TECHNICAL REPORT

PART II – SUMMATIVE ASSESSMENT

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

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

Grades K–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

Chapter 1. Test Administration

The summative tests 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 assessments involves four domain tests. Students can be exempted from as many as three domain tests.

1.1 Testing Window

The 2019–2020 summative testing windows for the seven states considered in this report are shown in Table 1.1. However, in March 2020, the states announced school closure until the end of school year due to the coronavirus (COVID-19) pandemic. Therefore, while test windows remained open, some students with English learner status did not complete the ELPA summative assessments.

Table 1.1 2019–2020 ELPA21 Summative Testing Windows by State

State	ELPA21 Summative
Arkansas	1/26/20–3/6/20
Iowa	2/3/20–4/13/20
Louisiana	2/2/20–3/20/20
Nebraska	2/2/20–3/27/20
Ohio	2/2/20–3/27/20
Washington	2/3/20–4/21/20
West Virginia	2/10/20–3/18/20

1.2 Test Design

The 2019–2020 summative assessment include one online form, one paper-pencil form, and one braille form for each of the 2019–2020 summative tests. Each form has separate tests for the four language domains. In addition to operational items, the online form consists of field-test items, which are embedded within each domain form.

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 in each test. Writing and speaking had fewer but comparable numbers of items in each test. Table 1.5 lists the number of field-test items in the pool for each domain and grade and grade band. Each student took either one discrete item or all the items in one entire passage, which consists of one to five items, in each domain. Each field-test item/passage was randomly administered to students.

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	28	28	24	24	24	26	27	30	33	36	24	27
Reading	23	23	30	30	29	34	25	27	26	31	34	35
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	80	96	83	99	76	109	73	117	74	122	73	117

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	19	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	69	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	27	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	91	63	104	61	102	66	105

Table 1.5 Number of Field Test Items by Domain and Grade-band—Online Summative

Domain	K	1	2–3	4–5	6–8	9–12
Listening	4	9	5	13	22	23
Reading	0	3	9	13	18	0
Speaking	0	8	6	0	8	5
Writing	0	9	10	7	15	10

1.3 TEST ADMINISTRATION MANUAL

1.3.1 Directions for Administration

For 2019–2020, the *Test Administration Manual* (TAM) was developed to guide test administrators (TAs) for the summative test. The TAM usually 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 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 TA session;
- monitoring students' progress throughout their tests; and
- stopping 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;
- starting 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 Tests

The instructions for summative tests include a brief direction for each domain test. They also provide detailed instructions for the following procedures:

- 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
- stopping the session and logging out.

1.4 Business Scoring Rules for the Summative Assessment

Business rules and instructions applied to the 2019–2020 ELPA21 summative tests include the following:

1. A domain test was considered "attempted" if a student was presented with the first operational item; it was not necessary for a 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 was 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 left out of 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 (CAI's) test delivery system (TDS), the test was considered attempted, even if an exemption was intended. Items in the domain were considered in the computation of scores.
5. If no domains were attempted (i.e., every domain is 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 1.6 for reporting of scenarios in which neither listening nor reading were attempted (i.e., each was either exempted or non-attempted).

Table 1.6 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. 2019–2020 Summary

The 2019–2020 student participation and performance statistics for each state and the pooled analysis in the summative assessment are presented in Sections 1–5 of the appendix. The figures and tables included in Sections 1–5 are listed below:

- Section 1. Summative Assessment—Student Participation
    - Table S1.1 displays the number and percentage of students in each test mode of 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, such as migrant, special education (SPED), Title I, or Section 504 Plan status. Subgroups vary across the states. The pooled analysis includes the summary by gender and ethnicity.
- Section 2. Summative Assessment—Scale Score Summary
    - Tables S2.1–S2.13 present the number of students, minimum, maximum, average, and standard deviation of domain, overall, and comprehension scores across the state and by subgroups in each grade. The pooled analysis includes the summary by gender and ethnicity.
    - Table S2.14 summarizes the number and percentage of students who were marked "non-attempt" or "exempt" in each domain and grade.
- Section 3. Summative Assessment—Percentage of Students by Domain Performance Level
    - Figure S3.1 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 S3.1–S3.13 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 4. Summative Assessment—Percentage of Students by Overall Proficiency Level
    - Figure S4.1 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 S4.1–S4.13 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 5. Summative Assessment—Testing Time
    - Table S5.1 summarizes testing time per grade or 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 summarizes student participation in each state. There were 310,930 students in total who participated in the 2019–2020 summative tests. The state of Washington had the most tested students, followed by the state of Ohio. School building closures had variable impacts on student participation. In some states, the testing windows were completed prior to closures, while in other states, some students were unable to be tested.

