TECHNICAL REPORT

PART II – SUMMATIVE ASSESSMENT

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

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

Grades K–12

2020–2021 Administration

Submitted to:
ELPA21

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

December 2021

Table of Contents

List of Tables

Chapter 1. Test Administrations

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

1.1 Testing Windows

The 2020–2021 summative assessment windows for the seven states discussed in this report are shown in Table 1.1. While testing windows remained open in the spring of 2021, some students were unable to complete the English Language Proficiency Assessment for the 21st Century (ELPA21) due to the ongoing impacts of the coronavirus (COVID-19) pandemic.

Table 1.1 2020–2021 ELPA21 Summative Testing Windows by State

State	ELPA21 Summative
Arkansas	1/25/2021–3/19/2021
Iowa	2/1/2021–4/9/2021
Louisiana	2/1/2021–3/12/2021
Nebraska	2/8/2021–4/2/2021
Ohio	2/1/2021–4/23/2021
Washington	3/22/2021–6/4/2021
West Virginia	2/2/2021–4/2/2021

1.2 Test Design

The 2020–2021 summative assessment included one online form, one paper-pencil form, and one braille form. Each form had separate tests for the four language domains.

Tables 1.2–1.4 list the number of operational items and score points in each online, paper-pencil, and braille form. The tables show that listening and reading had comparable numbers of items between online and paper forms in each test. Braille form has fewer items than the two other forms. Writing and speaking had fewer but comparable numbers of items in each test. No field-test items were included in the 2020–2021 summative assessments.

Table 1.2 Number of Items and Score Points by Domain and Grade Band—Online Summative

	Grade/Grade Band											
	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	29	29	24	24	25	26	29	32	34	38	23	26
Reading	23	23	30	30	30	35	27	30	29	33	38	40
Speaking	11	27	9	25	9	25	8	30	7	27	7	27
Writing	18	18	20	20	14	24	13	30	8	28	8	28
Total	81	97	83	99	78	110	77	122	78	126	76	121

Table 1.3 Number of Items and Score Points by Domain and Grade Band—Paper Summative

	Grade/Grade Band											
	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	28	28	22	22	23	24	24	27	30	31	21	21
Reading	23	23	29	29	26	28	26	28	28	32	35	38
Speaking	11	27	9	25	9	25	8	30	7	27	7	27
Writing	11	18	9	16	10	20	10	27	8	28	8	28
Total	73	96	69	92	68	97	68	112	73	118	71	114

Table 1.4 Number of Items and Score Points by Domain and Grade Band—Braille Summative

	Grade/Grade Band											
	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	17	19	21	21	20	20	23	26	22	23	19	21
Reading	13	13	22	22	23	25	23	23	25	29	34	37
Speaking	4	12	7	17	8	20	7	25	6	22	5	19
Writing	10	23	7	19	9	24	10	30	8	28	8	28
Total	44	67	57	79	60	89	63	104	61	102	66	105

1.3 TEST ADMINISTRATION MANUAL

1.3.1 Directions for Test Administration

For 2020–2021, the *Test Administration Manual* (TAM) was developed to guide test administrators (TAs) through the summative assessment. The TAM covers the following key points:

- Overview of the ELPA21 summative assessment
- TA qualifications
- Preliminary planning
- Materials required
- Administrative considerations
- Student preparation/guidance for practice tests
- Detailed instructions for preparing and administering the training tests and summative tests
- Test security instructions
- 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 were provided before and during the testing windows. Training/practice tests could be accessed through a non-secure or secure browser.

The summative assessment training tests have two components, one for TAs to create and manage the training/practice test sessions and the other for students to take an actual training/practice test.

The *Practice Test Administration* site introduces TAs to the following procedures:

- logging in;
- starting a test session;
- providing the session ID to the students who are signing into the test session;
- monitoring students' progress throughout their tests; and
- ending the test.

The *Practice Tests* site introduces students to the following procedures:

- 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 Instructions for Summative Assessments

The TA instructions for summative assessments include brief directions for each domain test. Detailed instructions for the following procedures are also provided:

- logging in to the Secure Browser;
- starting a test session;
- providing the session ID to the students;
- approving student test sessions, including reviewing and editing students' test settings and accommodations;
- monitoring students' progress throughout their tests by checking their testing statuses; and
- ending the test session and logging out.

