TECHNICAL REPORT

PART II—SUMMATIVE ASSESSMENT

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

English Language Proficiency Assessment for the 21st Century (ELPA21)—

Listening, Reading, Speaking, and Writing

Grades K–12

2024–2025 Test Administration

Submitted to
ELPA21

Submitted by
Cambium Assessment, Inc.
4200 Wilson Blvd., Ste. 610
Arlington, VA 22203-1800

October 2025

Table of Contents

Chapter 1. Test Administration 1
1.1 Testing Windows 1
1.2 Test Design 1
1.3 Test Administration Manual 3
1.3.1 Directions for Test Administration 3
1.3.2 Training/Practice Tests 3
1.3.3 Instructions for the Summative Assessment 4
1.4 Business Scoring Rules for the Summative Assessment 4
Chapter 2. 2024–2025 Summary 6
2.1 2024–2025 Student Participation 7
2.2 2024–2025 Student Scale Score and Performance-Level Summary 9
2.3 2024–2025 Testing Time for Online Summative Assessment 16
Chapter 3. Reliability 17
3.1 Internal Consistency 18
3.2 Marginal Standard Error of Measurement 18
3.3 Marginal Reliability and Conditional Standard Error of Measurement 19
3.4 Classification Accuracy and Consistency 19
3.5 Inter-Rater Analysis 23
Chapter 4. Validity 25
4.1 Dimensionality Analysis 26
4.2 Student Abilities Versus Item Difficulties 26
4.3 Relationship Between Summative and Screener Assessments 26
4.3.1 Correlation Between Summative and Screener Assessments 26
4.3.2 Student Progress from Screener to Summative 27
Chapter 5. Reporting 28
Chapter 6. Classical Item and Test Analysis Results 29
6.1 Item Analysis Results 29
6.2 Differential Item Functioning Results 30
6.3 Domain Intercorrelations 37
References 38
Appendix A 39

List of Tables

Table 1.1 2024–2025 ELPA21 Summative Testing Windows by State 1

Table 1.2 Number of Items and Score Points by Domain and Grade Band—Online Summative . 2

Table 1.3 Number of Items and Score Points by Domain and Grade Band—Paper Summative ... 2

Table 1.4 Number of Items and Score Points by Domain and Grade Band—Braille Summative . 2

Table 1.5 Number of Field-Test Items by Domain and Grade Band—Online Summative 3

Table 1.6 Scoring Outcome for the Comprehension Score 5

Table 2.1 Student Participation in Each State by Grade 8

Table 2.2 Scale Score Summary by Grade—Listening and Reading* 10

Table 2.3 Scale Score Summary by Grade—Speaking and Writing* 11

Table 2.4 Scale Score Summary by Grade—Comprehension and Overall* 12

Table 2.5 Percentage of Students in Each Performance Level by Grade—Listening and Reading* 13

Table 2.6 Percentage of Students in Each Performance Level by Grade—Speaking and Writing* 14

Table 2.7 Percentage of Students in Each Overall Proficiency Category by Grade 15

Table 3.1 Cronbach's Alpha by Domain and Grade 18

Table 3.2 Marginal Reliability by Score and Domain* 19

Table 3.3 Overall Classification Accuracy and Consistency for Domain Performance Levels by Grade and Domain* 20

Table 3.4 Classification Accuracy for Each Cut Score by Grade and Domain* 21

Table 3.5 Classification Consistency for Each Cut Score by Grade and Domain* 22

Table 3.6 Summative Classification Accuracy and Classification Consistency for Overall Proficiency Categories by Grade 22

Table 3.7 Summary of Kappa Coefficients by Grade Band 24

Table 6.1 Summary of Classical Item Analysis for Operational Items (Combined Data from Eight States) 29

Table 6.3 DIF Sample Sizes for DIF Groups 30

Table 6.4 2024–2025 Machine-Scored Field-Test Item DIF Analyses (Female vs. Male) 31

Table 6.5 2024–2025 Machine-Scored Field-Test Item DIF Analyses (Black vs. Non-Black) ... 32

Table 6.6 2024–2025 Machine-Scored Field-Test Item DIF Analyses (White vs. Non-White) .. 33

Table 6.7 2024–2025 Machine-Scored Field-Test Item DIF Analyses (Hispanic vs. Non-Hispanic) 34

Table 6.8 2024–2025 Machine-Scored Field-Test Item DIF Analyses (Asian vs. Non-Asian) ... 35

Table 6.9 2024–2025 Machine-Scored Field-Test Item DIF Analyses (SPED, IEP, or Section 504 Plan vs. Non-SPED, IEP, or Section 504 Plan) 36

Table 6.2 Intercorrelation Between the Domain Scale Scores by Grade Band (Eight States) 37

Table A.1 Classical Item Analysis of Field-Test Items (Combined Data from All States)..... 40

Chapter 1. Test Administration

The summative assessment was 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 in each grade band. Students can be exempted from as many as three domain tests. The summative assessment does not have a time limit.

1.1 Testing Windows

The 2024–2025 summative testing windows for the eight states discussed in this report are shown in Table 1.1.

Table 1.1 2024–2025 ELPA21 Summative Testing Windows by State

State	ELPA21 Summative
Arkansas	3/3–4/11/2025
Iowa	2/03–3/28/2025
Louisiana	2/10–3/19/2025
Mississippi	3/3-4/16/2025
Nebraska	1/27–3/14/2025
Ohio	2/03–3/28/2025
Tennessee	2/3-3/28/2025
West Virginia	2/4–3/21/2025

1.2 Test Design

The 2024–2025 summative assessment included one online form, one paper-pencil form, and one braille form. Each form had separate tests for the four language domains.

Table 1.2–Table 1.4 present the number of operational items and score points for the online, paper-pencil, and braille forms, respectively. Item counts are similar for the listening and reading tests, as well as for the writing and speaking tests; however, the number of items in the latter was fewer than that in the former. Braille forms had fewer items than those in the online and paper forms. Field-test items were also included in the 2024–2025 summative assessment (see details in Table 1.5).

Note that the 2024-2025 test administration represents the first year of a shorter test length for the online form since the start of ELPA21. The shorter test length was approved by the ELPA21 Governing Board in June 2024, following extensive analyses and input from the ELPA21 Technical Advisory Committee.

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	20	20	20	20	21	22	22	25	26	30	23	26
Reading	22	22	24	24	26	32	25	28	24	28	27	27
Speaking	7	16	6	15	6	15	6	22	5	19	5	19
Writing	16	16	17	17	12	20	8	15	5	19	5	19
Total	65	74	67	76	65	89	61	90	60	96	60	91

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	3	12	4	3	8	9	**39**
Reading	0	16	16	23	19	25	**99**
Speaking							**0**
Writing	21	15	5	10	7	5	**63**
Total	**24**	**43**	**25**	**36**	**34**	**39**	**201**

1.3 Test Administration Manual

1.3.1 Directions for Test Administration

For the 2024–2025 test 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 English Language Proficiency Assessment for the 21st Century (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 the Test Delivery System (TDS), 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 the Summative Assessment

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

- Logging in to the Cambium Assessment, Inc. (CAI) 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 2024–2025 summative assessment include the following:

1. A domain test was considered "attempted" if a student was presented with the first operational item; the student did not need to respond to an 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 score and performance level were set to "E" (Domain Exempted). If the exempted domain test was reading or listening, the test was left out of the computation of the comprehension score. However, if the domain test was started in CAI's 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 1.6 for the reporting of scenarios in which neither listening nor reading was attempted (i.e., each domain was either exempted or not 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. 2024–2025 Summary

This chapter presents the 2024–2025 student participation and performance summary statistics combining data from all eight states (Arkansas, Iowa, Louisiana, Mississippi, Nebraska, Ohio, Tennessee, and West Virginia). Additionally, summary statistics by subgroup and an analysis of testing times for the combined states can be found in Sections 1–7 of the pooled appendix for the summative assessment. State-specific results are detailed in each state-specific appendix of this technical report. The figures and tables included in Sections 1–7 of pooled appendix and state-specific appendices for the summative assessment are listed as follows:

- Section 1. Summative Assessment—Student Participation
    - Table S1.1 displays the number and percentage of students in each testing 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. In the pooled appendix, subgroups include grade, gender, and ethnicity; in the state-specific appendix, subgroups include additional characteristics (e.g., migrant, special education, Title I, or Section 504 Plan status) depending on the availability of each variable in the state.
- 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.
- 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, mean, and standard deviation of the domain scale scores; overall scale scores; and comprehension scale scores across the eight 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 “Not-Attempted” 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 by grade or grade band.

