TECHNICAL REPORT

PART III – SCREENER ASSESSMENT

(AR, IA, LA, NE, OH, WA, WV)

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

Grades PreK–12

2019–2020 Administration

Submitted to:
ELPA21

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

December 2020

Table of Contents

List of Tables

List of Figures

Chapter 1. Test Administration

The screener tests were administered to students in kindergarten, grade 1, grades 2–3, grades 4–5, grades 6–8, and grades 9–12. Some states administered the screener tests to pre-kindergarten students. Same as the summative assessment, each form of the screener assessments involves four domain tests. Students can be exempted from as many as three domain tests. The tests do not have a time limit.

1.1 Testing Window

The 2019–2020 screener testing windows are shown in Table 1.1. Although the screener testing windows extended into June or July, ELPA21 screener administrations essentially stopped in March 2020, due to the school building closures in response to the coronavirus (COVID-19) pandemic.

Table 1.1 2019–2020 ELPA21 Screener Testing Windows by State

State	ELPA21 Screener
Arkansas	8/1/19–6/19/20
Iowa	8/4/19–7/10/20
Louisiana	7/31/19–6/19/20
Nebraska	7/31/19–7/10/20
Ohio	8/5/19–6/31/20
Washington	8/5/19–6/19/20
West Virginia	8/6/19–6/22/20

1.2 Test Design

Each 2019–2020 screener test has one online form, one paper-pencil form, and one braille form. Pre-kindergarten (PreK) students could take kindergarten (K) tests.

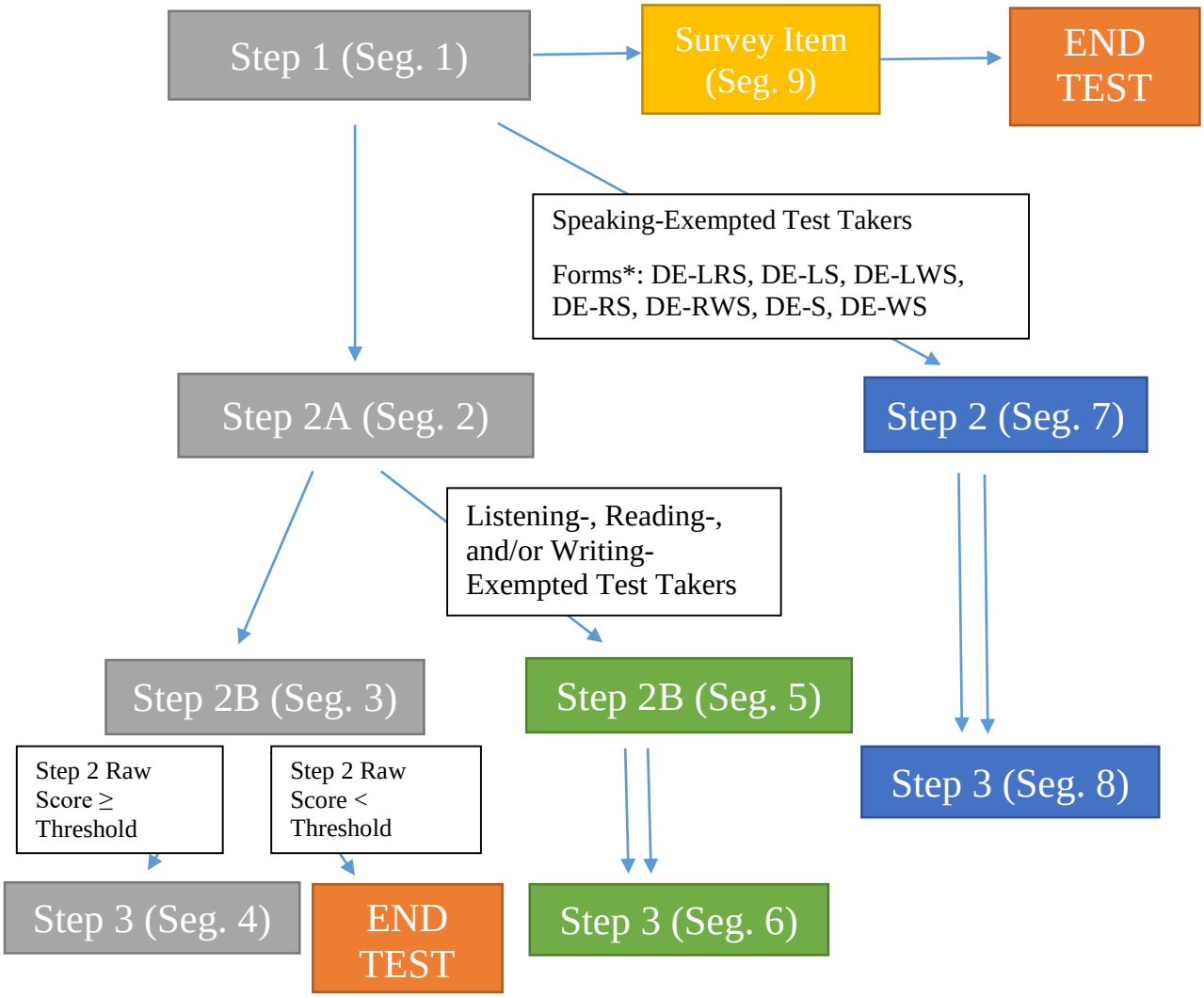
The online form has three steps. Step 1 consists of practice items, while Steps 2 and 3 include operational items. To allow for domain exemptions and because test administrator (TA) input is required (at the end of Step 1 and for the scoring of speaking items in Step 2), the three steps are administered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

