TECHNICAL REPORT

PART III – SCREENER ASSESSMENT

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

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

Grades Pre-K–12

2021–2022 Administration

Submitted to:
ELPA21

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

January 2023

Table of Contents

List of Tables

List of Figures

Chapter 1. Test Administration

The screener tests were administered to students in the following grade bands: kindergarten (K), grade 1, grades 2–3, grades 4–5, grades 6–8, and grades 9–12. Some states administered the screener tests to pre-kindergarten (pre-K) students. For the screener test, as with the summative test, each form involves four domains (listening, reading, speaking, and writing). Students can be exempted from as many as three domain tests. The assessments do not have a time limit.

1.1 TESTING WINDOW

The 2021–2022 screener testing windows for the six states discussed in this report are shown in Table 1.1. Although testing windows remained open in 2022, due to the continued impact of the COVID-19 pandemic, some students did not complete the English Language Proficiency Assessment (ELPA) screener assessments.

Table 1.1 2021–2022 ELPA21 Screener Testing Windows by State

State	ELPA21 Screener
Arkansas	8/2/2021–7/15/2022
Iowa	8/3/2021–7/15/2022
Louisiana	8/3/2021–7/15/2022
Nebraska	8/2/2021–7/15/2022
Ohio	8/23/2021–6/30/2022
West Virginia	8/3/2021–6/20/2022

1.2 TEST DESIGN

Each 2021–2022 screener test had one online form, one paper-pencil form, and one braille form. Pre-K students were permitted to take the kindergarten tests. However, Ohio is different from other states, administering two types of screeners, OELPS-BK and OELPS-K. The OELPS-BK is the screener administered to students before Dec. 31 of the kindergarten year. In Ohio, from end of before Kindergarten (BK) year and through Dec.31 of Kindergarten year, students taking the kindergarten screener before Dec. 31 of the kindergarten year were proficient if all non-exempt domains were 3 or better. The OELPS-K is the screener administered to kindergarteners after Jan. 1 of the kindergarten year. Students taking the OELPS-K will be proficient if they earn scores of 4 or higher in all non-exempt domains of the screener. For K and higher grades for all states, students need to obtain 4 or above in each domain for proficiency. The online form has three steps. Step 1 consists of practice items, while Steps 2 and 3 include 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 determines that the test should not proceed, the student is directed to Segment 9, and then the test ends. 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 test 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 on-the-fly, scored speaking items. After the student responds to these items, the TA assigns a score to each item. From 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) (see 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 in 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 all non-exempted domains in addition to the speaking domain.

- 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 2021–2022 ELPA21 Screener Online Test Design
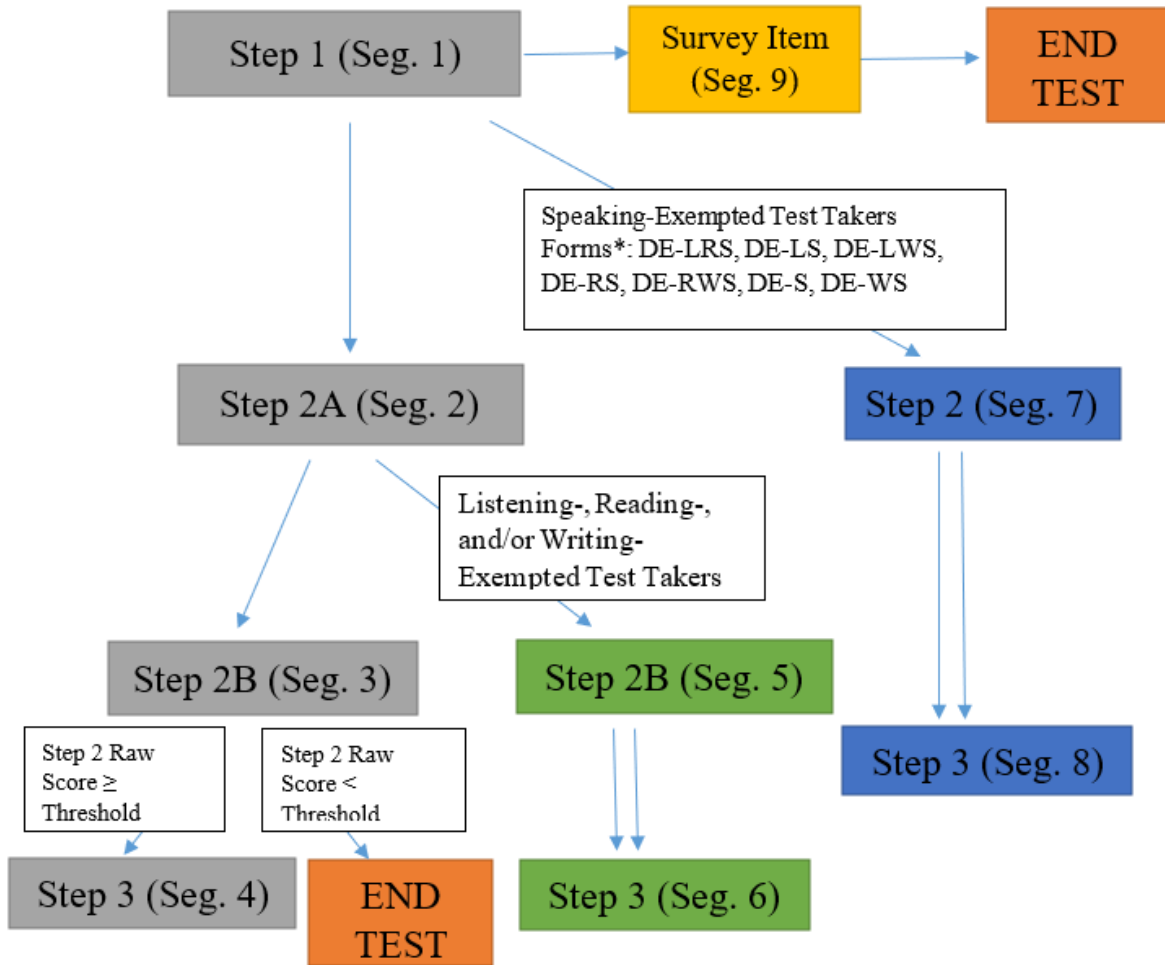
* DE-LRS (listening, reading, and speaking exempted), DE-LS (listening and speaking exempted), DE-LWS (listening, writing, and speaking exempted), DE-RS (reading and speaking exempted), DE-RWS (reading, writing, and speaking exempted), DE-S (speaking exempted), DE-WS (writing and speaking exempted).