2.1 2024–2025 Student Participation

Table 2.1 summarizes student participation in each state by grade in the past year and current year. There were 346,668 students in total who participated in the 2024–2025 summative assessment, which is an increase of 49.4% from 2023–2024. Tennessee and Mississippi are new states in 2025. The state of Ohio had the most tested students, followed by the state of Tennessee.

Table 2.1 Student Participation in Each State by Grade

Grade	Arkansas		Iowa		Louisiana		Mississippi	Nebraska		Ohio		Tennessee	West Virginia		Total		
	2024	2025	2024	2025	2024	2025	2025	2024	2025	2024	2025	2025	2024	2025	2024	2025	Diff
K	≥4,590	≥4,630	≥4,870	≥5,080	≥4,000	≥4,000	≥1,980	≥4,140	≥4,170	≥11,280	≥11,600	≥10,280	≥280	≥290	≥29,190	≥42,040	≥12,850
1	≥4,500	≥4,510	≥4,580	≥4,770	≥4,530	≥4,300	≥2,060	≥3,990	≥4,140	≥11,070	≥11,720	≥9,430	≥260	≥320	≥28,950	≥41,270	≥12,320
2	≥4,220	≥4,060	≥3,970	≥4,230	≥4,180	≥4,180	≥1,870	≥3,560	≥3,720	≥9,970	≥10,300	≥9,210	≥220	≥230	≥26,140	≥37,830	≥11,680
3	≥3,400	≥3,620	≥3,210	≥3,480	≥3,210	≥3,520	≥1,880	≥2,770	≥3,030	≥7,740	≥8,830	≥9,210	≥210	≥210	≥20,560	≥33,800	≥13,240
4	≥3,260	≥3,150	≥2,910	≥3,060	≥2,790	≥2,900	≥1,670	≥2,390	≥2,640	≥6,790	≥7,260	≥7,560	≥190	≥210	≥18,360	≥28,480	≥10,120
5	≥2,860	≥3,050	≥2,590	≥2,770	≥2,520	≥2,510	≥1,220	≥2,010	≥2,290	≥5,990	≥6,540	≥5,960	≥170	≥190	≥16,160	≥24,560	≥8,390
6	≥2,810	≥2,760	≥2,300	≥2,520	≥2,340	≥2,370	≥770	≥1,630	≥1,960	≥5,390	≥6,090	≥5,360	≥150	≥160	≥14,640	≥22,020	≥7,370
7	≥2,730	≥2,670	≥2,140	≥2,360	≥2,080	≥2,260	≥830	≥1,500	≥1,750	≥4,960	≥5,620	≥5,280	≥140	≥150	≥13,570	≥20,970	≥7,390
8	≥2,760	≥2,820	≥2,030	≥2,330	≥2,350	≥2,160	≥740	≥1,400	≥1,660	≥4,570	≥5,410	≥5,220	≥140	≥180	≥13,280	≥20,550	≥7,270
9	≥3,130	≥3,090	≥2,410	≥2,620	≥2,940	≥2,940	≥730	≥1,940	≥1,990	≥5,930	≥6,280	≥5,400	≥170	≥180	≥16,540	≥23,260	≥6,720
10	≥2,790	≥3,040	≥2,380	≥2,510	≥2,060	≥2,120	≥620	≥1,780	≥2,030	≥4,630	≥5,450	≥4,990	≥200	≥190	≥13,880	≥20,990	≥7,100
11	≥2,630	≥2,580	≥2,040	≥2,290	≥1,540	≥1,550	≥510	≥1,390	≥1,630	≥4,070	≥4,480	≥4,050	≥160	≥210	≥11,860	≥17,340	≥5,470
12	≥1,990	≥2,240	≥1,620	≥1,760	≥1,060	≥1,220	≥350	≥1,110	≥1,370	≥2,870	≥3,390	≥2,970	≥110	≥150	≥8,780	≥13,490	≥4,710
Total	≥41,740	≥42,270	≥37,110	≥39,830	≥35,640	≥36,090	≥15,300	≥29,680	≥32,420	≥85,320	≥93,030	≥84,970	≥2,460	≥2,720	≥231,970	≥346,660	≥114,690

Table S1.1 in Section 1 of the pooled appendix for the summative assessment presents student participation in each mode of testing. In the eight states combined, the most frequent mode of test administration was online (99.89%), followed by paper (0.11%) and braille (0.00%).

Table S1.2 in Section 1 of the pooled appendix for the summative assessment shows student participation by subgroups. For the pooled analysis, there were more male students (50.5%–54.6%) than female students (45.0%–48.9%) tested. In each test, most students were Hispanic or Latino (66.0%–71.3%), followed by Asian students (7.4%–12.1%) and Black or African American students (8.4%–10.7%).

Tables S2.1–S2.13 in Section 2 and Figures S3.1–S3.65 in Section 3 of the pooled appendix for the summative assessment present the raw score summary descriptive statistics and score distributions by domain performance level and by overall proficiency level in each grade.

2.2 2024–2025 Student Scale Score and Performance-Level Summary

Table 2.2–Table 2.4 summarize student performance in the 2024–2025 test administration across the eight 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 pooled appendix for the summative assessment.

Table 2.5 and Table 2.6 display the percentage of students in each performance level for each grade and domain. In addition, Sections 5 of the pooled appendix for the summative assessment further summarizes the percentage of students in each domain performance level by subgroup in each grade. For both reading and writing in the pooled analysis, Table 2.5 and Table 2.6 show that most students are in performance levels 1 and 3. The percentages of students in performance levels 4 and 5 in listening and speaking are higher than those in reading and writing.

Table 2.7 summarizes the percentage of students in each overall proficiency category in each grade. Additionally, Section 6 of the pooled appendix for the summative assessment further summarizes the percentage of student in each overall proficiency category by subgroup. Table 2.7 shows that most students (55.2%–67.0%) are in the Progressing category in all grades. The percentage of Proficient students is the lowest in grade 9 at 3.9%, followed by kindergarten at 5.2% and grade 10 at 5.8%. Grade 4 has the highest Proficient percentage at 22.1%. In general, the Proficient rate is higher in elementary schools (except for kindergarten) than that in middle schools and high schools.

