero.
- If a student has a domain exemption for a domain, the domain is reported as exempt if it is not attempted.
    - For online tests, any domain exemptions must be entered into the Test Information Distribution Engine (TIDE) before the student starts the test. Students taking the online screener will be presented with items in non-exempt domains only.
    - For paper-pencil tests, TAs are told which items to not administer if the student has any domain exemptions. However, if a student is exempt from a domain but responses to any items in the domain are entered into the DEI, the domain will be scored as though the student was not exempt.
- ELPA21 states decide whether to use the pre-K test on an individual basis.
- For the Ohio screener administration, handscored items are scored by local TAs.
- Tests in which the TA indicates that the student will not continue after the Step 1 practice items will be scored as follows:
    - Each domain will be scored zero. The score of zero will receive a label of "Performance Not Determined."
    - Proficiency status will be scored as "D" and reported as "Proficiency Not Demonstrated."

Chapter 2. 2024–2025 Summary

The 2024–2025 screener test results based on data pooling all eight states are presented in this chapter and in Sections 1–14 of the Appendix for Pooled Analysis—2024–2025 Summary Screener. Testing results based on individual states are presented in each state-specific Appendix. The figures and tables included in each section of the pooled appendix for the screener assessment are as follows:

- Section 1. Screener Assessment—Student Participation
    - Table S1.1 displays the number and percentage of students for each testing mode (braille, paper-pencil, and online) for each grade level (pre-K–12).
    - Table S1.2 lists the number and percentage of students taking each test by subgroups, including grade, gender, and ethnicity. In the state-specific appendices, additional subgroups are included, such as primary disabilities, migrant, special education (SPED), Title I, or Section 504 Plan. Subgroups may vary across states.
- Section 2. Screener Assessment—Raw Score Summary
    - Tables S2.1–S2.14 present the number of students, minimum, maximum, average, and standard deviation of raw scores by performance level and by the ending step for the four domain tests and the overall test in each grade.
    - Note that the multidimensional item response theory (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. For the screener, we also must consider whether a student advanced to Step 3 when interpreting raw scores.
- Section 3. Screener Assessment—Raw Score Distributions
    - Figures S3.1–S3.70 present the frequency distribution of raw scores by performance level for each domain test and the frequency distribution of overall raw scores by proficiency category in each grade.
- Section 4. Screener Assessment—Scale Score Summary
    - Tables S4.1–S4.14 present the number of students, minimum, average, maximum, and standard deviation of domain scale scores, overall scale scores, and comprehension scale scores, at the overall level and by subgroups of gender and ethnicity for each grade level. In the state-specific appendices, additional subgroups are included such as primary disabilities, migrant, special education (SPED), Title I, or Section 504 Plan. Subgroups may vary across states.
    - Table S4.15 summarizes the number and percentage of students who were exempted from or did not attempt a particular domain.

- Section 5. Screener Assessment—Percentage of Students by Domain Performance Level
  - Figure S5.1 plots the percentage of students by performance level in each domain test.
  - Tables S5.1–S5.14 present the total number of students taking each domain test and the percentage of students in each performance level at the overall level and by subgroups of gender and ethnicity for each grade level.
- Section 6. Screener Assessment—Percentage of Students by Overall Proficiency Category
  - Figure S6.1 plots the percentage of students by overall proficiency category.
  - Tables S6.1–S6.14 present the total number of students who are categorized in each of the proficiency categories: Emerging, Progressing, Proficient, and Proficiency Not Demonstrated, at the overall level and by subgroups of gender and ethnicity.
- Section 7. Screener Assessment—Testing Time
  - Table S7.1 shows the testing time by each step in each grade/grade band.

2.1 2024–2025 Student Participation

Table 2.1 presents the overall student participation for each state in the two most recent years. A total of 94,394 students from all eight states participated in the 2024–2025 screener assessment, an increase of 22% from the previous year. Ohio had the highest number of students, followed by Tennessee, Iowa, and Arkansas. Most students were from pre-K and kindergarten.

Table 2.2 presents the number of students who took the summative assessment, the screener assessment, and both assessments. It shows that kindergarten students represented the highest percentage of students taking both the screener and summative assessments in the 2024–2025 school year.

Section S1.1 of the pooled appendix for the screener assessment presents student participation in each mode. In the eight ELPA21 states combined, the most frequent mode of administration was online (99.97%), followed by paper-pencil (0.03%). One student took the braille form in the 2024-2025 test administration.

Section S1.2 of the appendix shows student participation by subgroups. For the pooled analysis from pre-K to grade 12, the number of test-takers generally decreased as the grade level increased, although this decrease was not monotonic. There were more male students than female students tested overall, and more male students tested than female students in every grade band except grade 1. In each test, the greatest number of participating students were in the group of Hispanic or Latino (55.6%–61.0%), followed by Asian students (9.9%–13.9%), and white students (8.3%–12.3%).

Table 2.1 Student Participation in 2023–2024 and 2024–2025 by State and Grade

Grade	Arkansas		Iowa		Louisiana		Mississippi	Nebraska		Ohio		Tennessee	West Virginia		Total		
	2023–24	2024–25	2023–24	2024–25	2023–24	2024–25	2024–25	2023–24	2024–25	2023–24	2024–25	2024–25	2023–24	2023–24	2023–24	2024–25	Diff
Pre-K	≥3,440	≥3,560	≥5,130	≥5,200	≥3,850	≥3,820	N/A	≥3,620	≥3,800	≥11,620	≥11,760	≥5,830	≥290	≥290	≥27,970	≥34,300	≥6,330
K	≥1,580	≥1,550	≥290	≥210	≥390	≥270	≥50	≥190	≥120	≥1,050	≥790	≥5,420	≥50	≥50	≥3,570	≥8,480	≥4,900
1	≥770	≥650	≥790	≥680	≥980	≥730	≥40	≥540	≥410	≥2,480	≥2,330	≥1,530	≥90	≥90	≥5,670	≥6,480	≥810
2	≥690	≥560	≥720	≥590	≥810	≥590	≥30	≥480	≥360	≥2,000	≥1,770	≥1,290	≥70	≥70	≥4,790	≥5,290	≥490
3	≥620	≥490	≥630	≥540	≥800	≥530	≥30	≥430	≥390	≥1,890	≥1,670	≥1,210	≥70	≥90	≥4,450	≥4,980	≥520
4	≥600	≥470	≥580	≥490	≥660	≥450	≥30	≥440	≥320	≥1,740	≥1,540	≥1,100	≥60	≥60	≥4,100	≥4,500	≥390
5	≥530	≥430	≥530	≥460	≥600	≥440	≥20	≥360	≥340	≥1,630	≥1,380	≥1,050	≥50	≥70	≥3,730	≥4,230	≥500
6	≥530	≥430	≥550	≥410	≥610	≥410	≥20	≥390	≥330	≥1,530	≥1,400	≥1,010	≥60	≥50	≥3,700	≥4,090	≥380
7	≥590	≥380	≥540	≥440	≥600	≥410	≥30	≥340	≥300	≥1,490	≥1,320	≥970	≥60	≥60	≥3,640	≥3,940	≥290
8	≥530	≥380	≥460	≥410	≥600	≥360	≥30	≥310	≥270	≥1,380	≥1,120	≥880	≥60	≥50	≥3,360	≥3,520	≥160
9	≥800	≥670	≥930	≥930	≥1,220	≥850	≥40	≥750	≥630	≥2,560	≥2,270	≥1,530	≥70	≥70	≥6,350	≥7,010	≥660
10	≥480	≥410	≥590	≥370	≥280	≥240	≥20	≥320	≥280	≥1,150	≥1,030	≥850	≥70	≥60	≥2,920	≥3,310	≥390
11	≥440	≥390	≥320	≥310	≥160	≥140	≥10	≥200	≥160	≥860	≥880	≥710	≥80	≥110	≥2,070	≥2,740	≥660
12	≥190	≥210	≥140	≥190	≥50	≥60	N/A	≥170	≥160	≥390	≥380	≥380	≥30	≥60	≥990	≥1,470	≥470
Total	≥11,850	≥10,640	≥12,260	≥11,280	≥11,660	≥9,350	≥390	≥8,580	≥7,940	≥31,840	≥29,710	≥23,830	≥1,170	≥1,220	≥77,380	≥94,390	≥17,010