Table 1.2 Threshold Step 2 Summed Scores for Proceeding to Step 3 by Grade Band

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	27	30

The paper-pencil form has five segments:

- 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.

- 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) (see 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) and 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.

Figure 1.2 displays the test design for the paper-pencil screener test. For the paper-pencil form, after test administration, student responses are entered into Cambium Assessment, Inc.'s (CAI) Data Entry Interface (DEI) on the state testing portal for all English Language Proficiency Assessment for the 21st Century (ELPA21) domain tests. Practice test items are not entered into the DEI and are not scored.

Figure 1.2 2021–2022 ELPA21 Screener Paper Test Design

```
DEI Step 1 (Seg. 1) ─────────────────────▶  END TEST
        │
        │           Domain Exempt
        ▼                    ▼
DEI Step 2 (Seg. 2)    DEI Step 2 (Seg. 4)
    │         │                │  │
    ▼         ▼                ▼  ▼
DEI Step 3   END TEST    DEI Step 3 (Seg. 5)
(Seg. 3)
```
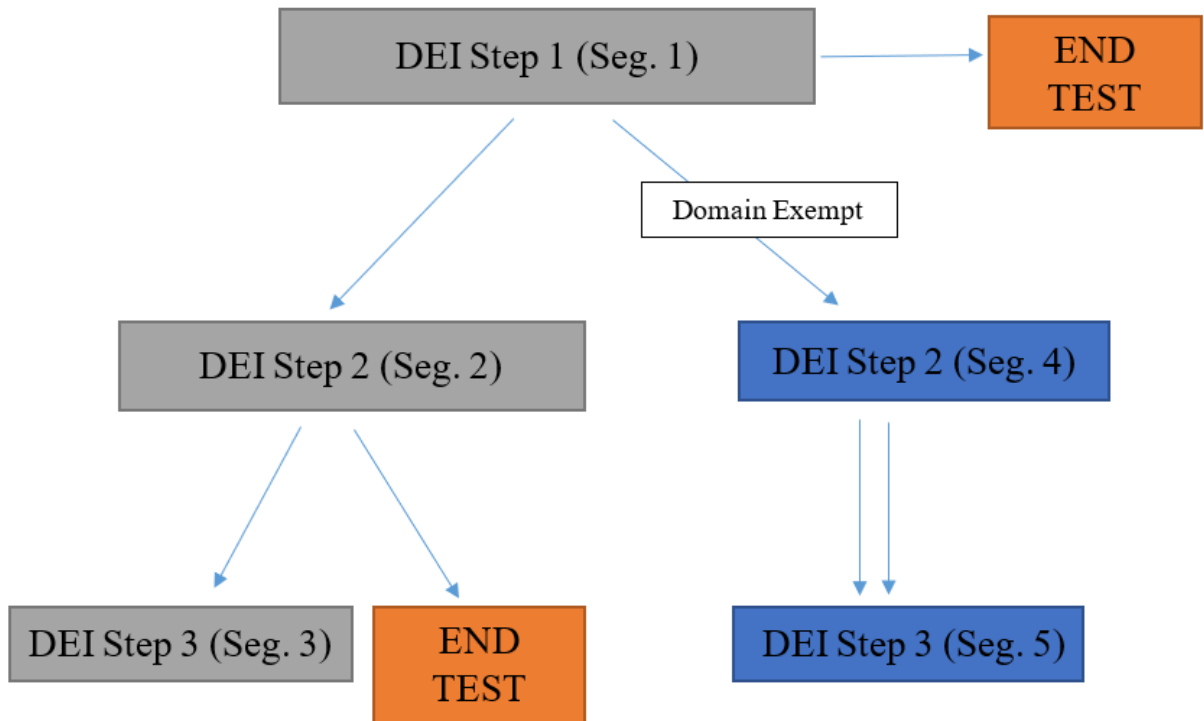
The braille form includes two segments. In Segment 1, the TA indicates whether the student should proceed to the operational items. If so, the student is routed to Segment 2, which contains operational items for all domains. If the TA indicates the student should not proceed, then the test ends.

The non-domain-exempted form summary of the screener tests is listed in Table 1.3–Table 1.1.5 Specifically, Table 1.3 includes items from Segments 2–4, Table 1.4 includes Segments 2–3, and Table 1.1.5 includes Segment 2 items.

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	15	18
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	46	82

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	15	18
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	46	82

Table 1.1.5 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 2021–2022 administration, a test administration manual (TAM) was developed for each state. The TAM guides TAs in test administration. The TAM for the screener tests usually includes the following key points:

- Overview of the ELPA21 screener test
- TA qualifications
- Preliminary planning
- Materials required
- 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, training/practice tests (Step 1 in screener tests) are 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 tests; and
- stopping the test.

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.

1.4 BUSINESS SCORING RULES FOR THE SCREENER ASSESSMENT

Business rules and instructions applied to the 2021–2022 screener assessment include the following:

1. All pending and expired test records in Step 2 should be scored.

2. 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.

3. 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 0.

4. If a student has a domain exemption for a domain, the domain is reported as exempt if it is not attempted.

   a. For online tests, any domain exemptions must be entered into the Test Information Distribution Engine (TIDE) prior to the student starting the test. Students taking the online screener will be presented with items in non-exempt domains only.

   b. 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.

5. ELPA21 states make the decision of whether to use the pre-K test on an individual basis.

6. For the Ohio screener administration, handscored items are scored by local TAs.

7. Tests in which the TA indicates that the student will not continue after the Step 1 practice items will be scored as follows:

   a. Each domain will be scored 0. The score of 0 will receive a label of "Performance Not Determined."

   b. Proficiency status will be scored as "D" and reported as "Proficiency Not Demonstrated."

Chapter 2. 2021–2022 Summary

The 2021–2022 screener results are presented in this chapter and in Sections 1–14 of the Appendix for Pooled Analysis – 2021–2022 Summary_Screener. The figures and tables included in each section are listed below:

- Section 1. Screener—Student Participation

    o Table S1.1 displays the number and percentage of students in each test mode of braille, paper-pencil, and online in each grade (pre-K–12) and across the state.

    o Table S1.2 lists the number and percentage of students taking each test by subgroup, including grade, gender, ethnicity, primary disabilities, and other groups such as migrant, special education (SPED), Title I, or Section 504 Plan. Subgroups can vary across states. The pooled analysis includes the summary by gender and ethnicity.