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	≥41,990	245	532.6	743	79.4	≥41,860	256	536.3	737	75.1
1	≥41,230	241	533.4	698	83.3	≥41,080	242	528.6	740	83.4
2	≥37,790	232	511.9	727	74.3	≥37,640	235	504.2	758	74.0
3	≥33,780	232	538.2	727	82.2	≥33,610	235	539.2	758	84.8
4	≥28,460	229	504.4	757	80.4	≥28,290	242	511.3	773	76.4
5	≥24,530	229	513.8	742	85.0	≥24,390	242	523.0	758	80.3
6	≥21,970	244	502.2	729	75.1	≥21,850	274	507.9	748	65.5
7	≥20,930	255	510.4	753	81.6	≥20,820	275	518.6	764	71.3
8	≥20,510	233	526.0	778	88.1	≥20,400	250	534.8	780	78.2
9	≥23,200	261	507.9	742	78.7	≥23,120	273	506.9	747	72.7
10	≥20,930	261	520.8	777	79.7	≥20,850	273	518.8	775	75.3
11	≥17,290	261	536.7	776	80.1	≥17,240	273	533.4	775	77.1
12	≥13,440	261	540.1	767	76.3	≥13,390	273	536.3	770	74.8

*Scores from domain tests marked as Exempt or Not Attempted are excluded.
*Results include all records with valid scale scores. No special filter was used to exclude invalidated records. If invalidated records had scale scores, they were included.
*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	≥41,610	301	537.4	713	88.2	≥41,810	315	528.7	716	74.0
1	≥40,880	277	552.1	725	82.2	≥41,050	245	520.3	727	90.7
2	≥37,520	267	515.3	722	78.4	≥37,620	242	498.2	757	82.1
3	≥33,570	267	543.4	722	84.0	≥33,630	242	536.3	757	90.8
4	≥28,300	249	521.0	775	85.4	≥28,330	238	505.5	767	83.5
5	≥24,400	249	526.8	764	88.3	≥24,420	238	517.3	751	87.3
6	≥21,850	278	522.7	729	81.8	≥21,860	259	499.1	753	79.3
7	≥20,790	285	528.1	751	85.1	≥20,840	273	510.6	763	85.4
8	≥20,370	270	538.5	766	88.2	≥20,410	247	526.9	784	90.7
9	≥22,890	313	522.3	721	77.9	≥23,110	275	504.8	750	83.9
10	≥20,620	313	535.1	742	75.1	≥20,830	275	518.7	780	83.1
11	≥17,030	313	549.3	742	73.6	≥17,220	275	535.4	779	81.9
12	≥13,240	313	552.9	732	70.4	≥13,380	275	539.8	772	77.0

*Scores from domain tests marked as Exempt or Not Attempted are excluded.
*Results include all records with valid scale scores. No special filter was used to exclude invalidated records. If invalidated records had scale scores, they were included.
*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	≥42,020	3419	5381.7	6776	550.6	≥42,040	3254	5358.6	6882	598.8
1	≥41,250	3437	5366.0	6649	567.7	≥41,270	3038	5363.0	6864	663.8
2	≥37,810	3325	5206.4	6801	528.9	≥37,830	3029	5148.4	7048	599.6
3	≥33,790	3325	5417.3	6665	595.1	≥33,800	3029	5409.1	7048	669.7
4	≥28,480	3372	5197.7	6801	562.3	≥28,480	3005	5185.6	7248	637.0
5	≥24,550	3372	5274.9	6801	595.3	≥24,560	3005	5265.0	7124	667.3
6	≥22,000	3392	5173.4	6944	502.3	≥22,020	3198	5162.2	6967	587.0
7	≥20,950	3535	5238.6	6944	546.9	≥20,970	3270	5234.8	7128	630.6
8	≥20,530	3392	5355.7	6944	601.4	≥20,550	3093	5352.6	7308	675.1
9	≥23,230	3553	5190.4	7060	542.1	≥23,260	3344	5179.7	6952	608.8
10	≥20,970	3553	5278.2	7060	563.9	≥20,990	3344	5283.6	7242	606.4
11	≥17,320	3553	5387.5	7060	578.5	≥17,340	3344	5407.2	7235	602.6
12	≥13,480	3553	5408.4	7060	561.2	≥13,490	3344	5436.7	7115	570.7

*Results include all records with valid scale scores. No special filter was used to exclude invalidated records. If invalidated records had scale scores, they were included.
*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	≥41,990	22.8	15.0	44.9	9.1	8.2	≥41,860	22.6	18.3	33.5	12.4	13.2
1	≥41,230	15.6	8.3	28.0	21.1	27.0	≥41,080	32.1	15.3	26.5	12.5	13.6
2	≥37,790	10.3	7.3	29.5	26.8	26.2	≥37,640	30.3	16.5	26.9	12.6	13.7
3	≥33,780	8.6	7.8	27.6	29.3	26.6	≥33,610	34.2	15.7	26.0	12.2	11.9
4	≥28,460	12.5	8.1	19.8	34.2	25.4	≥28,290	27.0	13.7	26.7	16.5	16.1
5	≥24,530	16.0	10.0	12.7	37.5	23.8	≥24,390	29.7	15.0	31.0	13.3	10.9
6	≥21,970	13.7	9.9	22.8	30.7	22.9	≥21,850	27.8	17.8	33.8	12.1	8.4
7	≥20,930	20.7	14.0	30.9	18.2	16.2	≥20,820	36.9	21.8	29.3	7.1	≤5.0
8	≥20,510	18.9	14.0	28.5	20.1	18.5	≥20,400	34.5	20.8	34.4	6.0	≤5.0
9	≥23,200	29.8	14.9	30.6	14.2	10.5	≥23,120	45.8	20.4	28.2	≤5.0	≤5.0
10	≥20,930	24.1	15.1	30.3	15.8	14.7	≥20,850	40.5	20.0	31.0	5.2	≤5.0
11	≥17,290	17.9	14.0	30.4	17.7	20.1	≥17,240	33.0	20.1	34.7	7.0	5.2
12	≥13,440	14.4	14.6	33.5	17.9	19.6	≥13,390	30.0	21.7	36.3	6.9	5.1
Total	≥346,110	16.9	11.2	28.9	22.6	20.3	≥344,610	31.9	17.7	29.9	10.6	9.9

*Scores from domain tests marked as Exempt or Not Attempted are excluded.
*Results include all records with valid scale scores. No special filter was used to exclude invalidated records. If invalidated records had scale scores, they were included.

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	≥41,610	27.2	15.6	29.9	11.6	15.8	≥41,810	38.1	32.4	21.4	≤5.0	≤5.0
1	≥40,880	32.3	23.8	9.0	13.4	21.5	≥41,050	41.2	17.9	23.6	7.3	10.0
2	≥37,520	32.1	19.4	14.7	16.6	17.1	≥37,620	31.7	15.4	25.6	12.9	14.4
3	≥33,570	24.8	14.2	18.2	22.3	20.5	≥33,630	33.5	14.5	25.2	12.8	14.0
4	≥28,300	21.9	13.2	18.6	22.3	24.0	≥28,330	24.0	12.2	38.4	12.6	12.8
5	≥24,400	26.0	15.0	24.5	17.7	16.8	≥24,420	22.5	11.8	46.4	10.6	8.7
6	≥21,850	23.0	13.3	27.1	18.7	17.9	≥21,860	21.9	13.4	41.5	12.3	10.9
7	≥20,790	25.1	15.8	29.8	14.4	14.9	≥20,840	33.2	15.4	35.3	8.6	7.5
8	≥20,370	23.3	13.9	30.5	14.9	17.3	≥20,410	31.7	15.0	37.3	8.8	7.3
9	≥22,890	30.1	19.7	31.4	10.3	8.5	≥23,110	41.7	16.2	32.5	5.3	≤5.0
10	≥20,620	24.0	20.2	32.4	11.6	11.8	≥20,830	35.8	16.9	34.3	7.1	6.0
11	≥17,030	18.3	18.0	34.1	13.6	16.0	≥17,220	28.8	16.4	36.6	9.0	9.1
12	≥13,240	15.4	18.7	35.7	14.6	15.7	≥13,380	25.3	17.8	39.6	9.0	8.4
Total	≥343,110	26.1	17.2	23.8	15.7	17.3	≥344,570	32.5	17.3	31.7	9.1	9.4

*Scores from domain tests marked as Exempt or Not Attempted are excluded.
*Results include all records with valid scale scores. No special filter was used to exclude invalidated records. If invalidated records had scale scores, they were included.