Table 2.1 Student Participation in Each State by Grade

Grade	Arkansas	Iowa	Louisiana	Nebraska	Ohio	Washington	West Virginia	Total
K	≥4,640	≥4,450	≥3,400	≥3,880	≥10,120	≥15,290	≥200	≥42,010
1	≥4,360	≥3,800	≥3,760	≥3,540	≥8,800	≥15,780	≥250	≥40,320
2	≥3,820	≥3,110	≥3,270	≥2,870	≥7,320	≥14,770	≥180	≥35,370
3	≥3,350	≥2,430	≥2,600	≥2,020	≥5,850	≥11,960	≥160	≥28,400
4	≥2,890	≥2,230	≥2,440	≥1,800	≥4,410	≥10,270	≥130	≥24,200
5	≥2,790	≥2,100	≥2,090	≥1,500	≥3,990	≥9,190	≥130	≥21,820
6	≥2,460	≥2,020	≥1,910	≥1,200	≥3,360	≥7,830	≥130	≥18,950
7	≥2,510	≥1,800	≥1,790	≥960	≥3,250	≥7,070	≥110	≥17,520
8	≥2,360	≥2,020	≥1,720	≥1,000	≥3,380	≥7,060	≥100	≥17,670
9	≥2,520	≥2,380	≥2,480	≥1,300	≥4,290	≥7,160	≥130	≥20,300
10	≥2,690	≥2,050	≥1,550	≥1,150	≥3,670	≥6,610	≥140	≥17,890
11	≥2,550	≥1,690	≥1,090	≥910	≥2,990	≥5,100	≥120	≥14,480
12	≥2,120	≥1,420	≥810	≥920	≥2,240	≥4,310	≥100	≥11,940
Total	≥39,120	≥31,550	≥28,980	≥23,100	≥63,720	≥122,460	≥1,960	≥310,930

Section S1.1 of the Appendix presents student participation in each mode. In the seven states combined, the most frequent mode of administration was online (99.86%), followed by paper (0.14%) and braille (<0.01%).

Section S1.2 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 8 and 9. There were more male students (50.8%–56.5%) than female students (43.3%–48.7%). In each test, most students were Hispanic or Latino (55%–65.4%), followed by Asian students (9.8%–17.7%), and White students (7.7%–11.4%).

2.2 2019–2020 STUDENT SCALE SCORE AND PERFORMANCE SUMMARY

Student performance in the 2019–2020 administration across the seven states is summarized by subgroup and for the students who completed the tests. Tables 2.2– 2.4 show the number of students, minimum, mean, maximum and standard deviation of domain, 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 (2–3, 4–5, 6–8, and 9–12). A disaggregated summary based on subgroups is also available in Section 2 of the Appendix.

Table 2.5 and Table 2.6 display the percentage of students in each performance level and for each grade and domain. In addition, Table 2.7 shows the percentage of student in each overall proficiency category in each grade. Sections 3 and 4 of the Appendix further summarize the percentage of students in each domain test by subgroups by performance level and overall proficiency category, respectively.

For both reading and writing in the pooled analysis, the plot presented in Figure S3.1 of the Appendix shows that most students are in Performance Level 3 except for writing in kindergarten. Middle school and high school students have higher percentage in levels 1 and 2 than in levels 4 and 5. In the listening and speaking domains, the greatest number of students is in level in grade 7 and higher. More students are in levels 4 and 5 than in levels 1 and 2 in grades 2 to grade 8.