1.4 Business Scoring Rules for the Summative Assessment

Business rules and instructions applicable to the 2020–2021 ELPA21 summative assessment included the following:

1. A domain test was considered "attempted" if a student was presented with the first operational item; it was not necessary for the student to respond to at least one item.
2. If a domain test was attempted, any items without a response (i.e., skipped, omitted, not reached) in that domain were assigned the minimum score (0 points).
3. If a domain test was not attempted and the student was not marked as "exempt" in that domain, the domain score and performance level were assigned the code "N" (Domain Not Attempted).
4. If any domain tests were exempted before a student started the first domain test, items from the exempted domains were excluded from the computation of the domain and composite scores. In this case, the domain score and performance level were assigned the code "E" (Domain Exempted). However, if the domain test was started in Cambium Assessment, Inc.'s Test Delivery System (TDS), the test was considered attempted even if an exemption was intended. In that case, items in the domain were included in the computation of scores.
5. If no domains were attempted (i.e., every domain was either not attempted or exempted), the overall composite score, domain score, and comprehension score were assigned the code "N."
6. If a student was exempted from reading or listening, the exempted domain was excluded from the computation of the comprehension score. For the comprehension score results, see Table 1.5 for reporting of scenarios in which neither listening nor reading were attempted (i.e., each domain was either exempted or non-attempted).

Table 1.5 Scoring Outcome for the Comprehension Score

If Listening is…	and Reading is…	Comprehension is reported as:
Exempt	Exempt	E
Exempt	Not Attempted	N
Not Attempted	Exempt	N
Not Attempted	Not Attempted	N

Chapter 2. 2020–2021 Summary

The 2020–2021 student participation and performance statistics for each state and the pooled analysis for the summative assessment are presented in Sections 1–5 of the Appendix. The figures and tables included in Sections 1–5 are listed here:

- Section 1. Summative Assessment—Student Participation
    - Table S1.1 displays the number and percentage of students in each test mode (braille, paper-pencil fixed form, and online) in each grade (K–12) and across the state (or states, in the case of the pooled analysis).
    - Table S1.2 lists the number and percentage of students taking each test by subgroups (including grade, gender, ethnicity, and primary disabilities) and by other characteristics (e.g., migrant, special education, Title I, or Section 504 Plan status). The pooled analysis includes the summary by gender and ethnicity. Subgroups vary across the states, for example, the female subgroups vary from 43.2% to 48.7% while male subgroups vary from 50.9% to 56.3% across the grade/grade bands
- Section 2. Summative Assessment—Raw Score Summary
    - Tables S2.1–S2.13 present the number of students; the minimum, mean, maximum, and standard deviation of domain raw scores by performance level in each grade and the overall raw scores by proficiency classification in each grade across the states.
- Section 3 Summative Assessment—Raw Score Distributions
    - Figures S3.1–S3.65 present the frequency distributions of raw scores by performance level for each domain in each grade and the frequency distributions of overall raw scores by proficiency classification (overall proficiency level) in each grade.
- Section 4. Summative Assessment—Scale Score Summary
    - Tables S4.1–S4.13 present the number of students; the minimum, maximum, average, and standard deviation of the domain scale scores, overall scale scores and comprehension scale scores across the seven states and by subgroups in each grade. The pooled analysis includes the summary by gender and ethnicity.
    - Table S4.14 summarizes the number and percentage of students who were marked "non-attempt" or "exempt" in each domain and grade.
- Section 5. Summative 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–S5.13 show the total number of students taking each domain test and the percentage of students in each performance level by domain test across the state

and by subgroups. The pooled analysis includes the summary by gender and ethnicity.

- Section 6. Summative 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–S6.13 show the total number of students who are categorized in each of the overall proficiency categories (i.e., Emerging, Progressing, and Proficient) across the state and by subgroups. The pooled analysis includes the summary by gender and ethnicity.
- Section 7. Summative Assessment—Testing Time
    - Table S7.1 summarizes testing time per grade or grade band.

2.1 2020–2021 STUDENT PARTICIPATION

In the 2020–2021 test administration, not all eligible students completed the tests due to the ongoing impacts of the COVID-19 pandemic. Table 2.1 summarizes student participation in each state. There were 272,131 students in total who participated in the 2020–2021 summative assessment. The state of Washington had the most tested students, followed by the state of Ohio.