- Section 2. Screener Assessment—Raw Score Summary

    o Tables S2.1–S2.14 present the number of students, minimum, maximum, average, and standard deviation of domain raw scores across the state and by each performance level in each grade. Tables S17.1–S17.14 also present the number of students, minimum, maximum, average, and standard deviation of the overall raw scores across the state and by each proficiency level in each grade.

    o 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 have to consider whether a student advanced to Step 3 when interpreting raw scores.

- Section 3. Screener Assessment—Raw Score Distributions

    o Figures S3.1–S3.70 present the frequency of raw score distributions by performance level for each domain in each grade, and the frequency of overall raw score distributions by proficiency level in each grade.

- Section 4. Screener Assessment—Scale Score Summary

    o Tables S4.1–S4.14 present the number of students, the minimum, average, maximum, and standard deviation of domain, overall, and comprehension scores across the state (or states, in the case of the pooled analysis), and by subgroups in each grade of pre-K–12. Subgroups can vary across the states. The pooled analysis includes the summary by gender and ethnicity.

    o Table S4.15 summarizes the number and percentage of students who were marked "exempt" in each domain and grade.

- Section 5. Screener Assessment—Percentage of Students by Domain Performance Level

- o Figure S5.1 shows the percentage of students in each performance level in each domain test across grades in the state (or states, in the case of the pooled analysis).

- o Tables S5.1–S5.14 present the total number of students taking each domain test and the percentage of students in each performance level by domain test across the state (or states, in the case of the pooled analysis) and by subgroups.

- Section 6. Screener Assessment—Percentage of Students by Overall Proficiency Category

  - o Figure S6.1 shows the percentage of students in each overall proficiency category across grades in the state (or states, in the case of the pooled analysis).

  - o Tables S6.1–S6.14 present the total number of students who are categorized in each of the overall proficiency categories: Emerging, Progressing, Proficient, and Proficiency Not Demonstrated by subgroups.

- Section 7. Screener Assessment—Testing Time

  - o Table S7.1 shows the testing time by end step in each grade/grade band.

2.1 2021–2022 STUDENT PARTICIPATION

During the 2021–2022 administration, the impact of COVID-19 pandemic has decreased significantly, almost all eligible students completed the assessment in each state.

Table 2.1 shows the overall student participation for each state. There were 61,899 students in total who took the 2021–2022 screener tests. Ohio had the most students, followed by Louisiana. Most students were from pre-K and kindergarten.

Table 2.2 presents the frequencies of students who took summative tests, screener tests, and both summative and screener tests. It shows that kindergarten students had the highest percentage of students taking both the screener and the summative tests in the 2021–2022 school year.

Section S1.1 of the Appendix presents student participation in each mode. In the six ELPA21 states combined, the most frequent mode of administration was online (99.95%), followed by paper (0.05%) and braille (<0.01%).

Section S1.2 of the Appendix shows student participation by subgroups. For the pooled analysis from K–12, the number of students tested decreases as the grade level increases. There were more male students (48.6%–53.4%) than female students (43.2%–48.6%) tested. In each test, the greatest number of participating students were in the group of Hispanic or Latino (49.3%–63.0%), followed by Asian students (7.4%–18.1%), and White students (8.6%–10.5%).

Table 2.1 Number of Students Who Participated in ELPA21 Screener in 2020–2021 and 2021–2022 by State and Grade

Grade	Arkansas	Arkansas	Iowa	Iowa	Louisiana	Louisiana	Nebraska	Nebraska	Ohio	Ohio	West Virginia	West Virginia	Total	Total	Total
	2020–2021	2021–2022	2020–2021	2021–2022	2020–2021	2021–2022	2020–2021	2021–2022	2020–2021	2021–2022	2020–2021	2021–2022	2020–2021	2021–2022	Two year N Diff
Pre-K	≥ 3,870	≥ 3,570	≥ 4,780	≥ 4,860	≥ 3,940	≥ 3,760	≥ 3,260	≥ 3,550	-	≥ 10,780	≥ 160	≥ 190	≥ 15,860	≥ 26,910	≥ 11,050
K	≥ 1,260	≥ 1,220	≥ 240	≥ 280	≥ 460	≥ 300	≥ 140	≥ 110	≥ 8,150	≥ 780	≥ 70	≥ 80	≥ 18,820	≥ 2.960	≥ -15,860
1	≥ 390	≥ 580	≥ 360	≥ 610	≥ 1,010	≥ 470	≥ 220	≥ 380	≥ 990	≥ 1,980	≥ 50	≥ 80	≥ 3,160	≥ 4,650	≥ 1,490
2	≥ 340	≥ 430	≥ 270	≥ 400	≥ 620	≥ 300	≥ 170	≥ 280	≥ 680	≥ 1,350	≥ 40	≥ 60	≥ 2,230	≥ 3,180	≥ 950
3	≥ 290	≥ 380	≥ 250	≥ 340	≥ 560	≥ 290	≥ 190	≥ 240	≥ 610	≥ 1,220	≥ 30	≥ 60	≥ 2,030	≥ 2,820	≥ 790
4	≥ 270	≥ 400	≥ 230	≥ 340	≥ 470	≥ 210	≥ 140	≥ 230	≥ 490	≥ 990	≥ 30	≥ 50	≥ 1,720	≥ 2,500	≥ 780
5	≥ 250	≥ 330	≥ 210	≥ 310	≥ 460	≥ 220	≥ 120	≥ 210	≥ 380	≥ 960	≥ 30	≥ 30	≥ 1,500	≥ 2,320	≥ 820
6	≥ 240	≥ 360	≥ 200	≥ 310	≥ 490	≥ 190	≥ 70	≥ 180	≥ 400	≥ 890	≥ 20	≥ 30	≥ 1,380	≥ 2,280	≥ 890
7	≥ 260	≥ 420	≥ 160	≥ 250	≥ 460	≥ 160	≥ 80	≥ 170	≥ 370	≥ 910	≥ 30	≥ 40	≥ 1,300	≥ 2,290	≥ 980
8	≥ 230	≥ 370	≥ 150	≥ 290	≥ 490	≥ 160	≥ 60	≥ 180	≥ 330	≥ 960	≥ 20	≥ 40	≥ 1,180	≥ 2,370	≥ 1,180
9	≥ 300	≥ 680	≥ 300	≥ 670	≥ 1,290	≥ 280	≥ 150	≥ 520	≥ 470	≥ 1,740	≥ 20	≥ 60	≥ 1,850	≥ 4,980	≥ 3,130
10	≥ 260	≥ 490	≥ 170	≥ 380	≥ 330	≥ 110	≥ 70	≥ 130	≥ 310	≥ 850	≥ 20	≥ 60	≥ 1,190	≥ 2,250	≥ 1,050
11	≥ 190	≥ 410	≥ 130	≥ 210	≥ 190	≥ 60	≥ 40	≥ 70	≥ 220	≥ 580	≥ 20	≥ 50	≥ 860	≥ 1,530	≥ 660
12	≥ 90	≥ 250	≥ 50	≥ 90	≥ 60	≥ 20	≥ 30	≥ 50	≥ 150	≥ 300	≥ 10	≥ 30	≥ 510	≥ 800	≥ 290
Total	≥ 8,310	≥ 9,960	≥ 7,570	≥ 9,390	≥ 10,890	≥ 6,610	≥ 4,790	≥ 6,360	≥ 13,600	≥ 24,350	≥ 580	≥ 920	≥ 53,640	≥ 61,890	≥ 8,250