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

Grade	N	Emerging	Progressing	Proficient
K	≥42,040	29.8	65.0	5.2
1	≥41,270	22.0	62.9	15.1
2	≥37,830	17.0	64.1	18.9
3	≥33,800	16.1	63.8	20.1
4	≥28,480	19.7	58.2	22.1
5	≥24,560	24.4	60.2	15.4
6	≥22,020	22.2	63.5	14.4
7	≥20,970	31.0	60.2	8.8
8	≥20,550	29.3	62.5	8.1
9	≥23,260	40.9	55.2	≤5.0
10	≥20,990	35.4	58.8	5.8
11	≥17,340	28.0	63.4	8.6
12	≥13,490	24.6	67.0	8.5
Total	≥346,660	25.3	62.1	12.7

2.3 2024–2025 Testing Time for Online Summative Assessment

Table S7.1 in the pooled appendix for the summative assessment shows the testing time at various percentiles 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. This analysis only included students who completed all four domain tests. Testing records 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, Chapter 4, of this technical report. The results for each method are included in Sections 8–12 of the pooled appendix for the summative assessment. The figures and the tables in each section of the pooled appendix for the summative assessment are as follows:

- Section 8. Summative Assessment—Cronbach's Alpha
    - Figure S8.1 presents the Cronbach's alpha for each domain test and for overall by grade.
- Section 9. Summative Assessment—Marginal Reliability
    - Figure S9.1 presents the ratio of MSEM to the standard deviation of scale scores at the test level (i.e., four domain tests, the overall test, and the comprehension test).
    - Figure S9.2 presents the marginal reliability of scores at the test level (i.e., four domain tests, the overall test, and the comprehension test).
    - Figures S9.3 and S9.4 present the marginal reliability of domain test scores by gender and ethnicity, respectively.
- Section 10. Summative Assessment—CSEM
    - Figures S10.1–S10.13 present the CSEM plots for each domain, overall, and comprehension tests in each grade. The CSEM plots use different colors to differentiate students who attempted all four domains from those who did not attempt or were exempted from one or more domains.
- Section 11. Summative Assessment—Classification Accuracy and Classification Consistency
    - Figures S11.1 and S11.2 show the CA and CC, respectively, for each domain test by grade.
    - 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 or grade band.

3.1 Internal Consistency

Due to a small sample size (see Section 1 of the pooled appendix for the summative assessment), students who took braille and paper-pencil tests were excluded from the reliability analysis. Table 3.1 shows the values of Cronbach's alpha derived from the pooled sample (across states), arranged by grade level. Values range from 0.79 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 the reliability estimates for all domains are acceptable based on this criterion. The results of Cronbach's alpha for all domains by grade are plotted in Figure S8.1 in the pooled appendix for the summative assessment.

Table 3.1 Cronbach's Alpha by Domain and Grade

Grade	Listening	Reading	Speaking	Writing	Overall
K	.82	.82	.87	.88	.94
1	.85	.79	.83	.93	.94
2	.84	.81	.81	.85	.93
3	.86	.85	.83	.86	.94
4	.85	.86	.85	.83	.94
5	.86	.87	.86	.83	.94
6	.86	.79	.86	.86	.93
7	.87	.82	.86	.86	.94
8	.88	.84	.86	.87	.94
9	.86	.84	.89	.83	.94
10	.87	.86	.88	.82	.94
11	.86	.87	.86	.81	.94
12	.85	.86	.85	.79	.93

3.2 Marginal Standard Error of Measurement

Another way to examine score reliability is with the marginal standard error of measurement (MSEM), or $\bar{\sigma}_{error}$. The ratio of the 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 over the total score standard deviation ($\frac{\bar{\sigma}_{error}}{\sigma_{total}}$). See details in Section 4.2 of Part I in this technical report for more information. The plot of this ratio is displayed in Figure S9.1 in the pooled appendix for the summative assessment.

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 pooled appendix for the summative assessment. See details in Section 4.3 of Part I in this technical report for more information. All the reliability indices are above 0.8, except for the comprehension composite score reliability in grade 1 at 0.73 and grade 2 at 0.79. In addition, Section 9 of the pooled appendix for the summative assessment presents marginal reliability by subgroups, and Section 10 of the pooled appendix for the summative assessment displays CSEM plots by grades.

Table 3.2 Marginal Reliability by Score and Domain*

Grade	N	Listening	Reading	Speaking	Writing	Comprehension	Overall
K	≥41,550	.85	.85	.86	.88	.80	.83
1	≥40,800	.81	.89	.80	.90	.73	.85
2	≥37,440	.84	.91	.83	.92	.79	.88
3	≥33,460	.84	.92	.83	.92	.80	.90
4	≥28,180	.88	.91	.86	.90	.85	.90
5	≥24,280	.89	.92	.87	.90	.86	.90
6	≥21,720	.90	.89	.86	.90	.84	.89
7	≥20,690	.91	.90	.86	.91	.86	.90
8	≥20,260	.91	.91	.87	.92	.87	.91
9	≥22,780	.92	.91	.88	.90	.89	.90
10	≥20,520	.92	.91	.88	.90	.89	.90
11	≥16,950	.91	.91	.87	.90	.89	.90
12	≥13,180	.91	.91	.86	.89	.88	.89

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

3.4 Classification Accuracy and Consistency

Table 3.3 shows the overall classification accuracy (CA) and classification consistency (CC) in each domain. The detailed description of CA and CC can be found in Section 4.4 of Part I in this technical report. Scores from paper-pencil and braille tests were excluded. CC rates can be lower than CA rates 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 of measurement.

The pooled analysis results for the CA and CC of each cut score, which can be found in Table 3.1 in Part I of this technical report, are presented in Table 3.4 and Table 3.5, as well as Figures S11.1 and S11.2 in the pooled appendix for the summative assessment. For each cut score, all CAs are

above 0.83 and all CCs are above 0.76. In speaking, CC indices at the Level 3 cut score are slightly lower than those for the other domains, although the indices remain close to .80. This pattern may suggest that speaking provides marginally less information around this cut score. If desired, this could potentially be addressed by incorporating speaking items that are better aligned with the cut score or that demonstrate stronger discrimination in this proficiency range.

The CA and CC results for overall proficiency categories are summarized in Table 3.6 and Figure S11.3 in the pooled appendix for the summative assessment. All CAs and CCs are above 0.83 for overall and above 0.89 for each cut score. The CA indices for the cut between Emerging and Progressing are equal to or higher than those for the cut between Progressing and Proficient in all grades except kindergarten and grades 9 and 10. The CC indices follow a similar pattern.

