TECHNICAL REPORT

PART III – SCREENER ASSESSMENT

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

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

Grades Pre-K–12

2020–2021 Administration

Submitted to:
ELPA21

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

September 2021

Table of Contents

List of Tables

List of Figures

Chapter 1. Test Administration

The screener tests were administered to students in the following groups: 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 assessment, each form of the screener assessments involves four domain (Listening, Reading, Speaking & Writing) tests. Students can be exempted from as many as three domain tests. The assessments do not have a time limit.

1.1 Testing Window

The 2020–2021 summative testing windows for the seven states discussed in this report are shown in Table 1.1. Although testing windows remained open in 2021, due to the continued impact of the Coronavirus (COVID-19) pandemic, some students did not complete the English Language Proficiency Assessment (ELPA) screener assessments.

Table 1.1 2020–2021 ELPA21 Screener Testing Windows by State

State	ELPA21 Screener
Arkansas	8/4/2020–7/16/2021
Iowa	8/3/2020–7/16/2021
Louisiana	8/3/2020–7/16/2021
Nebraska	8/4/2020–7/16/2021
Ohio	8/6/2020–7/16/2021
Washington	8/3/2020–6/30/2021
West Virginia	8/10/2020–6/21/2021

1.2 Test Design

Each 2020–2021 screener test has one online form, one paper-pencil form, and one braille form. Pre-K students were permitted to take the kindergarten tests.

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 non-scored 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 exceed 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 domain) 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.

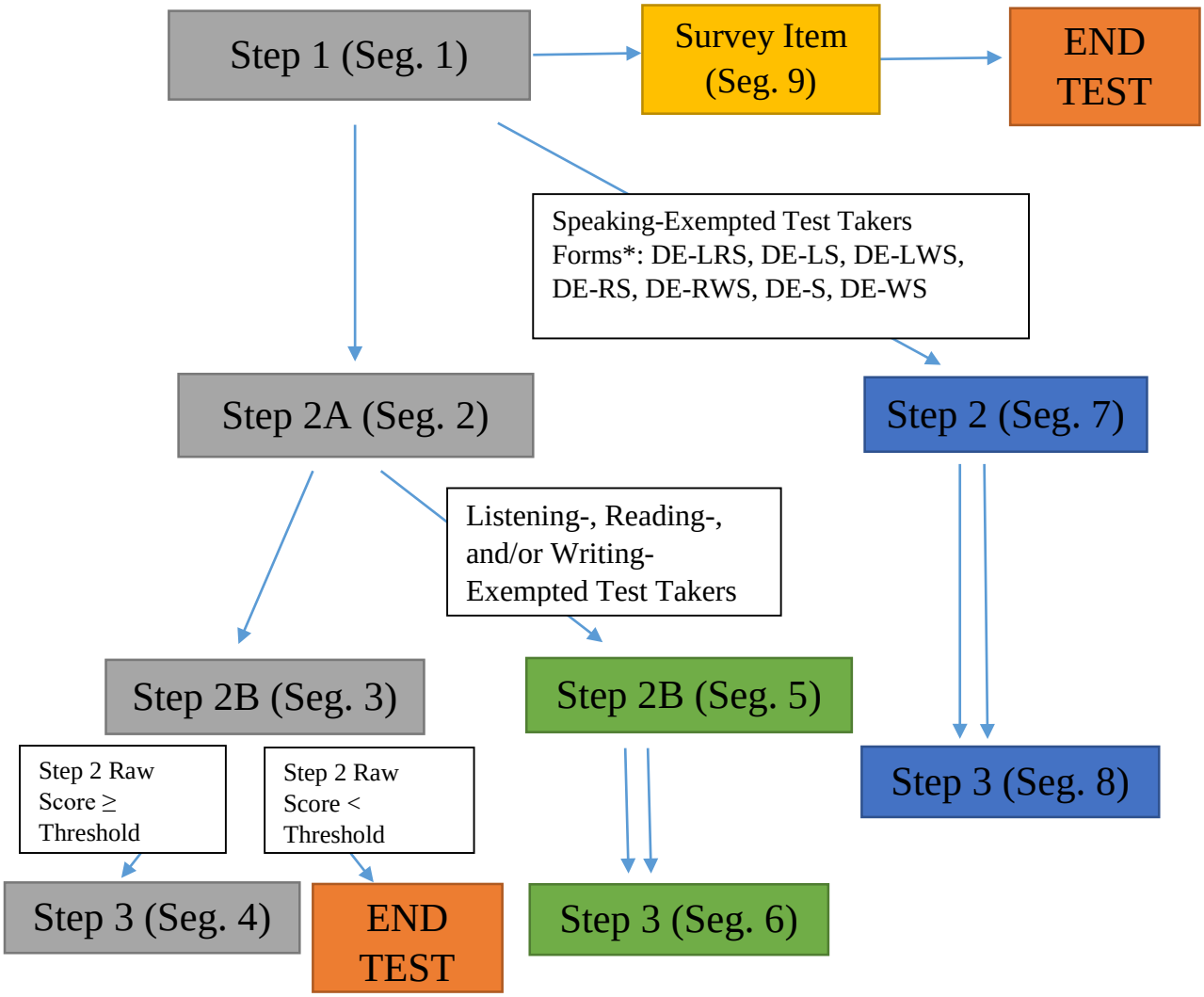
- Segment 6 (Step 3 for students who are exempted from the listening, reading, and/or writing domains) consists of items from all non-exempted domains.

