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

2022–2023 Administration

Submitted to:
ELPA21

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

April 2024

Table of Contents

List of Tables

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 2022–2023 screener testing windows for the six states discussed in this report are shown in Table 1.1.

Table 1.1 2022–2023 ELPA21 Screener Testing Windows by State

State	ELPA21 Screener
Arkansas	8/2/2022–7/14/2023
Iowa	8/1/2022–7/14/2023
Louisiana	8/1/2022–7/14/2023
Nebraska	8/2/2022–7/14/2023
Ohio	8/3/2022–6/30/2023
West Virginia	8/2/2022–6/19/2023

1.2 Test Design

Each 2022–2023 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. The difference between the two screeners is in the proficiency determination rules, not in the content. The OELPS-BK is the Ohio English Language Proficiency Screener for the Beginning of Kindergarten. Students enrolling in kindergarten in the first half of the kindergarten year (on or before December 31) are administered the OELPS-BK. Kindergarteners taking the OELPS-BK will be proficient (not an English learner) if they earn domain levels of 3 or higher in all nonexempt domains of the screener.

The OELPS-K is the Ohio English Language Proficiency Screener administered to kindergarteners enrolling in the latter half of the kindergarten year (after December 31). Kindergartners taking the OELPS-K will be proficient (not an English learner) if they earn domain levels of 4 or higher in all nonexempt domains of the screener.

The OELPS-BK and OELPS-K are the same assessment that only differ in the definition of proficiency. In the OELPS-BK, *Proficient* is defined as achieving Level 3 or above in all non-exempted domains, while in the OELPS-K, *Proficient* is defined as achieving Level 4 or above in all non-exempted domains.

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 student should not proceed, the student is directed to Segment 9, and then the test ends. Additional details will 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 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.

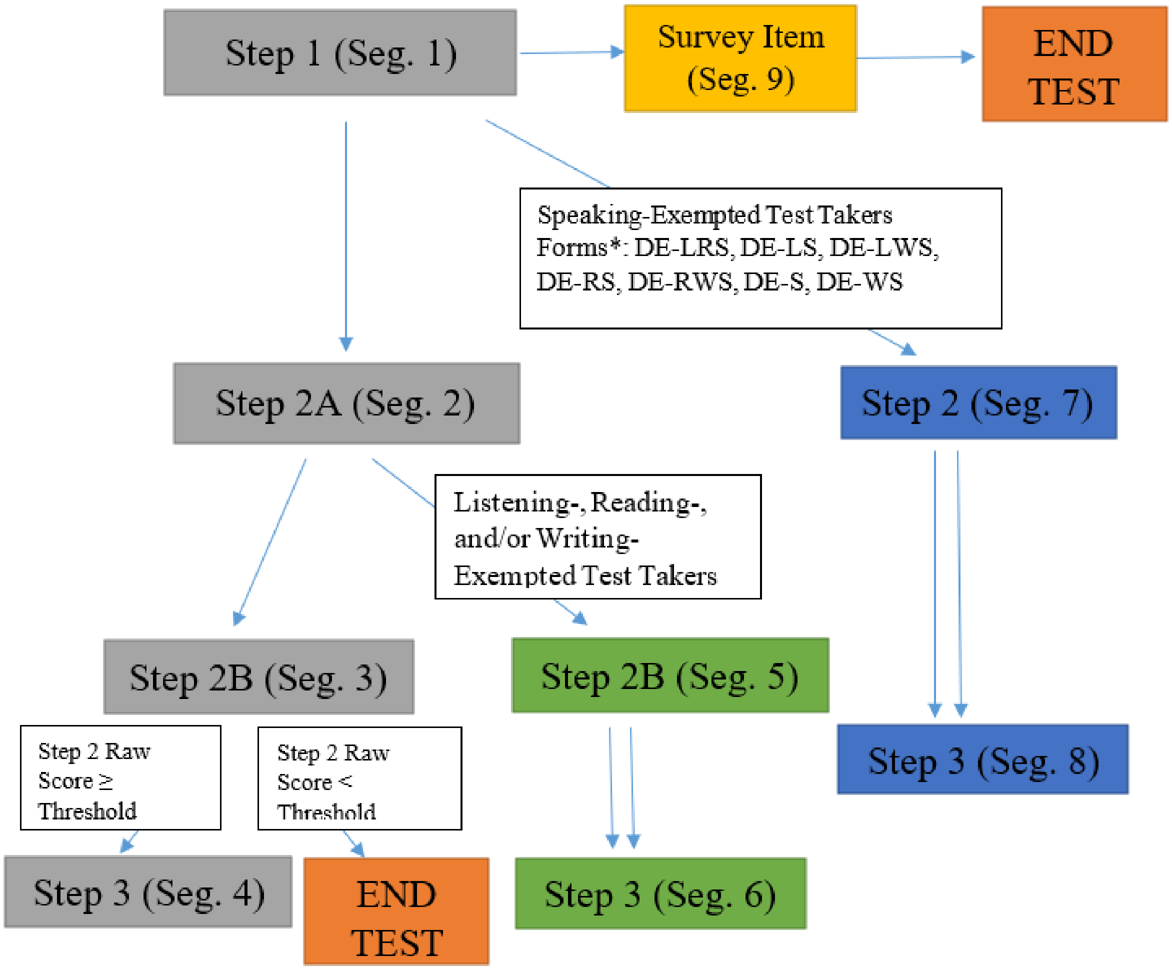
- Segment 6 (Step 3 for students who are exempted from the listening, reading, and/or writing domains) consists of items from all non-exempted domains.