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	.70	.66	.65	.76	.61	.57	.56	.68
1	.64	.70	.57	.73	.54	.61	.50	.65
2	.67	.72	.57	.73	.56	.63	.49	.64
3	.67	.72	.56	.72	.56	.64	.48	.63
4	.71	.72	.60	.72	.60	.63	.51	.63
5	.72	.74	.60	.75	.62	.66	.51	.67
6	.73	.71	.61	.73	.63	.61	.52	.64
7	.72	.75	.62	.73	.63	.66	.53	.65
8	.72	.77	.63	.75	.63	.69	.54	.66
9	.75	.81	.69	.77	.66	.74	.60	.69
10	.73	.79	.67	.74	.64	.71	.57	.66
11	.72	.77	.65	.72	.63	.69	.55	.63
12	.72	.76	.64	.71	.62	.68	.54	.61

*Scores for domain tests marked as Exempt 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	.94	.91	.90	.92	.93	.90	.89	.91	.93	.90	.87	.89	.89	.93	.95	.96
1	.96	.94	.86	.85	.93	.91	.91	.93	.88	.83	.83	.85	.94	.91	.91	.93
2	.96	.94	.88	.87	.92	.92	.93	.94	.89	.85	.85	.88	.94	.92	.92	.94
3	.97	.95	.88	.86	.94	.93	.91	.93	.91	.86	.83	.86	.94	.92	.91	.93
4	.96	.95	.90	.89	.93	.93	.92	.94	.94	.89	.85	.87	.95	.93	.90	.93
5	.96	.95	.92	.89	.94	.93	.92	.94	.93	.88	.85	.88	.96	.94	.90	.93
6	.96	.95	.92	.90	.92	.91	.92	.95	.94	.89	.85	.88	.95	.93	.90	.93
7	.96	.94	.90	.91	.92	.91	.94	.96	.94	.88	.86	.89	.94	.91	.91	.94
8	.96	.95	.90	.90	.94	.92	.94	.96	.94	.90	.86	.89	.95	.92	.91	.94
9	.94	.94	.92	.94	.94	.93	.96	.98	.93	.90	.90	.93	.93	.91	.94	.97
10	.94	.94	.92	.92	.94	.93	.95	.97	.94	.90	.88	.91	.94	.91	.92	.95
11	.95	.94	.91	.91	.94	.93	.94	.96	.95	.90	.87	.90	.94	.92	.91	.93
12	.95	.94	.91	.91	.94	.92	.93	.96	.95	.90	.86	.89	.93	.91	.91	.94

*Scores for domain tests marked as Exempt or Not Attempted are excluded.
*Cut scores 1–4 fall between performance levels 1 and 2, 2 and 3, 3 and 4, and 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	.91	.87	.86	.89	.90	.86	.84	.87	.91	.86	.82	.85	.85	.90	.94	.95
1	.93	.91	.80	.79	.89	.88	.88	.90	.82	.76	.77	.79	.92	.87	.87	.90
2	.94	.92	.82	.82	.89	.88	.90	.92	.84	.79	.79	.83	.91	.89	.89	.91
3	.95	.93	.82	.81	.91	.90	.88	.91	.87	.80	.77	.81	.92	.89	.87	.90
4	.95	.93	.86	.85	.90	.90	.88	.92	.91	.84	.79	.82	.93	.89	.86	.90
5	.94	.92	.88	.84	.92	.91	.89	.92	.89	.83	.80	.84	.94	.91	.86	.91
6	.94	.92	.88	.87	.88	.88	.89	.93	.91	.84	.79	.83	.93	.91	.86	.90
7	.94	.91	.86	.88	.89	.88	.92	.95	.91	.83	.81	.85	.92	.88	.88	.92
8	.94	.93	.86	.87	.91	.89	.92	.95	.92	.85	.81	.84	.93	.89	.88	.92
9	.91	.91	.89	.92	.91	.90	.95	.97	.91	.86	.86	.90	.91	.88	.92	.95
10	.92	.91	.88	.89	.91	.90	.93	.96	.91	.86	.84	.88	.91	.88	.89	.93
11	.93	.92	.88	.88	.92	.90	.91	.94	.93	.86	.82	.86	.91	.88	.87	.91
12	.94	.91	.87	.87	.91	.89	.91	.94	.93	.85	.81	.85	.91	.87	.87	.91

*Scores for domain tests marked as Exempt or Not Attempted are excluded.
*Cut scores 1–4 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	.90	.93	.97	.87	.91	.96
1	.87	.95	.92	.83	.93	.90
2	.88	.95	.92	.84	.94	.90
3	.87	.96	.91	.84	.94	.90
4	.87	.96	.91	.83	.94	.89
5	.88	.96	.92	.85	.94	.91
6	.88	.96	.93	.85	.94	.91
7	.90	.95	.95	.87	.93	.94
8	.90	.95	.95	.87	.94	.94
9	.92	.94	.97	.89	.92	.97
10	.90	.94	.96	.87	.92	.95
11	.89	.95	.95	.86	.93	.93
12	.89	.95	.95	.86	.93	.93

3.5 Inter-Rater Analysis

For the 2024–2025 summative assessment, consistency of handscoring was evaluated for a total of 45 items (7 items in kindergarten and grades 4–5, 6 items in grade 1, 9 items in grades 2–3, and 8 items in grades 6–8 and 9–12). Handscored items on paper-pencil and braille forms were not included in the results due to the small sample size.

Table 3.7 presents the summary of kappa coefficients by maximum score points possible in each grade or grade band for each summative assessment in the pooled analysis. The table combines items in the same grade/grade band with the same point value, regardless of domain (writing and speaking items are combined). The detailed item-level inter-rater consistencies are summarized in Section 12 of the pooled appendix for the summative assessment. The description of kappa coefficients can be found in Section 4.5, Inter-Rater Analysis, of Part I of this technical report. Table 3.7 shows that 53.3%–92.6% of handscores are consistent between the first rater and the second rater, and 0.9%–10.4% of handscores are off by two or more points. The weighted kappa coefficients ranged from 0.657 to 0.919. Overall, inter-rater consistency indices (including weighted kappa and rater agreement indices) for the majority of items indicate fair to strong agreement. The minimum weighted kappa values across items are above .70, with the exception of one item in grades 2–3, which showed a slightly lower value of .657. In contrast, three items in grades 4–5 at score points 4 and 5 exhibited lower exact agreement percentages (with both the minimum and maximum below 60%), suggesting a potential need for additional rater training to support more consistent and reliable scoring. ELPA21 is coordinating with Cambium Assessment, Inc. on more frequent monitoring of scoring activities with the handscoring vendor to improve inter-rater consistency results.

Table 3.7 Summary of Kappa Coefficients by Grade Band

Grade/ Grade Band	Score Points	Number of Items	Weighted Kappa		% Exact Agreement		% Within 1 Agreement		% Not Within 1 Agreement	
			Min	Max	Min	Max	Min	Max	Min	Max
K	2	5	0.747	0.890	71.3	91.7	96.9	98.5	1.5	3.1
	3	2	0.778	0.797	60.9	62.8	97.4	97.9	2.1	2.6
1	2	4	0.767	0.859	83.0	92.6	98.5	99.1	0.9	1.5
	3	1	0.919	0.919	86.3	86.3	99.0	99.0	1.0	1.0
	4	1	0.880	0.880	75.7	75.7	97.2	97.2	2.8	2.8
2–3	2	4	0.657	0.788	76.3	83.5	97.5	98.5	1.5	2.5
	3	2	0.802	0.844	67.3	69.8	96.9	98.0	2.0	3.1
	4	3	0.770	0.844	57.1	64.1	92.1	95.5	4.5	7.9
4–5	3	4	0.708	0.839	60.3	70.4	94.2	98.2	1.8	5.8
	4	1	0.774	0.774	53.5	53.5	94.3	94.3	5.7	5.7
	5	2	0.732	0.794	53.3	53.7	89.6	92.5	7.5	10.4
6–8	3	4	0.785	0.814	64.8	73.8	96.7	97.7	2.3	3.3
	5	4	0.805	0.849	58.4	62.3	92.8	96.2	3.8	7.2
9–12	3	4	0.786	0.866	66.2	74.7	97.0	98.3	1.7	3.0
	5	4	0.830	0.888	55.5	65.2	91.6	96.1	3.9	8.4

Chapter 4. Validity

This chapter presents the following validity evidence for the ELPA21 summative assessment:

- Internal Structure: Evidence from the domain test dimensionality analysis.
- Appropriateness for Student Population: Evidence assessed by comparing student abilities with item difficulties on the theta metric.
- Consequences of Testing: Evidence from the correlational analysis between the screener and summative assessments, as well as student progress measured by improvement from screener to summative test performance.