- 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 "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 (Step 2A), 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.
- 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 raw score meets or exceed the threshold score, the test is routed to Segment 4 (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 operational machine-scored items from all non-exempted domains. Upon completion of Segment 5, students proceed to Segment 6, 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 which their exempted domains are suppressed. Upon completion of Segment 7, students proceed to Segment 8 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 2019–2020 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

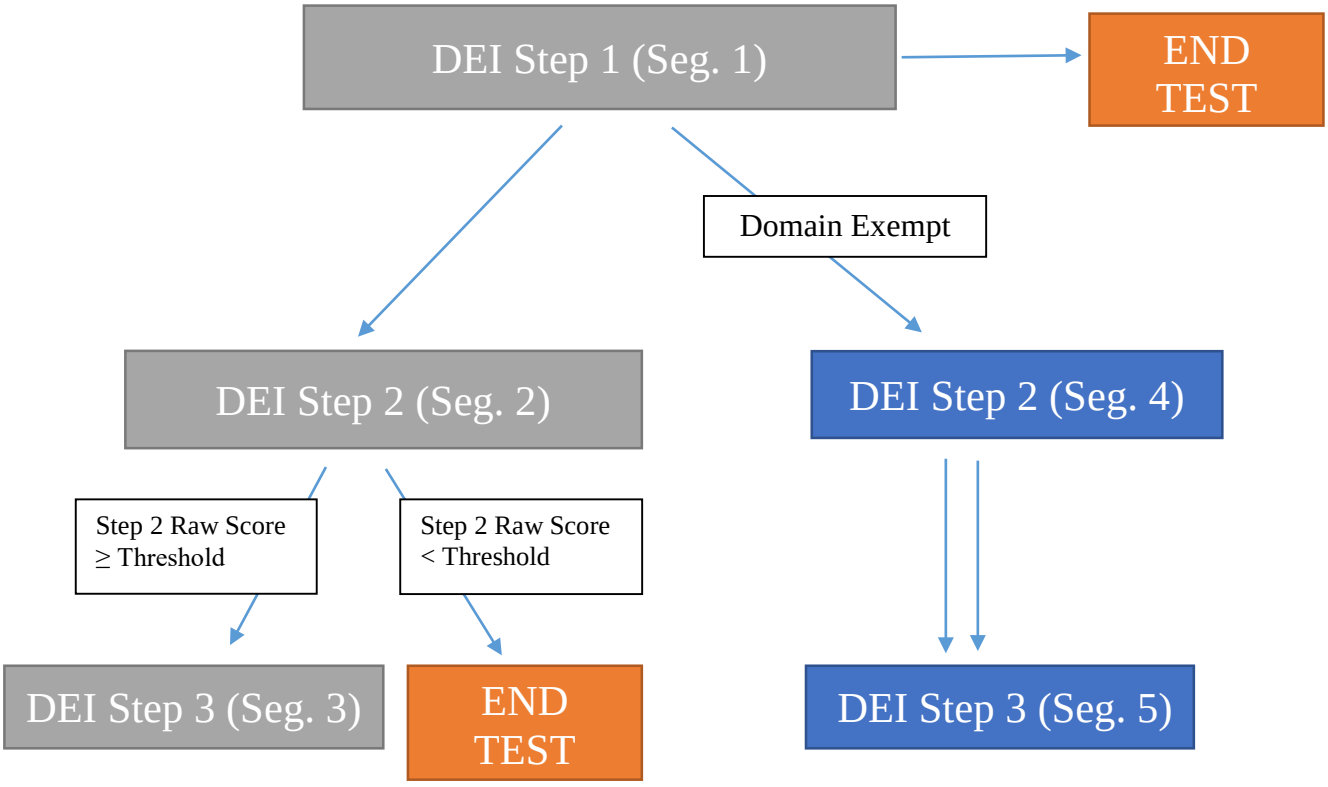
Grade Band	Threshold Score	Step 2 Max Score
PreK/K	23	26
1	24	27
2–3	25	28
4–5	26	31
6–8	28	33
9–12	27	30

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 exceed the threshold score, the test is routed to Segment 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 were entered into the Cambium Assessment, Inc.'s (CAI's) Data Entry Interface (DEI) on the state testing portal for all ELPA21 domain tests. Practice test items were not entered in the DEI and were not scored.