Table 2.2 Student Participation in 2024–2025 Summative and Screener Assessments

State	Grade/Grade Band	N Summative	N Screener	N Both
Arkansas	K	≥4,630	≥5,110	≥3,820
	1	≥4,510	≥650	≥510
	2-3	≥7,690	≥1,060	≥780
	4-5	≥6,210	≥900	≥640
	6-8	≥8,260	≥1,200	≥930
	9-12	≥10,960	≥1,690	≥1,230
Iowa	K	≥5,080	≥5,410	≥3,700
	1	≥4,770	≥680	≥530
	2-3	≥7,710	≥1,130	≥840
	4-5	≥5,840	≥950	≥690
	6-8	≥7,220	≥1,270	≥970
	9-12	≥9,200	≥1,810	≥1,380
Louisiana	K	≥4,000	≥4,100	≥2,850
	1	≥4,300	≥730	≥580
	2-3	≥7,710	≥1,120	≥850
	4-5	≥5,410	≥890	≥670
	6-8	≥6,800	≥1,180	≥950
	9-12	≥7,850	≥1,290	≥1,020
Mississippi	K	≥1,980	≥50	≥40
	1	≥2,060	≥40	≥30
	2-3	≥3,760	≥60	≥40
	4-5	≥2,890	≥60	≥40
	6-8	≥2,350	≥80	≥60
	9-12	≥2,240	≥80	≥60

State	Grade/Grade Band	N Summative	N Screener	N Both
Nebraska	K	≥4,170	≥3,930	≥2,500
	1	≥4,140	≥410	≥300
	2-3	≥6,750	≥760	≥570
	4-5	≥4,940	≥670	≥480
	6-8	≥5,380	≥900	≥690
	9-12	≥7,030	≥1,240	≥890
Ohio	K	≥11,600	≥12,550	≥10,460
	1	≥11,720	≥2,330	≥1,830
	2-3	≥19,130	≥3,440	≥2,620
	4-5	≥13,800	≥2,930	≥2,130
	6-8	≥17,130	≥3,850	≥2,880
	9-12	≥19,620	≥4,580	≥3,440
Tennessee	K	≥10,280	≥11,250	≥9,800
	1	≥9,430	≥1,530	≥1,200
	2-3	≥18,420	≥2,510	≥1,840
	4-5	≥13,530	≥2,160	≥1,550
	6-8	≥15,870	≥2,870	≥2,170
	9-12	≥17,420	≥3,480	≥2,550
West Virginia	K	≥290	≥340	≥270
	1	≥320	≥90	≥70
	2-3	≥440	≥160	≥100
	4-5	≥400	≥140	≥90
	6-8	≥500	≥170	≥110
	9-12	≥740	≥320	≥190

2.2 2024–2025 Student Scale Score and Performance-Level Summary

Table 2.3–Table 2.5 present the domain, comprehension, and overall scale score summary by grade level. The ELPA21 tests are not linked vertically across all grades. Scale scores can be compared only for students within the same grade band (grades 2–3, 4–5, 6–8, and 9–12). Scale score summary by subgroup for each grade is also presented in Section 4 of the pooled appendix for the screener assessment.

Table 2.6 and Table 2.7 present the total number of test takers and percentage of students by performance level in each domain and grade. The results indicate that performance Level 1 is the most frequent level achieved in all domains and all grade levels except for listening in pre-K where the most frequent level achieved is Level 3. All four domains displayed similar trends with respect to students with "Performance Not Determined", as indicated by the "0" column for each domain. The rates of "Performance Not Determined" were low for grades pre-K–5, higher for grades 6–10, and back down for grades 11–12. Disaggregated results by gender and ethnicity are provided in Section 5 of the pooled appendix for the screener assessment.

Table 2.8 and Figure S6.1 in the pooled appendix for the screener assessment display the percentage of students achieving each overall proficiency category by grade. Beginning in 2021–2022, all states recognize pre-K (or before kindergarten [BK] for Ohio) students as Proficient if they attain performance Level 3 or higher in each domain, rather than the previous requirement of Level 4. Students in kindergarten and higher grades must achieve performance Level 4 or above in each domain to be considered Proficient.

The results show that most students fall into the Emerging or Progressing categories. The percentages of Proficient students are above 10% in grades 4 to 6 and 11 to 12. Section 6 of the pooled appendix for the screener assessment displays the overall proficiency category for each grade by gender and ethnicity.

*Table 2.3 Scale Score Summary by Grade—Listening and Reading**

Grade	Listening					Reading				
	N	Min	Mean	Max	SD	N	Min	Mean	Max	SD
Pre-K	≥32,950	298	494.6	714	71.3	≥32,950	304	492.6	708	69.9
K	≥8,170	314	487.1	714	75.4	≥8,170	318	486.0	708	73.5
1	≥6,180	288	459.5	678	88.1	≥6,180	286	449.0	704	84.6
2	≥5,070	286	450.2	710	78.7	≥5,070	278	436.5	734	85.5
3	≥4,800	286	466.0	710	90.7	≥4,800	278	458.1	734	99.3
4	≥4,280	270	436.7	768	103.2	≥4,280	270	442.4	781	100.5
5	≥4,070	270	449.0	778	111.7	≥4,070	270	456.7	795	108.3
6	≥3,720	279	445.7	738	95.6	≥3,720	296	453.7	733	92.6
7	≥3,640	279	446.8	738	97.2	≥3,640	296	457.3	733	94.4
8	≥3,230	279	452.4	738	100.0	≥3,230	296	463.3	733	98.0
9	≥6,210	302	445.9	731	91.1	≥6,210	312	452.8	733	87.4
10	≥3,090	302	481.9	731	99.7	≥3,090	312	488.2	733	95.7
11	≥2,630	302	517.5	731	103.8	≥2,630	312	522.9	733	100.3
12	≥1,420	302	513.1	731	102.8	≥1,420	312	518.6	733	100.5

* Domains with Exempt or Not Attempted are excluded.
* Scale scores cannot be compared across grade bands.