Table 2.1 Student Participation in Each State by 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
K	4,194	4,644	4,415	4,451	3,242	3,407	3,678	3,888	8,991	10,123	12,042	15,294	205	207	36,767	42,014	-5,247
1	4,482	4,367	3,969	3,805	3,399	3,767	3,423	3,540	8,949	8,802	12,651	15,788	191	259	37,064	40,328	-3,264
2	3,870	3,826	3,205	3,113	3,112	3,277	2,665	2,877	7,068	7,322	11,379	14,772	203	188	31,502	35,375	-3,873
3	3,350	3,352	2,560	2,437	2,476	2,600	1,994	2,028	5,659	5,853	9,551	11,968	122	164	25,712	28,402	-2,690
4	3,061	2,892	2,276	2,231	2,135	2,446	1,577	1,800	4,757	4,419	8,447	10,279	135	134	22,388	24,201	-1,813
5	2,695	2,799	1,910	2,100	1,951	2,099	1,220	1,505	3,484	3,994	7,204	9,193	95	136	18,559	21,826	-3,267
6	2,647	2,464	1,832	2,022	1,709	1,917	1,113	1,209	3,317	3,365	6,278	7,838	103	139	16,999	18,954	-1,955
7	2,410	2,517	1,839	1,802	1,652	1,799	945	965	2,920	3,253	5,666	7,075	112	118	15,544	17,529	-1,985
8	2,496	2,368	1,820	2,026	1,598	1,724	854	1,006	3,039	3,382	5,414	7,064	102	103	15,323	17,673	-2,350
9	2,434	2,529	1,946	2,387	1,653	2,485	983	1,301	3,339	4,293	4,795	7,168	96	137	15,246	20,300	-5,054
10	2,439	2,693	2,036	2,058	1,734	1,559	1,072	1,151	3,197	3,673	4,547	6,614	127	149	15,152	17,897	-2,745
11	2,333	2,550	1,596	1,699	1,112	1,097	825	915	2,683	2,997	3,724	5,100	89	125	12,362	14,483	-2,121
12	1,860	2,121	1,246	1,425	762	812	713	923	2,089	2,247	2,750	4,312	93	108	9,513	11,948	-2,435
Total	38,271	39,122	30,650	31,556	26,535	28,989	21,062	23,108	59,492	63,723	94,448	122,465	1,673	1,967	272,131	310,930	-38,799

Table S1.1 in Section 1 of the Appendix presents student participation in each mode. In the seven states combined, the most frequent mode of test administration was online (99.85%), followed by paper (0.14%) and braille (<0.01%).

Table S1.2 in Section 1 of the Appendix shows student participation by subgroups. For the pooled analysis, the number of students tested decreases as the grade level increases. There were more male students tested (50.9%–56.3%) than female students (43.2%–48.7%). In each test, most students were Hispanic or Latino (57.6%–67.4%), followed by Asian students (8.8%–16.9%) and White students (7.0%–10.4%).

The results from Tables S2.1–S2.13 in Section 2 and Figures S3.1–S3.65 in Section 3 of the Appendix show that most of the students were in category 3 or 4 at the domain level in each grade. At the overall raw score level, most of the students were in the progressing category for all grades.

2.2 2020–2021 Student Scale Score and Performance Summary

Tables 2.2–2.4 summarize student performance in the 2020–2021 test administration across the seven states for the students who completed the tests. These tables show the number of students; the minimum, mean, maximum, and standard deviation of each domain scale scores; and the comprehension and overall scale scores in each grade for the pooled analysis. The ELPA21 tests are not vertically linked across all grades. Scale scores can be compared only within grade-band tests (i.e., grades 2–3, 4–5, 6–8, and 9–12). A disaggregated summary based on subgroups is also available in Section 4 of the Appendix.

Table 2.5 and Table 2.6 display the percentage of students in each performance level for each grade and domain. In addition, Table 2.7 shows the percentage of students in each overall proficiency category in each grade. Sections 5 and 6 of the Appendix further summarize the percentage of students in each domain test by subgroups, by performance level, and by overall proficiency category, respectively.

For both reading and writing in the pooled analysis, Table 2.5 and Table 2.6 show that most students are in performance level 3 except for grade 2 in reading and kindergarten and grade 1 in writing. Middle school and high school students have higher percentages in levels 1 and 2 than in levels 4 and 5. In the listening domain, the greatest number of level 3 students is in grade 7 and above. In the speaking domain, the greatest number of level 3 students is in grade 5 and above. In grades 2–12, more students are in levels 4 and 5 than in levels 1 and 2 in the listening and speaking domains.

The percentage of students in each proficiency category is summarized in Table 2.7 and Figure S6.1 in the Appendix. Table 2.7 shows that most students (70.6%–77.3%) are in the Progressing category in all grades. The percentage of students who are Progressing is relatively stable from kindergarten to grade 2 and the largest increase occurs from grade 2 to 3. The largest drop occurs from grade 3 to grade 4 and remains stable to grade 8, decreases until grade 10, and then increases to grade 12. The percentage of students in the Emerging category decreases from kindergarten to grade 3, then increases until grade 10, and thereafter drops consistently.