Figure 1.2 2019–2020 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 Tables 1.3– 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											
	PreK/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											
	PreK/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											
	PreK/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 2019–2020 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) were 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 the 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 2019–2020 screener assessment include:

- 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.
- If a single item in Step 2 is attempted, all domains without domain exemptions are considered attempted, and all non-attempted items in Step 2 should be given a score of zero.
- If the student's test is stopped by the automatic stopping rule after Step 2, items in Step 3 should be treated as not presented. If the student's test continues to Step 3, all items in Step 3 that the student does not respond to should be scored as zero.
- If a student has a domain exemption for a domain, the domain is reported as exempt if it is

not attempted.

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

- ELPA21 states made the decision of whether to use the PreK test on an individual basis.
- For Ohio screener administration, handscored items are scored by local TAs.
- Tests in which the TA indicates that the student will not continue after the Step 1 practice items will be scored as follows:
  - Each domain will be scored 0. The score of 0 will receive a label of "Performance Not Determined."
  - Proficiency status will be scored as "D" and reported as "Proficiency Not Demonstrated."

Chapter 2. 2019–2020 Summary

The 2019–2020 screener results are presented in this chapter and in Sections 14–18 of the appendix. The figures and tables included in each section are listed below:

- Section 14. Screener—Student Participation
  - Table S14.1 displays the number and percentage of students in each test mode of braille, paper-pencil, and online in each grade (PreK–12) and across the state.
  - Table S14.2 lists the number and percentage of students taking each test by subgroups 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 the states. The pooled analysis includes the summary by gender and ethnicity.
- Section 15. Screener Assessment—Scale Score Summary
  - Tables S15.1–S15.14 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 PreK–12. Subgroups can vary across the states. The pooled analysis includes the summary by gender and ethnicity.
  - Table S15.15 summarizes the number and percentage of students who were marked "non-attempt" or "exempt" in each domain and grade.
- Section 16. Screener Assessment—Percentage of Students by Domain Performance Level
  - Figure S16.1 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 S16.1–S16.14 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 17. Screener Assessment—Percentage of Students by Overall Proficiency Level
  - Figure S17.1 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 S17.1–S17.14 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 18. Screener Assessment—Testing Time
  - Table S18.1 shows testing time by end step in each grade/grade band.

2.1 2019–2020 STUDENT PARTICIPATION

In the 2019-2020 administration, not all the eligible students have completed the tests due to the COVID-related school closure beginning in March 2020. Table 2.1 shows the overall student participation for each state. There were 78,713 students in total who took the 2019–2020 screener tests. The state of Washington had the most students, followed by the state of Ohio. Most students were from pre-kindergarten 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 2019–2020 school year.

Section S14.1 of the appendix presents student participation in each mode. In the seven states combined, the most frequent mode of administration was online (99.88%), followed by paper (0.12%) and braille (<0.01%).

Section S14.2 of the appendix shows student participation by subgroups. For the pooled analysis, the number of students tested decreases as the grade level increases, with some fluctuation at grades 7 and 9. There were more male students (45.3%–50.2%) than female students (43%–49.6%). In each test, the greatest number of students were in the group of Hispanic or Latino (24.9%–70.5%), followed by Asian students (11.4%–18.3%), and White students (4.8%–9.6%).

Table 2.1 Number of Students who Participated in ELPA21 Screener in 2019–2020 by State and Grade

Grade	Arkansas	Iowa	Louisiana	Nebraska	Ohio	Washington	West Virginia	Total
PreK	≥2,150	≥3,100	≥1,920	≥2,710	<10	≥10	≥180	≥10,090
K	≥1,320	≥170	≥200	≥ 60	≥9,960	≥14,310	≥50	≥26,100
1	≥540	≥430	≥810	≥310	≥1,610	≥1,970	≥60	≥5,760
2	≥440	≥380	≥630	≥210	≥1,240	≥1,420	≥40	≥4,390
3	≥390	≥370	≥580	≥180	≥1,080	≥1,300	≥80	≥4,010
4	≥310	≥360	≥540	≥210	≥930	≥1,200	≥80	≥3,650
5	≥380	≥280	≥480	≥150	≥930	≥1,100	≥70	≥3,410
6	≥320	≥280	≥490	≥90	≥780	≥1,130	≥40	≥3,150
7	≥340	≥250	≥480	≥90	≥830	≥1,060	≥120	≥3,200
8	≥310	≥260	≥440	≥ 90	≥680	≥980	≥30	≥2,810
9	≥430	≥530	≥940	≥220	≥1,300	≥1,600	≥60	≥5,100
10	≥450	≥270	≥240	≥90	≥680	≥1,140	≥70	≥2,960
11	≥440	≥160	≥140	≥50	≥410	≥1,110	≥50	≥2,390
12	≥240	≥110	≥60	≥40	≥250	≥870	≥30	≥1,620
Total	≥8,130	≥7,010	≥8,010	≥4,540	≥20,720	≥29,250	≥1,020	≥78,710