Table 2.2 Number of Students Participating in 2021–2022 ELPA21 Summative, Screener Tests, and Both by State and Grade Band

State	Grade/Grade Band	N Summative	N Screener	N Both
Arkansas	Pre-K and K	≥ 4,550	≥ 4,800	≥ 3,670
	1	≥ 4,250	≥ 580	≥ 460
	2–3	≥ 7,740	≥ 820	≥ 610
	4–5	≥ 5,760	≥ 730	≥ 490
	6–8	≥ 7,730	≥ 1,170	≥ 850
	9–12	≥ 9,700	≥ 1,840	≥ 1,350
Iowa	Pre-K and K	≥ 4,610	≥ 5,140	≥ 3,640
	1	≥ 4,100	≥ 610	≥ 450
	2–3	≥ 6,440	≥ 750	≥ 480
	4–5	≥ 4,480	≥ 650	≥ 410
	6–8	≥ 5,610	≥ 860	≥ 600
	9–12	≥ 7,690	≥ 1,360	≥ 1,010
Louisiana	Pre-K and K	≥ 3,930	≥ 4,410	≥ 2,950
	1	≥ 3,880	≥ 1,010	≥ 830
	2–3	≥ 6,240	≥ 1,190	≥ 940
	4–5	≥ 4,520	≥ 940	≥ 710
	6–8	≥ 5,740	≥ 1,460	≥ 1,170
	9–12	≥ 6,370	≥ 1,880	≥ 1,420
Nebraska	Pre-K and K	≥ 3,920	≥ 3,660	≥ 2,750
	1	≥ 3,680	≥ 380	≥ 270
	2–3	≥ 5,510	≥ 530	≥ 310
	4–5	≥ 3,260	≥ 440	≥ 260
	6–8	≥ 3,440	≥ 540	≥ 330
	9–12	≥ 4,550	≥ 790	≥ 490
Ohio	K	≥ 10,230	≥ 11,570	≥ 9,340
	1	≥ 9,380	≥ 1,980	≥ 1,550
	2–3	≥ 15,120	≥ 2,570	≥ 1,780
	4–5	≥ 9,970	≥ 1,960	≥ 1,230
	6–8	≥ 10,820	≥ 2,770	≥ 1,800
	9–12	≥ 13,950	≥ 3,480	≥ 2,490
West Virginia	Pre-K and K	≥ 230	≥ 280	≥ 210

State	Grade/Grade Band	N Summative	N Screener	N Both
	1	≥ 230	≥ 80	≥ 60
	2–3	≥ 390	≥ 130	≥ 70
	4–5	≥ 260	≥ 80	≥ 50
	6–8	≥ 360	≥ 120	≥ 70
	9–12	≥ 490	≥ 210	≥ 130

2.2 2021–2022 STUDENT SCALE SCORE AND PERFORMANCE-LEVEL SUMMARY

Table 2.3–Table 2.5 show the domain, comprehension, and overall scale score summary by grade level. The ELPA21 tests are not vertically linked across all grades. Scale scores can be compared only for tests or students within a 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 Appendix.

Table 2.6 and Table 2.7 present the number and percentage of students by grade and performance level in each domain test. The results indicate that performance level 1 is the most frequent level achieved in speaking and writing in grades pre-K–11, in reading in grades 1–10, and in listening in grades 1–11. Reading, writing, and speaking follow a similar pattern; the percentage of students in level 1 decreases from pre-K to grade 6 (with an increase in grade 1), then slightly increase to grade 9 and decrease in the remaining grades. For listening, the percentage of students who reach level 1 decreases from pre-K to grade 3 (with a slight increase in grade 1), then increases until grade 9 (with a slight decrease in grade 6), and thereafter decreases consistently. Disaggregated results by gender and ethnicity are provided in Section 5 of the Appendix.

Table 2.8 and Figure S5.1 in the Appendix present the percentage of students reaching each overall proficiency category, by grade. Starting 2021–22 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 K and higher grades, students need to obtain 4 or above in each domain for proficiency.

The results show that the majority of students have reached the Emerging or Progressing category. The percentages of students who are proficient decrease from grades pre-K to kindergarten, consistently increase from grade 1 (5.9%) to grade 4 (19.5%), and slightly decrease to grade 9 (4.3%), and thereafter increase consistently. The percentages of students in the Emerging category are relatively stable until grade 4 (41.1%), increase from grade 4 to grade 5 (46.4%), and then decrease from grade 5 to grade 6 (41.1%), and then consistently increase to grade 9 (55.9%), and thereafter decreases consistently. Section 6 of the Appendix 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	≥ 25,890	314	512.2	714	66.9	≥ 25,880	318	509.4	708	66.1
K	≥ 2,730	314	509.6	714	78.3	≥ 2,730	318	507.3	708	77.8
1	≥ 4,380	288	483.9	678	91.3	≥ 4,380	286	464.2	704	91.7
2	≥ 2,970	286	467.3	710	86.9	≥ 2,970	278	450.6	734	94.5
3	≥ 2,660	286	486.8	710	98.8	≥ 2,660	278	478.6	734	108.0
4	≥ 2,330	270	465.7	768	110.6	≥ 2,330	270	470.6	781	108.6
5	≥ 2,160	270	471.5	778	120.0	≥ 2,160	251	477.9	795	117.9
6	≥ 1,980	279	470.1	738	102.4	≥ 1,980	296	475.8	733	100.0
7	≥ 1,980	279	473.3	738	110.6	≥ 1,980	296	481.8	733	108.1
8	≥ 2,020	279	469.3	738	110.5	≥ 2,020	296	479.5	733	108.0
9	≥ 3,930	297	443.5	731	97.8	≥ 3,930	309	451.2	733	93.5
10	≥ 2,000	297	482.5	731	105.0	≥ 2,000	309	487.3	733	101.3
11	≥ 1,430	297	522.8	731	103.9	≥ 1,430	309	526.9	733	100.7
12	≥ 740	297	544.3	731	97.9	≥ 740	309	547.1	733	94.3

