TECHNICAL REPORT

PART II – SUMMATIVE ASSESSMENT

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

English Language Proficiency Assessment for the 21st Century—

Listening, Reading, Speaking, and Writing

Grades K–12

2021–2022 Administration

Submitted to:
ELPA21

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

January 2023

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. Each form of the summative assessment involves four domain tests. Students can be exempted from as many as three domain tests. The assessments do not have a time limit.

1.1 Testing Windows

The 2021–2022 summative testing windows for the six states discussed in this report are shown in Table 1.1. While testing windows remained open in spring 2022, some students were unable to complete the English Language Proficiency Assessment for the 21st Century (ELPA21) due to the ongoing impacts of the COVID-19 pandemic.

Table 1.1 2021–2022 ELPA21 Summative Testing Windows by State

State	ELPA21 Summative
Arkansas	2/22/2022–4/8/2022
Iowa	2/1/2022–3/25/2022
Louisiana	2/14/2022–3/18/2022
Nebraska	2/7/2022–3/18/2022
Ohio	1/31/2022–3/25/2022
West Virginia	2/8/2022–3/25/2022

1.2 Test Design

The 2021–2022 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 forms had fewer items than the two other forms. Writing and speaking had fewer but comparable numbers of items in each test. Field-test items were also included in the 2021–2022 summative assessments (see details in Table 1.5). Table S7.1 in the Appendix shows testing time for each grade or grade band.

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	21	23
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	116

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

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	Total
Listening				7	12	13	32
Reading	20		21	16	16	35	108
Writing	25	10	7	10			52
Total	45	10	28	33	28	48	192

1.3 TEST ADMINISTRATION MANUAL

1.3.1 Directions for Test Administration

For the 2021–2022 administration, a 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 test
- 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 System, training/practice tests are provided before and during the testing windows. Training/practice tests can be accessed through a nonsecure browser or a secure browser.

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

The *Practice Test Administration* site introduces TAs to

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

The *Practice Tests* site introduces students to

- signing in;
- verifying student information;

- selecting a test;
- waiting for the TA to check the test settings and approve participation;
- preparing to begin the test (adjusting the audio level, checking the microphone for recording speaking responses, and reviewing test instructions);
- taking the test; and
- submitting the test.

1.3.3 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 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
- Ending the test session and logging out

1.4 Business Scoring Rules for the Summative Assessment

Business rules and instructions applicable to the 2021–2022 summative assessment 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 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 (CAI) 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 1.6 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. 2021–2022 Summary

The 2021–2022 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 below:

- 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 grades/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 six 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 2021–2022 STUDENT PARTICIPATION

In the 2021–2022 administration, not all eligible students completed the tests due to the ongoing impacts of the COVID-19 pandemic.

summarizes student participation in each state. There were 199,318 students in total who participated in the 2021–2022 summative assessment. The state of Ohio had the most tested students, followed by the state of Arkansas.