Table 2.2 Number of Students Participating in 2019–2020 ELPA21 Summative, Screener Tests, and Both; by State and Grade-band

State	Grade/Grade Band	N Summative	N Screener	N Both
Arkansas	PreK and K	≥4,640	≥3,480	≥3,050
	1	≥4,360	≥540	≥440
	2-3	≥7,170	≥840	≥620
	4-5	≥5,690	≥690	≥500
	6-8	≥7,340	≥980	≥740
	9-12	≥9,890	≥1,580	≥1,180
Iowa	PreK and K	≥4,450	≥3,280	≥2,860
	1	≥3,800	≥430	≥330
	2-3	≥5,550	≥760	≥510
	4-5	≥4,330	≥640	≥420
	6-8	≥5,850	≥800	≥570
	9-12	≥7,560	≥1,090	≥800
Louisiana	PreK and K	≥3,400	≥2,120	≥1,910
	1	≥3,760	≥810	≥680
	2-3	≥5,870	≥1,210	≥1,010
	4-5	≥4,540	≥1,030	≥830
	6-8	≥5,440	≥1,420	≥1,220
	9-12	≥5,950	≥1,390	≥1,170
Nebraska	PreK and K	≥3,880	≥2,770	≥2,470
	1	≥3,540	≥310	≥240
	2-3	≥4,900	≥400	≥290
	4-5	≥3,300	≥360	≥240
	6-8	≥3,180	≥280	≥220
	9-12	≥4,290	≥410	≥270
Ohio	K	≥10,120	≥9,960	≥8,970
	1	≥8,800	≥1,610	≥1,230
	2-3	≥13,170	≥2,320	≥1,740
	4-5	≥8,410	≥1,860	≥1,230
	6-8	≥10,000	≥2,300	≥1,600
	9-12	≥13,210	≥2,650	≥1,990
Washington	PreK and K	≥15,290	≥14,320	≥9,890
	1	≥15,780	≥1,970	≥1,240
	2-3	≥26,740	≥2,720	≥1,470
	4-5	≥19,470	≥2,310	≥1,120
	6-8	≥21,970	≥3,180	≥1,620
	9-12	≥23,190	≥4,730	≥2,250
West Virginia	PreK and K	≥200	≥240	≥190
	1	≥250	≥60	≥40
	2-3	≥350	≥130	≥60
	4-5	≥270	≥150	≥40
	6-8	≥360	≥200	≥60
	9-12	≥510	≥210	≥140

2.2 2019–2020 Student Scale Score and Performance Level Summary

Tables 2.3– 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 (2–3, 4–5, 6–8, and 9–12). Scale score summary by subgroup for each grade is also presented in Section 15 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 was the most frequent level achieved in speaking and writing in grades PreK–10, and in listening and reading in grades 1–10. Reading and speaking followed a similar pattern: the percentage of students who reached level 1 increased from kindergarten to grade 1, then decreased until grade 6 (with slight increase in grade 5), slightly increased again up to grade 9, and consistently decreased afterwards. For writing, the percentage of students in level 1 decreased from pre-kindergarten to grade 6 (with slight increase in grade 3), then slightly increased to grade 9 and decreased in the remaining grades. Disaggregated results by gender and ethnicity are provided in Section 16 of the Appendix.

Table 2.8 and Figure S17.1 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 K–4, consistently decrease from grade 4 to grade 9, and slightly increase above grade 9. 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 17 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
PreK	≥9,480	314	512.8	714	65.8	≥9,480	318	509.5	708	65.0
K	≥25,010	314	519.8	714	66.2	≥25,020	318	516.5	708	65.5
1	≥5,310	288	497.2	678	94.0	≥5,310	286	480.6	704	96.5
2	≥4,030	286	482.8	710	89.3	≥4,030	278	471.3	734	97.9
3	≥3,670	286	499.3	710	101.7	≥3,670	278	494.2	734	111.0
4	≥3,320	270	478.7	778	112.0	≥3,320	270	482.1	795	110.9
5	≥3,070	270	496.2	772	121.3	≥3,070	270	500.5	786	120.4
6	≥2,700	279	496.2	738	103.9	≥2,700	296	500.1	733	101.2
7	≥2,660	279	489.8	738	109.3	≥2,660	296	496.7	733	106.7
8	≥2,360	274	497.5	738	108.6	≥2,360	296	506.4	733	105.9
9	≥4,160	297	475.4	731	107.7	≥4,160	309	480.4	733	103.5
10	≥2,680	297	518.1	731	99.4	≥2,680	309	521.4	733	96.8
11	≥2,260	297	547.8	731	94.9	≥2,260	309	551.3	733	92.3
12	≥1,550	297	555.6	731	89.4	≥1,550	309	559.6	733	87.3