The percentage of students in each proficiency category is summarized in Figure S4.1 in the Appendix. The figure shows that most students (65.8%–77.3%) are in the Progressing category in all grades. The percentage of students who are Proficient increases from kindergarten to grade 2, then decreases until grade 9, and slightly increases after grade 9. The percentage of students in the Emerging category is relatively stable until grade 6, increases until grade 9, and then consistently drops afterwards.

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	≥41,720	237	560.0	775	76.4	≥41,210	247	559.7	770	74.8
1	≥40,120	239	558.1	712	71.4	≥39,770	241	548.9	744	78.1
2	≥35,150	229	528.6	742	66.1	≥34,910	228	521.9	766	68.2
3	≥28,250	229	550.8	742	70.9	≥28,030	228	550.9	766	72.2
4	≥24,050	213	528.1	734	72.5	≥23,800	228	521.2	753	67.5
5	≥21,640	213	544.1	762	78.8	≥21,500	228	542.3	772	73.8
6	≥18,640	232	526.4	756	68.6	≥18,550	247	522.0	763	63.0
7	≥17,250	232	533.6	743	74.8	≥17,190	247	531.9	761	68.3
8	≥17,280	232	550.9	758	82.0	≥17,210	247	552.6	780	76.5
9	≥20,010	253	530.1	777	78.0	≥19,950	258	527.7	790	71.5
10	≥17,620	253	548.2	762	75.9	≥17,540	258	546.5	767	72.6
11	≥14,290	253	559.6	794	73.4	≥14,210	258	558.2	797	72.9
12	≥11,790	253	557.1	809	71.7	≥11,720	258	555.6	817	72.4

*Domains tests with 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	≥40,500	291	575.2	756	84.7	≥40,700	309	538.5	723	76.3
1	≥39,110	265	573.7	736	75.4	≥39,300	245	543.9	733	83.7
2	≥34,160	252	546.4	747	71.6	≥34,490	235	520.0	765	70.9
3	≥27,600	252	569.2	747	75.8	≥27,790	235	550.1	765	73.0
4	≥23,390	237	544.6	745	78.3	≥23,570	221	518.8	743	73.5
5	≥21,140	237	557.2	761	82.8	≥21,200	221	539.2	776	79.5
6	≥18,110	268	549.1	752	72.9	≥18,200	243	518.5	768	71.2
7	≥16,720	268	552.5	753	78.2	≥16,920	243	527.6	750	77.9
8	≥16,830	268	564.4	766	84.0	≥16,970	243	545.9	776	84.8
9	≥19,640	297	544.7	728	80.0	≥19,790	263	520.7	782	79.7
10	≥17,310	297	562.5	745	75.3	≥17,420	263	540.2	770	74.8
11	≥13,950	297	572.4	762	72.6	≥14,120	263	551.9	795	70.9
12	≥11,480	297	570.3	770	74.1	≥11,600	263	549.6	808	69.0

*Domains tests with 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	≥41,870	3377	5563.0	6865	539.6	≥42,010	3185	5562.3	7178	572.2
1	≥40,230	3428	5524.3	6640	505.2	≥40,320	3021	5554.9	6998	586.5
2	≥35,260	3300	5315.9	6729	483.3	≥35,370	2968	5327.9	7156	528.6
3	≥28,340	3300	5497.6	6805	518.8	≥28,400	2968	5539.2	7156	562.4
4	≥24,130	3298	5317.4	6878	510.2	≥24,200	2892	5325.6	7049	560.3
5	≥21,760	3298	5452.2	6878	560.0	≥21,820	2892	5465.5	7221	608.9
6	≥18,800	3361	5312.2	6938	479.2	≥18,950	3052	5334.0	7094	523.1
7	≥17,390	3361	5374.9	6938	519.5	≥17,520	3052	5393.6	7046	570.5
8	≥17,460	3361	5517.6	6938	580.7	≥17,670	3052	5531.4	7242	626.0
9	≥20,160	3505	5360.4	7177	550.1	≥20,300	3235	5350.8	7084	593.2
10	≥17,750	3505	5492.9	7177	559.6	≥17,890	3235	5497.9	7138	571.3
11	≥14,380	3505	5578.7	7177	560.4	≥14,480	3235	5587.1	7382	551.1
12	≥11,870	3505	5560.5	7177	553.9	≥11,940	3235	5568.4	7483	541.3