* Domains with Exemption 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	≥ 25,880	339	501.6	711	85.3	≥ 25,880	345	475.8	684	57.7
K	≥ 2,720	339	488.7	711	92.9	≥ 2,730	347	487.1	684	70.1
1	≥ 4,380	310	461.5	669	98.7	≥ 4,380	283	459.7	698	92.2
2	≥ 2,970	292	443.9	703	105.2	≥ 2,970	276	445.7	737	96.3
3	≥ 2,660	292	461.2	703	117.7	≥ 2,660	276	475.6	737	110.5
4	≥ 2,330	270	467.1	780	136.9	≥ 2,330	268	465.9	787	114.9
5	≥ 2,160	248	469.4	786	143.8	≥ 2,160	248	473.9	797	124.2
6	≥ 1,980	296	469.7	732	117.2	≥ 1,980	281	469.0	741	104.1
7	≥ 1,980	296	471.9	732	123.6	≥ 1,980	281	473.5	741	112.2
8	≥ 2,020	296	464.8	732	122.9	≥ 2,020	281	470.5	741	112.3
9	≥ 3,930	332	453.1	722	100.9	≥ 3,930	315	451.1	732	91.0
10	≥ 2,000	332	490.2	722	108.8	≥ 2,000	315	486.2	732	97.4
11	≥ 1,430	332	533.6	722	103.0	≥ 1,430	315	523.2	732	97.5
12	≥ 740	332	553.7	722	96.1	≥ 740	315	541.7	732	91.7

* Domains with Exemption 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	≥ 25,890	3978	5314.1	6375	495.5	≥ 25,890	3646	5065.4	6763	523.0
K	≥ 2,730	3978	5294.3	6375	568.0	≥ 2,730	3646	5049.6	6763	623.0
1	≥ 4,380	3785	5022.3	6387	627.5	≥ 4,380	3364	4812.9	6629	742.1
2	≥ 2,970	3756	4902.7	6439	643.6	≥ 2,970	3326	4689.8	6880	763.1
3	≥ 2,660	3756	5055.0	6439	717.4	≥ 2,660	3326	4882.1	6880	872.6
4	≥ 2,330	3649	4917.3	6700	717.0	≥ 2,330	3237	4826.5	7323	943.6
5	≥ 2,160	3649	4956.0	6700	775.0	≥ 2,160	3151	4873.5	7401	1014.1
6	≥ 1,980	3803	4965.8	6476	690.8	≥ 1,980	3388	4860.2	6974	842.7
7	≥ 1,980	3803	4998.5	6476	745.7	≥ 1,980	3388	4892.2	6974	905.7
8	≥ 2,020	3803	4974.6	6476	739.6	≥ 2,020	3388	4858.7	6974	903.9
9	≥ 3,930	3787	4768.6	6524	677.4	≥ 3,930	3605	4691.7	6923	758.4
10	≥ 2,000	3787	5038.1	6524	735.4	≥ 2,000	3605	4989.8	6923	819.2
11	≥ 1,430	3787	5320.3	6524	736.3	≥ 1,430	3605	5314.2	6923	800.1
12	≥ 740	3787	5479.6	6524	699.9	≥ 740	3605	5476.4	6923	747.2

* Scale scores cannot be compared across grade bands.

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

Grade	Listening							Reading						
	N	0	1	2	3	4	5	N	0	1	2	3	4	5
Pre-K	≥ 26,880	3.7	22.3	18.2	49.5	3.4	2.9	≥ 26,880	3.7	26.1	21.3	40.0	4.7	4.1
K	≥ 2,940	7.4	27.4	14.0	43.6	3.2	4.4	≥ 2,950	7.4	30.9	14.8	36.7	4.3	5.9
1	≥ 4,640	5.5	32.6	7.8	30.8	10.3	13.0	≥ 4,640	5.5	59.6	10.2	12.3	5.5	6.9
2	≥ 3,160	6.0	28.8	11.5	25.0	14.6	14.0	≥ 3,160	6.0	56.4	7.3	15.6	5.1	9.5
3	≥ 2,810	5.5	27.4	13.5	21.3	16.6	15.7	≥ 2,810	5.5	54.0	11.5	15.1	5.7	8.2
4	≥ 2,480	6.4	33.9	7.6	13.7	17.3	21.1	≥ 2,480	6.4	47.6	8.5	14.1	8.6	14.9
5	≥ 2,320	6.8	39.2	7.6	7.6	17.4	21.4	≥ 2,320	6.8	49.9	8.5	14.4	6.9	13.5
6	≥ 2,260	12.3	34.6	7.9	10.1	15.9	19.3	≥ 2,260	12.3	46.6	5.9	15.4	8.4	11.3
7	≥ 2,260	12.5	41.3	6.7	13.9	9.7	15.9	≥ 2,260	12.5	48.4	9.0	14.9	7.6	7.6
8	≥ 2,310	12.6	43.7	8.3	13.7	9.9	11.7	≥ 2,310	12.6	51.9	9.5	17.5	4.2	4.2
9	≥ 4,970	20.8	51.5	6.1	10.1	4.6	6.9	≥ 4,960	20.8	56.9	6.6	10.8	2.5	2.4
10	≥ 2,240	10.7	42.2	7.6	16.8	9.4	13.1	≥ 2,240	10.7	48.7	12.3	18.1	5.6	4.5
11	≥ 1,520	6.5	28.3	9.0	21.5	12.8	22.0	≥ 1,520	6.5	35.5	13.7	26.2	8.9	9.3
12	≥ 780	5.1	19.6	9.1	22.1	16.5	27.7	≥ 780	5.1	26.7	14.6	32.8	10.2	10.6
Total	≥ 61,640	7.1	30.1	13.0	32.7	7.8	9.2	≥ 61,640	7.1	39.5	14.7	27.1	5.3	6.1