* Domains with Exemption 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
PreK	≥9,480	339	505.2	711	80.2	≥9,480	347	475.8	684	55.2
K	≥25,000	339	514.9	711	80.6	≥25,020	347	483.9	684	59.9
1	≥5,310	310	477.5	669	100.6	≥5,310	283	476.7	698	97.2
2	≥4,030	292	460.3	703	104.1	≥4,030	276	467.7	737	100.2
3	≥3,670	292	474.8	703	116.6	≥3,670	276	492.3	737	113.1
4	≥3,310	270	480.0	786	132.9	≥3,320	268	478.5	797	116.0
5	≥3,070	270	493.2	782	139.6	≥3,070	268	498.3	791	125.2
6	≥2,700	296	496.2	732	111.1	≥2,700	281	493.4	741	104.3
7	≥2,660	296	486.0	732	115.5	≥2,660	281	489.5	741	109.4
8	≥2,360	284	493.8	732	114.2	≥2,360	281	498.4	741	108.7
9	≥4,160	332	484.8	722	106.5	≥4,160	315	479.9	732	100.2
10	≥2,680	332	525.5	722	96.6	≥2,680	314	518.7	732	93.3
11	≥2,260	332	555.8	722	88.6	≥2,260	315	544.7	732	88.0
12	≥1,550	332	564.5	722	82.2	≥1,550	315	552.6	732	83.3

* Domains with Exemption 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
PreK	≥9,480	3978	5324.7	6375	500.9	≥9,480	3646	5073.6	6763	506.7
K	≥25,020	3978	5363.1	6375	493.7	≥25,020	3646	5140.8	6763	515.4
1	≥5,310	3785	5120.7	6387	649.6	≥5,310	3364	4943.8	6629	773.4
2	≥4,030	3756	5027.7	6439	664.9	≥4,030	3326	4842.0	6880	780.5
3	≥3,670	3756	5153.0	6439	738.3	≥3,670	3326	5002.0	6880	888.3
4	≥3,320	3649	5008.2	6700	741.9	≥3,320	3237	4929.1	7401	944.8
5	≥3,070	3649	5121.6	6700	800.0	≥3,070	3237	5068.2	7352	1,014.0
6	≥2,700	3803	5152.1	6476	717.6	≥2,700	3388	5065.5	6974	835.2
7	≥2,660	3803	5116.8	6476	753.5	≥2,660	3388	5016.0	6974	875.6
8	≥2,360	3753	5178.3	6476	756.7	≥2,360	3388	5084.3	6974	867.7
9	≥4,160	3787	4989.5	6524	755.0	≥4,160	3605	4937.2	6923	828.8
10	≥2,680	3787	5292.8	6524	716.7	≥2,680	3605	5266.4	6923	761.2
11	≥2,260	3787	5515.7	6524	693.1	≥2,260	3605	5502.3	6923	714.2
12	≥1,550	3787	5573.3	6524	658.8	≥1,550	3605	5569.6	6923	669.2

* Scale scores cannot be compared across grade bands.

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

Grade	Listening							Reading						
	N	0	1	2	3	4	5	N	0	1	2	3	4	5
PreK	≥10,090	6.0	21.0	15.5	54.2	1.7	1.7	≥10,090	6.0	24.4	19.7	44.3	3.2	2.4
K	≥26,080	4.1	19.9	15.0	55.6	2.4	3.0	≥26,090	4.1	23.2	18.8	45.9	4.0	4.0
1	≥5,750	7.7	27.4	6.8	28.9	12.8	16.4	≥5,750	7.7	50.2	10.4	15.8	7.2	8.8
2	≥4,380	8.1	24.3	9.3	22.0	17.8	18.4	≥4,390	8.1	45.7	7.4	19.6	7.2	12.0
3	≥3,980	7.8	23.8	10.8	20.1	18.4	19.1	≥3,980	7.8	45.9	12.7	16.8	7.1	9.8
4	≥3,620	8.3	28.5	6.7	12.0	20.2	24.3	≥3,620	8.3	39.8	9.1	15.4	10.2	17.1
5	≥3,380	9.0	29.6	6.7	7.3	21.1	26.2	≥3,380	9.0	39.8	9.8	16.0	8.1	17.2
6	≥3,140	13.9	24.3	6.5	11.3	18.2	25.8	≥3,140	13.9	34.7	6.7	19.1	9.8	15.7
7	≥3,150	15.5	32.8	7.0	17.0	9.8	18.1	≥3,150	15.5	41.4	9.7	17.0	7.7	8.8
8	≥2,800	15.7	31.9	7.7	16.7	11.3	16.6	≥2,800	15.7	40.6	9.9	22.9	6.6	4.3
9	≥5,060	17.9	41.3	7.0	14.3	7.8	11.8	≥5,060	17.9	47.1	9.8	16.0	4.7	4.6
10	≥2,940	8.8	27.7	9.2	21.7	13.9	18.7	≥2,940	8.8	35.6	14.4	26.8	7.9	6.6
11	≥2,380	5.1	18.1	8.0	23.6	16.2	29.1	≥2,380	5.1	24.2	14.8	35.2	10.9	9.8
12	≥1,610	3.8	14.0	8.6	25.3	19.0	29.2	≥1,610	3.8	19.9	15.4	38.4	12.3	10.2