The analysis results for each state and the pooled analysis combining all states are summarized in the following sections of the pooled appendix for the summative assessment:

- 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.
    - Tables S13.1–S13.13 summarize the results of the principal component analysis for each domain and 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.

The correlational analysis between the screener and summative assessments, and student progress are presented in the pooled appendix for the screener assessment:

- Section 12. Screener Assessment—Correlations Between Summative and Screener Tests
    - Table S12.1 presents the correlation between the scale scores from summative and screener tests assessed using Pearson correlations.
    - Table S12.2 presents the correlation between the performance levels from both tests assessed using Goodman and Kruskal's Gamma correlation.
- Section 13. Screener Assessment—Student Progress from Screener to Summative
    - Figures S13.1–S13.10 summarize the results of progress analysis for each domain, comprehension, and overall using a box plot, and for each grade band using a scatterplot.
    - Tables S13.1–S13.6 summarize the results of progress analysis for each domain, comprehension, and overall.

4.1 Dimensionality Analysis

The graded response model (Samejima, 1969) used for operational scoring of the ELPA21 tests 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 pooled appendix for the summative assessment. 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.

Additionally, domain intercorrelations based on the scale scores of the four domains (speaking, listening, reading, and writing) are presented in Section 6.3, Domain Intercorrelations, of this technical report.

4.2 Student Abilities Versus Item Difficulties

The appropriateness of the assessment for the student population is assessed by comparing student abilities with item difficulties in the test. When student abilities are well matched to item difficulties, the measurement errors are reduced. Therefore, it is desired that the item difficulty matches the student's ability. To examine this aspect of the test, item difficulties were plotted against student abilities for each domain. Specifically, the density plots of students' ability estimates ($\widehat{\boldsymbol{\theta}}_i$) and item location parameter estimates were plotted and compared in each domain.

The results, which are included in Section 14 of the pooled appendix for the summative assessment, show that student abilities are generally higher than the item difficulties in listening and speaking especially for lower grades, while item difficulties match with student abilities in reading and writing for all grades except reading in kindergarten where items are relatively easy.

4.3 Relationship Between Summative and Screener Assessments

Students who took the ELPA21 screener and were classified as English learners (ELs) (i.e., Proficiency Not Demonstrated, Emerging, or Progressing) would, in general, be expected to also take the ELPA21 summative assessment. The test items on the screener and summative assessments were drawn from the same item pools and assessed the same English Language Proficiency (ELP) standards adopted by the ELPA21 member states. We identified the students who completed both the screener and summative assessments and compared their performance across the two assessments.

4.3.1 Correlation Between Summative and Screener Assessments

The correlation between the scale scores from summative and screener assessments was assessed using Pearson correlations. The correlation between the performance levels from both assessments was assessed using Goodman and Kruskal's Gamma correlation (Goodman & Kruskal, 1954). The correlation results are presented in Tables S12.1 and S12.2 in the pooled appendix for screener assessment.

The Pearson correlations between screener and summative scale scores for kindergarten are the lowest, ranging from 0.42 in writing to 0.72 for the overall test; Pearson correlations for grades 1 and above exceed 0.73 for all tests. The Gamma correlations reveal a similar pattern but are generally higher in magnitude. In kindergarten, Gamma correlations range from 0.51 in writing to 0.91 for the overall test, while they exceed 0.81 in all tests for grades 1 and above.

These correlation values demonstrate predictive validity between the two ELPA21 assessments taken by the same students at different points in time.

4.3.2 Student Progress from Screener to Summative

Student progress from the time they took the screener assessment to the time they took the summative assessment was evaluated by the changes in scale scores and performance levels. Section 13 of the pooled appendix for the screener assessment summarizes the results of progress analysis. Only students who had valid scores on both the screener and summative assessments were included in each analysis. Results indicate that students' scale scores were higher on the summative assessment than on the screener assessment, and more students were in the higher performance level on the summative assessment.

Chapter 5. Reporting

A detailed introduction to the Centralized Reporting System (CRS) can be found in Chapter 6 of Part I in this technical report. The reporting mock-ups for the summative assessment of each state are presented in Section 15 of the state-specific appendix. The mock-up for score reports is not included in the pooled appendix for the summative assessment.

Chapter 6. Classical Item and Test Analysis Results

6.1 Item Analysis Results

Cambium Assessment, Inc. (CAI) employed classical item analysis procedures to ensure that items functioned as intended with respect to the underlying scales. The summary of classical item statistics for the spring 2024–2025 operational items is presented in Table 6.1. The operational data file used for this analysis was based on the online tests combining data from eight states. The summary statistics include item difficulty and discrimination statistics for each domain and grade band, presented in Table 6.1). Item-level statistics for the 2024–2025 field-test items are presented in Table A.1 in Appendix A of Part II, the Summative Assessment, of this technical report.

Table 6.1 Summary of Classical Item Analysis for Operational Items (Combined Data from Eight States)

Grade/ Grade Band	Domain	Number of Items	Sample Size		[1]Item Difficulty		Item Discrimination	
			Min	Max	Mean	SD	Mean	SD
K	Listening	20	≥41,030	≥41,650	0.77	0.15	0.61	0.14
	Reading	22	≥40,910	≥41,560	0.76	0.16	0.63	0.16
	Speaking	7	≥40,740	≥41,280	0.64	0.10	0.72	0.05
	Writing	16	≥41,050	≥41,400	0.50	0.17	0.60	0.12
1	Listening	20	≥40,920	≥41,010	0.83	0.09	0.62	0.13
	Reading	24	≥40,510	≥40,870	0.63	0.22	0.49	0.22
	Speaking	6	≥40,530	≥40,700	0.80	0.07	0.69	0.03
	Writing	17	≥40,810	≥40,840	0.70	0.11	0.75	0.08
2–3	Listening	21	≥71,070	≥71,350	0.80	0.07	0.59	0.14
	Reading	26	≥70,610	≥71,070	0.64	0.18	0.58	0.19
	Speaking	6	≥70,590	≥70,830	0.75	0.11	0.70	0.06
	Writing	12	≥70,120	≥71,070	0.57	0.21	0.76	0.09
4–5	Listening	22	≥52,640	≥52,880	0.75	0.12	0.59	0.15
	Reading	25	≥52,460	≥52,620	0.55	0.15	0.55	0.17
	Speaking	6	≥52,310	≥52,520	0.67	0.07	0.68	0.03
	Writing	8	≥52,300	≥52,670	0.68	0.14	0.78	0.04
6–8	Listening	26	≥62,420	≥63,280	0.75	0.14	0.61	0.18
	Reading	24	≥62,310	≥62,990	0.52	0.16	0.48	0.19
	Speaking	5	≥62,390	≥62,630	0.65	0.13	0.74	0.02
	Writing	5	≥62,360	≥63,020	0.52	0.09	0.77	0.10
9–12	Listening	23	≥74,030	≥74,650	0.66	0.15	0.57	0.14
	Reading	27	≥72,470	≥74,520	0.47	0.14	0.48	0.19
	Speaking	5	≥72,400	≥72,910	0.64	0.09	0.76	0.03
	Writing	5	≥72,560	≥74,460	0.46	0.06	0.71	0.17

Note. For polytomous items, item raw scores are adjusted to be on the 0-1 scale.