- 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 2020–2021 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

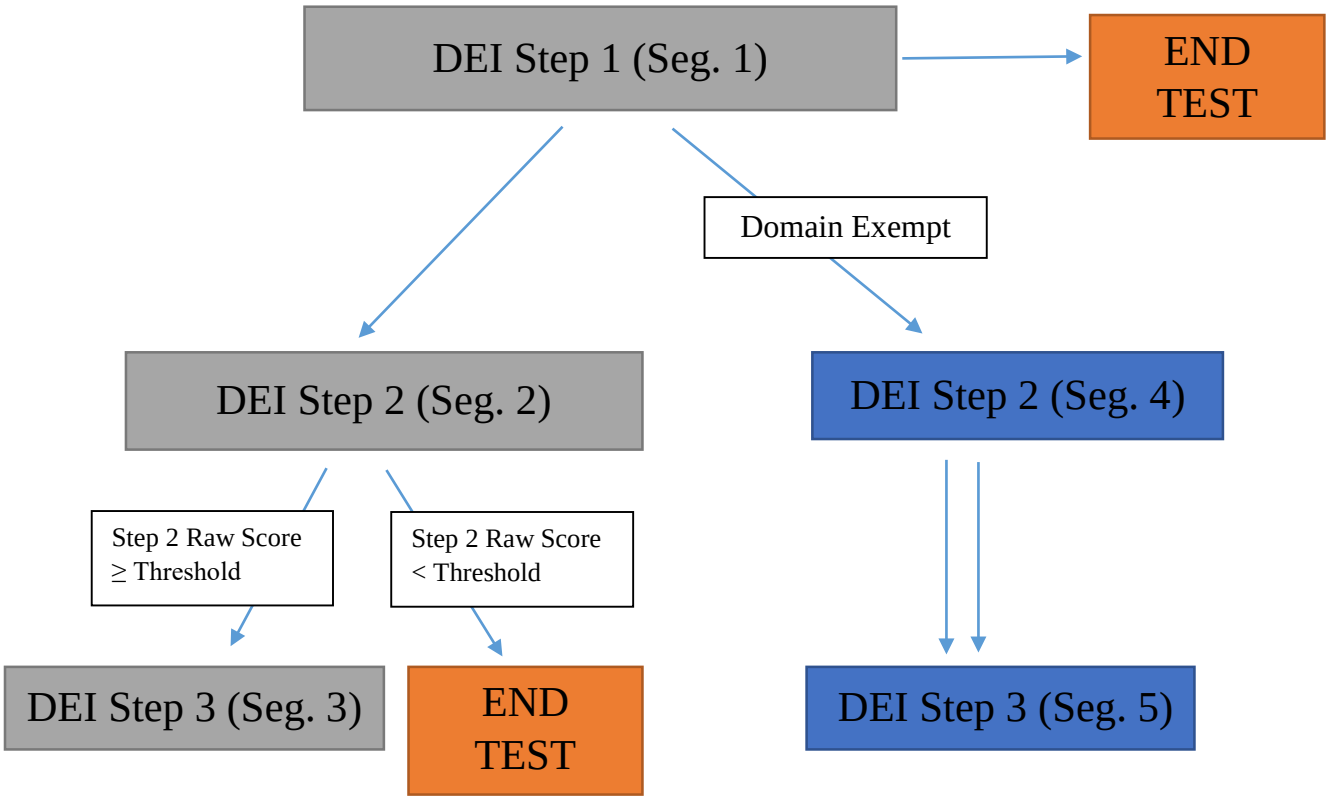
Grade Band	Threshold Score	Step 2 Max Score
Pre-K/K	23	26
1	24	27
2–3	25	28
4–5	26	31
6–8	28	33
9–12	27	30

The paper-pencil form has five segments:

- Segment 1 (Step 1) includes non-scored, 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 the Cambium Assessment, Inc.'s (CAI) Data Entry Interface (DEI) on the state testing portal for all ELPA21 domain tests. Practice test items are not entered in the DEI and are not scored.

Figure 1.2 2020–2021 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 1.5. 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	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 1.5 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	9	21	7	23	8	26
Total	34	42	32	42	34	49	41	76	37	71	38	77

1.3 Test Administration Manual

1.3.1 Directions for Administration

For the 2020–2021 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
- 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 systems, training or practice tests (Step 1 in screener tests) are provided before and during the testing windows. Training/practice tests can be accessed through a non-secure browser or a secure browser. For screener assessments, the tests become secure automatically when students proceed to Step 2.

The 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 TA 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;
- starting the test (adjusting the audio sound, 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 2020–2021 screener assessment include the following:

1. All pending and expired test records in Step 2 should be scored. Exception: Expired tests in Washington are not scored due to an existing state rule.