*Table 2.2 Scale Score Summary by Grade—Listening and Reading**

Grade	Listening					Reading				
	N	Min	Mean	Max	SD	N	Min	Mean	Max	SD
K	36,730	233	554.2	745	77.7	36,603	247	555.0	740	74.9
1	37,021	233	551.5	711	71.9	36,912	235	530.6	759	82.5
2	31,459	221	530.6	728	63.8	31,354	224	509.8	762	69.3
3	25,679	221	555.6	737	67.0	25,563	224	545.9	770	73.2
4	22,359	216	514.1	722	66.6	22,218	227	511.8	737	66.3
5	18,521	216	531.9	758	69.5	18,416	227	532.8	774	70.0
6	16,948	222	517.6	737	64.4	16,825	239	517.4	752	60.1
7	15,481	222	530.2	768	69.4	15,409	239	532.5	777	65.1
8	15,248	222	543.2	782	76.3	15,198	239	548.5	783	71.2
9	15,137	249	538.8	770	72.6	15,090	257	537.5	782	69.9
10	15,029	249	543.8	758	75.6	15,008	257	543.4	772	74.2
11	12,281	249	557.4	775	72.6	12,245	257	555.1	783	73.4
12	9,441	249	555.5	735	68.9	9,402	257	553.0	753	70.0

*Scores from domain tests marked as Exemption or Not Attempted are excluded.
*Scale scores cannot be compared across grade bands.

*Table 2.3 Scale Score Summary by Grade—Speaking and Writing**

Grade	Speaking					Writing				
	N	Min	Mean	Max	SD	N	Min	Mean	Max	SD
K	36,519	285	569.7	744	90.1	36,564	302	529.8	718	81.1
1	36,857	263	562.5	736	74.1	36,880	238	518.6	741	89.4
2	31,307	251	536.8	749	68.8	31,322	230	502.3	760	76.5
3	25,543	251	562.0	753	72.0	25,544	230	541.4	768	77.6
4	22,228	235	534.7	754	71.8	22,231	222	507.8	725	72.1
5	18,425	235	545.2	782	73.1	18,412	222	529.2	771	73.4
6	16,804	260	536.9	739	70.4	16,791	235	509.9	750	69.0
7	15,340	260	543.8	735	73.9	15,377	235	525.0	775	73.1
8	15,116	260	551.4	773	77.9	15,158	235	538.6	787	79.2
9	14,949	300	555.4	742	75.4	15,009	261	531.2	751	74.6
10	14,868	300	561.3	736	74.9	14,963	261	535.7	741	76.1
11	12,132	300	574.1	732	70.2	12,169	261	548.5	778	70.7
12	9,302	300	574.2	724	68.9	9,350	261	547.7	726	66.7

*Scores from domain tests marked as Exemption or Not Attempted are excluded.
*Scale scores cannot be compared across grade bands.

*Table 2.4 Scale Score Summary by Grade—Comprehension and Overall**

Grade	Comprehension					Overall				
	N	Min	Mean	Max	SD	N	Min	Mean	Max	SD
K	36,752	3361	5522.2	6776	536.6	36,767	3160	5512.2	7023	598.2
1	37,048	3387	5451.9	6698	534.3	37,064	2967	5423.7	7032	611.4
2	31,488	3260	5298.9	6801	483.3	31,502	2930	5252.4	7097	532.4
3	25,699	3260	5517.6	6654	515.2	25,712	2930	5508.5	7174	557.5
4	22,374	3273	5237.4	6817	487.9	22,388	2877	5239.2	6911	532.7
5	18,541	3273	5382.7	6817	520.0	18,559	2877	5384.3	7262	549.7
6	16,986	3323	5269.2	6967	459.7	16,999	2993	5264.3	6915	504.1
7	15,527	3323	5373.2	6967	500.2	15,544	2993	5366.2	7150	538.3
8	15,302	3323	5484.9	6967	552.7	15,323	2993	5466.6	7337	585.0
9	15,208	3470	5423.3	7171	531.8	15,246	3220	5425.0	7187	560.8
10	15,115	3470	5465.5	7171	565.6	15,151	3220	5468.2	7116	576.5
11	12,336	3470	5559.8	7171	561.4	12,362	3220	5570.8	7110	546.4
12	9,484	3470	5541.3	7171	535.0	9,513	3220	5562.1	6935	518.9

*Scale scores cannot be compared across grade bands.

*Table 2.5 Percentage of Students in Each Performance Level by Grade—Listening and Reading**

Grade	Listening						Reading					
	N	1	2	3	4	5	N	1	2	3	4	5
K	36,730	13.6	13.7	49.0	10.8	12.9	36,603	14.3	15.6	37.4	14.5	18.1
1	37,021	7.0	6.1	30.6	26.7	29.6	36,912	30.6	16.9	27.0	11.5	13.9
2	31,459	4.1	4.2	26.4	33.8	31.4	31,354	24.1	18.5	30.7	13.6	13.1
3	25,679	3.5	4.0	24.9	39.6	28.0	25,563	25.9	19.5	33.8	12.2	8.6
4	22,359	5.7	5.8	21.5	43.4	23.6	22,218	19.5	16.5	34.7	18.0	11.3
5	18,521	6.6	6.8	13.7	47.8	25.0	18,416	19.0	17.6	40.2	14.9	8.3
6	16,948	6.3	6.5	22.1	41.2	24.0	16,825	18.6	18.4	41.4	13.8	7.8
7	15,481	9.7	10.8	38.0	25.2	16.2	15,409	25.2	24.4	38.2	7.9	4.3
8	15,248	10.3	10.2	34.7	26.9	18.0	15,198	23.6	23.2	43.8	5.9	3.4
9	15,137	14.6	10.6	37.3	22.6	15.0	15,090	26.7	21.6	42.8	5.8	3.2
10	15,029	14.2	11.2	33.7	21.4	19.5	15,008	26.4	19.9	40.8	7.7	5.2
11	12,281	9.2	10.2	33.7	22.2	24.7	12,245	21.3	19.5	42.7	9.4	7.1
12	9,441	7.8	10.6	36.3	23.9	21.4	9,402	20.0	21.6	43.8	8.8	5.8