*Table 2.4 Scale Score Summary by Grade—Speaking and Writing**

Grade	Speaking					Writing				
	N	Min	Mean	Max	SD	N	Min	Mean	Max	SD
Pre-K	≥32,930	339	479.9	711	90.3	≥32,950	326	471.0	684	57.7
K	≥8,170	339	466.4	711	90.6	≥8,170	347	477.1	684	61.2
1	≥6,180	310	435.9	669	96.6	≥6,180	283	444.1	698	85.6
2	≥5,070	292	421.2	703	99.9	≥5,070	276	431.9	737	86.6
3	≥4,800	292	437.5	703	113.2	≥4,800	276	454.9	737	100.4
4	≥4,280	270	428.4	780	129.1	≥4,280	268	435.3	787	106.9
5	≥4,070	270	439.3	786	137.0	≥4,070	268	450.4	797	114.9
6	≥3,720	296	443.6	732	111.1	≥3,720	281	445.9	741	97.0
7	≥3,640	296	443.6	732	112.6	≥3,640	281	448.5	741	99.3
8	≥3,230	296	450.5	732	112.6	≥3,230	281	454.6	741	102.1
9	≥6,210	334	460.0	722	98.5	≥6,210	318	453.9	732	85.7
10	≥3,090	334	494.1	722	103.9	≥3,090	318	487.6	732	93.9
11	≥2,630	334	530.8	722	104.8	≥2,630	318	519.4	732	97.5
12	≥1,420	334	525.1	722	104.4	≥1,420	318	515.0	732	95.8

* Domains with Exempt or Not Attempted are excluded.
* Scale scores cannot be compared across grade bands.

*Table 2.5 Scale Score Summary by Grade—Comprehension and Overall**

Grade	Comprehension					Overall				
	N	Min	Mean	Max	SD	N	Min	Mean	Max	SD
Pre-K	≥32,950	3777	5179.1	6375	521.8	≥32,950	3609	4938.2	6763	556.4
K	≥8,180	3978	5123.0	6375	550.0	≥8,180	3646	4894.1	6763	584.0
1	≥6,180	3785	4879.0	6387	588.8	≥6,180	3364	4643.3	6629	703.7
2	≥5,070	3756	4785.8	6439	589.7	≥5,070	3326	4550.1	6880	698.1
3	≥4,800	3756	4905.3	6439	666.8	≥4,800	3326	4707.4	6880	810.2
4	≥4,280	3649	4730.3	6700	665.3	≥4,280	3237	4567.5	7323	880.7
5	≥4,070	3649	4814.4	6700	715.4	≥4,070	3237	4675.0	7401	945.3
6	≥3,720	3803	4800.0	6476	637.8	≥3,720	3388	4665.4	6974	787.1
7	≥3,640	3803	4813.0	6476	646.6	≥3,640	3388	4680.4	6974	801.1
8	≥3,230	3803	4852.5	6476	673.2	≥3,230	3388	4729.7	6974	818.9
9	≥6,210	3818	4774.7	6522	624.0	≥6,210	3628	4720.4	6922	716.1
10	≥3,090	3818	5025.7	6522	687.2	≥3,090	3628	5002.0	6922	778.2
11	≥2,630	3818	5281.6	6522	729.0	≥2,630	3628	5282.3	6922	803.2
12	≥1,420	3818	5256.9	6522	735.5	≥1,420	3628	5243.2	6922	797.6

* Scale scores cannot be compared across grade bands.

*Table 2.6 Percentage of Students by Performance Level—Listening and Reading**

Grade	Listening							Reading						
	N	0	1	2	3	4	5	N	0	1	2	3	4	5
Pre-K	≥34,260	≤5.0	33.5	17.8	40.0	≤5.0	≤5.0	≥34,260	≤5.0	37.7	19.9	31.3	≤5.0	≤5.0
K	≥8,460	≤5.0	40.6	16.1	35.5	≤5.0	≤5.0	≥8,460	≤5.0	44.5	17.3	28.4	≤5.0	≤5.0
1	≥6,470	≤5.0	44.8	9.1	25.3	7.7	8.7	≥6,460	≤5.0	68.7	8.2	9.6	≤5.0	≤5.0
2	≥5,280	≤5.0	36.6	16.5	21.5	11.7	9.9	≥5,280	≤5.0	65.7	6.1	13.5	≤5.0	6.3
3	≥4,960	≤5.0	33.5	19.3	20.9	12.0	11.2	≥4,960	≤5.0	65.5	10.1	10.7	≤5.0	5.9
4	≥4,470	≤5.0	46.6	9.2	12.4	13.6	13.8	≥4,480	≤5.0	61.8	7.6	10.4	5.8	10.1
5	≥4,220	≤5.0	49.2	9.5	8.2	13.8	15.6	≥4,220	≤5.0	62.7	7.2	11.3	≤5.0	10.0
6	≥4,050	8.1	46.4	9.4	10.7	11.1	14.3	≥4,050	8.1	59.7	5.4	12.5	5.6	8.5
7	≥3,910	7.0	56.1	8.5	11.1	6.6	10.7	≥3,910	6.9	65.2	6.9	11.0	≤5.0	5.6
8	≥3,500	7.7	54.8	8.3	12.2	7.5	9.5	≥3,500	7.7	64.4	6.9	14.4	≤5.0	≤5.0
9	≥6,980	11.0	58.4	7.7	11.3	≤5.0	6.7	≥6,980	11.0	64.8	8.6	10.3	≤5.0	≤5.0
10	≥3,290	6.2	45.3	10.5	16.8	7.9	13.2	≥3,290	6.2	52.8	13.6	17.0	5.2	5.2
11	≥2,720	≤5.0	32.4	10.0	19.3	12.2	22.5	≥2,730	≤5.0	39.9	13.5	25.2	8.7	9.2
12	≥1,460	≤5.0	33.5	10.0	20.3	12.6	21.1	≥1,460	≤5.0	40.4	15.1	24.9	8.5	8.6
Total	≥94,090	≤5.0	40.9	13.8	26.4	6.5	7.5	≥94,090	≤5.0	51.5	13.4	21	≤5.0	≤5.0

* Level 0: Performance Not Determined.
* Domains with Exempt or Not Attempted are excluded.

*Table 2.7 Percentage of Students by Performance Level—Speaking and Writing**

Grade	Speaking							Writing						
	N	0	1	2	3	4	5	N	0	1	2	3	4	5
Pre-K	≥34,230	≤5.0	51.6	15.9	17.9	≤5.0	6.5	≥34,250	≤5.0	70.4	19.5	5.1	≤5.0	≤5.0
K	≥8,450	≤5.0	57.9	13.2	15.4	5.8	≤5.0	≥8,460	≤5.0	65.8	22.2	6.8	≤5.0	≤5.0
1	≥6,460	≤5.0	74.0	12.5	≤5.0	≤5.0	≤5.0	≥6,460	≤5.0	73.6	8.4	7.7	≤5.0	≤5.0
2	≥5,280	≤5.0	69.2	10.9	≤5.0	≤5.0	7.4	≥5,280	≤5.0	65.1	9.4	10.9	≤5.0	6.5
3	≥4,960	≤5.0	65.8	8.7	5.7	6.7	9.8	≥4,960	≤5.0	67.4	8.3	9.9	≤5.0	6.7
4	≥4,470	≤5.0	59.3	7.7	8.4	5.2	15.0	≥4,480	≤5.0	59.1	7.7	15.0	≤5.0	9.8
5	≥4,220	≤5.0	61.2	7.1	7.4	≤5.0	16.2	≥4,220	≤5.0	55.7	9.5	17.0	≤5.0	10.0
6	≥4,050	8.1	55.9	7.0	12.4	5.7	10.9	≥4,050	8.1	51.9	10.3	15.3	≤5.0	9.8
7	≥3,910	7.0	59.5	8.7	10.7	≤5.0	10.5	≥3,910	6.9	64.5	6.6	10.4	≤5.0	7.3
8	≥3,500	7.7	57.1	9.5	12.2	≤5.0	9.3	≥3,500	7.7	63.1	7.3	13.1	≤5.0	5.6
9	≥6,980	11.0	56.7	10.5	11.3	≤5.0	7.4	≥6,980	11.0	64.8	8.2	9.5	≤5.0	≤5.0
10	≥3,290	6.2	46.5	12.6	15.3	5.6	13.7	≥3,290	6.2	53.2	11.9	16.3	≤5.0	8.0
11	≥2,720	≤5.0	33.7	12.8	18.0	8.2	23.7	≥2,730	≤5.0	39.8	13.5	22.5	7.5	13.1
12	≥1,460	≤5.0	35.1	12.3	19.2	9.8	20.9	≥1,460	≤5.0	41.4	13.8	23.7	7.2	11.3
Total	≥94,050	≤5.0	56.4	12.4	13	≤5.0	8.7	≥94,090	≤5.0	64.9	14.1	9.5	≤5.0	≤5.0