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 in 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 in 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:

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

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

Chapter 2. 2020-2021 Summary

The 2020–2021 screener results are presented in this chapter and in Sections 16–22 of the appendix. The figures and tables included in each section are listed below:

- Section 16. Screener—Student Participation
    - Table S16.1 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.
    - Table S16.2 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 17. Screener Assessment—Raw Score Summary
    - Tables S17.1–S17.14 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.
    - Note that the MIRT model precludes one-to-one correspondence between domain raw and scale scores and allows the same domain raw score to fall into different performance levels depending on performance on the off-domain items. This is important in interpreting the raw score statistics in the Appendices. For the screener, we also have to consider whether a student advanced to Step 3 when interpreting raw scores.
- Section 18 Screener Assessment—Raw Score Distributions
    - Figures S18.1–S18.65 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 19. Screener Assessment—Scale Score Summary
    - Tables S19.1–S19.14 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.
    - Table S19.15 summarizes the number and percentage of students who were marked “non-attempt” or “exempt” in each domain and grade.
- Section 20. Screener Assessment—Percentage of Students by Domain Performance Level

  - Figure S20.1 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 S20.1–S20.14 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 21. Screener Assessment—Percentage of Students by Overall Proficiency Level
  - Figure S21.1 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 S21.1–S21.14 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 22. Screener Assessment—Testing Time
  - Table S22.1 shows the testing time by end step in each grade/grade band.

2.1 2020–2021 Student Participation

Due to the COVID-19 pandemic, not all eligible students completed the assessments during the 2020–2021 administration. Section S16.2 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 (47.7%–50.9%) than female students (44.9%–48.9%) tested. In each test, the greatest number of participating students were in the group of Hispanic or Latino (43.7%–71.4%), followed by Asian students (10.3%–19.3%), and White students (4.5%–11.6%).

Table 2.1 shows the overall student participation for each state. There were 53,644 students in total who took the 2020–2021 screener tests. Washington had the most students, followed by Ohio. 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 2020–2021 school year.

Section S16.1 of the Appendix presents student participation in each mode. In the seven ELPA21 states combined, the most frequent mode of administration was online (99.94%), followed by paper (0.06%) and braille (<0.01%).

Section S16.2 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 (47.7%–50.9%) than female students (44.9%–48.9%) tested. In each test, the greatest number of participating students were in the group of Hispanic or Latino (43.7%–71.4%), followed by Asian students (10.3%–19.3%), and White students (4.5%–11.6%).

Table 2.6 Number of Students Who Participated in ELPA21 Screener in 2019–2020 and 2020–2021 by State and Grade

Grade	Arkansas	Arkansas	Iowa	Iowa	Louisiana	Louisiana	Nebraska	Nebraska	Ohio	Ohio	Washington	Washington	West Virginia	West Virginia	Total	Total	Total
	2020-21	2019-20	2020-21	2019-20	2020-21	2019-20	2020-21	2019-20	2020-21	2019-20	2020-21	2019-20	2020-21	2019-20	2020-21	2019-20	Two Year N Diff
Pre-K	≥3870	≥2150	≥4780	≥3100	≥3760	≥1920	≥3260	≥2710				≥10	≥160	≥180	≥15860	≥10090	≥5760
K	≥1260	≥1320	≥240	≥170	≥300	≥200	≥140	≥60	≥8150	≥9960	≥8630	≥14310	≥70	≥50	≥18820	≥26100	≥-7280
1	≥390	≥540	≥360	≥430	≥470	≥810	≥220	≥310	≥990	≥1610	≥650	≥1970	≥50	≥60	≥3160	≥5760	≥-2610
2	≥340	≥440	≥270	≥380	≥300	≥630	≥170	≥210	≥680	≥1240	≥410	≥1420	≥40	≥40	≥2230	≥4390	≥-2170
3	≥290	≥390	≥250	≥370	≥290	≥580	≥190	≥180	≥610	≥1080	≥340	≥1300	≥30	≥80	≥2030	≥4010	≥-1990
4	≥270	≥310	≥230	≥360	≥210	≥540	≥140	≥210	≥490	≥930	≥320	≥1200	≥30	≥80	≥1720	≥3650	≥-1940
5	≥250	≥380	≥210	≥280	≥220	≥480	≥120	≥150	≥380	≥930	≥270	≥1100	≥30	≥70	≥1500	≥3410	≥-1910
6	≥240	≥320	≥200	≥280	≥190	≥490	≥70	≥90	≥400	≥780	≥240	≥1130	≥20	≥40	≥1380	≥3150	≥-1770
7	≥260	≥340	≥160	≥250	≥160	≥480	≥80	≥90	≥370	≥830	≥210	≥1060	≥30	≥120	≥1300	≥3200	≥-1900
8	≥230	≥310	≥150	≥260	≥160	≥440	≥60	≥90	≥330	≥680	≥210	≥980	≥20	≥30	≥1180	≥2810	≥-1630
9	≥300	≥430	≥300	≥530	≥280	≥940	≥150	≥220	≥470	≥1300	≥290	≥1600	≥20	≥60	≥1850	≥5100	≥-3260
10	≥260	≥450	≥170	≥270	≥110	≥240	≥70	≥90	≥310	≥680	≥220	≥1140	≥20	≥70	≥1190	≥2960	≥-1770
11	≥190	≥440	≥130	≥160	≥60	≥140	≥40	≥50	≥220	≥410	≥180	≥1110	≥20	≥50	≥860	≥2390	≥-1540
12	≥90	≥240	≥50	≥110	≥20	≥60	≥30	≥40	≥150	≥250	≥130	≥870	≥10	≥30	≥510	≥1620	≥-1110
Total	≥8310	≥8130	≥7570	≥7010	≥6610	≥8010	≥4790	≥4540	≥13600	≥20720	≥12160	≥29250	≥580	≥1020	≥53640	≥78710	≥-25070