*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	≥41,720	11.8	13.0	49.9	11.2	14.1	≥41,210	12.7	14.6	38.3	14.8	19.5
1	≥40,120	5.7	6.0	29.0	26.2	33.1	≥39,770	19.8	18.2	30.3	13.6	18.1
2	≥35,150	5.7	3.9	25.4	33.6	31.5	≥34,910	19.3	15.3	31.1	18.4	15.8
3	≥28,250	5.6	3.9	24.3	40.1	26.1	≥28,030	23.3	16.4	38.1	14.1	8.2
4	≥24,050	6.1	5.2	16.3	37.1	35.4	≥23,800	17.4	14.4	32.5	20.5	15.2
5	≥21,640	8.2	6.5	9.8	38.8	36.7	≥21,500	18.2	14.3	37.2	17.8	12.4
6	≥18,640	7.5	5.4	17.3	37.8	32.0	≥18,550	17.6	18.2	38.3	15.4	10.5
7	≥17,250	12.4	8.5	33.0	25.6	20.5	≥17,190	26.8	24.2	35.1	8.8	5.1
8	≥17,280	12.2	7.8	28.3	27.2	24.5	≥17,210	24.8	20.5	41.2	8.2	5.2
9	≥20,010	19.9	11.1	33.3	20.5	15.1	≥19,950	31.9	23.5	36.6	5.1	2.9
10	≥17,620	13.4	10.6	32.1	21.6	22.3	≥17,540	24.2	21.6	40.6	8.2	5.4
11	≥14,290	8.8	10.3	32.3	22.1	26.5	≥14,210	18.8	22.2	41.1	9.8	8.1
12	≥11,790	7.8	11.4	34.8	21.8	24.2	≥11,720	19.6	22.7	40.7	9.5	7.5

*Domains tests with 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	≥40,500	14.8	11.3	28.1	15.4	30.4	≥40,700	35.3	27.8	26.7	4.3	5.9
1	≥39,110	21.1	24.2	10.3	16.4	28.0	≥39,300	27.3	21.5	30.1	8.6	12.5
2	≥34,160	16.5	16.3	15.9	21.9	29.3	≥34,490	17.9	15.3	31.9	18.9	16.0
3	≥27,600	13.9	10.7	16.9	28.5	30.0	≥27,790	21.8	16.3	37.1	16.0	8.8
4	≥23,390	13.8	9.2	16.0	26.7	34.3	≥23,570	14.7	11.4	47.3	15.2	11.4
5	≥21,140	15.6	9.5	22.6	23.5	28.7	≥21,200	12.5	9.2	55.1	12.9	10.3
6	≥18,110	12.8	9.2	27.4	23.6	26.8	≥18,200	11.6	9.6	52.4	14.8	11.6
7	≥16,720	15.5	12.1	31.7	19.2	21.6	≥16,920	20.7	17.4	44.9	9.8	7.2
8	≥16,830	15.1	9.8	29.3	18.4	27.3	≥16,970	20.4	15.6	45.3	10.3	8.5
9	≥19,640	20.5	15.9	32.2	15.4	16.0	≥19,790	28.3	19.3	43.5	6.1	2.8
10	≥17,310	14.1	14.3	31.9	17.4	22.4	≥17,420	21.3	18.1	46.1	8.9	5.6
11	≥13,950	10.7	12.5	31.5	18.2	27.0	≥14,120	16.1	19.1	46.3	10.3	8.2
12	≥11,480	10.9	12.7	32.6	18.1	25.8	≥11,600	15.5	20.7	46.9	9.8	7.1