- 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 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 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 2022–2023 ELPA21 Screener Online Test Design


Step 1 (Seg. 1)
Survey Item (Seg. 9)
END TEST
Speaking-Exempted Test Takers Forms*: DE-LRS, DE-LS, DE-LWS, DE-RS, DE-RWS, DE-S, DE-WS
Step 2A (Seg. 2)
Step 2 (Seg. 7)
Listening-, Reading-, and/or Writing-Exempted Test Takers
Step 2B (Seg. 3)
Step 2B (Seg. 5)
Step 3 (Seg. 8)
Step 2 Raw Score ≥ Threshold
Step 2 Raw Score < Threshold
Step 3 (Seg. 4)
END TEST
Step 3 (Seg. 6)

* 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

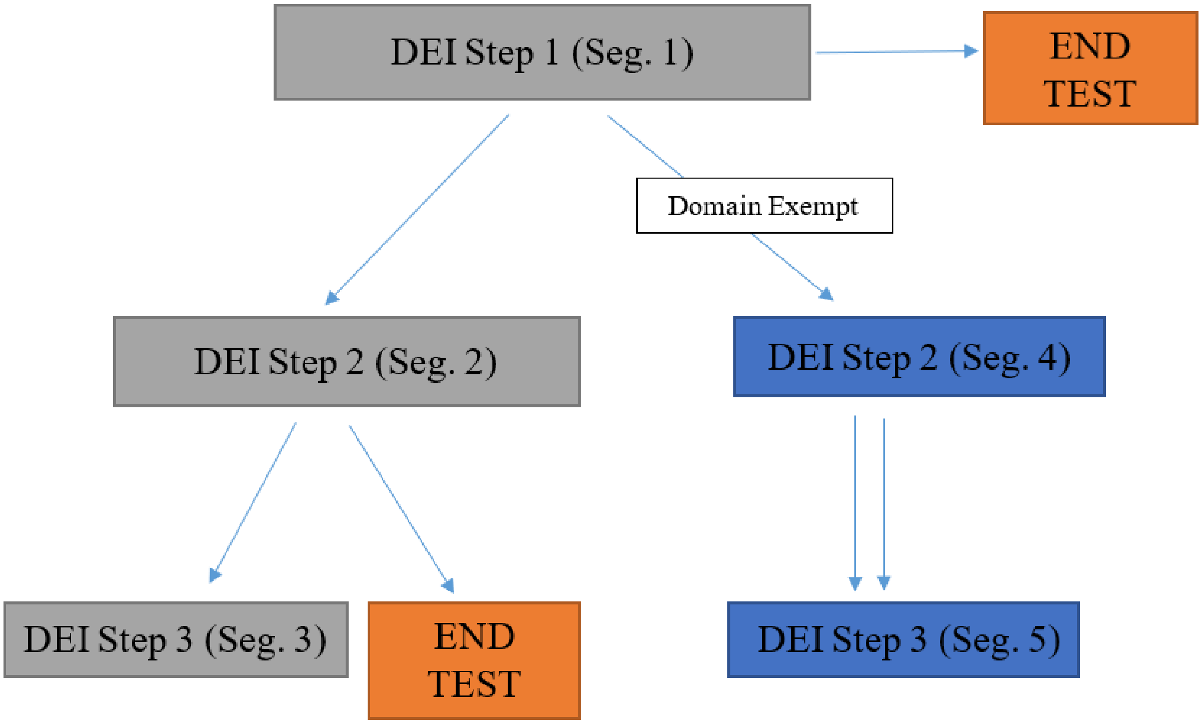
Grade Band	Threshold Score	Step 2 Max Score
Pre-K	20	26
K	23	26
1	24	27
2–3	25	28
4–5	26	31
6–8	28	33
9–12	26	30