Table 2.7 Number of Students Participating in 2020–2021 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	≥4,190	≥5,140	≥3,900
	1	≥4,480	≥390	≥290
	2–3	≥7,220	≥630	≥430
	4–5	≥5,750	≥530	≥320
	6–8	≥7,550	≥730	≥470
	9–12	≥9,060	≥870	≥600
Iowa	Pre-K and K	≥4,410	≥5,030	≥3,940
	1	≥3,960	≥360	≥260
	2–3	≥5,760	≥520	≥350
	4–5	≥4,180	≥450	≥270
	6–8	≥5,490	≥510	≥330
	9–12	≥6,820	≥670	≥430
Louisiana	Pre-K and K	≥3,240	≥4,060	≥2,910
	1	≥3,390	≥470	≥380
	2–3	≥5,580	≥600	≥410
	4–5	≥4,080	≥430	≥290
	6–8	≥4,950	≥520	≥400
	9–12	≥5,260	≥500	≥340
Nebraska	Pre-K and K	≥3,670	≥3,410	≥2,690
	1	≥3,420	≥220	≥150
	2–3	≥4,650	≥360	≥220
	4–5	≥2,790	≥260	≥120
	6–8	≥2,910	≥220	≥120
	9–12	≥3,590	≥300	≥160
Ohio	K	≥8,990	≥8,150	≥7,130
	1	≥8,940	≥990	≥720
	2–3	≥12,720	≥1,290	≥890
	4–5	≥8,240	≥880	≥500
	6–8	≥9,270	≥1,110	≥690
	9–12	≥11,300	≥1,160	≥780
Washington	K	≥12,040	≥8,630	≥6,590
	1	≥12,650	≥650	≥410
	2–3	≥20,930	≥750	≥420
	4–5	≥15,650	≥590	≥300
	6–8	≥17,350	≥680	≥340
	9–12	≥15,810	≥830	≥390
West Virginia	Pre-K and K	≥200	≥240	≥190
	1	≥190	≥50	≥30
	2–3	≥320	≥70	≥40

State	Grade/Grade Band	N Summative	N Screener	N Both
	4–5	≥230	≥60	≥20
	6–8	≥310	≥70	≥30
	9–12	≥400	≥80	≥50

2.2 2020–2021 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 19 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–10, in reading in grades 1–10, and in speaking in grades 7–10. Reading and writing follow a similar pattern; the percentage of students in level 1 decrease from pre-K to grade 6 (with slight 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 slight increase in grade 1), then increases until grade 9 (with slight decrease in grade 6), and then decreases afterwards. Disaggregated results by gender and ethnicity are provided in Section 20 of the Appendix.

Table 2.8 and Figure S21.1 in the Appendix present the percentage of students achieving each overall proficiency category, by grade. The results show that the majority of students have achieved the Emerging or Progressing category. The percentages of students who are proficient increase from grades pre-K to kindergarten, consistently decrease from grade 1 to grade 5, and slightly increase to grade 7, and then decrease to grade 9, and go up afterwards. The percentages of students in the Emerging category are relatively stable until grade 6, increase from grade 6 to grade 9, and then consistently decrease above grade 9. Section 21 of the Appendix displays the overall proficiency category for each grade by gender and ethnicity.

*Table 2.8 Scale Score Summary by Grade–Listening and Reading**

Grade	Listening					Reading				
	N	Min	Mean	Max	SD	N	Min	Mean	Max	SD
Pre-K	≥15,290	314	517.4	714	61.4	≥15,300	318	514.0	708	61.1
K	≥18,340	314	528.9	714	67.6	≥18,330	318	525.7	708	67.1
1	≥3,040	288	512.3	678	81.9	≥3,040	286	488.1	704	89.9
2	≥2,160	286	492.8	710	81.2	≥2,160	278	478.9	734	89.2
3	≥1,960	286	516.2	710	91.3	≥1,960	278	509.0	734	102.2
4	≥1,630	270	493.1	778	102.8	≥1,630	270	494.4	795	104.2
5	≥1,440	270	518.4	778	113.3	≥1,440	270	523.2	795	112.2
6	≥1,260	279	509.1	738	96.9	≥1,260	296	512.3	733	96.1
7	≥1,200	279	513.6	738	101.3	≥1,200	296	520.6	733	99.0
8	≥1,070	279	505.6	738	108.0	≥1,070	296	513.1	733	105.6
9	≥1,600	297	499.2	731	108.2	≥1,600	309	501.9	733	104.6
10	≥1,080	297	513.3	731	100.6	≥1,080	309	517.0	733	96.9
11	≥800	297	541.5	731	96.9	≥800	309	544.9	733	94.0
12	≥480	297	549.4	731	97.0	≥470	309	551.8	733	94.7