* Level 0: Performance Not Determined.
* Domains with Exemption 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
PreK	≥10,080	6.0	35.8	20.6	24.7	9.8	3.1	≥10,090	6.0	63.6	24.7	4.7	0.7	0.3
K	≥26,070	4.1	32.5	21.0	27.0	10.0	5.4	≥26,090	4.1	60.7	26.1	7.0	1.2	0.8
1	≥5,750	7.7	57.3	19.1	3.8	4.8	7.2	≥5,750	7.7	56.4	11.8	13.9	4.6	5.8
2	≥4,380	8.1	50.1	14.9	8.7	7.6	10.6	≥4,390	8.1	45.3	11.0	15.6	7.0	13.0
3	≥3,980	7.8	48.0	11.3	10.1	10.4	12.5	≥3,980	7.8	47.5	11.4	14.5	7.3	11.5
4	≥3,620	8.3	38.0	7.9	12.2	9.8	23.8	≥3,620	8.3	37.3	8.4	22.8	7.3	15.8
5	≥3,380	9.0	39.0	8.5	11.4	8.8	23.4	≥3,380	9.0	33.9	8.3	25.4	6.4	17.0
6	≥3,130	13.9	30.9	8.2	21.2	9.0	16.7	≥3,140	13.9	28.5	8.2	24.8	7.8	16.8
7	≥3,150	15.5	37.4	11.2	16.3	6.5	13.1	≥3,150	15.5	39.7	10.0	17.6	6.8	10.6
8	≥2,800	15.7	35.1	10.1	19.5	8.1	11.5	≥2,800	15.7	39.0	11.2	19.5	7.0	7.6
9	≥5,060	17.9	41.8	10.0	15.0	5.3	10.0	≥5,060	17.9	46.9	9.0	15.8	3.8	6.5
10	≥2,940	8.8	28.9	13.8	23.6	9.7	15.2	≥2,940	8.8	34.7	14.2	25.6	7.3	9.3
11	≥2,380	5.1	19.2	13.5	25.4	12.5	24.4	≥2,380	5.1	23.8	15.2	33.1	9.7	13.1
12	≥1,610	3.8	15.5	12.5	29.2	14.7	24.3	≥1,610	3.8	19.4	16.6	35.9	10.8	13.4

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

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

Grade	N	Emerging	Progressing	Proficient	Proficiency Not Demonstrated
PreK	≥10,090	32.1	59.7	2.3	6.0
K	≥26,090	30.0	63.9	1.9	4.1
1	≥5,750	33.7	51.2	7.4	7.7
2	≥4,390	33.2	44.0	14.7	8.1
3	≥3,980	34.5	43.4	14.4	7.8
4	≥3,620	34.8	35.0	21.9	8.3
5	≥3,380	36.0	33.4	21.5	9.0
6	≥3,140	29.6	37.4	19.1	13.9
7	≥3,150	38.5	32.9	13.2	15.5
8	≥2,800	37.9	36.5	9.8	15.7
9	≥5,060	46.5	27.8	7.9	17.9
10	≥2,940	34.6	44.8	11.8	8.8
11	≥2,380	24.0	54.2	16.8	5.1
12	≥1,610	19.9	58.4	17.8	3.8

2.3 2019–2020 TESTING TIME FOR ONLINE SCREENER TESTS

In the 2019–2020 online screener tests, students who did not have domain exemption were proceeded to Segments 2 and 3 (Step 2) and were proceeded 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 S18.1 of the appendix summarizes testing time by end step in each grade/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 the 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 4 in Part I of the technical report, the reliability for screener tests is assessed using

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

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

- Section 19. Screener Assessment—Marginal Reliability
    - Figure S19.1 shows the ratio of marginal standard error of measurement to the standard deviation of scale scores at the test level, by domain and grade.
    - Figure S19.2 presents the marginal reliability for each domain test across grades.
- Section 20. Screener Assessment—Conditional Standard Error of Measurement (CSEM)
    - Figures S20.1–S20.14 show the CSEM plots for each domain, overall, and comprehension scores. 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 completed Step 3.
- Section 21. Screener Assessment—Classification Accuracy and Consistency
    - Figure S21.1 shows the classification accuracy for each domain test.
    - Figure S21.2 shows the classification consistency for each domain test.
    - Figure S21.3 presents the classification accuracy and consistency for the overall proficiency.
- Section 22. Screener Assessment—Inter-Rater Analysis