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

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

Grade	Speaking							Writing						
	N	0	1	2	3	4	5	N	0	1	2	3	4	5
Pre-K	≥ 26,880	3.7	40.8	20.7	21.8	5.0	7.9	≥ 26,880	3.7	67.9	21.4	5.3	1.1	0.6
K	≥ 2,940	7.4	46.6	14.7	17.4	7.7	6.3	≥ 2,950	7.4	56.7	22.2	10.5	1.9	1.3
1	≥ 4,640	5.5	65.8	16.4	2.8	3.4	6.0	≥ 4,640	5.5	66.4	9.4	10.4	3.9	4.4
2	≥ 3,160	6.0	58.6	13.5	6.9	5.0	10.0	≥ 3,160	6.0	56.6	10.7	11.6	5.0	10.1
3	≥ 2,810	5.5	54.9	9.9	8.5	9.1	12.1	≥ 2,810	5.5	55.6	9.6	13.8	5.9	9.6
4	≥ 2,480	6.4	45.0	7.8	10.6	8.0	22.4	≥ 2,480	6.4	44.5	8.4	19.6	6.5	14.7
5	≥ 2,320	6.8	49.8	7.5	8.5	5.9	21.5	≥ 2,320	6.8	44.5	7.7	21.5	5.6	13.9
6	≥ 2,260	12.3	42.3	8.0	16.6	6.4	14.4	≥ 2,260	12.3	39.8	8.2	20.1	7.0	12.7
7	≥ 2,260	12.5	44.7	9.4	13.5	5.4	14.5	≥ 2,260	12.5	47.5	8.3	16.2	5.2	10.3
8	≥ 2,310	12.6	47.3	9.2	14.6	4.8	11.5	≥ 2,310	12.6	51.1	9.0	16.2	4.3	6.8
9	≥ 4,970	20.8	52.1	7.6	10.0	3.3	6.2	≥ 4,970	20.8	56.8	6.6	9.7	2.3	3.7
10	≥ 2,240	10.7	45.2	9.1	16.3	5.9	12.8	≥ 2,240	10.7	48.5	11.2	18.2	4.9	6.4
11	≥ 1,520	6.5	30.9	11.0	21.6	9.3	20.8	≥ 1,520	6.5	34.7	13.7	25.6	7.6	12.0
12	≥ 780	5.1	22.1	12.8	21.7	11.9	26.4	≥ 780	5.1	26.8	14.3	30.7	9.1	14.0
Total	≥ 61,630	7.1	46.1	15.1	15.9	5.5	10.3	≥ 61,640	7.1	58.9	15.1	10.8	3.1	4.8

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

Table 2.8 Percentage of Students in Each Overall Proficiency Category by Grade

Grade	N	Emerging	Progressing	Proficient	Proficiency Not Demonstrated
Pre-K	≥ 26,880	35.4	55.1	5.8	3.7
K	≥ 2,950	37.7	51.9	2.9	7.4
1	≥ 4,640	39.9	48.7	5.9	5.5
2	≥ 3,160	40.1	42.1	11.8	6.0
3	≥ 2,810	40.8	41.6	12.1	5.5
4	≥ 2,480	41.1	33.0	19.5	6.4
5	≥ 2,320	46.4	29.2	17.7	6.8
6	≥ 2,260	41.1	31.8	14.8	12.3
7	≥ 2,260	47.1	28.6	11.7	12.5
8	≥ 2,310	50.5	29.4	7.6	12.6
9	≥ 4,970	55.9	19.0	4.3	20.8
10	≥ 2,240	47.9	33.2	8.2	10.7
11	≥ 1,520	34.7	43.6	15.2	6.5
12	≥ 780	26.4	51.1	17.3	5.1
Total	≥ 61,650	40.2	44.5	8.2	7.1

2.3 2021–2022 TESTING TIME FOR ONLINE SCREENER TESTS

In the 2021–2022 online screener tests, students who did not have domain exemption were advanced to Segments 2 and 3 (Step 2) and were advanced to Segment 4 (Step 3) if their raw scores met or exceeded the threshold score for Step 2 (Table 1.2). Therefore, students who completed Step 3 took more items than those who stopped at Step 2. Table S7.1 in the Appendix summarizes testing time by end step in each grade and grade band. Students who had any non-attempted or exempted domains or had Proficiency Not Demonstrated are excluded. As expected, students who ended the test at Step 3 had longer testing times than those who ended at Step 2. In addition, upper-grade tests had longer testing times than lower-grade tests due to the tests being longer and the items being more complex.

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.

The results for each state are illustrated in the following sections of the Appendix:

- Section 8. Screener Assessment—Marginal Reliability

  o Figure S8.1 shows the ratio of MSEM to the standard deviation of scale scores at the test level by domain and grade.

  o Figure S8.2 presents the marginal reliability for each domain test across grades.

- Section 9. Screener Assessment—Conditional Standard Error of Measurement (CSEM)

  o Figures S9.1–S9.14 show the CSEM plots for each domain, overall, and comprehension score. If an ELPA21 test applies to multiple grades, the CSEM plots are broken down by grade. Scores can be computed from tests that end at Step 2 or Step 3. Because students stopping after Step 2 completed a shorter test, it is expected that these students' scores would have a greater error. The CSEM plots use different colors to differentiate the students who ended the test after Step 2 from those who completed Step 3.

- Section 10. Screener Assessment—Classification Accuracy and Consistency

  o Figure S10.1 shows the CA for each domain test.

  o Figure S10.2 shows the CC for each domain test.

  o Figure S10.3 presents the CA and CC for the overall proficiency.

- Section 11. Screener Assessment—Inter-Rater Analysis

  o Tables S11.1–S11.7 display the inter-rater analysis result for each handscored item in each grade.

3.1 MARGINAL STANDARD ERROR OF MEASUREMENT

As described in Part I of this technical report, the MSEM is a way to examine score reliability. The ratio of MSEM to the standard deviation of scale scores can also indicate the measure errors. The analysis for the ratio is displayed in Figure S8.1 in the Appendix.

3.2 MARGINAL RELIABILITY

The marginal reliability for the pooled analysis is presented in Table 3.1 and is plotted in Figure S8.2 in the Appendix. Pre-K and kindergarten have lower marginal reliability than the other grades. Writing has lower marginal reliability at pre-K and grades 9–12 but has higher reliability for grades 3 and 4. Listening has relatively lower reliability than the other domains in grades 1–5. In addition, Section 9 of the Appendix displays CSEM plots by domain and grade.