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

*Table 2.9 Scale Score Summary by Grade–Speaking and Writing**

Grade	Speaking					Writing				
	N	Min	Mean	Max	SD	N	Min	Mean	Max	SD
Pre-K	≥15,290	339	506.9	711	77.9	≥15,300	347	480.5	684	56.1
K	≥18,330	339	518.7	711	82.7	≥18,330	334	495.0	684	65.8
1	≥3,040	310	493.6	669	86.5	≥3,040	283	483.7	698	90.3
2	≥2,160	292	476.6	703	95.3	≥2,160	276	474.4	737	90.9
3	≥1,960	292	499.1	703	107.1	≥1,960	276	506.3	737	104.1
4	≥1,630	270	502.7	786	125.3	≥1,630	268	491.9	797	108.5
5	≥1,440	270	525.0	786	131.4	≥1,440	268	522.4	797	116.3
6	≥1,260	296	515.6	732	107.0	≥1,260	281	506.2	741	99.0
7	≥1,200	296	518.6	732	108.0	≥1,200	281	512.8	741	101.7
8	≥1,070	296	505.6	732	116.0	≥1,070	281	506.1	741	108.3
9	≥1,600	332	509.9	722	107.1	≥1,600	315	502.0	732	101.2
10	≥1,080	332	524.9	722	97.9	≥1,080	315	514.5	732	93.9
11	≥800	332	550.5	722	93.6	≥800	315	539.6	732	91.0
12	≥480	330	560.5	722	88.7	≥480	315	547.9	732	93.3

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

*Table 2.10 Scale Score Summary by Grade–Comprehension and Overall**

Grade	Comprehension					Overall				
	N	Min	Mean	Max	SD	N	Min	Mean	Max	SD
Pre-K	≥15,300	3978	5356.0	6375	468.3	≥15,300	3646	5106.8	6763	481.3
K	≥18,340	3936	5426.0	6375	489.9	≥18,340	3646	5209.7	6763	537.9
1	≥3,040	3785	5203.8	6387	586.6	≥3,040	3364	5039.8	6629	684.8
2	≥2,160	3756	5098.5	6439	615.4	≥2,160	3326	4926.6	6880	707.3
3	≥1,960	3756	5269.4	6439	677.8	≥1,960	3326	5147.3	6880	810.1
4	≥1,630	3649	5092.2	6700	681.1	≥1,630	3237	5058.2	7401	881.2
5	≥1,440	3649	5261.0	6700	743.9	≥1,440	3237	5273.6	7401	944.6
6	≥1,260	3803	5226.2	6476	665.9	≥1,260	3388	5183.1	6974	790.9
7	≥1,200	3803	5279.2	6476	703.0	≥1,200	3388	5228.6	6974	812.5
8	≥1,070	3803	5223.2	6476	745.3	≥1,070	3388	5155.4	6974	870.0
9	≥1,600	3787	5144.2	6524	757.6	≥1,600	3605	5125.6	6923	834.3
10	≥1,080	3787	5254.9	6524	719.8	≥1,080	3605	5240.5	6923	766.7
11	≥800	3787	5463.3	6524	698.2	≥800	3605	5455.4	6923	737.7
12	≥480	3787	5499.2	6524	688.4	≥480	3605	5520.6	6923	731.1

* Scale scores cannot be compared across grade bands.

*Table 2.11 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	≥15,850	3.5	19.2	18.0	55.3	2.0	2.1	≥15,850	3.5	23.2	21.3	45.6	3.6	2.9
K	≥18,810	2.5	17.0	15.3	56.3	3.6	5.3	≥18,810	2.5	20.1	19.0	46.4	5.2	6.8
1	≥3,140	3.2	18.6	7.9	40.2	12.7	17.3	≥3,140	3.2	50.8	13.2	15.7	7.5	9.6
2	≥2,230	3.1	17.6	9.5	29.8	21.2	18.8	≥2,230	3.1	44.4	10.1	23.5	7.5	11.5
3	≥2,010	2.8	16.3	10.6	26.4	22.4	21.6	≥2,010	2.8	42.2	16.1	19.9	9.5	9.6
4	≥1,710	4.2	20.9	7.8	15.7	25.9	25.4	≥1,710	4.2	35.5	10.4	20.3	12.0	17.5
5	≥1,490	3.7	22.9	7.4	9.2	24.7	32.2	≥1,490	3.7	33.0	12.4	19.4	10.8	20.7
6	≥1,370	8.2	20.0	7.3	12.2	22.0	30.4	≥1,370	8.2	31.7	7.8	22.0	11.9	18.4
7	≥1,290	7.1	25.8	8.9	20.6	15.9	21.6	≥1,290	7.1	35.5	13.3	24.0	8.4	11.8
8	≥1,170	8.9	30.4	10.0	17.3	15.0	18.3	≥1,170	8.9	41.1	11.5	26.7	6.8	5.0
9	≥1,830	12.9	34.1	8.3	18.5	9.9	16.3	≥1,830	12.9	41.3	12.8	20.6	6.4	5.9
10	≥1,170	7.5	30.4	9.0	22.1	11.6	19.5	≥1,170	7.5	38.4	13.6	26.6	7.8	6.1
11	≥850	5.4	20.3	10.0	22.7	14.6	27.0	≥850	5.4	27.6	15.2	31.4	9.8	10.7
12	≥500	4.6	17.3	9.7	23.5	15.5	29.4	≥500	4.6	24.5	16.5	30.7	11.0	12.7
Total	≥53,490	5.2	20.8	9.4	24.9	14.7	19.3	≥53,490	5.2	32.7	13.0	25.1	8.1	10.3