6.2 Differential Item Functioning Results

Differential item functioning (DIF) analysis was conducted for the field-test items in the 2024-25 test administration. The DIF analysis included only online tests. Paper tests were not included due to a small sample size. Table 6.3 provides sample sizes used for each DIF analysis group. Due to a small sample size in some ethnic subgroups, data from all states were combined for the DIF analysis. Table 6.4–Table 6.9 summarize the number of moderate (B) and large (C) DIF items by grade band and domains. The percentages of items identified as large (C) DIF items among the 201 FT items in spring 2025 range from 1% (Female vs. Male) to 5% (Asian vs. Non-Asian).

Table 6.2 DIF Sample Sizes for DIF Groups

		K	1	2–3	4–5	6–8	9–12
Gender	**Female**	270 ~ 1812	275 ~ 720	848 ~ 2260	621 ~ 6536	988 ~ 3746	1269 ~ 2729
	Male	283 ~ 1765	291 ~ 731	907 ~ 2532	706 ~ 7372	1225 ~ 4463	1557 ~ 3367
Ethnicity	**AA**	18 ~ 132	11 ~ 62	51 ~ 167	38 ~ 464	78 ~ 316	127 ~ 288
	Non-AA	570 ~ 3444	560 ~ 1393	1693 ~ 4635	1288 ~ 13453	2168 ~ 7903	2703 ~ 5794
	WH	36 ~ 337	47 ~ 145	122 ~ 383	98 ~ 1081	140 ~ 549	205 ~ 445
	Non-WH	548 ~ 3238	534 ~ 1317	1602 ~ 4410	1232 ~ 12837	2102 ~ 7688	2625 ~ 5660
	HS	456 ~ 2766	449 ~ 1156	1418 ~ 3897	1095 ~ 11411	1876 ~ 6846	2331 ~ 5025
	Non-HS	115 ~ 809	115 ~ 318	315 ~ 919	226 ~ 2506	365 ~ 1373	499 ~ 1074
	AS	40 ~ 296	39 ~ 116	104 ~ 322	71 ~ 822	106 ~ 431	133 ~ 303
	Non-AS	536 ~ 3279	523 ~ 1328	1656 ~ 4472	1266 ~ 13095	2130 ~ 7788	2697 ~ 5789
SPED	**SPED**	2 ~ 36	3 ~ 21	23 ~ 83	24 ~ 318	47 ~ 198	32 ~ 93
	Non-SPED	591 ~ 3539	571 ~ 1418	1732 ~ 4705	1311 ~ 13599	2192 ~ 8021	2798 ~ 6005

Note. AA=African American; WH=White; HS=Hispanic; AS=Asian; SPED=Special Education, IEP, or Section 504 Plan.

Table 6.3 2024–2025 Machine-Scored Field-Test Item DIF Analyses (Female vs. Male)

Grade Band	Domain	Number of Items		
		All Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	3	0	0
	Reading			
	Writing	21	0	0
1	Listening	12	1	0
	Reading	16	0	0
	Writing	15	1	0
2–3	Listening	4	1	0
	Reading	16	0	0
	Writing	5	0	0
4–5	Listening	3	0	0
	Reading	23	0	0
	Writing	10	0	1
6–8	Listening	8	0	1
	Reading	19	0	0
	Writing	7	0	0
9–12	Listening	9	1	0
	Reading	25	0	0
	Writing	5	0	0

Table 6.4 2024–2025 Machine-Scored Field-Test Item DIF Analyses (Black vs. Non-Black)

Grade Band	Domain	Number of Items		
		All Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	3	0	1
	Reading			
	Writing	21	0	0
1	Listening	12	0	1
	Reading	16	0	2
	Writing	15	0	3
2–3	Listening	4	0	0
	Reading	16	0	0
	Writing	5	0	0
4–5	Listening	3	0	0
	Reading	23	2	0
	Writing	10	0	0
6–8	Listening	8	0	0
	Reading	19	0	0
	Writing	7	0	0
9–12	Listening	9	1	0
	Reading	25	2	1
	Writing	5	2	0

Table 6.5 2024–2025 Machine-Scored Field-Test Item DIF Analyses (White vs. Non-White)

Grade Band	Domain	Number of Items		
		All Items	Moderate (B) IF Items	Large (C) DIF Items
K	Listening	3	0	0
	Reading			
	Writing	21	0	4
1	Listening	12	0	0
	Reading	16	0	2
	Writing	15	0	1
2–3	Listening	4	0	1
	Reading	16	1	0
	Writing	5	0	0
4–5	Listening	3	0	0
	Reading	23	0	0
	Writing	10	1	1
6–8	Listening	8	0	0
	Reading	19	2	0
	Writing	7	0	0
9–12	Listening	9	0	0
	Reading	25	0	0
	Writing	5	0	0

Table 6.6 2024–2025 Machine-Scored Field-Test Item DIF Analyses (Hispanic vs. Non-Hispanic)

Grade Band	Domain	Number of Items		
		All Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	3	0	0
	Reading			
	Writing	21	4	2
1	Listening	12	0	0
	Reading	16	1	0
	Writing	15	1	2
2–3	Listening	4	0	1
	Reading	16	0	1
	Writing	5	0	0
4–5	Listening	3	0	0
	Reading	23	0	0
	Writing	10	2	1
6–8	Listening	8	0	0
	Reading	19	0	0
	Writing	7	0	0
9–12	Listening	9	0	0
	Reading	25	0	0
	Writing	5	0	0

Table 6.7 2024–2025 Machine-Scored Field-Test Item DIF Analyses (Asian vs. Non-Asian)

Grade Band	Domain	Number of Items		
		All Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	3	0	0
	Reading			
	Writing	21	0	5
1	Listening	11	0	0
	Reading	16	0	0
	Writing	15	0	1
2–3	Listening	4	0	0
	Reading	16	1	1
	Writing	5	0	0
4–5	Listening	3	0	0
	Reading	23	1	0
	Writing	10	1	1
6–8	Listening	8	1	0
	Reading	19	2	0
	Writing	7	0	0
9–12	Listening	9	0	0
	Reading	25	2	0
	Writing	5	0	2

Table 6.8 2024–2025 Machine-Scored Field-Test Item DIF Analyses (SPED, IEP, or Section 504 Plan vs. Non-SPED, IEP, or Section 504 Plan)

Grade Band	Domain	Number of Items		
		All Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	3	0	0
	Reading			
	Writing	20	0	1
1	Listening	12	0	0
	Reading	16	0	0
	Writing	14	0	0
2–3	Listening	4	1	0
	Reading	16	0	0
	Writing	5	0	0
4–5	Listening	3	0	0
	Reading	23	0	0
	Writing	10	0	0
6–8	Listening	8	1	1
	Reading	19	1	2
	Writing	7	0	0
9–12	Listening	9	1	1
	Reading	25	0	2
	Writing	5	0	1

6.3 Domain Intercorrelations

Domain intercorrelations were estimated using the Pearson correlation coefficient based on the scale scores of the four domains (speaking, listening, reading, and writing). Table 6.2 shows the intercorrelation of the four domains by grade band. In kindergarten, for example, the correlations range from 0.63–0.96. The between-domain correlations are generally high, with most values larger than 0.80. The lowest correlations are observed in kindergarten between writing and three other domains, ranging from 0.63 to 0.75.