Table 3.1 Marginal Reliability by Score and Grade*

Grade	N	Listening	Reading	Speaking	Writing	Comprehension	Overall
Pre-K	≥ 25,880	.77	.74	.81	.68	.70	.75
K	≥ 2,720	.81	.79	.82	.74	.74	.80
1	≥ 4,380	.82	.87	.85	.87	.74	.87
2	≥ 2,970	.85	.91	.88	.91	.80	.91
3	≥ 2,660	.87	.92	.90	.93	.83	.92
4	≥ 2,330	.91	.92	.92	.93	.86	.94
5	≥ 2,160	.92	.93	.93	.93	.88	.94
6	≥ 1,980	.92	.91	.92	.91	.88	.93
7	≥ 1,980	.93	.92	.92	.92	.89	.94
8	≥ 2,020	.93	.92	.92	.92	.89	.94
9	≥ 3,930	.93	.91	.89	.86	.90	.91
10	≥ 2,000	.93	.92	.91	.88	.90	.92
11	≥ 1,430	.92	.91	.91	.88	.89	.92
12	≥ 740	.91	.90	.90	.87	.87	.91

* Domains with Exemption or Not Attempted are excluded.

3.3 CLASSIFICATION ACCURACY AND CONSISTENCY

Table 3.2 presents overall CA and 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 results for each cut score are presented in Table 3.3 and Table 3.4 as well as Figure S10.1 and Figure S10.2 in the Appendix. Across the four performance cut scores, the CA indices were all above 0.8, denoting that the degree to which we can reliably differentiate students between adjacent performance levels is typically above or close to 0.8. In terms of CC, the indices were all above 0.74 in all cut scores and all grades. The reliability indices in the middle school tests were above 0.87 for all domains. Table 3.5 and Figure S10.3 in the Appendix display the CA and CC for overall proficiency categories. The plot shows that all the accuracy and consistency indices were above 0.79. Both accuracy and consistency indices for between Emerging and Progressing were

lower than those for between Progressing and Proficient in pre-K to grade 3 and are comparable with those for between Progressing and Proficient in the other grades.

Table 3.2 Overall Classification Accuracy and Consistency for Domain Performance Levels by Domain and Grade*

Grade	Accuracy				Consistency			
	Listening	Reading	Speaking	Writing	Listening	Reading	Speaking	Writing
Pre-K	.68	.61	.65	.76	.58	.51	.57	.67
K	.71	.64	.68	.72	.61	.54	.62	.64
1	.66	.78	.77	.82	.56	.72	.72	.77
2	.65	.81	.75	.81	.55	.75	.71	.75
3	.66	.79	.75	.81	.56	.74	.70	.76
4	.73	.79	.77	.80	.65	.73	.72	.73
5	.77	.81	.80	.81	.69	.75	.75	.75
6	.76	.79	.77	.76	.68	.73	.71	.68
7	.79	.80	.80	.81	.73	.75	.74	.75
8	.80	.83	.81	.83	.73	.78	.75	.78
9	.85	.88	.83	.85	.80	.84	.78	.79
10	.79	.81	.78	.77	.72	.76	.71	.71
11	.75	.75	.72	.71	.67	.67	.64	.64
12	.72	.71	.69	.69	.63	.62	.61	.60

* Domains with Exemption 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	.90	.84	.94	.97		.87	.82	.91	.95		.87	.87	.92	.94		.82	.95	.98	.99
K	.91	.87	.93	.96		.89	.86	.89	.94		.89	.89	.91	.94		.83	.93	.97	.98
1	.91	.90	.88	.91		.91	.93	.95	.96		.88	.92	.93	.95		.94	.94	.95	.96
2	.89	.91	.89	.92		.94	.94	.96	.96		.90	.91	.92	.95		.93	.94	.96	.97
3	.90	.93	.90	.91		.94	.93	.94	.95		.93	.92	.92	.93		.95	.95	.95	.96
4	.93	.94	.92	.92		.94	.94	.95	.95		.95	.92	.92	.93		.94	.94	.95	.95
5	.94	.95	.94	.93		.95	.95	.95	.95		.95	.94	.94	.93		.95	.95	.95	.95
6	.93	.96	.94	.93		.95	.95	.94	.94		.95	.92	.92	.94		.91	.95	.94	.95
7	.95	.96	.94	.93		.96	.95	.93	.94		.95	.94	.94	.95		.96	.95	.94	.95
8	.95	.96	.94	.94		.96	.95	.95	.96		.96	.94	.94	.95		.95	.95	.95	.96
9	.95	.97	.96	.97		.96	.96	.97	.98		.94	.96	.95	.96		.94	.95	.97	.97
10	.95	.95	.94	.94		.95	.94	.95	.96		.94	.95	.93	.94		.92	.93	.95	.96
11	.95	.95	.92	.91		.95	.92	.92	.93		.94	.93	.90	.91		.92	.91	.92	.93
12	.95	.95	.91	.89		.94	.92	.90	.92		.94	.93	.89	.90		.92	.90	.91	.92

* Domains with Exemption or Not Attempted are excluded.
* Cuts 1 to 4 fall between performance levels 1 and 2, 2 and 3, 3 and 4, 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	.77	.91	.95		.81	.75	.87	.93		.82	.82	.88	.91		.75	.93	.98	.98
K	.87	.82	.89	.94		.84	.80	.85	.91		.84	.85	.88	.91		.76	.89	.96	.97
1	.87	.85	.84	.87		.87	.90	.93	.94		.84	.88	.90	.93		.91	.92	.93	.95
2	.85	.87	.85	.89		.91	.91	.93	.95		.87	.87	.89	.92		.90	.91	.95	.95
3	.86	.90	.86	.88		.92	.91	.92	.94		.90	.88	.88	.90		.93	.92	.93	.94
4	.89	.91	.89	.90		.91	.92	.92	.93		.92	.90	.89	.90		.92	.92	.92	.93
5	.91	.93	.91	.90		.93	.93	.93	.93		.92	.92	.91	.91		.92	.93	.92	.93
6	.89	.93	.92	.90		.92	.93	.91	.92		.93	.90	.89	.92		.87	.92	.92	.93
7	.93	.94	.91	.90		.94	.93	.91	.92		.93	.91	.91	.92		.94	.93	.91	.93
8	.93	.95	.91	.92		.94	.93	.93	.94		.94	.91	.92	.93		.94	.93	.93	.94
9	.93	.95	.95	.95		.94	.94	.96	.97		.92	.94	.93	.95		.90	.93	.96	.96
10	.93	.93	.92	.92		.93	.92	.93	.94		.92	.92	.90	.92		.88	.90	.93	.94
11	.93	.93	.89	.88		.92	.90	.89	.91		.91	.91	.86	.87		.89	.87	.89	.90
12	.93	.93	.87	.85		.92	.88	.87	.89		.92	.90	.85	.85		.89	.86	.87	.89

* Domains with Exemption or Not Attempted are excluded.