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

*Table 2.12 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	≥15,850	3.5	37.2	20.8	25.1	9.9	3.5	≥15,850	3.5	63.3	25.7	6.1	1.0	0.5
K	≥18,810	2.5	33.2	20.8	25.2	10.5	7.8	≥18,810	2.5	56.1	27.3	10.0	2.6	1.4
1	≥3,140	3.2	58.7	22.6	4.1	4.2	7.1	≥3,140	3.2	59.8	12.4	13.2	4.8	6.6
2	≥2,230	3.1	48.7	18.4	10.0	7.8	11.9	≥2,230	3.1	44.8	14.1	18.6	7.4	12.0
3	≥2,010	2.8	43.6	13.4	12.0	12.8	15.4	≥2,010	2.8	44.9	13.8	18.5	8.1	11.9
4	≥1,710	4.2	32.6	11.0	12.6	11.4	28.2	≥1,710	4.2	32.2	9.8	29.2	8.7	16.0
5	≥1,490	3.7	34.1	9.2	11.9	9.5	31.6	≥1,490	3.7	28.0	9.1	30.1	8.1	21.0
6	≥1,370	8.2	28.6	8.2	20.9	10.9	23.2	≥1,370	8.2	24.8	9.8	27.7	9.7	20.0
7	≥1,290	7.1	30.7	11.2	22.3	8.9	19.9	≥1,290	7.1	33.6	13.4	23.5	8.3	14.1
8	≥1,170	8.9	35.5	11.0	19.5	8.3	16.8	≥1,170	8.9	39.6	10.9	24.6	7.7	8.2
9	≥1,830	12.9	35.9	10.7	18.0	8.2	14.4	≥1,830	12.9	40.8	12.7	18.4	6.1	9.1
10	≥1,170	7.5	30.2	13.0	24.9	8.4	16.1	≥1,170	7.5	37.6	15.0	23.2	7.0	9.9
11	≥850	5.4	22.7	12.3	24.3	11.9	23.4	≥850	5.4	27.4	16.6	27.3	10.3	13.0
12	≥500	4.6	17.5	11.9	28.0	13.7	24.3	≥500	4.6	25.2	16.7	26.4	9.5	17.5
Total	≥53,480	5.2	32.7	13.1	17.5	9.4	16.6	≥53,480	5.2	37.3	14.0	20.0	6.9	11.1

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

Table 2.13 Percentage of Students in Each Overall Proficiency Category by Grade

Grade	N	Emerging	Progressing	Proficient	Proficiency Not Demonstrated
Pre-K	≥15,850	32.3	61.5	2.7	3.5
K	≥18,810	28.0	65.8	3.8	2.5
1	≥3,140	26.0	63.2	7.5	3.2
2	≥2,230	26.8	55.5	14.6	3.1
3	≥2,010	26.7	54.3	16.2	2.8
4	≥1,710	28.2	44.9	22.6	4.2
5	≥1,490	29.6	39.9	26.7	3.7
6	≥1,370	26.1	42.3	23.5	8.2
7	≥1,290	32.7	43.1	17.1	7.1
8	≥1,170	38.8	41.4	10.9	8.9
9	≥1,830	40.5	35.8	10.7	12.9
10	≥1,170	35.9	44.8	11.9	7.5
11	≥850	27.4	49.8	17.4	5.4
12	≥500	22.9	51.5	21.1	4.6
Total	≥53,490	30.1	49.6	14.8	5.5

2.3 2020–2021 Testing Time for Online Screener Tests

In the 2020–2021 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 S22.1 of 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 Chapter 3 in Part I of the *ELPA21 2020–2021 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 consistency (CC)
- inter-rater analysis

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

- Section 23. Screener Assessment—Marginal Reliability
    - Figure S23.1 shows the ratio of MSEM to the standard deviation of scale scores at the test level, by domain and grade
    - Figure S23.2 presents the marginal reliability for each domain test across grades
- Section 24. Screener Assessment—Conditional Standard Error of Measurement (CSEM)
    - Figures S24.1–S24.14 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 25. Screener Assessment—Classification Accuracy and Consistency
    - Figure S25.1 shows the CA for each domain test
    - Figure S25.2 shows the CC for each domain test
    - Figure S25.3 presents the CA and CC for the overall proficiency
- Section 26. Screener Assessment—Inter-Rater Analysis
    - Tables S26.1–S26.7 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, 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 S23.1 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 S23.2 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 5. Listening has relatively lower reliability than the other domains in grades 1–5. In addition, Section 24 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	≥15,290	.72	.70	.77	.66	.66	.71
K	≥18,330	.75	.72	.79	.72	.67	.75
1	≥3,040	.77	.86	.81	.86	.71	.85
2	≥2,160	.82	.90	.86	.90	.78	.89
3	≥1,960	.83	.91	.88	.92	.79	.91
4	≥1,630	.89	.92	.91	.92	.84	.93
5	≥1,440	.90	.92	.91	.93	.85	.93
6	≥1,260	.90	.90	.91	.90	.86	.92
7	≥1,200	.91	.90	.91	.91	.86	.92
8	≥1,070	.92	.91	.92	.92	.88	.93
9	≥1,600	.93	.92	.91	.89	.90	.92
10	≥1,080	.92	.91	.90	.87	.89	.91
11	≥800	.91	.90	.90	.87	.87	.91
12	≥470	.90	.90	.88	.87	.87	.90