    Tables S22.1–S22.7 display the inter-rater analysis result for each handscore 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, and the analysis for the ratio is displayed in Figure S19.1 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 S19.2 in the appendix. Pre-kindergarten and kindergarten have lower marginal reliability than the other grades. Writing has lower marginal reliability at pre-kindergarten, kindergarten, and high school grades, but has higher reliability from grades 1–5. Listening has relatively lower reliability than the other domains in grades 1–5. In addition, Section 20 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
PreK	≥9,480	.75	.72	.77	.65	.69	.73
K	≥25,000	.75	.72	.77	.68	.68	.73
1	≥5,310	.82	.88	.85	.88	.75	.87
2	≥4,030	.85	.91	.88	.91	.81	.91
3	≥3,670	.87	.92	.90	.93	.83	.92
4	≥3,310	.90	.93	.92	.93	.87	.93
5	≥3,070	.91	.93	.92	.93	.87	.94
6	≥2,700	.92	.91	.91	.91	.87	.93
7	≥2,660	.93	.91	.92	.92	.88	.93
8	≥2,360	.92	.91	.92	.92	.88	.93
9	≥4,160	.93	.92	.91	.88	.91	.92
10	≥2,680	.92	.91	.90	.87	.89	.91
11	≥2,260	.90	.90	.89	.86	.87	.90
12	≥1,550	.89	.88	.87	.85	.85	.89

* Domains with Exemption are excluded.

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. Classification consistency rates can be lower than classification accuracy 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 are presented in Tables 3.3–3.4 as well as Figures S21.1–S21.2 in the appendix. Across the four performance cut scores, the classification accuracy indices are all above .8, denoting that the degree to which we can reliably differentiate students between adjacent performance levels is typically above or close to .8. In terms of classification consistency, the indices are all above .7 in all cuts and all grades. The reliability indices in the middle school tests are above .9 for all domains. Table 3.5 and Figure S21.3 in the appendix display the classification

accuracy and consistency for overall proficiency categories. The plot shows that all the accuracy and consistency indices are above .8. The accuracy indices for Between Emerging and Progressing are lower than those for Between Progressing and Proficient in pre-kindergarten and kindergarten and are comparable than 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
PreK	.70	.61	.59	.74	.58	.50	.52	.65
K	.69	.60	.58	.72	.57	.49	.51	.63
1	.64	.74	.72	.78	.55	.67	.67	.72
2	.64	.77	.70	.78	.55	.70	.65	.71
3	.66	.76	.71	.78	.56	.70	.66	.72
4	.73	.77	.73	.78	.64	.70	.68	.71
5	.76	.78	.74	.79	.68	.72	.68	.72
6	.75	.76	.72	.74	.67	.68	.64	.66
7	.77	.77	.75	.78	.70	.71	.69	.71
8	.76	.78	.74	.77	.68	.72	.67	.71
9	.81	.82	.78	.79	.74	.77	.71	.72
10	.74	.75	.70	.71	.65	.68	.62	.63
11	.72	.70	.67	.67	.63	.62	.58	.58
12	.70	.69	.65	.66	.61	.60	.56	.57

* Domains with Exemption 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
PreK	.91	.84	.93	.97	.88	.82	.89	.95	.87	.85	.89	.93	.80	.94	.99	.99
K	.91	.85	.93	.96	.88	.82	.89	.94	.87	.84	.88	.92	.81	.93	.98	.99
1	.92	.91	.87	.89	.91	.92	.94	.95	.87	.89	.91	.93	.93	.93	.94	.95
2	.91	.92	.88	.90	.94	.93	.93	.95	.90	.89	.90	.93	.93	.93	.95	.96
3	.91	.94	.89	.89	.94	.93	.93	.94	.92	.90	.90	.92	.94	.94	.94	.95
4	.94	.95	.92	.91	.94	.94	.94	.94	.94	.92	.91	.92	.95	.94	.94	.94
5	.95	.95	.93	.91	.95	.95	.94	.93	.94	.92	.91	.91	.95	.95	.94	.94
6	.94	.96	.94	.91	.95	.94	.92	.93	.95	.91	.90	.92	.93	.94	.92	.93
7	.95	.96	.93	.92	.95	.94	.93	.93	.95	.92	.93	.94	.95	.94	.93	.94
8	.95	.96	.93	.92	.95	.94	.92	.94	.95	.92	.92	.93	.94	.94	.93	.94
9	.95	.96	.94	.95	.95	.94	.95	.96	.94	.94	.93	.94	.93	.93	.95	.96
10	.95	.95	.92	.91	.94	.92	.92	.94	.94	.92	.90	.91	.91	.91	.92	.94
11	.96	.95	.91	.88	.95	.92	.89	.92	.94	.92	.88	.89	.92	.90	.90	.91
12	.96	.95	.90	.88	.95	.91	.89	.91	.95	.92	.86	.88	.93	.89	.89	.91