* Level 0: Performance Not Determined.
* Domains with Exempt or Not Attempted are excluded.

Table 2.8 Percentage of Students by Overall Proficiency Category

Grade	N	Proficiency Not Demonstrated	Emerging	Progressing	Proficient
Pre-K	≥34,270	≤5.0	47.3	43.9	≤5.0
K	≥8,460	≤5.0	53.4	41.5	≤5.0
1	≥6,470	≤5.0	53.5	37.8	≤5.0
2	≥5,280	≤5.0	52.8	34.9	8.4
3	≥4,960	≤5.0	52.7	34.9	9.1
4	≥4,480	≤5.0	55.3	27.2	13.1
5	≥4,230	≤5.0	58.3	25.0	12.9
6	≥4,050	8.1	54.8	25.9	11.2
7	≥3,910	6.9	63.3	21.4	8.4
8	≥3,500	7.7	61.7	24.3	6.3
9	≥6,980	11.0	63.3	21.1	≤5.0
10	≥3,290	6.2	53.2	31.2	9.3
11	≥2,730	≤5.0	39.8	40.6	16.0
12	≥1,460	≤5.0	40.8	42.1	14.5
Total	≥94,110	≤5.0	52.3	36.0	6.9

2.3 2024–2025 Testing Time for Online Screener Tests

In the 2024–2025 online screener tests, students without domain exemptions advanced to Segments 2 and 3 (Step 2) and progressed to Segment 4 (Step 3) if their raw scores met or exceeded the threshold score for Step 2 (refer to Table 1.2). Consequently, students who completed Step 3 answered more items than those who stopped at Step 2. Table S7.1 in the pooled appendix for the screener assessment summarizes testing times based on the last step for each grade. Students who had any non-attempted or exempted domains or received Proficiency Not Demonstrated designation were excluded from this analysis. As anticipated, students who concluded the test at Step 3 had longer testing durations than those who finished at Step 2. In addition, upper-grade tests had longer testing times than lower-grade tests due to their increased length and complexity.

Chapter 3. Reliability

In the same procedure as the summative assessment described in Part I, Chapter 3, of this technical report, the reliability for screener tests is assessed using

- marginal standard error of measurement (MSEM);
- marginal reliability;
- conditional standard error of measurement (CSEM);
- classification accuracy (CA) and classification consistency (CC); and
- inter-rater analysis.

Results for individual states are provided in the state-specific appendices. Results based on the combined data across states are illustrated in the following sections of the pooled appendix for the screener assessment.

- Section 8. Screener Assessment—Marginal Reliability
    - Figure S8.1 shows the ratio of MSEM to the standard deviation of scale scores at the test level by grade.
    - Figure S8.2 presents the marginal reliability at the test level by grade.
- Section 9. Screener Assessment—Conditional Standard Error of Measurement (CSEM)
    - Figures S9.1–S9.14 display the CSEM of domain scale scores, overall scale scores, and comprehension scale scores in each grade. Tests ending at Step 2 (represented by black dots) are shorter than tests ending at Step 3 (represented by red dots), and thus, their CSEMs are expected to be higher.
- Section 10. Screener Assessment—Classification Accuracy and Consistency
    - Figure S10.1 shows the CA for each domain test.
    - Figure S10.2 shows the CC for each domain test.
    - Figure S10.3 presents the CA and CC for the overall proficiency.
- Section 11. Screener Assessment—Inter-Rater Analysis

3.1 Marginal Standard Error of Measurement

As described in Part I of this technical report, the MSEM serves as a method for evaluating score reliability. The ratio of the MSEM to the standard deviation of scale scores also indicates measurement errors. These ratios are displayed in Figure S8.1 in the pooled appendix for the screener assessment.

3.2 Marginal Reliability

The marginal reliability for the pooled analysis is shown in Table 3.1 and illustrated in Figure S8.2 in the pooled appendix for the screener assessment. Pre-K and kindergarten exhibit lower marginal reliability compared to other grades. Additionally, the reliability of comprehension scores tends to

be lower than that of other score types in the lower grade levels. Section 9 of the pooled appendix for the screener assessment displays CSEM plots by test in each grade.

*Table 3.1 Marginal Reliability by Score and Grade**

Grade	N	Listening	Reading	Speaking	Writing	Comprehension	Overall
Pre-K	≥32,930	.81	.77	.82	.68	.74	.77
K	≥8,170	.82	.78	.82	.69	.76	.79
1	≥6,180	.83	.86	.85	.86	.73	.86
2	≥5,070	.84	.90	.87	.89	.79	.90
3	≥4,800	.86	.92	.90	.91	.82	.92
4	≥4,280	.90	.92	.91	.91	.86	.93
5	≥4,070	.91	.93	.92	.93	.87	.94
6	≥3,720	.92	.90	.91	.90	.87	.92
7	≥3,640	.93	.90	.91	.90	.87	.92
8	≥3,230	.93	.91	.91	.91	.88	.93
9	≥6,210	.92	.90	.89	.84	.88	.90
10	≥3,090	.93	.91	.91	.87	.89	.92
11	≥2,630	.92	.91	.91	.88	.89	.92
12	≥1,420	.92	.91	.91	.88	.89	.92

* Scores from domain tests marked as Exempt or Not Attempted are excluded. Step 2 and Step 3 are combined for this analysis.

3.3 Classification Accuracy and Consistency

Table 3.2 presents overall classification accuracy (CA) and classification consistency (CC) by domain and grade. The paper-pencil and braille forms were excluded. CC rates can be lower than CA rates because consistency was based on two tests with measurement errors, while accuracy was based on one test with a measurement error and the true score.

The CA and CC rates for each cut score are presented in Table 3.3 and Table 3.4 below, as well as Figure S10.1 and Figure S10.2 in the pooled appendix for the screener assessment. Across the four performance cut scores, the CA indices were all above 0.83, denoting that the degree to which we can reliably differentiate students between adjacent performance levels is 0.83 or above. In terms of CC, the indices were all above 0.76 in all cut scores and all grades.