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 2022–2023 ELPA21 Screener Paper Test Design



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 Specifically, Table 1.3 includes items from Segments 2–4, Table 1.4 includes Segments 2–3, and Table 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	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 2022–2023 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 included 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) 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 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.

1.3.3 Business Scoring Rules for the Screener Assessment

Business rules and instructions applied to the 2022–2023 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. 2022–2023 Summary

The 2022–2023 screener results are presented in this chapter and in Sections 1–14 of the Appendix for Pooled Analysis – 2022–2023 Summary Screener. The figures and tables included in each section are listed below:

- Section 1. Screener Assessment—Student Participation
  - Table S1.1 displays the number and percentage of students in each testing mode of braille, paper-pencil, and online in each grade (pre-K–12) and across the state.
  - 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 grade, gender, and ethnicity.
- Section 2. Screener Assessment—Raw Score Summary
  - 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 S2.1–S2.14 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.
  - 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
  - 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
  - Tables S4.1–S4.14 present the number of students, 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.
  - 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
  - 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).
  - 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
  - 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).
  - 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
  - Table S7.1 shows the testing time by each step in each grade/grade band.

2.1 2022–2023 Student Participation

Table 2.1 shows the overall student participation for each state. There were 66,323 students who took the 2022–2023 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 2022–2023 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 pre-K–12, the number of students tested decreased as the grade level increased though the increase is not monotonic. There were more male students (48.5%–51.3%) than female students (45.0%–47.8%) tested. In each test, the greatest number of participating students were in the group of Hispanic or Latino (47.0%–57.7%), followed by Asian students (7.7%–17.1%), and white students (9.1%–12.7%).

Table 2.1 Number of Students Who Participated in ELPA21 Screener in 2021–2022 and 2022–2023 by State and Grade

Grade	Arkansas		Iowa		Louisiana		Nebraska		Ohio		West Virginia		Total		
	2021–22	2022–23	2021–22	2022–23	2021–22	2022–23	2021–22	2022–23	2021–22	2022–23	2021–22	2022–23	2021-22	2022–23	Diff
Pre-K	≥3,570	≥4,150	≥4,860	≥4,940	≥3,760	≥3,860	≥3,550	≥3,790	≥10,780	≥10,840	≥190	≥250	≥26,910	≥27,850	938
K	≥1,220	≥890	≥280	≥240	≥300	≥350	≥110	≥130	≥780	≥760	≥80	≥40	≥2.960	≥2,430	-528
1	≥580	≥510	≥610	≥600	≥470	≥880	≥380	≥380	≥1,980	≥2, 180	≥80	≥60	≥4,650	≥4,630	-16
2	≥430	≥440	≥400	≥490	≥300	≥580	≥280	≥350	≥1,350	≥1,600	≥60	≥80	≥3,180	≥3,560	384
3	≥380	≥390	≥340	≥450	≥290	≥550	≥240	≥310	≥1,220	≥1,520	≥60	≥60	≥2,820	≥3,310	483
4	≥400	≥370	≥340	≥420	≥210	≥480	≥230	≥300	≥990	≥1,360	≥50	≥50	≥2,500	≥3,010	506
5	≥330	≥340	≥310	≥420	≥220	≥470	≥210	≥250	≥960	≥1,360	≥30	≥50	≥2,320	≥2,920	598
6	≥360	≥390	≥310	≥370	≥190	≥400	≥180	≥220	≥890	≥1,180	≥30	≥60	≥2,280	≥2,640	360
7	≥420	≥360	≥250	≥320	≥160	≥470	≥170	270	≥910	≥1,170	≥40	≥50	≥2,290	≥2,660	374
8	≥370	≥380	≥290	≥340	≥160	≥480	≥180	≥250	≥960	≥1,150	≥40	≥50	≥2,370	≥2,670	303
9	≥680	≥660	≥670	≥680	≥280	≥1,230	≥520	≥750	≥1,740	≥1,990	≥60	≥80	≥4,980	≥5,420	433
10	≥490	≥460	≥380	≥410	≥110	≥340	≥130	≥270	≥850	≥1,000	≥60	≥70	≥2,250	≥2,570	324
11	≥410	≥450	≥210	≥240	≥60	≥180	≥70	≥160	≥580	≥610	≥50	≥50	≥1,530	≥1,720	190
12	≥250	≥210	≥90	≥110	≥20	≥40	≥50	≥120	≥300	≥340	≥30	≥40	≥800	≥870	75
Total	≥9,960	≥10,060	≥9,390	≥10,070	≥6,610	≥10,350	≥6,360	≥7,610	≥24,350	≥27,140	≥920	≥1,060	≥61,890	≥66,320	4424