Table 2.1 Student Participation in Each State by Grade

Grade	Arkansas	Arkansas	Iowa	Iowa	Louisiana	Louisiana	Nebraska	Nebraska	Ohio	Ohio	West Virginia	West Virginia	Total	Total	Total
	2020–2021	2021–2022	2020–2021	2021–2022	2020–2021	2021–2022	2020–2021	2021–2022	2020–2021	2021–2022	2020–2021	2021–2022	2020–2021	2021–2022	Two year N Diff
K	≥ 4,190	≥ 4,550	≥ 4,410	≥ 4,610	≥ 3,240	≥ 3,930	≥ 3,670	≥ 3,920	≥ 8,990	≥ 10,230	≥ 200	≥ 230	≥ 24,720	≥ 27,500	≥ 2,780
1	≥ 4,480	≥ 4,250	≥ 3,960	≥ 4,100	≥3,390	≥ 3,880	≥ 3,420	≥ 3,680	≥ 8,940	≥ 9,380	≥ 190	≥ 230	≥ 24,410	≥ 25,550	≥ 1,140
2	≥ 3,870	≥ 4,260	≥ 3,200	≥ 3,640	≥ 3,110	≥ 3,380	≥ 2,660	≥ 3,190	≥ 7,060	≥ 8,530	≥ 200	≥ 190	≥ 20,120	≥ 23,210	≥ 3,090
3	≥ 3,350	≥ 3,480	≥ 2,560	≥ 2,800	≥ 2,470	≥ 2,860	≥ 1,990	≥ 2,320	≥ 5,650	≥ 6,580	≥ 120	≥ 190	≥ 16,160	≥ 18,250	≥ 2,090
4	≥ 3,060	≥ 3,030	≥ 2,270	≥ 2,380	≥ 2,130	≥ 2,460	≥ 1,570	≥ 1,820	≥ 4,750	≥ 5,320	≥ 130	≥ 130	≥ 13,940	≥ 15,160	≥ 1,220
5	≥ 2,690	≥ 2,720	≥ 1,910	≥ 2,100	≥ 1,950	≥ 2,050	≥ 1,220	≥ 1,440	≥ 3,480	≥ 4,650	≥ 90	≥ 130	≥ 11,350	≥ 13,110	≥ 1,760
6	≥ 2,640	≥ 2,610	≥ 1,830	≥ 1,890	≥ 1,700	≥ 2,080	≥ 1,110	≥ 1,180	≥ 3,310	≥ 3,720	≥ 100	≥ 100	≥ 10,720	≥ 11,590	≥ 870
7	≥ 2,410	≥ 2,620	≥ 1,830	≥ 1,780	≥ 1,650	≥ 1,830	≥ 940	≥ 1,140	≥ 2,920	≥ 3,610	≥ 110	≥ 120	≥ 9,870	≥ 11,120	≥ 1,250
8	≥ 2,490	≥ 2,490	≥ 1,820	≥ 1,930	≥ 1,590	≥ 1,830	≥ 850	≥ 1,110	≥ 3,030	≥ 3,490	≥ 100	≥ 130	≥ 9,900	≥ 10,990	≥ 1,080
9	≥ 2,430	≥2,840	≥ 1,940	≥ 2,200	≥ 1,650	≥ 2,610	≥ 980	≥ 1,570	≥ 3,330	≥ 4,780	≥ 90	≥ 140	≥ 10,450	≥ 14,170	≥ 3,720
10	≥ 2,430	≥ 2,510	≥ 2,030	≥ 2,110	≥ 1,730	≥ 1,480	≥ 1,070	≥ 1,130	≥ 3,190	≥ 3,550	≥ 120	≥ 120	≥ 10,600	≥ 10,910	≥ 310
11	≥ 2,330	≥ 2,280	≥ 1,590	≥ 1,990	≥ 1,110	≥ 1,390	≥ 820	≥ 1,010	≥ 2,680	≥ 3,100	≥ 80	≥ 120	≥ 8,630	≥ 9,910	≥ 1,270
12	≥ 1,860	≥ 2,060	≥ 1,240	≥ 1,380	≥ 760	≥ 880	≥ 710	≥ 830	≥ 2,080	≥ 2,510	≥ 90	≥ 90	≥ 6,760	≥ 7,770	≥ 1,010
Total	≥ 38,270	≥ 39,760	≥ 30,650	≥ 32,960	≥ 26,530	≥ 30,710	≥ 21,060	≥ 24,390	≥ 59,490	≥ 69,500	≥ 1,670	≥ 1,980	≥ 177,680	≥ 199,310	≥ 21,630

Table S1.1 in Section 1 of the Appendix presents student participation in each mode. In the six states combined, the most frequent mode of test administration was online (99.82%), followed by paper (0.18%) 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 (50.7%–55.8%) than female students (43.9%–48.7%) tested. In each test, most students were Hispanic or Latino (58.6%–65.9%), followed by Asian students (9.0%–14.9%), and White students (6.5%–9.2%).

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 students were in category 3 or 4 at the domain level in each grade. At the overall raw score level, most students were in the progressing category for all grades.