Table 3.5 and Figure S10.3 in the pooled appendix for the screener assessment display the CA and CC for overall proficiency categories. The results show that the overall CA and CC indices were 0.80 or higher in all grades. Both CA and CC indices for the cut between Emerging and Progressing were lower than those for the cut between Progressing and Proficient in pre-K to grade 10 and are comparable with those between Progressing and Proficient in the other grades.

Table 3.2 Overall Classification Accuracy and Consistency for Domain Performance Levels

Grade	Accuracy				Consistency			
	Listening	Reading	Speaking	Writing	Listening	Reading	Speaking	Writing
Pre-K	.71	.65	.71	.77	.61	.55	.64	.69
K	.72	.66	.72	.73	.63	.57	.66	.65
1	.68	.81	.82	.85	.59	.75	.78	.80
2	.64	.83	.80	.82	.55	.77	.76	.77
3	.66	.83	.79	.84	.56	.78	.75	.79
4	.74	.83	.81	.83	.66	.78	.77	.78
5	.77	.84	.83	.83	.69	.79	.79	.77
6	.78	.83	.80	.78	.70	.76	.75	.71
7	.81	.85	.82	.84	.75	.80	.77	.80
8	.81	.86	.81	.85	.75	.81	.76	.80
9	.83	.87	.81	.83	.78	.82	.75	.77
10	.78	.80	.76	.76	.71	.74	.69	.69
11	.74	.74	.72	.71	.66	.66	.64	.63
12	.73	.73	.70	.69	.64	.65	.61	.61

* Scores from domain tests marked as Exempt or Not Attempted are excluded.

*Table 3.3 Classification Accuracy for Each Cut Score by Domain and Grade**

Grade	Listening				Reading				Speaking				Writing			
	Cut 1	Cut 2	Cut 3	Cut 4	Cut 1	Cut 2	Cut 3	Cut 4	Cut 1	Cut 2	Cut 3	Cut 4	Cut 1	Cut 2	Cut 3	Cut 4
Pre-K	.89	.86	.95	.97	.87	.85	.93	.96	.89	.89	.93	.95	.83	.95	.99	.99
K	.91	.88	.94	.96	.88	.87	.91	.95	.90	.91	.93	.95	.84	.93	.97	.98
1	.90	.90	.90	.93	.92	.94	.96	.97	.91	.93	.94	.96	.94	.95	.96	.97
2	.87	.90	.91	.94	.94	.94	.96	.97	.92	.93	.94	.96	.93	.94	.96	.97
3	.87	.92	.92	.93	.95	.95	.96	.96	.93	.93	.94	.94	.95	.95	.96	.96
4	.91	.94	.94	.95	.94	.95	.96	.97	.94	.93	.94	.95	.94	.95	.96	.97
5	.92	.94	.94	.95	.95	.96	.96	.96	.95	.94	.95	.95	.94	.95	.96	.97
6	.92	.95	.95	.95	.94	.96	.95	.96	.95	.94	.94	.95	.90	.95	.96	.96
7	.94	.96	.96	.95	.96	.96	.96	.96	.95	.94	.95	.96	.95	.96	.96	.96
8	.94	.96	.95	.95	.96	.96	.96	.97	.94	.94	.95	.96	.95	.96	.96	.96
9	.94	.96	.96	.97	.95	.96	.97	.98	.93	.95	.95	.96	.92	.95	.97	.97
10	.93	.95	.95	.95	.94	.94	.95	.96	.93	.94	.93	.94	.91	.93	.95	.95
11	.94	.95	.93	.92	.94	.93	.92	.93	.93	.93	.91	.91	.91	.91	.92	.93
12	.94	.94	.92	.91	.94	.92	.92	.93	.94	.93	.89	.90	.91	.90	.91	.93

* Scores from domain tests marked as Exempt or Not Attempted are excluded.
* Cut scores 1 to 4 fall between performance levels 1 and 2, 2 and 3, 3 and 4, and 4 and 5, respectively.

*Table 3.4 Classification Consistency for Each Cut Score by Domain and Grade**

Grade	Listening				Reading				Speaking				Writing			
	Cut 1	Cut 2	Cut 3	Cut 4	Cut 1	Cut 2	Cut 3	Cut 4	Cut 1	Cut 2	Cut 3	Cut 4	Cut 1	Cut 2	Cut 3	Cut 4
Pre-K	.85	.80	.93	.96	.81	.79	.90	.94	.85	.85	.90	.93	.76	.93	.98	.99
K	.86	.82	.91	.95	.83	.81	.87	.92	.85	.87	.90	.92	.77	.89	.96	.97
1	.85	.85	.87	.90	.88	.92	.94	.96	.88	.90	.92	.94	.91	.93	.95	.96
2	.81	.85	.87	.91	.91	.92	.94	.96	.89	.90	.92	.94	.90	.92	.95	.96
3	.83	.88	.89	.91	.93	.92	.94	.95	.91	.91	.91	.92	.93	.93	.94	.95
4	.86	.91	.91	.93	.91	.93	.95	.95	.92	.91	.92	.93	.91	.93	.95	.95
5	.88	.92	.92	.93	.92	.94	.95	.95	.93	.92	.93	.93	.91	.93	.95	.95
6	.88	.93	.93	.93	.91	.94	.93	.94	.92	.91	.92	.93	.86	.92	.94	.94
7	.91	.94	.94	.93	.94	.94	.94	.95	.93	.92	.93	.94	.93	.94	.94	.95
8	.91	.95	.93	.93	.94	.94	.94	.95	.92	.91	.93	.94	.93	.94	.94	.95
9	.91	.94	.95	.95	.93	.94	.96	.97	.90	.93	.93	.94	.88	.93	.96	.96
10	.90	.92	.92	.92	.91	.91	.93	.94	.90	.91	.90	.91	.87	.90	.93	.93
11	.92	.92	.90	.88	.92	.90	.89	.90	.91	.91	.87	.87	.88	.88	.89	.90
12	.92	.92	.88	.87	.91	.89	.88	.90	.91	.90	.85	.85	.88	.87	.88	.90

* Scores from domain tests marked as Exempt or Not Attempted are excluded.
* Cut scores 1–4 fall between performance levels 1 and 2, 2 and 3, 3 and 4, and 4 and 5, respectively.

Table 3.5 Classification Accuracy and Consistency for Overall Proficiency Categories

Grade	Accuracy			Consistency		
	Overall	Between Emerging and Progressing	Between Progressing and Proficient	Overall	Between Emerging and Progressing	Between Progressing and Proficient
Pre-K	.85	.87	.97	.80	.83	.96
K	.87	.89	.98	.83	.85	.98
1	.87	.90	.97	.83	.86	.96
2	.87	.91	.97	.82	.86	.96
3	.88	.92	.96	.83	.89	.95
4	.90	.94	.96	.87	.91	.95
5	.91	.95	.96	.88	.93	.95
6	.91	.95	.96	.87	.93	.95
7	.91	.95	.96	.89	.93	.95
8	.91	.95	.97	.89	.93	.96
9	.92	.95	.98	.90	.93	.97
10	.89	.94	.95	.86	.92	.94
11	.87	.94	.92	.83	.92	.91
12	.87	.94	.92	.83	.93	.91

3.4 Inter-Rater Reliability Analysis

In the 2024–2025 screener assessment, a total of 29 constructed-response items (10 in writing and 19 in speaking) were scored using Automated Scoring (AS) across all grade bands for all states except Ohio, where scoring was conducted locally. The AS system produced either a rubric score accompanied by a confidence level, or a condition code with a score of 0.