Table 2.2 Number of Students Participating in 2022–2023 ELPA21 Summative, Screener Tests, and Both by State and Grade Band

State	Grade Band	N Summative	N Screener	N Both
Arkansas	PreK and K	≥4,380	≥5,040	≥3,580
	1	≥4,500	≥510	≥390
	2-3	≥7,400	≥840	≥600
	4-5	≥5,680	≥720	≥490
	6-8	≥7,780	≥1,150	≥860
	9-12	≥9,860	≥1,800	≥1,280
Iowa	PreK and K	≥4,830	≥5,180	≥3,640
	1	≥4,300	≥600	≥420
	2-3	≥6,700	≥940	≥620
	4-5	≥4,870	≥840	≥500
	6-8	≥5,700	≥1,040	≥680
	9-12	≥7,980	≥1,450	≥1,010
Louisiana	PreK and K	≥4,030	≥4,210	≥2,840
	1	≥4,360	≥880	≥690
	2-3	≥6,580	≥1,130	≥880
	4-5	≥4,850	≥960	≥680
	6-8	≥6,090	≥1,350	≥1,080
	9-12	≥7,070	≥1,790	≥1,320
Nebraska	PreK and K	≥3,890	≥3,930	≥2,660
	1	≥3,850	≥380	≥240
	2-3	≥5,830	≥660	≥398
	4-5	≥3,750	≥560	≥330
	6-8	≥3,900	≥760	≥480
	9-12	≥5,300	≥1,310	≥850
Ohio	K	≥10,580	≥11,610	≥9,510
	1	≥10,570	≥2,180	≥1,690
	2-3	≥15,880	≥3,130	≥2,260
	4-5	≥11,180	≥2,730	≥1,760

State	Grade Band	N Summative	N Screener	N Both
	6-8	≥12,610	≥3,500	≥2,400
	9-12	≥15,230	≥3,970	≥2,800
West Virginia	PreK and K	≥230	≥300	≥220
	1	≥230	≥60	≥40
	2-3	≥430	≥150	≥90
	4-5	≥290	≥110	≥50
	6-8	≥400	≥160	≥90
	9-12	≥610	≥260	≥170

2.2 2022–2023 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 percentage of students in each performance level for each of the grade levels. They also provide the total number of test-takers per grade level. These results are presented separately for all four domains. The results indicate that performance level 1 is the most frequent level achieved in speaking and writing for grades pre-K–12, in reading for grades 1–12, and in listening for grades 1–11. All four domains displayed similar trends with respect to students who failed to reach level 1, as indicated by the "0" column for each domain: Failure rates were low for grades pre-K–5, relatively higher for grades 6–9, and back down for grades 10–12. 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 in 2021–2022 for all states, Pre-K (or BK for Ohio) students are considered overall proficient with scores of all 3 or above in each domain rather than all 4 or above. For kindergarten and higher grades, students need to obtain a score of 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 (4.7%) to grade 4 (17.7%), slightly decrease in grade 9 (5.0%), and thereafter increase consistently. The percentages of students in the Emerging category are relatively stable until grade 4 (46.4%), increase from grade 4 to grade 5 (49.3%), then decrease from grade 5 to grade 6 (45.4%), consistently increase in grade 9 (58.1%), and thereafter decrease 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	≥26,750	314	509.1	714	67.1	≥26,740	318	506.1	708	66.4
K	≥2,260	314	503.6	714	79.2	≥2,260	318	502.1	708	78.6
1	≥4,390	288	478.8	678	89.3	≥4,390	286	462.2	704	88.0
2	≥3,390	276	462.7	710	83.0	≥3,390	278	447.5	734	90.2
3	≥3,140	286	479.3	710	96.0	≥3,140	278	471.0	734	105.2
4	≥2,810	270	457.3	778	110.7	≥2,810	270	461.7	795	108.1
5	≥2,740	270	468.3	778	119.9	≥2,740	270	474.1	795	117.3
6	≥2,350	279	464.8	738	102.8	≥2,350	296	471.6	733	99.7
7	≥2,360	279	465.5	738	103.1	≥2,360	296	473.9	733	100.9
8	≥2,370	279	464.0	738	108.5	≥2,370	296	474.5	733	105.7
9	≥4,430	302	444.2	731	97.7	≥4,430	305	450.3	733	94.2
10	≥2,330	302	485.0	731	103.5	≥2,330	312	490.3	733	99.2
11	≥1,620	302	520.0	731	103.6	≥1,620	312	524.5	733	99.5
12	≥840	302	534.6	731	101.1	≥840	312	538.2	733	98.4