* 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 is based on two tests with measurement errors, while accuracy is 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 Figures S25.1–S25.2 in the Appendix. Across the four performance cut scores, the CA indices are 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 are all above 0.7 in all cut scores and all grades. The reliability indices in the middle school tests are above 0.85 for all domains. Table 3.5 and Figure S25.3 in the Appendix display the CA and CC for overall proficiency categories. The plot shows that all the accuracy and consistency indices are above 0.79. The accuracy indices for between Emerging and Progressing are lower than those for between

Progressing and Proficient in pre-K to grade 2 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	.59	.59	.72	.56	.48	.52	.64
K	.68	.59	.59	.70	.56	.48	.52	.61
1	.61	.72	.70	.77	.50	.65	.64	.71
2	.62	.74	.67	.75	.52	.67	.62	.67
3	.64	.72	.67	.75	.53	.66	.61	.68
4	.69	.74	.70	.75	.60	.66	.64	.68
5	.73	.75	.73	.76	.64	.68	.66	.69
6	.74	.73	.71	.73	.65	.65	.63	.64
7	.72	.73	.71	.73	.64	.65	.63	.66
8	.74	.77	.74	.77	.66	.71	.67	.70
9	.78	.78	.74	.74	.70	.72	.67	.67
10	.75	.76	.71	.71	.66	.69	.62	.64
11	.72	.72	.68	.67	.63	.64	.59	.59
12	.72	.70	.67	.68	.62	.62	.57	.59

* 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	.83	.93	.97	.87	.81	.89	.95	.87	.85	.89	.93	.80	.93	.99	.99
K	.91	.85	.92	.95	.88	.83	.88	.93	.87	.85	.88	.92	.81	.92	.97	.97
1	.92	.88	.85	.89	.89	.91	.94	.95	.84	.89	.91	.94	.92	.93	.94	.95
2	.92	.91	.86	.89	.92	.92	.93	.95	.88	.87	.89	.93	.91	.92	.94	.96
3	.93	.93	.87	.88	.93	.91	.92	.94	.91	.89	.89	.90	.93	.93	.93	.94
4	.94	.94	.90	.89	.93	.93	.92	.94	.93	.91	.90	.91	.94	.93	.92	.94
5	.95	.94	.92	.90	.94	.94	.93	.93	.94	.92	.91	.90	.95	.94	.93	.93
6	.95	.96	.93	.90	.95	.94	.91	.92	.95	.91	.90	.91	.93	.94	.92	.92
7	.95	.95	.90	.90	.95	.93	.91	.92	.94	.91	.91	.92	.94	.92	.92	.93
8	.95	.96	.91	.91	.95	.94	.92	.94	.95	.92	.92	.93	.95	.94	.92	.94
9	.95	.95	.93	.93	.95	.93	.94	.95	.94	.94	.91	.93	.92	.92	.94	.94
10	.94	.94	.92	.92	.94	.93	.93	.94	.93	.92	.90	.92	.91	.91	.93	.94
11	.95	.95	.91	.90	.94	.92	.91	.92	.94	.93	.88	.89	.92	.90	.91	.92
12	.95	.94	.92	.90	.94	.92	.90	.91	.95	.92	.87	.89	.91	.90	.91	.91

* 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	.76	.90	.95	.81	.74	.85	.92	.81	.79	.85	.89	.73	.90	.98	.99
K	.87	.78	.87	.93	.83	.76	.83	.90	.82	.79	.84	.88	.74	.88	.96	.97
1	.88	.83	.79	.84	.85	.88	.91	.93	.79	.84	.87	.91	.88	.90	.92	.93
2	.89	.87	.80	.85	.88	.88	.90	.93	.84	.83	.85	.89	.87	.89	.92	.94
3	.90	.90	.82	.83	.90	.88	.89	.91	.87	.84	.84	.87	.90	.90	.90	.92
4	.92	.91	.86	.85	.91	.90	.89	.91	.90	.87	.86	.87	.92	.90	.89	.92
5	.92	.92	.89	.86	.92	.91	.90	.89	.91	.89	.87	.86	.92	.92	.90	.90
6	.92	.93	.90	.86	.92	.91	.88	.89	.92	.88	.85	.87	.90	.91	.89	.89
7	.93	.93	.87	.86	.92	.89	.88	.90	.92	.87	.87	.89	.91	.89	.89	.90
8	.93	.94	.88	.87	.93	.91	.89	.92	.92	.89	.88	.90	.93	.91	.89	.91
9	.93	.93	.90	.91	.92	.91	.92	.93	.91	.91	.88	.90	.88	.89	.91	.92
10	.92	.92	.89	.89	.91	.90	.90	.92	.91	.89	.86	.88	.87	.87	.90	.91
11	.93	.92	.88	.86	.92	.89	.87	.89	.92	.90	.83	.85	.88	.86	.87	.89
12	.93	.91	.88	.85	.92	.89	.86	.88	.93	.88	.82	.84	.88	.87	.87	.88