For essays not receiving a condition code, Autoscore^{T} (scoring engine for Writing items) and Autoscore^{S} (scoring engine for Speaking items) also produce a confidence value. This value reflects the degree of confidence the engine has that the score it predicted matches the score a well-trained and experienced human scorer would assign. The confidence value is based upon a statistical method that uses the patterns of the score probabilities to predict whether the engine score matches the human score.

In Autoscore^{T}, the confidence model inputs are the logits for each score point arising out of each model and the maximum probability of the ensemble model. In Autoscore^{S}, the confidence model inputs confidence values arising out of each transcriber and the word error rate between the two transcribers. These inputs were chosen because we assume that transcriber error is the likely driver of scoring errors in speech scoring. The dependent variable is whether the engine-predicted score matches the human score (1 = Match, 0 = Non-match). Logistic regression is used to predict the

dependent variable and a model is trained for each item using the held-out validation sample. The probability values are then converted to percentiles, again using the held-out validation sample.

For each engine, the confidence score is reported on a percentile scale and ranges from 0 to 100. A low value indicates that the engine has low confidence that the score it assigned will match the score a human would assign; a high value indicates that the engine has high confidence that the score would match a score assigned by a human. While each item has its own confidence model, all models use the same threshold: responses with a confidence score below the 20th percentile are flagged and routed for human review. The 20th percentile confidence cutoff was established based on data illustrating that, below this threshold, agreement between the engine score and human raters tends to be lower than the agreement between two human raters. The confidence model and percentiles are built from the held-out validation sample.

Responses scored by AS were routed to human scoring under three conditions, described in the order in which they were applied:

1. Random routing for approximately 10% of all responses,
2. Presence of specific condition codes (see Table 2.2 in Part I of this technical report), and
3. Low confidence level assigned by AS (less than the 20th percentile).

Of the responses scored by humans, 10% were also randomly selected for a second human rating to support reliability analyses. Starting from 2024–2025, MetriTech (MT) replaced Measurement Inc. (MI) as the provider of human hand scoring for all states except Ohio and Oregon.

This section presents the inter-rater reliability analysis between AS and Human Score 1 (H1), and between Human Score 1 (H1) and Human Score 2 (H2) for the states scored by MT. Quadratic weighted kappa (QWK), exact agreement rate, and exact-or-adjacent agreement rate were generated for each comparison based on combined data across all states except Ohio.

AS-H1 Agreement

The AS–H1 inter-rater reliability analysis excluded responses routed from AS to human scoring due to condition codes or low confidence level; only the 10% of responses randomly routed for quality control were included. Table 3.6 summarizes AS-H1 agreement. The data shown in the table combine items in the same grade/grade band with the same point value, regardless of domain (writing and speaking items are combined). Item-level inter-rater analysis results are provided in Section 11 of the pooled appendix for the screener assessment.

Overall, inter-rater agreement between AS and H1 is low for many items. This can be attributed to several factors. First, the AS model was trained using data rated by the previous vendor (MI), while the 2024-2025 human scoring was conducted by a new vendor (MT). Switching to a new vendor can introduce discrepancies as they refine their processes and rater training. Secondly, ELPA tasks involve open-ended responses and linguistic features that can present challenges for hand scoring and may contribute to variability among human raters, particularly during the early phase of scoring activities. Some variability in agreement rates between H1 and H2 were observed (see next section). These patterns suggest that continued calibration and refinement of scorer

training materials during early implementation with MT, along with ongoing training and alignment of the AS model, may help improve scoring consistency over time.

The AS engine was trained and validated prior to the 2024–2025 test administration using historical operational screener data from multiple years (2021–2022, 2022–2023, and 2023–2024 through mid-February) and multiple states (Arkansas, Iowa, Louisiana, Nebraska, Oregon, and West Virginia). The report of *Automated and Hybrid Scoring Methods and Results: 2024-2025 ELPA21 Screener* (hereafter referred to as the AS Study Report) demonstrated reasonable inter-rater reliability between the AS and human ratings. Table 3.8, adapted from Table 11 in the AS Study Report, shows that the AS-H1 agreement closely aligns with H1-H2 agreement. Specifically, differences in QWK between AS-H1 and H1-H2 are within 0.1 for all items except for two; one of these two items still has a strong kappa value of 0.71. Differences in exact agreement rates between AS-H1 and H1-H2 are within the predetermined target of 5.25% (PARCC, 2015) for all items except four; for these four items, exact agreement rates remain high–close to or above 80%.

In the 2024-2025 test administration, AS served as the primary scoring mode, with approximately 70% of responses using AS scores as final. For the remaining 30%, H1 was used as final. The AS Study Report also compared means and standard deviations of item-level final scores from the 2024-2025 administration to AS scores from the held-out validation set, disaggregated by year, see Table 3.9 (adopted from Table 26 in the AS Study Report). While some differences are expected (e.g., 2024-2025 data include approximately 30% human rater scores, a change in vendor, and differences in the samples), results are generally comparable across administrations. There are two exceptions: Speaking item 3400 has a higher-than-expected standard deviation, and Writing item 3530 has a lower-than-expected standard deviation.

The close alignment of AS–H1 and H1–H2 inter-rater reliability in the AS training and validation data, combined with comparable item-level score distributions between 2024-2025 operational data and data from the prior three years, supports that AS, as the primary scoring method, maintains consistency with prior years and ensures reliable operational scoring.

H1-H2 Agreement

Table 3.7 presents the inter-rater reliability analysis for the H1-H2 comparison. The table combines items in the same grade/grade band with the same point value, regardless of domain (writing and speaking items are combined). Item-level inter-rater analysis results are provided in Section 11 of the pooled appendix for the screener assessment. The analysis included all responses that received both H1 and H2 scores, regardless of routing reason (condition code, low confidence level, or random selection). Limiting the analysis to only randomly routed responses would result in sample sizes too small to support valid conclusions (ranging from 2 to 38 across items). Even with this broader inclusion, the sample sizes remain relatively low – under 100 for all items except one. The lower sample sizes compared to previous years can be attributed to the switch from full human scoring to auto scoring.

Low agreement between H1 and H2 is also observed, particularly for items in grade bands 6–8 and 9–12. Including responses routed due to low confidence may negatively impact agreement, as these

are responses the AS engine identified as difficult to score. Additionally, smaller sample sizes can reduce stability and increase sensitivity to outliers.

Meanwhile, efforts have been made to monitor scoring activity and enhance rating consistency for MT since the 2024–2025 administration. Key measures include:

- Enhanced Training: Utilization of additional training materials from the previous hand-scoring vendor and participation in rangefinding sessions to strengthen scorer calibration.
- Experienced Scorers: All 2025–2026 Screener scorers previously scored the 2024–2025 Screener, increasing familiarity with item types and expectations.
- Ongoing Monitoring: More frequent internal inter-rater reliability analyses conducted by MT to identify potential issues and implement retraining as needed.

A mid-window inter-rater reliability analysis for the 2025-2026 Screener is planned for February 2026 to assess the impact of these measures and further evaluate improvements in inter-rater agreement.