* 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	≥26,740	338	498.2	711	86.7	≥26,740	336	474.2	684	56.8
K	≥2,260	339	482.1	711	93.6	≥2,260	347	488.3	684	71.0
1	≥4,390	310	456.4	669	97.0	≥4,390	283	457.3	698	88.8
2	≥3,390	280	437.6	703	102.7	≥3,390	272	442.9	737	91.7
3	≥3,140	292	453.6	703	117.6	≥3,140	276	468.6	737	106.7
4	≥2,810	270	455.1	786	137.2	≥2,810	268	456.3	797	115.4
5	≥2,740	270	461.8	786	143.8	≥2,740	268	469.9	797	124.0
6	≥2,350	296	463.4	732	116.8	≥2,350	281	465.1	741	105.4
7	≥2,360	296	462.2	732	116.6	≥2,360	281	466.1	741	105.7
8	≥2,370	296	459.2	732	119.1	≥2,370	281	466.5	741	110.3
9	≥4,430	300	455.1	722	101.9	≥4,430	304	451.8	732	91.7
10	≥2,330	333	495.6	722	106.5	≥2,330	318	489.4	732	97.4
11	≥1,620	334	529.3	722	104.0	≥1,620	318	521.3	732	97.1
12	≥840	334	546.1	722	101.0	≥840	318	533.7	732	96.0

* 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	≥26,750	3978	5288.9	6375	497.4	≥26,750	3646	5041.2	6763	526.3
K	≥2,260	3978	5243.1	6454	567.4	≥2,260	3646	5016.4	6763	633.6
1	≥4,390	3785	4997.3	6387	606.1	≥4,390	3364	4781.7	6629	717.6
2	≥3,390	3756	4870.4	6439	615.4	≥3,390	3325	4654.7	6880	731.4
3	≥3,140	3756	4994.8	6439	697.6	≥3,140	3326	4821.9	6880	854.7
4	≥2,810	3649	4858.6	6700	704.1	≥2,810	3237	4746.2	7401	944.5
5	≥2,740	3649	4933.3	6700	769.2	≥2,740	3237	4834.4	7401	1011.7
6	≥2,350	3803	4927.1	6476	683.9	≥2,350	3388	4819.4	6974	845.0
7	≥2,370	3803	4938.5	6476	694.3	≥2,360	3388	4824.6	6974	846.7
8	≥2,370	3803	4933.7	6476	725.6	≥2,370	3388	4818.0	6974	881.8
9	≥4,430	3818	4765.3	6522	674.9	≥4,430	3542	4696.6	6922	762.4
10	≥2,330	3818	5049.7	6522	715.7	≥2,330	3628	5018.3	6922	803.8
11	≥1,620	3818	5301.1	6522	724.5	≥1,620	3628	5291.0	6922	800.6
12	≥840	3818	5399.3	6522	707.4	≥840	3628	5406.7	6922	783.6