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

*Table 2.6 Percentage of Students in Each Performance Level by Grade—Speaking and Writing**

Grade	Speaking						Writing					
	N	1	2	3	4	5	N	1	2	3	4	5
K	36,519	17.1	10.4	27.6	15.4	29.5	36,564	42.2	26.0	21.1	3.6	7.1
1	36,857	26.3	28.1	9.6	14.6	21.4	36,880	40.7	20.1	23.5	6.4	9.2
2	31,307	20.2	19.2	17.1	21.3	22.2	31,322	26.4	17.5	29.5	13.7	12.9
3	25,543	15.0	13.1	20.2	28.5	23.2	25,544	26.2	18.4	33.1	13.5	8.9
4	22,228	13.6	12.1	21.4	28.5	24.4	22,231	16.7	13.2	50.6	12.4	7.1
5	18,425	15.6	14.4	30.2	22.6	17.2	18,412	12.4	10.3	62.1	9.6	5.5
6	16,804	13.6	11.9	33.7	23.3	17.5	16,791	12.4	10.6	57.1	12.2	7.7
7	15,340	14.8	14.8	38.1	18.1	14.2	15,377	19.8	19.4	47.3	8.2	5.3
8	15,116	14.7	12.9	38.5	17.8	16.1	15,158	20.2	19.2	48.0	7.3	5.3
9	14,949	16.2	12.8	36.8	17.4	16.9	15,009	22.2	19.9	48.3	6.3	3.2
10	14,868	14.7	13.5	33.1	17.6	21.1	14,963	22.7	18.6	45.9	7.7	5.2
11	12,132	10.1	12.5	32.2	19.0	26.3	12,169	17.0	19.2	46.9	9.7	7.1
12	9,302	9.4	11.6	34.5	19.3	25.2	9,350	15.3	21.4	49.2	8.4	5.8

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

Table 2.7 Percentage of Students in Each Overall Proficiency Category by Grade

Grade	N	Emerging	Progressing	Proficient
K	36,767	17.6	74.4	8.0
1	37,064	12.2	74.6	13.3
2	31,502	8.0	72.2	19.7
3	25,712	7.2	76.4	16.4
4	22,388	10.7	72.9	16.5
5	18,559	11.9	76.3	11.7
6	16,999	11.7	76.3	12.0
7	15,544	17.4	75.7	6.9
8	15,323	17.6	76.2	6.2
9	15,246	21.5	73.5	5.0
10	15,151	21.6	70.6	7.8
11	12,362	16.0	73.4	10.7
12	9,513	14.0	77.3	8.6

2.3 2020–2021 TESTING TIME FOR ONLINE SUMMATIVE TESTS

Table S7.1 in the Appendix shows testing time for each grade or grade band. In general, tests for upper grades show longer testing times than the tests for lower grades. Testing time was computed by taking the sum of the total time spent on all pages (cumulative across all visits to each page) in the test. In this analysis, only valid scores from students who took online tests (i.e., students who answered all items and earned a score) were included. Scores from students who had domain exemptions or skipped any item were not included in the analysis.

Chapter 3. Reliability

In this section, test reliability for the summative assessment is provided using

- Cronbach's alpha;
- 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 methods used in the computation of test reliability are described in Part I of Chapter 4. The results for each method are included in Sections 8–12 of the Appendix. The figures and the tables in each section of the Appendix are illustrated below:

- Section 8. Summative Assessment—Cronbach's Alpha
    - Figure S8.1 shows the Cronbach's alpha for each domain test across grades.
- Section 9. Summative Assessment—Marginal Reliability
    - Figure S9.1 shows the ratio of MSEM to the standard deviation of scale scores at the test level.
    - Figure S9.2 presents the marginal reliability for each domain test across grades.
    - Figures S9.3 and S9.4 present the marginal reliability by gender and by ethnicity for each domain test across grades, respectively.
- Section 10. Summative Assessment—CSEM
    - Figures S10.1–S10.13 show the CSEM plots for each domain, overall, and comprehension tests.
- Section 11 Summative Assessment—Classification Accuracy and Classification Consistency
    - Figures S11.1 and S11.2 show the CA and CC for each domain test across grades, respectively.
    - Figure S11.3 shows the CA and CC for each overall proficiency category.
- Section 12. Summative Assessment—Inter-Rater Analysis
    - Tables S12.1–12.6 display the inter-rater analysis result for each handscored item in each grade.