*Domains tests with 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	≥42,010	15.7	77.3	7.0
1	≥40,320	9.7	73.0	17.3
2	≥35,370	9.1	65.8	25.0
3	≥28,400	9.3	73.2	17.5
4	≥24,200	10.8	67.0	22.2
5	≥21,820	13.4	68.1	18.4
6	≥18,950	11.3	72.4	16.2
7	≥17,520	17.8	73.2	9.0
8	≥17,670	17.5	73.1	9.4
9	≥20,300	27.8	67.6	4.6
10	≥17,890	20.8	70.6	8.6
11	≥14,480	15.5	72.8	11.7
12	≥11,940	14.8	74.4	10.7

2.3 2019–2020 TESTING TIME FOR ONLINE SUMMATIVE TESTS

Table S5.1 of the Appendix shows testing time per each grade or grade band. In general, tests for upper grades show longer testing time 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 students who took online tests and had valid scores (i.e., students answered the item and earned a score) on all items were included. Specifically, 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 tests is provided using

- Cronbach's alpha;
- marginal SEM;
- marginal reliability;
- conditional SEM (CSEM);
- classification accuracy (CA) and classification consistency (CC); and
- inter-rater analysis.

The methods used in the computation are described in Part I Chapter 4. The results for each are included in Sections 6–10 of the Appendix. The figures and the tables in each section of the Appendix are illustrated below:

- Section 6. Summative Assessment—Cronbach's Alpha
    - Figure S6.1 shows the Cronbach's alpha for each domain tests across grades.
- Section 7. Summative Assessment—Marginal Reliability
    - Figure S7.1 shows the ratio of marginal SEM (MSEM) and the standard deviation of scale scores at the test level.
    - Figure S7.2 presents the marginal reliability for each domain test across grades.
    - Figures S7.3 and S7.4 present the marginal reliability by gender and by ethnicity for each domain test across grades, respectively.
    - Figure S7.5 or after (if any) present the marginal reliability by other subgroups for each domain test across grades. Depending on the state, the subgroups may vary.
- Section 8. Summative Assessment—Conditional Standard Error of Measurement (CSEM)
    - Figures S8.1–S8.13 show the CSEM plots for each domain, overall, and comprehension tests.
- Section 9. Summative Assessment—Classification Accuracy and Consistency
    - Figures S9.1 and S9.2 show the classification accuracy (CA) and consistency (CC) for each domain tests of each across grades, respectively.
    - Figure S9.3 shows the classification accuracy and consistency for each overall proficiency category.
- Section 10. Summative Assessment—Inter-Rater Analysis
    - Tables S10.1–10.6 display the inter-rater analysis result for each handscored item in each grade.

3.1 INTERNAL CONSISTENCY

Due to smaller sample size (see Section 1 of the Appendix), 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, by grade level. Values range from .76 to .93. Nunnally (1978) suggested .70 as a minimally acceptable value for the alpha coefficient. All domain tests have alpha coefficients that exceed .70. The results of Cronbach's alpha for all domains and grades are plotted in Figure S6.1 in the Appendix.

Table 3.1 Cronbach's Alpha by Domain and Grade

Grade	Listening	Reading	Speaking	Writing
K	.82	.76	.89	.91
1	.80	.86	.82	.93
2	.82	.82	.82	.85
3	.84	.83	.83	.85
4	.85	.83	.86	.86
5	.86	.85	.88	.87
6	.90	.80	.84	.88
7	.92	.82	.86	.90
8	.93	.86	.87	.90
9	.89	.82	.89	.88
10	.88	.84	.87	.86
11	.87	.85	.86	.83
12	.86	.86	.86	.82

3.2 Marginal Standard Error of Measurement

The plot of this ratio is displayed in Figure S7.1 of the Appendix.