* Cuts 1 to 4 fall between performance levels 1 and 2, 2 and 3, 3 and 4, 4 and 5, respectively.

Table 3.5 Screener Classification for Overall Proficiency Classifications by Grade

Grade	Accuracy				Consistency		
	Overall	Between Emerging and Progressing	Between Progressing and Proficient		Overall	Between Emerging and Progressing	Between Progressing and Proficient
Pre-K	.83	.87	.96		.79	.82	.97
K	.87	.89	.98		.83	.85	.98
1	.87	.90	.96		.82	.86	.95
2	.88	.92	.96		.83	.88	.95
3	.88	.93	.95		.84	.90	.94
4	.89	.95	.95		.86	.93	.93
5	.90	.95	.95		.87	.94	.93
6	.90	.95	.95		.86	.93	.93
7	.90	.96	.95		.88	.94	.93
8	.91	.95	.96		.89	.94	.95
9	.94	.96	.98		.92	.95	.97
10	.91	.95	.96		.88	.94	.95
11	.88	.95	.93		.85	.93	.91
12	.87	.95	.92		.83	.93	.90

3.4 INTER-RATER ANALYSIS

In the 2021–2022 screener tests, two to four handscored items in kindergarten to grade band 4–5 online tests and nine handscored items in each of the middle school (grade band 6–8) and high school (grade band 9–12) online tests had second rater scores. Around 10% of the responses to the handscored items were scored by a second rater. Table 3.6 contains the number of items in each grade or grade band, the ranges of Cohen's kappa (for items with a max score of 1 point) or quadratic weighted kappa (QWK) (for items with a max score of 2 or more points), the percentage of exact matches, the percentage of within one agreement, and the percentage of more than one agreement for the pooled analysis. The weighted kappa coefficients were all above 0.62, except for one item in grade 1, four items in grade band 6–8, and four items in grade band 9–12. Overall, 60.6%–91.1% of handscores were consistent (exact agreement) between the first rater and the second rater, and 100% of handscores agreed within one score point.

The inter-rater consistencies were also assessed by item and are summarized in Section 11 of the Appendix.

Table 3.6 Summary of Kappa Coefficients by Grade Band

Grade/Grade Band	Number of Items	Weighted Kappa		% Exact Agreement		% within 1 Agreement		% Not within 1 Agreement	
		Min	Max	Min	Max	Min	Max	Min	Max
Pre-K	2	.823	.870	72.9	80.5	100.0	100.0	0.0	0.0
K	2	.741	.870	68.2	79.2	100.0	100.0	0.0	0.0
1	2	.650	.789	66.1	72.9	100.0	100.0	0.0	0.0
2–3	3	.798	.841	67.5	78.4	100.0	100.0	0.0	0.0
4–5	4	.758	.913	60.6	83.0	100.0	100.0	0.0	0.0
6–8	9	.548	.923	67.5	90.6	100.0	100.0	0.0	0.0
9–12	9	.622	.950	60.7	91.1	100.0	100.0	0.0	0.0

Chapter 4. Validity

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

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

- Section 12. Correlations Between Summative and Screener Tests

    o Table S12.1 shows the correlations between domain, overall, and comprehension scores.

    o Table S12.2 summarizes the correlations by between domain performance level and overall proficiency categories.

- Section 13. Student Progress from Screener to Summative

    o Figures S13.1–S13.2 display within-year average differences in domain, overall, and comprehension scale score.

    o Figures S13.3–S13.4 present changes in domain performance level and overall proficiency.

    o Figures S13.5–S13.10 show scatter plots of scale scores for the screener and summative assessments.

    o Tables S13.1–S13.6 summarize the comparison of scale score summary statistics between domain, overall, and comprehension scores.

4.1 COMPARISONS OF PERFORMANCE FROM SCREENER TO SUMMATIVE

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

The correlation between the scale scores from summative and screener tests was assessed using Pearson correlations. The correlation between the performance levels from both tests was assessed using Goodman and Kruskal's Gamma correlation (Goodman & Kruskal, 1954). 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 Appendix.

Table S12.1 shows the Pearson correlation between the screener and the summative assessments in domain and composite scores. Correlations of all types of scores were the lowest in the kindergarten test, followed by the grade 1 test except for speaking; the correlations were above 0.79 in listening, reading, writing, comprehension, and overall scale scores in grades 2 and above. The speaking tests had relatively lower correlations than the other three domains at the grades 6–12 levels.

Table S12.2 shows the Gamma correlations between domain performance levels and test proficiency categories. Similar to the correlations between scale scores presented in Table S12.1, kindergarten had the lowest correlations in all domain performance levels and overall proficiency categories. For grade 2 and above, the correlations were about 0.8. In addition, the correlations between overall proficiency categories were generally higher than those between domain performance levels. This is because there are three levels in overall proficiency while there are five levels in domain performance. These correlations show predictive validity between the two ELPA21 tests because they were given to the same students at different times.

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. The major confounding factor in this result is measurement error in both assessments. Given the acceptable marginal reliability indices described in Part I and Part II of this technical report, we can still see the trend of student progress. Section 13 of the Appendix summarizes the results of progress analysis. Only students who had valid scores on both the screener and summative assessments were included in each of the analyses.

Figures S13.1 and S13.2 in the Appendix show the growth of the average domain scores and composite scores, respectively. The average scale scores in the summative assessment were, in general, 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 shows the screener test and the right bar 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. In addition, Figures S13.5–S13.10 in the Appendix present scatter plots of scale score changes from screener to summative for each grade, and Tables 13.1–S13.6 summarize comparisons of scale scores between screener and summative assessments.

Chapter 5. Reporting

A detailed introduction to the Centralized Reporting System can be found in Part I, Chapter 6, of this technical report. The reporting mock-ups for the screener tests of each state appear in Section 14 of the Appendix. It is noted that the mock-up for score reports is not included in the Appendix for the pooled analysis.

References

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

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

Goodman, L, & Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*(268), 732–764.