3.1 Internal Consistency

Due to the smaller sample size (see Section 1 of the Appendix), scores earned by students who took braille and paper-pencil tests were excluded from the analysis. Table 3.1 shows the values of

Cronbach's alpha for the pooled sample (across states) based on the items in each domain test, arranged by grade level. Values range from 0.81 to 0.94. Nunnally (1978) suggested 0.70 as a minimally acceptable value for the alpha coefficient. All domain tests have alpha coefficients that exceed 0.70, indicating that reliability for all domain assessments is acceptable based on this criterion. The results of Cronbach's alpha for all domains and grades are plotted in Figure S8.1 in the Appendix.

Table 3.1 Cronbach's Alpha by Domain and Grade

Grade	Listening	Reading	Speaking	Writing	Overall
K	.85	.81	.90	.91	.94
1	.83	.84	.82	.94	.94
2	.81	.81	.81	.86	.93
3	.82	.83	.82	.86	.93
4	.83	.84	.84	.88	.94
5	.84	.85	.85	.88	.94
6	.85	.82	.87	.89	.93
7	.86	.84	.88	.89	.94
8	.87	.86	.88	.89	.95
9	.84	.88	.91	.88	.95
10	.85	.89	.91	.88	.95
11	.84	.89	.89	.86	.95
12	.83	.88	.88	.84	.94

3.2 Marginal Standard Error of Measurement

Another way to examine score reliability is with the marginal standard error of measurement (MSEM) (or $\bar{\sigma}^2_{error}$). The ratio of MSEM and the standard deviation of scale scores (i.e., signal-noise ratio) can also indicate the measurement errors. In other words, it shows the ratio of the error and total score ($\frac{\bar{\sigma}_{error}}{\sigma_{total}}$). See details in 4.2 (p.13) in "ELPA21_2020-21_Technical_Report_Part I_Assessment Overview". The plot of this ratio is displayed in Figure S9.1 in the Appendix.

3.3 Marginal Reliability and Conditional Standard Error of Measurement

The marginal reliability for the pooled analysis is presented in Table 3.2 and is plotted in Figure S9.2 in the Appendix. The results show that the listening tests for grades 1–5 have the lowest reliabilities, followed by the speaking tests. The reliability for the speaking domain in the

middle and high school tests are lower than the other domains. All the reliability indexes are above .8, except for the listening test in grades 1–3 and the comprehension test in grades K–3. In addition, Section 9 of the Appendix presents marginal reliability by subgroups, and Section 10 of the Appendix displays CSEM plots by grades.

*Table 3.2 Marginal Reliability by Score and Domain**

Grade	N	Listening	Reading	Speaking	Writing	Comprehension	Overall
K	36,432	.86	.84	.90	.89	.79	.83
1	36,763	.76	.91	.81	.91	.71	.84
2	31,214	.79	.91	.83	.92	.75	.86
3	25,453	.77	.90	.83	.91	.75	.86
4	22,117	.85	.90	.85	.91	.81	.88
5	18,323	.85	.90	.85	.90	.82	.88
6	16,683	.87	.89	.85	.90	.82	.87
7	15,235	.88	.89	.87	.90	.84	.88
8	15,014	.89	.90	.87	.91	.85	.89
9	14,817	.90	.92	.90	.91	.88	.89
10	14,763	.91	.93	.89	.91	.89	.90
11	12,048	.89	.92	.88	.90	.88	.88
12	9,219	.88	.92	.87	.88	.87	.87

*Scores for domain tests marked as Exemption or Not Attempted are excluded.

3.4 Classification Accuracy and Consistency

Table 3.3 shows the overall CA and CC in each domain. The detail description of CA and CC can be found on p.12 in Section 4.4 of ELPA21_2020-21_Technical_Report_Part I. Scores from paper-pencil and braille tests were excluded. CC rates can be lower than CA because CC is based on two tests with measurement errors, while CA is based on one test with a measurement error and the true score. The CA and CC rates for each performance level are higher for the levels with a smaller standard error.

The pooled analysis results for each cut score (cut scores can be found in Table 3.1 in ELPA21 2021-21 Technical Report Part I) are presented in Table 3.4 and Table 3.5, as well as Figures S11.1 and S11.2 in the Appendix. For each cut score, all CAs are above 0.83 and all CCs are above 0.77. In listening and speaking, both indexes for cut score 3 and/or cut score 4 are relatively low in elementary and middle school grades, which indicates a lack of difficult items.