2.2 2021–2022 STUDENT SCALE SCORE AND PERFORMANCE-LEVEL SUMMARY

Table 2.2–

Table 2.4 summarize student performance in the 2021–2022 administration across the six 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 score; 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 grades 1 and 9 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 2–8 and 11–12, 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**Error! Reference source not found.**. Table 2.7 shows that most students (60.5%–74.0%) are in the Progressing category in all grades. The percentage of students who are Progressing is relatively stable from kindergarten to grade 1, and the largest increase occurs from grade 9 to grade 10. The largest drop occurs from grade 8 to grade 9 and from grade 1 to grade 2 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 9, and thereafter decreases 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	≥ 27,480	237	546.9	775	78.5	≥ 27,360	247	549.4	770	75.2
1	≥ 25,530	239	549.3	712	74.4	≥ 25,400	241	537.6	744	80.7
2	≥ 23,190	229	527.9	742	71.1	≥ 23,070	228	511.8	766	71.7
3	≥ 18,230	229	550.6	742	74.4	≥ 18,090	228	543.5	766	75.4
4	≥ 15,140	213	515.2	735	73.4	≥ 15,010	228	509.6	733	67.9
5	≥ 13,100	213	529.5	734	79.0	≥ 12,950	228	530.5	740	73.6
6	≥ 11,560	232	514.4	728	69.6	≥ 11,470	247	513.8	749	62.4
7	≥ 11,100	232	521.6	753	76.3	≥ 10,980	247	525.5	770	68.7
8	≥ 10,960	232	536.1	784	84.6	≥ 10,870	247	542.0	796	77.2
9	≥ 14,070	253	512.3	723	81.0	≥ 14,020	258	513.8	740	72.6
10	≥ 10,850	253	539.3	781	77.5	≥ 10,800	258	538.9	793	73.1
11	≥ 9,850	253	550.4	777	76.1	≥ 9,810	258	549.4	789	73.6
12	≥ 7,710	253	556.7	774	72.3	≥ 7,660	258	555.5	790	71.3

*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	≥ 27,270	291	547.2	756	87.6	≥ 27,330	309	524.6	727	75.3
1	≥ 25,370	265	559.3	736	80.1	≥ 25,390	245	527.7	733	89.1
2	≥ 23,050	252	540.0	747	74.8	≥ 23,060	235	508.2	765	75.6
3	≥ 18,110	252	563.6	747	77.2	≥ 18,110	235	541.5	765	77.1
4	≥ 15,030	237	537.3	746	81.6	≥ 15,020	221	505.9	747	75.7
5	≥ 13,000	237	547.5	758	85.1	≥ 12,960	221	525.9	747	80.3
6	≥ 11,480	268	536.9	740	73.1	≥ 11,470	243	508.1	724	72.2
7	≥ 11,010	268	542.2	760	79.2	≥ 11,000	243	517.7	748	78.6
8	≥ 10,890	268	551.6	776	85.7	≥ 10,870	243	531.5	792	86.1
9	≥ 13,980	297	522.8	717	82.1	≥ 14,010	263	503.1	713	86.9
10	≥ 10,760	297	549.4	729	74.0	≥ 10,800	263	531.8	775	79.2
11	≥ 9,770	297	559.8	738	70.6	≥ 9,780	263	543.8	772	75.2
12	≥ 7,610	297	565.0	720	68.9	≥ 7,650	263	549.9	782	69.7

*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	≥ 27,490	3377	5497.9	6865	561.3	≥ 27,500	3185	5425.1	7178	590.2
1	≥ 25,550	3428	5460.7	6633	526.0	≥ 25,550	3021	5446.2	6998	620.2
2	≥ 23,200	3300	5283.0	6729	516.6	≥ 23,210	2968	5267.6	7156	565.2
3	≥ 18,250	3300	5476.4	6729	544.9	≥ 18,250	2968	5494.5	7156	589.9
4	≥ 15,150	3298	5226.2	6878	510.4	≥ 15,160	2892	5232.2	7001	576.2
5	≥ 13,110	3298	5355.5	6878	559.4	≥ 13,110	2892	5363.9	6881	617.2
6	≥ 11,580	3361	5239.2	6938	481.7	≥ 11,590	3052	5244.6	6907	532.1
7	≥ 11,120	3361	5308.6	6938	531.7	≥ 11,120	3052	5311.4	7161	586.2
8	≥ 10,980	3361	5428.4	6938	597.1	≥ 10,990	3052	5421.9	7370	647.2
9	≥ 14,130	3505	5249.5	7177	554.6	≥ 14,170	3235	5204.8	6783	625.2
10	≥ 10,890	3505	5435.1	7177	561.8	≥ 10,910	3235	5420.7	7203	584.1
11	≥ 9,890	3505	5514.0	7177	569.8	≥ 9,910	3235	5508.9	7160	565.6
12	≥ 7,750	3505	5560.0	7148	554.1	≥ 7,770	3235	5555.4	7143	536.4