* 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	≥27,840	≤5	24.1	18.0	48.3	≤5	≤5	≥27,830	≤5	28.5	20.7	38.6	≤5	≤5
K	≥2,420	6.8	31.2	14.1	41.4	≤5	≤5	≥2,420	6.8	34.3	15.3	34.3	≤5	≤5
1	≥4,630	5.1	34.0	8.9	30.6	10.4	11.1	≥4,630	5.1	61.1	10.7	11.5	6.1	5.6
2	≥3,550	≤5	31.0	12.9	24.7	15.0	12.0	≥3,550	≤5	59.1	8.0	15.3	≤5	8.2
3	≥3,290	≤5	29.1	15.5	22.7	13.8	14.4	≥3,290	≤5	58.3	11.7	12.8	5.2	7.5
4	≥2,980	5.7	37.8	9.1	12.2	16.2	19.0	≥2,980	5.7	51.7	8.7	13.4	6.7	13.9
5	≥2,900	5.5	41.3	8.4	8.1	16.1	20.6	≥2,900	5.5	53.2	8.6	13.1	6.4	13.1
6	≥2,620	10.5	38.1	8.4	10.0	13.8	19.2	≥2,630	10.5	49.5	6.5	14.1	8.6	10.8
7	≥2,640	10.4	45.9	7.9	14.1	8.1	13.5	≥2,640	10.4	54.8	8.8	13.6	≤5	7.4
8	≥2,650	10.5	47.7	8.2	12.5	9.3	11.7	≥2,650	10.5	56.5	8.1	16.4	≤5	≤5
9	≥5,400	17.9	53.8	6.5	9.8	≤5	7.4	≥5,400	17.9	59.2	7.4	9.6	≤5	≤5
10	≥2,550	8.7	43.5	8.7	16.1	9.3	13.8	≥2,550	8.7	50.5	11.5	19.0	5.2	5.1
11	≥1,710	5.1	30.1	8.3	21.5	13.0	22.0	≥1,710	5.1	36.7	15.3	25.9	7.9	9.0
12	≥870	≤5	24.5	8.5	24.8	13.1	26.1	≥870	≤5	31.5	18.1	26.4	9.4	11.5
Total	≥66,120	6.43	32.76	13.14	31.14	7.56	8.93	≥66,100	6.43	42.88	14.43	25.26	5.23	5.75

* 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	≥27,830	≤5	42.3	19.7	21.4	5.1	7.6	≥27,830	≤5	68.7	20.8	5.1	≤5	≤5
K	≥2,420	6.8	49.6	14.6	16.1	6.9	6.1	≥2,420	6.8	56.5	23.2	10.0	≤5	≤5
1	≥4,630	5.1	68.0	15.7	≤5	≤5	≤5	≥4,620	5.1	67.8	9.7	10.6	≤5	≤5
2	≥3,550	≤5	62.6	12.9	6.1	≤5	9.0	≥3,550	≤5	58.9	10.9	12.5	≤5	8.7
3	≥3,290	≤5	59.4	9.5	6.7	7.9	12.1	≥3,290	≤5	60.3	10.0	11.2	5.3	8.7
4	≥2,980	5.7	50.0	7.8	9.1	6.6	20.9	≥2,980	5.7	49.1	8.8	17.5	5.1	13.8
5	≥2,900	5.5	53.7	7.2	7.2	5.5	20.9	≥2,900	5.5	47.5	8.2	19.9	6.0	13.0
6	≥2,620	10.5	46.3	7.8	14.7	6.7	14.1	≥2,630	10.5	42.6	9.6	17.0	6.9	13.3
7	≥2,640	10.4	49.5	10.4	12.4	5.6	11.7	≥2,640	10.4	53.8	8.4	13.4	≤5	9.3
8	≥2,650	10.5	52.2	8.3	12.8	5.1	11.1	≥2,650	10.5	55.1	8.2	15.1	≤5	7.0
9	≥5,400	17.9	52.9	8.9	9.8	≤5	7.2	≥5,400	17.9	58.8	7.2	9.3	≤5	≤5
10	≥2,550	8.7	44.3	11.9	14.4	7.1	13.8	≥2,550	8.7	50.0	11.9	16.9	≤5	7.9
11	≥1,710	5.1	32.8	11.2	20.0	9.6	21.3	≥1,710	5.1	36.0	14.5	25.4	6.9	12.2
12	≥870	≤5	27.3	12.3	20.8	10.2	26.3	≥870	≤5	31.4	17.8	25.7	6.4	15.5
Total	≥66,100	6.43	48.50	14.47	14.93	5.47	10.22	≥66,110	6.43	60.37	14.84	10.38	≤5	4.95