* Domains with Exemption 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
PreK	.86	.77	.90	.95	.82	.75	.85	.93	.82	.79	.85	.89	.73	.91	.98	.99
K	.87	.78	.89	.94	.82	.75	.84	.91	.82	.78	.84	.88	.73	.89	.97	.98
1	.88	.86	.82	.85	.87	.89	.91	.93	.83	.85	.87	.90	.90	.90	.91	.93
2	.87	.88	.84	.86	.91	.90	.91	.93	.86	.85	.86	.90	.90	.90	.92	.94
3	.88	.91	.85	.85	.91	.90	.90	.92	.89	.86	.86	.89	.92	.91	.92	.93
4	.91	.92	.89	.87	.92	.91	.91	.92	.92	.89	.87	.88	.92	.91	.91	.92
5	.92	.93	.90	.87	.93	.93	.91	.91	.92	.90	.88	.88	.93	.92	.91	.91
6	.92	.94	.91	.87	.92	.92	.89	.90	.93	.88	.86	.89	.90	.91	.89	.90
7	.93	.94	.90	.89	.93	.91	.90	.91	.92	.89	.89	.91	.92	.91	.91	.92
8	.93	.94	.89	.88	.93	.91	.89	.92	.92	.89	.88	.91	.92	.91	.90	.91
9	.93	.94	.92	.92	.93	.92	.93	.95	.91	.92	.90	.92	.89	.91	.93	.94
10	.92	.92	.88	.88	.92	.89	.90	.92	.91	.89	.86	.88	.88	.87	.89	.91
11	.94	.93	.87	.84	.92	.89	.85	.89	.92	.89	.83	.84	.89	.86	.86	.88
12	.95	.92	.85	.84	.92	.88	.85	.88	.93	.88	.81	.83	.89	.85	.85	.88

* Domains with Exemption 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
PreK	.86	.87	.98	.81	.83	.98
K	.86	.88	.99	.82	.83	.98
1	.87	.91	.95	.82	.88	.94
2	.87	.93	.94	.82	.89	.93
3	.88	.94	.94	.84	.91	.92
4	.89	.95	.94	.85	.93	.92
5	.89	.96	.93	.86	.94	.91
6	.89	.96	.93	.85	.94	.91
7	.89	.96	.94	.86	.94	.92
8	.89	.95	.94	.86	.93	.93
9	.91	.95	.96	.89	.94	.95
10	.88	.95	.94	.85	.93	.92
11	.86	.95	.91	.83	.93	.89
12	.86	.95	.91	.82	.93	.89

3.4 Inter-Rater Analysis

In the 2019–2020 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 and high school 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 (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 .70, except for one item in grade 1, three items in grade band 6–8, and five items in grade band 9–12. Overall, 73%–90.1% 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 22 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
PreK	2	.842	.897	77.8	82.5	100.0	100.0	0.0	0.0
K	2	.882	.898	80.7	82.3	100.0	100.0	0.0	0.0
1	2	.669	.878	75.0	84.8	100.0	100.0	0.0	0.0
2–3	3	.846	.902	80.6	86.1	100.0	100.0	0.0	0.0
4–5	4	.796	.929	75.0	85.6	100.0	100.0	0.0	0.0
6–8	9	.596	.944	75.6	86.9	100.0	100.0	0.0	0.0
9–12	9	.529	.933	73.0	90.1	100.0	100.0	0.0	0.0

Chapter 4. Validity

Discussions on the test development, form construction, scaling, equating, and standard setting can be found in the related documents from ELPA21.

Since the items and the 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 the student progress from the time they took screener tests to the time they took summative tests. The statistical methods and the results are presented in this chapter and Sections 23–24 in the appendix:

- Section 23. Correlations Between Summative and Screener Tests
    - Table S23.1 shows the correlations between domain, overall and comprehension scores.
    - Table S23.2 summarizes the correlations by between domain performance level and overall proficiency categories.
- Section 24. Student Progress from Screener to Summative
    - Figures S24.1–S24.2 display within-year average differences in domain, overall, and comprehension scale score.
    - Figures S24.3–S24.4 present changes domain performance level and overall proficiency.
    - Figures S24.5–S24.10 show scatter plots of scale scores for the screener and summative assessment.
    - Tables S24.1–S24.6 summarize the comparison of scale score summary statistics between for domain, overall and comprehension scores.

4.1 Comparisons of Performance from Screener to Summative

Students who took the ELPA21 Screener and were classified as an English Learner (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 are 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 S23.1 and S23.2.

Table S23.1 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 .8 in listening, reading, writing, comprehension, and overall scale scores in grade 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 S23.2 shows the Gamma correlations between domain performance levels and test proficiency categories. Similar to the correlations between scale scores presented in Table S23.1, kindergarten has the lowest correlations in all domain performance levels and overall proficiency categories. For grade 2 and above, the correlations are about .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 is 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 Chapter 3 of the technical report Part II and Part III, we can still see the trend of student progress. Section 24 of the appendix summarizes the results of progress analysis. In each of the analyses, only students who had valid scores on both the screener and summative tests were included.

Figures S24.1 and S24.2 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 S24.3 and S24.4 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. In addition, Figures S24.5–S24.10 in the appendix present scatter plots of scale score change from screener to summative assessments for each grade, and Tables S24.1–S24.6 summarize comparisons of scale scores between screener and summative assessments.

Chapter 5. Reporting

The detailed introduction for the ORS can be found in Chapter 6 in Part I of the technical report. The reporting mockups for the screener tests of each state are included in Section 25 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

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