The CA and CC results for overall proficiency categories are summarized in Table 3.6 and Figure S11.3 in the Appendix. All CAs and CCs are above 0.86 for overall and above 0.90 for each category. The CA indexes for between Emerging and Progressing are higher than those for between Progressing and Proficient in all grades except for kindergarten and grades 7–9. The CC

indexes for between Emerging and Progressing are higher than those for between Progressing and Proficient in all grades except for kindergarten and grades 8–10.

*Table 3.3 Overall Classification Accuracy and Consistency for Domain Performance Levels, by Grade and Domain**

Grade	Accuracy				Consistency			
	Listening	Reading	Speaking	Writing	Listening	Reading	Speaking	Writing
K	.71	.66	.69	.77	.63	.56	.60	.69
1	.62	.73	.57	.75	.53	.64	.49	.68
2	.67	.71	.57	.73	.56	.62	.48	.64
3	.66	.71	.56	.70	.55	.62	.47	.61
4	.72	.71	.60	.76	.62	.62	.50	.67
5	.72	.73	.59	.79	.62	.64	.49	.72
6	.74	.72	.61	.76	.64	.62	.51	.68
7	.70	.75	.62	.73	.61	.65	.52	.64
8	.71	.77	.64	.75	.62	.69	.54	.66
9	.72	.80	.67	.75	.62	.73	.58	.66
10	.72	.79	.67	.75	.62	.72	.58	.66
11	.72	.78	.67	.72	.62	.70	.57	.63
12	.71	.77	.66	.72	.61	.69	.57	.63

*Scores for domain tests marked as Exemption or Not Attempted are excluded.

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

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
K	.96	.92	.89	.91	.95	.91	.87	.89	.96	.93	.88	.88	.90	.95	.95	.95
1	.97	.95	.84	.83	.93	.92	.93	.94	.88	.84	.84	.86	.94	.92	.93	.94
2	.98	.96	.87	.84	.92	.91	.93	.94	.91	.86	.85	.87	.94	.92	.92	.94
3	.99	.97	.86	.83	.94	.92	.91	.94	.94	.88	.83	.85	.94	.91	.90	.93
4	.98	.96	.90	.88	.94	.92	.91	.94	.95	.90	.85	.86	.96	.93	.90	.95
5	.98	.96	.91	.87	.95	.92	.91	.94	.95	.89	.84	.87	.98	.95	.91	.95
6	.98	.96	.91	.88	.93	.91	.92	.95	.96	.90	.84	.88	.97	.94	.90	.94
7	.98	.95	.87	.90	.94	.91	.93	.96	.96	.89	.85	.89	.95	.89	.92	.95
8	.98	.96	.88	.89	.94	.91	.94	.96	.96	.90	.85	.88	.95	.90	.92	.96
9	.96	.95	.89	.91	.95	.92	.95	.97	.97	.93	.86	.89	.95	.90	.93	.96
10	.96	.95	.90	.91	.95	.93	.94	.96	.97	.93	.87	.88	.95	.91	.92	.95
11	.97	.95	.90	.90	.95	.93	.93	.95	.97	.93	.86	.87	.95	.91	.91	.94
12	.97	.95	.89	.90	.95	.93	.93	.96	.98	.93	.85	.87	.95	.90	.91	.95

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

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

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
K	.94	.89	.85	.88	.93	.87	.83	.85	.95	.91	.83	.83	.86	.92	.93	.94
1	.96	.92	.78	.77	.90	.89	.90	.92	.83	.77	.78	.81	.92	.88	.90	.92
2	.97	.95	.81	.79	.89	.88	.90	.92	.87	.80	.79	.82	.92	.89	.89	.91
3	.98	.96	.80	.77	.91	.88	.87	.91	.91	.83	.77	.80	.91	.87	.86	.91
4	.97	.95	.85	.83	.91	.88	.88	.92	.93	.86	.79	.80	.95	.90	.87	.92
5	.97	.94	.87	.82	.93	.89	.87	.92	.92	.84	.78	.83	.96	.92	.87	.93
6	.97	.95	.87	.84	.90	.87	.88	.93	.94	.86	.78	.83	.96	.91	.86	.92
7	.96	.93	.83	.86	.91	.87	.91	.95	.94	.84	.79	.85	.93	.85	.88	.93
8	.97	.94	.83	.85	.92	.88	.91	.95	.95	.86	.80	.84	.93	.86	.89	.94
9	.95	.92	.85	.88	.93	.89	.93	.96	.96	.89	.81	.85	.93	.86	.90	.95
10	.94	.93	.86	.87	.93	.90	.91	.95	.96	.90	.81	.84	.93	.87	.89	.93
11	.95	.93	.86	.86	.93	.90	.90	.93	.96	.90	.80	.82	.93	.87	.87	.92
12	.96	.92	.84	.86	.93	.89	.91	.94	.97	.90	.80	.82	.92	.86	.88	.93

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

Table 3.6 Summative Classification Accuracy and Classification Consistency for Overall Proficiency Categories by Grade

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

3.5 INTER-RATER ANALYSIS

For the 2020–2021 summative assessment, consistency of handscoring was evaluated for a total of 72 items (11 items in kindergarten, 9 items in grade 1, and 13 items in each of the other four grade bands). Handscored items on paper-pencil and braille forms were not included in the results due to the small sample size.