3.3 Marginal Reliability and Conditional Standard Error of Measurement

The marginal reliability for the pool analysis is presented in Table 3.2 and is plotted in Figure S7.2 in the Appendix. The results show that the listening tests at 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 listening tests in grade 1 and the comprehension test in grades 1–3. In addition, Section 7 of the Appendix presents marginal reliability by subgroups, and Section 8 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	≥39,730	.86	.83	.89	.89	.80	.82
1	≥38,520	.76	.90	.82	.89	.69	.83
2	≥33,680	.82	.90	.84	.90	.76	.86
3	≥27,200	.82	.90	.84	.90	.77	.87
4	≥22,980	.87	.90	.87	.90	.82	.88
5	≥20,770	.87	.90	.88	.91	.83	.89
6	≥17,640	.89	.88	.87	.90	.83	.88
7	≥16,340	.90	.89	.88	.91	.85	.89
8	≥16,380	.91	.90	.89	.92	.86	.90
9	≥19,280	.92	.91	.90	.92	.88	.90
10	≥16,980	.91	.91	.89	.90	.88	.89
11	≥13,720	.90	.90	.87	.89	.88	.88
12	≥11,270	.89	.90	.88	.89	.87	.88

*Domains tests with Exemption or Not Attempted are excluded.

3.4 Classification Accuracy and Consistency

Table 3.3 shows overall CA and CC in each domain. The paper-pencil and braille forms 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 are presented in Table 3.4 and Table 3.5, as well as Figure S9.1 and Figure S9.2 in the Appendix. For each cut score, all CAs are above .80 and all

CCs are above .75. In listening and speaking, both indexes for cut score 3 and/or cut score 4 are relatively lower 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 S9.3 in the Appendix. All CAs and CCs are above .80 for overall and above .85 for each category. The CC indexes for Between Emerging and Progressing are higher than those for Between Progressing and Proficient in all grades except for kindergarten and grade 9.

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	.62	.56	.60	.69
1	.63	.72	.58	.70	.54	.62	.50	.62
2	.68	.70	.58	.70	.58	.60	.50	.60
3	.68	.70	.59	.69	.57	.61	.50	.59
4	.73	.71	.64	.73	.64	.61	.55	.65
5	.74	.72	.62	.76	.65	.62	.53	.69
6	.76	.69	.63	.75	.67	.59	.54	.67
7	.73	.73	.64	.73	.64	.64	.55	.64
8	.73	.75	.67	.74	.64	.67	.58	.66
9	.73	.78	.68	.76	.64	.70	.58	.68
10	.72	.75	.66	.73	.63	.67	.57	.64
11	.73	.74	.66	.71	.63	.65	.56	.62
12	.72	.74	.66	.71	.63	.65	.56	.62

*Domains tests with Exemption or Not Attempted are excluded.

Table 3.4 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	.95	.92	.89	.91	.95	.91	.87	.89	.96	.93	.88	.88	.92	.93	.94	.95
1	.97	.95	.85	.83	.93	.92	.92	.93	.91	.84	.83	.85	.95	.90	.90	.92
2	.98	.97	.87	.85	.93	.93	.91	.92	.93	.87	.85	.85	.95	.92	.90	.92
3	.98	.98	.87	.84	.95	.91	.89	.93	.95	.90	.85	.84	.95	.91	.89	.93
4	.98	.97	.92	.87	.94	.93	.91	.93	.96	.93	.87	.85	.97	.94	.89	.92
5	.98	.96	.93	.87	.95	.93	.90	.92	.96	.92	.85	.85	.98	.95	.89	.92
6	.98	.97	.93	.88	.93	.91	.91	.94	.97	.92	.85	.86	.97	.95	.89	.93
7	.98	.96	.89	.89	.92	.91	.93	.96	.96	.91	.86	.88	.96	.90	.91	.95
8	.98	.97	.89	.88	.94	.92	.92	.95	.97	.93	.87	.87	.96	.92	.91	.94
9	.96	.95	.91	.92	.93	.92	.95	.97	.96	.92	.88	.90	.95	.90	.93	.96
10	.96	.95	.91	.90	.93	.92	.93	.96	.96	.92	.87	.88	.95	.91	.91	.95
11	.97	.95	.91	.89	.94	.92	.92	.94	.97	.93	.86	.87	.95	.91	.90	.93
12	.97	.94	.90	.90	.93	.92	.93	.95	.97	.92	.86	.87	.94	.90	.91	.94

*Domains tests with Exemption or Not Attempted are excluded.