*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	≥ 27,480	16.3	14.2	48.9	9.6	11.1	≥ 27,360	15.8	16.1	38.7	13.5	15.9
1	≥ 25,530	7.8	6.7	31.0	25.2	29.3	≥ 25,400	26.4	19.3	25.9	12.0	16.3
2	≥ 23,190	6.9	4.4	24.8	30.7	33.3	≥ 23,070	26.0	16.1	27.9	15.6	14.4
3	≥ 18,230	6.3	4.1	24.4	37.3	28.0	≥ 18,090	28.0	16.4	35.0	13.0	7.6
4	≥ 15,140	7.4	6.2	20.0	38.9	27.5	≥ 15,010	22.1	16.0	32.5	17.8	11.5
5	≥ 13,100	9.7	8.1	12.9	40.8	28.5	≥ 12,950	21.9	16.0	37.8	15.1	9.1
6	≥ 11,560	9.6	6.7	20.4	37.9	25.4	≥ 11,470	21.1	19.3	38.2	13.1	8.3
7	≥ 11,100	15.0	11.0	33.9	23.4	16.7	≥ 10,980	30.1	24.4	33.4	7.9	4.3
8	≥ 10,960	15.6	9.4	30.7	25.1	19.2	≥ 10,870	28.8	22.2	38.7	6.5	3.9
9	≥ 14,070	28.4	11.0	32.2	17.7	10.8	≥ 14,020	39.1	22.7	32.5	3.8	1.9
10	≥ 10,850	16.9	10.9	32.2	21.0	18.9	≥ 10,800	27.3	21.3	40.1	7.1	4.1
11	≥ 9,850	12.3	12.1	30.8	20.9	23.9	≥ 9,810	23.5	21.1	40.2	8.7	6.5
12	≥ 7,710	8.9	10.7	33.3	22.2	24.9	≥ 7,660	19.1	22.4	42.5	9.1	6.9

*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	≥ 27,270	23.0	15.3	29.8	12.9	19.0	≥ 27,330	42.9	27.1	22.6	3.2	4.2
1	≥ 25,370	28.2	25.1	9.6	14.7	22.3	≥ 25,390	36.0	20.2	25.8	7.1	11.0
2	≥ 23,050	19.0	17.5	15.8	21.2	26.6	≥ 23,060	24.2	15.9	29.7	16.0	14.2
3	≥ 18,110	14.8	11.9	19.4	27.6	26.3	≥ 18,110	26.1	17.0	34.8	14.4	7.7
4	≥ 15,030	15.2	10.6	17.2	26.5	30.5	≥ 15,020	18.8	12.8	48.2	12.3	7.9
5	≥ 13,000	17.7	11.4	24.5	22.5	23.9	≥ 12,960	15.4	10.3	57.6	10.1	6.7
6	≥ 11,480	14.6	11.8	31.4	23.1	19.1	≥ 11,470	13.9	10.6	54.3	13.0	8.1
7	≥ 11,010	17.2	14.4	34.2	17.9	16.3	≥ 11,000	23.2	19.0	44.9	8.2	4.7
8	≥ 10,890	17.2	12.1	33.1	17.1	20.5	≥ 10,870	24.2	18.0	45.0	7.6	5.2
9	≥ 13,980	28.9	17.4	33.4	12.1	8.3	≥ 14,010	36.0	18.4	38.9	4.6	2.1
10	≥ 10,760	17.5	16.3	35.5	15.8	14.9	≥ 10,800	24.5	17.9	45.1	7.9	4.7
11	≥ 9,770	13.1	16.2	34.9	17.0	18.8	≥ 9,780	20.4	18.1	44.4	10.2	7.0
12	≥ 7,610	10.7	14.6	36.3	18.7	19.7	≥ 7,650	16.0	20.0	47.4	9.9	6.7