Table 3.7 contains the summary of Kappa coefficients for each summative assessment in the pooled analysis. The description about Kappa coefficients can be found in Chapter 4 (p.10) of the ELPA21_2020-21_Technical_Report_Part I. The table shows that 58.2–94.1% of handscores are consistent between the first rater and the second rater, and 0.3%–5.8% of handscores are off by two or more points across the six tests. The weighted Kappa coefficients ranged from 0.612 to 0.910. In 2019-2020, the weighted Kappa coefficients ranged from 0.656 to 0.909. The inter-rater consistencies are also assessed by item and are summarized in Section 12 of the Appendix. In general, the inter-rater consistency values (weighted kappa; rater agreement) are reasonable and are in the similar range as those in the previous years. Some items in the Speaking domain (e.g., see grade band 4-5 in Table S12.4) have relatively lower exact agreement (e.g., 58.8, 63.0), this may be due to the higher score points (e.g., score point=5).

Table 3.5 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
K	11	.0.759	.0.863	69.8	93.8	96.8	99.5	0.5	3.2
1	9	0.612	0.881	58.2	94.1	97.1	99.3	0.7	2.9
2–-3	13	0.689	0.896	63.0	93.5	97.8	99.7	0.3	2.2
4–-5	13	0.684	0.878	58.8	86.1	94.2	99.4	0.6	5.8
6–-8	13	0.730	0.908	64.3	91.8	98.1	99.4	0.6	1.9
9–-12	13	0.729	0.910	65.3	91.3	97.8	99.5	0.5	2.2

Chapter 4. Validity

In this chapter, validity for the summative assessment is measured by examining the internal structure of the items and the comparison of student abilities versus the difficulty of the items. The domain test internal structure is measured using domain dimensionality. The appropriateness of the assessment for the student population is assessed by comparing student abilities with test difficulties.

The analysis results for each state and the pooled analysis are summarized in the following sections of the Appendix:

- Section 13. Summative Assessment—Dimensionality
    - Figures S13.1–S13.6 present the scree plots for each domain test. If a test involves multiple grades, the results are broken down by grade.
- Section 14. Summative Assessment—Ability versus Difficulty
    - Figures S14.1–S14.6 present the comparison of student ability versus test difficulty on the logit scale for each domain test for each grade band of students, respectively.

4.1 Dimensionality Analysis

The graded response model (Samejima, 1969) used for operational scoring of ELPA21 assumes that the domain tests are essentially unidimensional. For ELPA21, a principal component analysis with an orthogonal rotation (Cook, Kallen, & Amtmann, 2009; Jolliffe, 2002) was used to investigate the dimensionality for each domain test and the overall test.

The dimensionality analysis results are presented in the scree plots in Section 13 of the Appendix. The graphs show that the magnitude of the first eigenvalue is always noticeably larger than the magnitude of the second factor in all tests, which indicates that each domain test has one dominant factor, consistent with the assumption of essential unidimensionality within domains and the overall test.

4.2 Student Abilities versus Test Difficulties

When student abilities are well matched to test difficulties, the measurement errors are reduced. Therefore, it is desired that the test difficulty matches student ability. To examine this aspect of the test, item difficulties were plotted versus student abilities for each domain. Specifically, the density plots of students' abilities (θ) and item location parameters were plotted and compared in each domain.

The results, which are included in Section 14 in the Appendix, show that student abilities are generally higher than the test difficulties in all domain tests, except for the reading tests in grade 1, grades 2–3, grades 4–5, grades 6–8 and grades 9–12 and the writing test in kindergarten, where the test difficulties match student abilities well.

Chapter 5. Reporting

A detailed introduction to the Centralized Reporting System can be found in Chapter 6 of Part I of the technical report. Reporting mockups for the summative assessment in each state appear in Section 15 of the Appendix. It is noted that the mockup for score reports is not included in the Appendix for the pooled analysis.

References

Cook, K. F., Kallen, M., & Amtmann, D. (2009). Having a fit: Impact of number of items and non-normality on tests of IRT's unidimensionality assumption. *Quality of Life Research, 18*(4), 447–460.

Jolliffe, I. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (Series 17) *Psychometric Monographs*. Richmond, VA: Psychometric Society.