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

Table 3.5 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	.93	.89	.85	.88	.93	.87	.83	.85	.95	.90	.83	.83	.88	.90	.92	.93
1	.96	.93	.78	.77	.90	.89	.89	.91	.86	.78	.77	.79	.92	.86	.86	.89
2	.97	.95	.82	.79	.91	.89	.87	.89	.90	.82	.79	.80	.93	.89	.86	.89
3	.98	.96	.81	.79	.93	.88	.85	.90	.93	.86	.79	.78	.93	.87	.84	.90
4	.97	.95	.89	.82	.92	.90	.87	.90	.95	.89	.82	.79	.95	.91	.85	.89
5	.97	.95	.91	.81	.94	.90	.86	.89	.94	.88	.80	.80	.97	.93	.85	.89
6	.97	.96	.90	.83	.89	.87	.88	.92	.95	.89	.80	.81	.96	.92	.84	.90
7	.97	.95	.84	.85	.89	.87	.90	.94	.95	.88	.80	.83	.94	.87	.88	.92
8	.97	.96	.85	.83	.92	.89	.89	.93	.95	.90	.81	.82	.94	.88	.87	.91
9	.94	.93	.87	.89	.90	.88	.93	.96	.94	.88	.83	.86	.92	.86	.91	.95
10	.95	.93	.87	.86	.91	.89	.91	.94	.95	.89	.82	.83	.93	.87	.88	.92
11	.96	.93	.87	.85	.91	.89	.89	.92	.96	.89	.81	.82	.93	.87	.86	.91
12	.95	.92	.86	.86	.91	.88	.90	.93	.96	.89	.81	.82	.92	.86	.87	.92

*Domains tests with Exemption or Not Attempted are excluded.

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

Table 3.6 Summative Classification 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	.89	.94	.95
1	.88	.96	.92	.85	.95	.90
2	.87	.97	.90	.84	.96	.88
3	.89	.98	.91	.86	.97	.89
4	.88	.97	.91	.84	.96	.88
5	.88	.98	.91	.85	.97	.88
6	.90	.98	.92	.87	.97	.90
7	.91	.97	.95	.89	.96	.93
8	.91	.97	.94	.89	.96	.93
9	.92	.95	.97	.89	.94	.96
10	.90	.96	.95	.87	.94	.93
11	.89	.96	.93	.86	.94	.92
12	.89	.95	.94	.86	.94	.92

3.5 INTER-RATER ANALYSIS

For the 2019–2020 summative assessments, consistency of handscoring was evaluated for a total of 71 items (11 handscored items in kindergarten, 9 items in grade 1, and 13 handscored items in each of the other five grade bands). Handscored items on paper 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 test in the pooled analysis. The table shows that 55.3–95% of handscores are consistent between the first rater and the second rater, and 0.2%–5.7% of handscores are off by two or more points across the six tests. The weighted Kappa coefficients ranged from .656 to .909. The inter-rater consistencies are also assessed by item and are summarized in Section 10 of the Appendix.

Table 3.7 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	.779	.869	68.0	93.2	97.6	99.2	0.8	2.4
1	9	.656	.886	55.7	95.0	95.6	99.5	0.5	4.4
2–3	13	.660	.868	62.7	92.6	94.8	99.8	0.2	5.2
4–5	13	.668	.909	55.3	91.1	94.3	99.2	0.8	5.7
6–8	13	.763	.882	61.6	83.6	96.4	99.3	0.7	3.6
9–12	13	.737	.895	56.9	81.3	95.3	99.2	0.8	4.7

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 11. Summative Assessment—Dimensionality
    - Figures S11.1–S11.6 present the scree plots for each domain test. If a test involves multiple grades, the plots are broken down by grade.
- Section 12. Summative Assessment—Ability vs. Difficulty
    - Figures S12.1–S12.6 present the comparison of student ability vs. 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 (PCA) 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 11 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 vs. 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 are plotted and compared in each domain.

The results are included in Section 12 in the Appendix. It shows that the student abilities are generally higher than the test difficulties in all domain tests, except the reading tests in grade bands 6–8 and 9–12, where the test difficulties match student abilities well.

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 summative tests of each state are included in Section 13 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

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.). Springer.

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

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