*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	≥ 27,500	22.6	72.4	5.0
1	≥ 25,550	12.8	72.5	14.8
2	≥ 23,210	10.7	66.2	23.1
3	≥ 18,250	10.0	73.7	16.3
4	≥ 15,160	12.9	69.6	17.6
5	≥ 13,110	16.0	70.1	13.9
6	≥ 11,590	14.4	72.4	13.2
7	≥ 11,120	21.7	71.1	7.1
8	≥ 10,990	21.7	71.3	6.9
9	≥ 14,170	36.7	60.5	2.8
10	≥ 10,910	24.9	68.7	6.4
11	≥ 9,910	21.0	69.6	9.4
12	≥ 7,770	16.4	74.0	9.6

2.3 2021–2022 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) an14.1d classification consistency (CC); and
- inter-rater analysis.

The methods used in the computation of test reliability are described in Part I, Chapter 4, of this technical report. 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 0.77 to 0.96. 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	.83	.94	.77	.90	.90
1	.83	.95	.87	.85	.94
2	.84	.94	.84	.83	.86
3	.86	.94	.85	.84	.86
4	.85	.94	.83	.87	.87
5	.86	.95	.85	.88	.88
6	.91	.94	.80	.85	.89
7	.92	.95	.83	.87	.90
8	.93	.96	.86	.88	.91
9	.90	.95	.81	.91	.90
10	.89	.95	.84	.88	.88
11	.88	.94	.85	.87	.86
12	.87	.94	.85	.86	.83

3.2 Marginal Standard Error of Measurement

Another way to examine score reliability is with the MSEM (or $\bar{\sigma}_{error}$). 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 Section 4.2 of Part I of this technical report for more information. 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	≥ 27,220	.87	.85	.91	.89	.82	.83
1	≥ 25,310	.78	.91	.84	.91	.71	.85
2	≥ 22,990	.83	.91	.85	.91	.78	.88
3	≥ 18,030	.83	.91	.85	.91	.79	.88
4	≥ 14,960	.87	.90	.88	.91	.82	.89
5	≥ 12,900	.88	.91	.88	.91	.84	.90
6	≥ 11,400	.90	.88	.87	.91	.84	.88
7	≥ 10,930	.91	.90	.89	.92	.86	.90
8	≥ 10,800	.92	.91	.90	.92	.87	.91
9	≥ 13,880	.93	.91	.91	.93	.89	.91
10	≥ 10,690	.91	.91	.89	.91	.88	.90
11	≥ 9,700	.91	.91	.88	.90	.88	.89
12	≥ 7,540	.90	.90	.87	.89	.87	.87

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

3.4 Classification Accuracy and Consistency

Error! Reference source not found. shows the overall CA and CC in each domain. The detailed description of CA and CC can be found in Section 4.4 of Part I of this technical report. 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 Part I of this technical report) are presented in Table 3.3 and

Table 3.4, as well as Figure S11.1 and Figure S11.2 in the Appendix. For each cut score, all CAs are above 0.84 and all CCs are above 0.78. 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 **Error! Reference source not found.** and Figure S11.3 in the Appendix. All CAs and CCs are above 0.85 for overall and above 0.89 for each category. The CA indexes for between Emerging and Progressing are equal or higher than those for between Progressing and Proficient in all grades except for kindergarten and grades 9 and 10. 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 9 and 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	.72	.66	.69	.79	.63	.56	.60	.71
1	.64	.73	.59	.74	.54	.64	.52	.66
2	.69	.72	.58	.72	.59	.62	.50	.62
3	.68	.72	.58	.70	.58	.63	.49	.61
4	.72	.71	.63	.75	.62	.62	.54	.67
5	.73	.73	.62	.79	.63	.64	.53	.71
6	.76	.70	.62	.76	.67	.60	.52	.68
7	.73	.74	.64	.74	.64	.65	.55	.65
8	.74	.77	.67	.75	.65	.68	.57	.67
9	.76	.80	.70	.79	.67	.73	.61	.71
10	.73	.77	.66	.75	.64	.68	.57	.66
11	.72	.75	.65	.72	.63	.66	.55	.63
12	.72	.74	.64	.71	.62	.65	.55	.62

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

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

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

3.5 INTER-RATER ANALYSIS

For the 2021–2022 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.