* 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	.84	.86	.98	.79	.82	.97
K	.85	.88	.98	.81	.84	.97
1	.85	.89	.95	.79	.85	.94
2	.87	.92	.95	.82	.89	.93
3	.87	.94	.93	.82	.91	.91
4	.88	.95	.93	.84	.92	.91
5	.88	.95	.92	.84	.94	.90
6	.87	.95	.92	.84	.94	.90
7	.87	.95	.93	.84	.93	.91
8	.89	.95	.94	.86	.93	.92
9	.90	.95	.95	.87	.93	.93
10	.88	.95	.94	.85	.92	.93
11	.86	.94	.92	.83	.93	.90
12	.85	.94	.91	.82	.92	.89

3.4 Inter-Rater Analysis

In the 2020–2021 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 max score of 1 point) or quadratic weighted Kappa (QWK) (for items with 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 are all above 0.70, except for one item in grade 1, four items in grade band 6–8, and four items in grade band 9–12. Overall, 63%–92.9% of handscores are consistent (exact agreement) between the first rater and the second rater, and 100% of handscores agreed within one score point.

The inter-rater consistencies are also assessed by item and are summarized in Section 26 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	.819	.932	74.3	87.8	100.0	100.0	0.0	0.0
K	2	.909	.912	84.9	86.7	100.0	100.0	0.0	0.0
1	2	.630	.873	63.9	85.9	100.0	100.0	0.0	0.0
2–3	3	.731	.849	73.9	75.2	100.0	100.0	0.0	0.0
4–5	4	.829	.857	63.1	82.9	100.0	100.0	0.0	0.0
6–8	9	.473	.929	68.9	88.5	100.0	100.0	0.0	0.0
9–12	9	.344	.917	63.0	92.9	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 from the item pool for summative tests, and the purpose of the screener is for the prediction of students' English overall proficiency categories. Instead of evaluating the validity aspects as those for the summative tests, we evaluate the relationships between the screener and 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 27–28 in the Appendix:

- Section 27. Correlations Between Summative and Screener Tests
    - Table S27.1 shows the correlations between domain, overall, and comprehension scores.
    - Table S27.2 summarizes the correlations by between domain performance level and overall proficiency categories.
- Section 28. Student Progress from Screener to Summative
    - Figures S28.1–S28.2 display within-year average differences in domain, overall, and comprehension scale score.
    - Figures S28.3–S28.4 present changes domain performance level and overall proficiency.
    - Figures S28.5–S28.10 show scatter plots of scale scores for the screener and summative assessment.
    - Tables S28.1–S28.6 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 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 S27.1 and S27.2 in the Appendix.

Table S27.1 shows the Pearson correlation between the screener and the summative tests in domain and composite scores. Correlations of all types of scores are the lowest in the kindergarten test, followed by the grade 1 test; the correlations are above 0.79 in listening, reading, writing, comprehension, and overall scale scores in grades 2 and above. The speaking tests have relatively lower correlations than the other three domains except those taken at the kindergarten and grade 1 levels.

Table S27.2 shows the Gamma correlations between domain performance levels and test proficiency categories. Similar to the correlations between scale scores presented in Table S27.1, kindergarten has the lowest correlations in all domain performance levels and overall proficiency categories. For grade 2 and above, the correlations are about 0.8 except for the speaking domain. In addition, the correlations between overall proficiency categories are 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 screener tests to the time they took summative tests was evaluated by the changes in scale scores and performance levels. The major confounding factor in this result is the measurement error in both assessments. Given the acceptable marginal reliability indices described in 0 of this document, as well as the Part II of the *ELPA21 2020–2021 Technical Report*, we can still see the trend of student progress. Section 28 of the Appendix summarizes the results of progress analysis. Only students who had valid scores on both the screener and summative tests were included in each of the analyses.

Figures S28.1 and S28.2 in the Appendix show the growth of the average domain scores and composite scores, respectively. The average scale scores in the summative assessment are, in general, higher than those in the screener assessment. Figures S28.3 and S28.4 display the percentage of students in each domain performance level and overall proficiency category, respectively. In each pair of bars, the left bar is from the screener test and the right bar is from the corresponding summative test. The plots indicate that more students are in higher domain performance levels and overall proficiency categories in the summative tests than in the screener tests. In addition, Figures S28.5–S28.10 in the Appendix present scatter plots of scale score change from screener to summative assessments for each grade, and Tables S28.1–S28.6 summarize comparisons of scale scores between screener and summative assessments.

Chapter 5. Reporting

A detailed introduction for the Online Reporting System (ORS) can be found in Chapter 5 in Part I of the *ELPA21 2020–2021 Technical Report*. The reporting mockups for the screener tests of each state are included in Section 29 of the Appendix for each state. It is noted that the mockup for score reports is not included in the Appendix for pooled analysis.

References

Center for Research on Evaluation, Standards, and Student Testing (CRESST). (2019). *ELPA21 Scoring Specification: School Year 2019–2020.* Center for Research on Evaluation, Standards, and Student Testing.

Center for Research on Evaluation, Standards, and Student Testing (CRESST) & Pacific Metrics (2016). *ELPA21 Standard Setting Technical Report*. Center for Research on Evaluation, Standards, and Student Testing.

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