* 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	≥27,840	37.7	53.0	5.4	≤5
K	≥2,420	42.3	47.5	≤5	6.8
1	≥4,630	42.6	47.7	≤5	5.1
2	≥3,550	43.6	42.1	9.9	≤5
3	≥3,290	44.5	39.6	11.5	≤5
4	≥2,980	46.4	30.2	17.7	5.7
5	≥2,900	49.3	27.9	17.3	5.5
6	≥2,630	45.4	28.9	15.1	10.5
7	≥2,640	52.6	26.8	10.2	10.4
8	≥2,650	54.8	26.8	7.9	10.5
9	≥5,400	58.1	19.0	≤5	17.9
10	≥2,550	49.4	33.0	9.0	8.7
11	≥1,710	35.7	44.7	14.5	5.1
12	≥870	30.6	48.0	18.3	≤5
Total	≥66,120	43.3	42.1	8.1	6.4

2.3 2022–2023 Testing Time for Online Screener Tests

In the 2022–2023 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 were 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
  - Figure S8.1 shows the ratio of MSEM to the standard deviation of scale scores at the test level by domain and grade.
  - Figure S8.2 presents the marginal reliability for each domain test across grades.
- Section 9. Screener Assessment—Conditional Standard Error of Measurement (CSEM)
  - Figures S9.1–S9.14 show the CSEM plots for each domain, overall, and comprehension score in each 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 greater error. However, the difference between Step 2 and Step 3 reliability did not differ substantially despite the greater number of items attempted by Step 3 students, due to the mismatch between item difficulty (most screener items are quite easy) and student ability for the high ability students who reached Step 3. See the CSEM plots in the appendix. 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
  - 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
  - Tables S11.1–S11.7 display the inter-rater analysis result for each handscored item in each grade or grade band.

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 the MSEM to the standard deviation of scale scores can also indicate the measure errors. The computed ratios are 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 other grades. Writing has lower marginal reliability for pre-K, kindergarten, and grades 1 and 9–12, but has higher reliability for grades 4 and 5. 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	≥26,730	.78	.75	.81	.68	.71	.75
K	≥2,260	.82	.79	.83	.74	.75	.81
1	≥4,390	.82	.86	.85	.86	.73	.86
2	≥3,390	.84	.90	.88	.90	.79	.90
3	≥3,140	.87	.92	.90	.92	.82	.92
4	≥2,810	.91	.92	.92	.93	.86	.94
5	≥2,740	.92	.93	.93	.93	.88	.94
6	≥2,350	.93	.91	.92	.91	.88	.93
7	≥2,360	.93	.91	.92	.91	.88	.93
8	≥2,370	.93	.91	.92	.92	.89	.93
9	≥4,430	.93	.91	.90	.86	.90	.91
10	≥2,320	.93	.92	.91	.88	.90	.92
11	≥1,620	.92	.91	.91	.88	.89	.92
12	≥840	.92	.91	.91	.88	.88	.92

* Domains with Exemption or Not Attempted are excluded. Step 2 and Step 3 were 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 results for each cut score are presented in Table 3.3 and Table 3.4, as well as in Figure S10.1 and Figure S10.2 in the appendix. Across the four performance cut scores, the CA indices were all above 0.81, denoting that the degree to which we can reliably differentiate students between adjacent performance levels is 0.81 or above. 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 0.86 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 4, 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	.69	.62	.66	.76	.59	.52	.58	.68
K	.72	.65	.70	.72	.62	.55	.64	.63
1	.66	.78	.78	.82	.56	.72	.74	.78
2	.64	.81	.77	.81	.55	.75	.73	.75
3	.66	.81	.77	.83	.56	.76	.73	.78
4	.74	.81	.79	.82	.65	.75	.75	.76
5	.77	.82	.81	.81	.69	.77	.77	.76
6	.77	.80	.78	.76	.69	.73	.72	.68
7	.80	.83	.80	.82	.73	.77	.74	.77
8	.81	.84	.82	.84	.74	.80	.76	.79
9	.85	.88	.83	.85	.80	.84	.78	.79
10	.79	.82	.77	.78	.72	.76	.70	.71
11	.75	.75	.73	.72	.67	.68	.65	.64
12	.75	.73	.71	.71	.66	.65	.63	.62