Error! Reference source not found. 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 of Part I of this technical report. The table shows that 55.9%–93.4% of handscores are consistent between the first rater and the second rater, and 0.3%–5.6% of handscores are off by two or more points across the six tests. The weighted kappa coefficients ranged from 0.649 to 0.925. In 2020–2021, the weighted kappa coefficients ranged from 0.612 to 0.910. 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 in the Appendix) have relatively lower exact agreement (e.g., 58.7, 56.6), which 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	.747	.839	66.0	90.8	96.6	99.2	0.8	3.4
1	9	.666	.878	58.5	93.4	96.2	99.6	0.4	3.8
2–3	13	.649	.879	59.5	90.7	94.5	99.6	0.4	5.5
4–5	13	.700	.925	56.6	91.5	94.4	99.5	0.5	5.6
6–8	13	.755	.881	61.3	80.3	96.8	99.7	0.3	3.2
9–12	13	.769	.902	55.9	79.1	95.2	99.1	0.9	4.8

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 of 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.

4.3 SUMMARY OF CLASSICAL ITEM DIFFICULTY AND ITEM DISCRIMINATION

This section contains the summary of classical statistics for the spring 2021-2022 operational Forms. The operational data file used for this analysis was the 100% (all schools) student data file. CAI employs classical item analysis procedures to ensure that items function as intended with respect to the underlying scales. The summary statistics are based on Classical Test Theory (CTT) and include information such as the item difficulty and the discrimination mean statistics for each modality and grade band.

Table 4.1. Operational Summary of Classical Item Difficulty and Item Discrimination Indices by Grade Band (All Schools)

Grade Band	Modality	N-C ount	Item Difficulty		Item Discrimination	
			Mean	SD	Mean	SD
K	Listening	27037	0.71	0.41	0.53	0.12
	Speaking	26760	0.62	0.88	0.68	0.13
	Reading	26877	0.73	0.40	0.51	0.11
	Writing	26878	0.51	0.48	0.60	0.19
1	Listening	25260	0.84	0.34	0.54	0.07
	Speaking	25096	0.76	0.82	0.62	0.10
	Reading	25053	0.65	0.45	0.53	0.14
	Writing	25170	0.72	0.44	0.76	0.14
2-3	Listening	41090	0.82	0.38	0.55	0.09
	Speaking	40800	0.75	0.83	0.63	0.10
	Reading	40795	0.67	0.47	0.55	0.13
	Writing	40799	0.57	0.63	0.68	0.17
4-5	Listening	28025	0.77	0.42	0.53	0.10
	Speaking	27758	0.71	1.04	0.65	0.11
	Reading	27791	0.55	0.49	0.50	0.17
	Writing	27759	0.65	0.76	0.66	0.13
6-8	Listening	33349	0.79	0.40	0.64	0.10
	Speaking	32869	0.65	1.09	0.67	0.12
	Reading	33153	0.51	0.53	0.45	0.18
	Writing	32962	0.63	1.02	0.72	0.13
9-12	Listening	42170	0.69	0.48	0.61	0.14
	Speaking	41297	0.65	1.20	0.70	0.12
	Reading	42035	0.49	0.48	0.42	0.19
	Writing	41506	0.58	1.04	0.69	0.14

Chapter 5. Reporting

A detailed introduction to the Centralized Reporting System can be found in Part I, Chapter 6, of this technical report. The reporting mock-ups for the summative tests of each state appear in Section 15 of the Appendix. It is noted that the mock-up 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.). 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.