Table 3.6 Summary of AS-H1 Agreement by Grade Band

Grade/ Grade Band	Score Points	Number of Items	Weighted Kappa		% Exact Agreement		% within 1 Agreement		% Not within 1 Agreement	
			Min	Max	Min	Max	Min	Max	Min	Max
Pre-K	3	2	0.747	0.767	67.4	68.8	97.9	98.3	1.7	2.1
K	3	2	0.556	0.763	50.9	66.1	93.0	98.3	1.7	7.0
1	3	1	0.857	0.857	83.6	83.6	100	100	0	0
	4	1	0.336	0.336	51.4	51.4	97.2	97.2	2.8	2.8
2–3	3	2	0.601	0.688	61.2	71.3	98.4	100	0	1.6
	4	1	0.538	0.538	49.6	49.6	90.8	90.8	9.2	9.2
4–5	3	1	0.588	0.588	76.1	76.1	97.2	97.2	2.8	2.8
	5	3	0.505	0.743	44.4	53.5	87.3	97.9	2.1	12.7
6–8	3	6	0.142	0.593	43.1	81.0	90	98.3	1.7	10.0
	5	3	0.609	0.831	45.5	58.0	87.6	98.3	1.7	12.4
9–12	3	6	0.139	0.520	58.2	87.5	87.0	99.6	0.4	13.0
	5	3	0.606	0.655	42.9	52.3	85.8	97.9	2.1	14.2

Table 3.7 Summary of H1-H2 Agreement by Grade Band

Grade/ Grade Band	Score Points	Number of Items	Weighted Kappa		% Exact Agreement		% within 1 Agreement		% Not within 1 Agreement	
			Min	Max	Min	Max	Min	Max	Min	Max
Pre-K	3	2	0.717	0.751	56.1	60.2	96.5	100	0	3.5
K	3	2	0.676	0.802	62.5	73.1	93.8	96.2	3.8	6.3
1	3	1	0.905	0.905	81.3	81.3	100	100	0	0
	4	1	0.890	0.890	69.2	69.2	100	100	0	0
2–3	3	2	0.722	0.802	75.9	80.6	96.6	97.2	2.8	3.4
	4	1	0.805	0.805	59.5	59.5	94.6	94.6	5.4	5.4
4–5	3	1	0.883	0.883	88.0	88.0	100	100	0	0
	5	3	0.606	0.921	39.4	85.7	91.7	97.0	3.0	8.3
6–8	3	6	0.377	0.685	50.0	75.0	92.7	100	0	7.3
	5	3	0.682	0.848	51.4	75.0	91.4	96.9	3.1	8.6
9–12	3	6	0.240	0.737	61.8	88.3	91.0	100	0	9.0
	5	3	0.373	0.702	47.8	56.1	86.6	93.2	6.8	13.4

Table 3.8 Autoscore Performance Compared to Human-Human Agreement on Exact Agreement and Quadratic Weighted Kappa on the Held-out Validation Sample (Condition Codes Removed)

Modality	Grade Band	Item ID	N		QWK			Exact Agreement		
			Total	Held-out	H1H2	H1AS	diff	H1H2	H1AS	diff
Writing	2-3	1466	≥7970	≥1080	0.84	0.80	-0.04	75%	72%	-3%
	4-5	2310	≥3970	≥400	0.82	0.78	-0.05	68%	67%	0%
	6-8	2794	≥5300	≥530	0.79	0.71	-0.08	65%	61%	-4%
	6-8	2800	≥5360	≥520	0.62	0.61	0.00	86%	88%	2%
	6-8	2802	≥5350	≥530	0.54	0.60	0.06	85%	88%	3%
	6-8	2804	≥5340	≥530	0.62	0.48	-0.15	86%	85%	-1%
	9-12	3502	≥7030	≥710	0.76	0.73	-0.03	64%	62%	-2%
	9-12	3528	≥7110	≥660	0.61	0.58	-0.03	88%	88%	0%
	9-12	3530	≥7110	≥680	0.68	0.71	0.03	89%	91%	2%
	9-12	3532	≥7100	≥680	0.66	0.61	-0.05	88%	88%	0%
Speaking	K	6545	≥18000	≥2070	0.86	0.87	0.02	77%	80%	3%
	K	6546	≥15100	≥1520	0.82	0.82	0.00	72%	72%	0%
	1	694	≥2290	≥230	0.85	0.85	0.00	80%	81%	0%
	1	6547	≥2340	≥240	0.71	0.62	-0.09	73%	73%	0%
	2-3	1378	≥4110	≥420	0.86	0.80	-0.06	82%	74%	-8%
	2-3	6548	≥4110	≥410	0.83	0.71	-0.12	79%	67%	-13%
	4-5	2078	≥3930	≥400	0.92	0.84	-0.08	86%	78%	-8%
	4-5	2080	≥3930	≥400	0.90	0.81	-0.09	80%	70%	-10%
	4-5	2108	≥4570	≥530	0.78	0.75	-0.03	79%	77%	-2%
	6-8	2662	≥5220	≥530	0.91	0.90	-0.01	74%	70%	-4%
	6-8	2664	≥5180	≥520	0.92	0.91	0.00	75%	73%	-1%
	6-8	2694	≥5250	≥530	0.74	0.71	-0.04	76%	74%	-2%
	6-8	2696	≥5210	≥530	0.82	0.78	-0.03	76%	74%	-2%
	6-8	2698	≥5220	≥530	0.84	0.77	-0.08	86%	81%	-5%
	9-12	3344	≥6910	≥700	0.92	0.91	-0.01	80%	78%	-2%
	9-12	3346	≥6810	≥690	0.87	0.83	-0.04	74%	70%	-5%
	9-12	3400	≥6970	≥700	0.64	0.64	0.00	77%	80%	3%
	9-12	3402	≥6910	≥700	0.70	0.73	0.02	78%	82%	4%
	9-12	3404	≥6890	≥690	0.72	0.69	-0.02	83%	82%	-1%

Note. Adapted from Table 11 in the AS Study Report. Target performance for Exact Agreement is a difference of less than 5.25%. Target performance for QWK is a difference of less than 0.10.