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

* 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, 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	.78	.92	.96	.81	.76	.88	.94	.83	.82	.88	.91	.75	.93	.98	.99
K	.87	.82	.90	.94	.84	.81	.86	.92	.85	.86	.89	.92	.77	.89	.96	.97
1	.86	.85	.84	.88	.87	.90	.93	.95	.85	.89	.90	.93	.91	.92	.94	.95
2	.84	.86	.86	.90	.90	.91	.93	.95	.87	.88	.90	.93	.89	.91	.95	.96
3	.85	.88	.88	.89	.92	.92	.93	.94	.90	.90	.90	.91	.92	.93	.94	.95
4	.88	.91	.90	.91	.91	.92	.93	.94	.92	.91	.91	.91	.92	.92	.94	.94
5	.90	.92	.92	.91	.93	.94	.93	.93	.93	.92	.92	.91	.92	.93	.93	.93
6	.89	.93	.93	.91	.91	.94	.92	.93	.93	.91	.90	.92	.87	.92	.92	.92
7	.91	.94	.92	.92	.93	.93	.93	.94	.92	.91	.92	.93	.93	.93	.93	.94
8	.92	.95	.92	.92	.94	.94	.93	.95	.93	.92	.92	.94	.94	.94	.93	.94
9	.92	.95	.95	.95	.94	.95	.96	.97	.91	.94	.94	.95	.90	.94	.96	.96
10	.91	.94	.92	.92	.92	.92	.93	.95	.91	.92	.90	.91	.88	.91	.93	.94
11	.93	.93	.89	.88	.92	.90	.89	.91	.92	.91	.87	.88	.89	.88	.89	.91
12	.93	.92	.89	.88	.92	.89	.88	.90	.91	.91	.86	.87	.88	.87	.89	.90

* 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, and 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	.80	.82	.98
K	.87	.89	.98	.83	.85	.98
1	.87	.90	.97	.82	.86	.96
2	.87	.91	.96	.82	.87	.95
3	.88	.92	.96	.83	.89	.94
4	.90	.94	.96	.87	.92	.94
5	.90	.95	.95	.87	.93	.94
6	.89	.95	.95	.86	.93	.93
7	.91	.95	.96	.88	.94	.95
8	.92	.96	.96	.89	.94	.95
9	.93	.96	.97	.91	.94	.97
10	.91	.95	.96	.87	.93	.95
11	.88	.95	.93	.85	.93	.92
12	.87	.95	.92	.83	.93	.91

3.4 Inter-Rater Analysis

In the 2022–2023 screener tests, two to four handscored items in the elementary school (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 maximum score of 1 point) or quadratic weighted kappa (QWK) (for items with a maximum 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, 54.4%–94.1% 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	.808	.842	72.2	76.9	100.0	100.0	0.0	0.0
K	2	.897	.967	84.2	94.1	100.0	100.0	0.0	0.0
1	2	.389	.855	72.2	81.5	100.0	100.0	0.0	0.0
2–3	3	.777	.905	74.5	80.6	100.0	100.0	0.0	0.0
4–5	4	.718	.894	64.6	83.2	100.0	100.0	0.0	0.0
6–8	9	.473	.908	54.4	89.3	100.0	100.0	0.0	0.0
9–12	9	.569	.937	60.3	88.0	100.0	100.0	0.0	0.0

Chapter 4. Validity

Discussions on test development, form construction, scaling, equating, and standard setting can be found in related documents from the English Language Proficiency Assessment of the 21st Century (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 since 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. We also summarize student progress from the time they took the screener tests to the time they took the summative tests. The statistical methods and results are presented in this chapter and Sections 12–13 of the appendix for screener assessment.

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

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) 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.

4.1.1 Correlation between Screener and Summative Tests

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 for screener assessment. These correlations show predictive validity between the two ELPA21 tests because they were given to the same students at different times.

Table S12.1 shows the Pearson correlation between the screener and the summative assessments for 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 in grades 4-12.

Table S12.2 shows the Gamma correlations between domain performance levels and test proficiency categories for the screener and the summative assessments. Similar to the Pearson correlations between scale scores presented in Table S12.1, kindergarten had the lowest Gamma correlations in all domain performance levels and overall proficiency categories. For grade 2 and above, the correlations were 0.8 or above. In addition, the correlations between overall proficiency categories were generally higher than those between domain performance levels. This is likely because there are three levels in overall proficiency while there are five levels in domain performance.

4.1.2 Student Progress from Screener to Summative

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 pooled Appendix of the screener assessment 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 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. 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 for each grade band, 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 state's Appendix. It is noted that the mock-up for score reports is not included in the pooled Appendix for the screener's 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.