Table 3.9 Means and Standard Deviations of Final Scores for the Past Four Years

Modality	Grade Span	Item ID	Means				Standard Deviations			
			2021-2022	2022-2023	2023-2024	2024-2025	2021-2022	2022-2023	2023-2024	2024-2025
Writing	2-3	1466	2.00	1.87	1.89	2.08	1.02	1.10	1.08	1.03
	4-5	2310	2.60	2.84	2.76	2.83	0.86	0.87	0.95	0.91
	6-8	2794	2.93	2.88	2.98	3.05	0.80	0.82	0.81	0.90
	6-8	2800	2.81	2.86	2.78	2.83	0.50	0.40	0.60	0.42
	6-8	2802	2.83	2.80	2.87	2.83	0.41	0.45	0.46	0.43
	6-8	2804	2.85	2.84	2.79	2.84	0.45	0.45	0.55	0.42
	9-12	3502	3.42	3.41	3.53	3.52	0.87	0.80	0.77	0.88
	9-12	3528	2.83	2.78	2.85	2.85	0.45	0.47	0.35	0.39
	9-12	3530	2.80	2.71	2.78	2.87	0.47	0.60	0.53	0.36
	9-12	3532	2.77	2.74	2.79	2.87	0.57	0.55	0.49	0.37
Speaking	K	6545	1.54	1.65	1.88	1.73	0.94	0.92	0.87	0.89
	K	6546	1.51	1.59	1.56	1.52	0.87	0.91	0.87	0.88
	1	694	2.54	2.36	2.38	2.47	0.77	0.83	0.87	0.78
	1	6547	3.15	3.15	3.23	3.26	0.69	0.58	0.68	0.71
	2-3	1378	2.26	2.21	2.12	2.29	0.82	0.80	0.83	0.83
	2-3	6548	2.39	2.29	2.22	2.36	0.72	0.79	0.72	0.78
	4-5	2078	3.48	3.63	3.65	3.57	0.98	0.82	0.78	0.96
	4-5	2080	3.28	3.31	3.34	3.57	0.93	0.93	0.84	1.06
	4-5	2108	2.46	2.49	2.53	2.52	0.70	0.63	0.65	0.68
	6-8	2662	3.09	3.05	3.31	3.26	1.34	1.24	1.24	1.34
	6-8	2664	3.14	3.27	3.20	3.26	1.31	1.28	1.26	1.27
	6-8	2694	2.16	2.16	2.18	2.23	0.65	0.70	0.69	0.72
	6-8	2696	2.34	2.42	2.38	2.38	0.86	0.78	0.80	0.87
	6-8	2698	2.44	2.50	2.45	2.45	0.64	0.62	0.67	0.70
	9-12	3344	3.98	3.86	4.04	3.98	1.13	1.34	1.20	1.26
	9-12	3346	2.97	3.13	3.08	3.27	0.98	0.97	0.98	1.10
	9-12	3400	2.67	2.62	2.70	2.59	0.54	0.57	0.48	0.80
	9-12	3402	2.55	2.55	2.63	2.63	0.57	0.56	0.59	0.61
	9-12	3404	2.74	2.74	2.77	2.84	0.55	0.50	0.51	0.49

Note. Adapted from Table 26 in the AS Study Report. Final-assigned condition codes are removed from the analysis.

Chapter 4. Validity

Discussions on test development, form construction, scaling, equating, and standard setting can be found in related documents from the ELPA21. Refer to *ELPA21 Scoring Specification: School Year 2019–2020* (CRESST, 2019) and *ELPA21 Standard Setting Technical Report* (CRESST, 2016).

Since the items and item parameters in the screener tests are drawn from the item pool for summative tests, and because the purpose of the screener assessment is to predict students' overall English proficiency categories, we evaluate the relationship between the screener and summative assessments instead of evaluating the validity aspects as we do for the summative assessment. We also summarize student progress from the time they took the screener assessment to the time they took the summative assessment. The statistical methods and results are presented in this chapter and Sections 12–13 of the pooled appendix for the screener assessment.

- Section 12. Screener Assessment—Correlations Between Summative and Screener Tests
    - Table S12.1 presents the correlations between screener and summative tests in domain test scores, overall test scores, and comprehension test scores.
    - Table S12.2 summarizes the correlations between screener and summative tests in domain performance levels and overall proficiency categories.
- Section 13. Screener Assessment— Student Progress from Screener to Summative Assessment
    - Figures S13.1–S13.2 display within-year average differences in domain, overall, and comprehension scale scores.
    - Figures S13.3–S13.4 present changes in domain performance level and overall proficiency.
    - Figures S13.5–S13.10 present scatter plots showing scale score changes from the screener to the summative assessment for each grade level.
    - Tables S13.1–S13.6 present the comparison of scale score summary statistics of domain, overall, and comprehension scores between the screener and summative tests.

4.1 Relationship Between Screener and Summative Tests

Students who took the ELPA21 screener and were classified as English learners (ELs) (Proficiency Not Demonstrated, Emerging, or Progressing) are expected to also take the ELPA21 summative assessment. The test questions on the screener and summative assessments are drawn from the same item pools and assess the same English Language Proficiency (ELP) standards adopted by the ELPA21 member states. We identified students who completed both the screener and summative assessments and compared their performance across the two occasions.

4.1.1 Correlation between Screener and Summative Tests

Pearson correlation was used to evaluate the relationship between the screener and summative scale scores, while Goodman and Kruskal's Gamma correlation (Goodman & Kruskal, 1954) was used to evaluate the relationship in domain performance levels and overall proficiency categories between the two assessments. The gamma correlation, or gamma statistics, is for ordinal-level data with a small number of response categories. It is designed to determine how effectively a researcher can use the information about an individual measured on one variable to predict the measure of the individual on another variable. The correlation results are presented in Tables S12.1 and S12.2 in the pooled appendix for the screener assessment. These correlations show predictive validity between the two ELPA21 assessments because they were given to the same students at various times.

Table S12.1 shows the Pearson correlation between the screener and the summative assessments for domain and composite scores. Correlations were the lowest in kindergarten tests ranging from 0.42 in writing to 0.72 for the overall test. Correlations were above 0.70 in all four domains, comprehension, and overall scale scores in grades 1 and above. The speaking tests had lower correlations than the other three domains in grade bands 6–8 and 9–12.

The Gamma correlations reveal a similar pattern but are generally higher in magnitude. In kindergarten, Gamma correlations range from 0.51 in writing to 0.91 for the overall test, while they exceed 0.80 in all tests for grades 1 and above. Generally speaking, the screener assessment can serve as a strong indicator of students' performance on the summative assessment.

4.1.2 Student Progress from Screener to Summative Assessments

Student progress from the time they took the screener to the time they took the summative was evaluated by the changes in scale scores and performance levels. A potential confounding factor in this result is measurement error in both assessments. Given the acceptable marginal reliability indices described in the marginal reliability of summative (Part II) and screener (Part III) assessments, respectively, in this technical report, we can still derive reasonable conclusions by observing the trend of student progress. Section 13 of the screener pooled appendix summarizes the results of progress analysis. Only students who had valid scores on both the screener and summative assessments were included in each analysis.

Figures S13.1 and S13.2 in the pooled appendix of the screener assessment show the growth of the average domain scores and composite scores, respectively. The average scale scores in the summative assessment were higher than those in the screener assessment. Figures S13.3 and S13.4 display the percentage of students in each domain performance level and overall proficiency category, respectively. In each pair of bars, the left bar (marked as A) shows the screener test, and the right bar (marked as B) shows the corresponding summative test. The plots indicate that more students were in higher domain performance levels and overall proficiency categories in the summative than in the screener tests. In addition, Figures S13.5–S13.10 in the pooled appendix of the screener assessment present scatter plots of scale score changes from screener to summative tests in each grade band, and Tables S13.1–S13.6 summarize comparisons of scale scores between screener and summative tests.

Chapter 5. Reporting

A comprehensive introduction to the Centralized Reporting System (CRS) can be found in Part I, Chapter 6 of this technical report. Mock-up score reports for the screener assessment in each state are presented in Section 14 of the state-specific appendix. Please note that the mock-up score reports are not included in the pooled appendix for the screener assessment.

References

Center for Research on Evaluation, Standards, and Student Testing (CRESST). (2019). *ELPA21 scoring specification: school year 2019–2020.*

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Goodman, L., & Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*(268), 732–764. https://doi.org/10.2307/2281536

PARCC (2015, March 9). *Research Results of PARCC Automated Scoring Proof of Concept Study*. Retrieved from: http://www.parcconline.org/images/Resources/Educator-resources/PARCC_AI_Research_Report.pdf.