Table 6.9 Intercorrelation Between the Domain Scale Scores by Grade Band (Eight States)

Grade Band	Domain	Listening	Reading	Speaking	Writing
K	Listening	1.00			
	Reading	0.96	1.00		
	Speaking	0.83	0.81	1.00	
	Writing	0.75	0.74	0.63	1.00
G1	Listening	1.00			
	Reading	0.83	1.00		
	Speaking	0.81	0.78	1.00	
	Writing	0.83	0.95	0.78	1.00
G2–3	Listening	1.00			
	Reading	0.89	1.00		
	Speaking	0.84	0.83	1.00	
	Writing	0.87	0.97	0.84	1.00
G4–5	Listening	1.00			
	Reading	0.92	1.00		
	Speaking	0.86	0.82	1.00	
	Writing	0.94	0.96	0.87	1.00
G6–8	Listening	1.00			
	Reading	0.93	1.00		
	Speaking	0.86	0.81	1.00	
	Writing	0.94	0.92	0.88	1.00
G9–12	Listening	1.00			
	Reading	0.95	1.00		
	Speaking	0.84	0.80	1.00	
	Writing	0.95	0.93	0.85	1.00

References

Cook, K. F., Kallen, M. A., & 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. https://doi.org/10.1007/s11136-009-9464-4

Goodman, L., & Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*(268), 732–764. https://doi.org/10.2307/2281536

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 (Psychometric Monograph No. 17)*. Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pd

Appendix A

Table A.1 presents the classical item statistics of 2024–2025 field-test items. There are two major item formats: multiple-choice (MC) items and Question and Test Interoperability (QTI) items, which are either machine-scored technology-enhanced items or handscored items.

The columns under "Proportion at Each Point/Option" represent the proportion of students who scored at each point (0/1/2/3/4) on QTI items, or the proportion of students who selected each response option (i.e., A/B/C/D) on MC items. For MC items, the bolded value indicates the key of the MC item.

The *p*-value column presents the proportion of students who answered the items correctly on one-point items, or the average proportion correct on multiple-point items.

Point-biserial is the Pearson correlation between the item score and the overall scale score. Biserial (for one-point items) or polyserial (for multiple-point items) are the correlation between item score and overall scale score assuming the discrete item scores are categorized based on a continuous underlying normal distribution.

Table A.1 Classical Item Analysis of Field-Test Items (Combined Data from All States)

Grade	Domain	Item #	Max. Points	N- count	Omit N-count	Proportion at Each Point/Option					*p*-value	Point Biserial	Biserial/ Polyserial
						0	1/A	2/B	3/C	4/D			
KG	Listening	7195	1	≥14260	≥80	0.06	0.94				0.94	0.35	0.74
KG	Listening	7196	1	≥14190	≥100	0.05	0.95				0.95	0.26	0.63
KG	Listening	7198	1	≥14080	≥80	0.11	0.89				0.89	0.34	0.59
KG	Writing	7204	1	≥2890	≥20	0.26	0.74				0.74	0.21	0.30
KG	Writing	7205	1	≥2940	≥10	0.26	0.74				0.74	0.30	0.42
KG	Writing	7249	1	≥2850	≥30	0.45	0.55				0.55	0.14	0.18
KG	Writing	7250	1	≥2860	≥20	0.70	0.30				0.30	0.44	0.55
KG	Writing	7251	1	≥2860	≥20	0.52	0.48				0.48	0.12	0.14
KG	Writing	7252	1	≥2850	≥30	0.39	0.61				0.61	0.35	0.45
KG	Writing	7264	1	≥2920	≥20	0.70	0.30				0.30	0.48	0.58
KG	Writing	7288	1	≥2990	≥30	0.63	0.37				0.37	0.04	0.05
KG	Writing	7289	1	≥2990	≥30	0.67	0.33				0.33	0.16	0.20
KG	Writing	7290	1	≥2950	≥20	0.67	0.33				0.33	0.34	0.42
KG	Writing	7291	1	≥2950	≥20	0.28	0.72				0.72	0.21	0.30
KG	Writing	7292	1	≥2850	≥20	0.70	0.30				0.30	0.59	0.72
KG	Writing	7293	1	≥2810	≥20	0.62	0.38				0.38	0.46	0.56
KG	Writing	7294	1	≥2840	≥20	0.72	0.28				0.28	0.43	0.54
KG	Writing	7295	1	≥2880	≥10	0.42	0.58				0.58	0.32	0.39
KG	Writing	7296	1	≥2930	≥30	0.88	0.12				0.12	0.37	0.55
KG	Writing	7297	1	≥2880	≥10	0.77	0.23				0.23	0.28	0.37
KG	Writing	7298	1	≥2900	≥20	0.86	0.14				0.14	0.46	0.64
KG	Writing	7299	1	≥2830	≥10	0.86	0.14				0.14	0.46	0.71
KG	Writing	7300	1	≥2830	≥10	0.86	0.14				0.14	0.43	0.70
KG	Writing	7301	1	≥2860	≥20	0.88	0.12				0.12	0.46	0.73
1	Listening	7197	1	≥6040	≥10	0.08	0.92				0.92	0.35	0.74
1	Listening	7201	1	≥6090	≥20	0.05	0.95				0.95	0.26	0.69

Grade	Domain	Item #	Max. Points	N- count	Omit N-count	Proportion at Each Point/Option					*p*-value	Point Biserial	Biserial/ Polyserial
						0	1/A	2/B	3/C	4/D			
1	Listening	7302	1	≥6050	≥10	0.55	0.45				0.45	0.55	0.66
1	Listening	7303	1	≥6030	≥10	0.35	0.65				0.65	0.40	0.50
1	Listening	7304	1	≥5990	≥50	0.58	0.42				0.42	0.15	0.18
1	Listening	7305	1	≥6030	≥10	0.56	0.44				0.44	0.12	0.16
1	Listening	7306	1	≥6210	≥10	0.14	0.86				0.86	0.47	0.72
1	Listening	7351	1	≥6030	≥10	0.69	0.31				0.31	0.07	0.09
1	Listening	7352	1	≥6050	≥10	0.46	0.54				0.54	0.34	0.42
1	Listening	7355	1	≥5900	≥10	0.45	0.55				0.55	0.30	0.38
1	Listening	7361	1	≥5890	≥10	0.25	0.75				0.75	0.37	0.49
1	Listening	7365	1	≥5890	≥20	0.57	0.43				0.43	0.53	0.64
1	Reading	7058	1	≥3550	≥10	0.70	0.30				0.30	0.31	0.39
1	Reading	7093	1	≥3590	≥10	0.59	0.41				0.41	0.42	0.50
1	Reading	7094	1	≥3610	≥10	0.61	0.39				0.39	0.33	0.40
1	Reading	7095	1	≥3490	<10	0.52	0.48				0.48	0.56	0.66
1	Reading	7096	1	≥3630	<10	0.46	0.54				0.54	0.38	0.48
1	Reading	7097	1	≥3590	≥10	0.31	0.69				0.69	0.43	0.55
1	Reading	7098	1	≥3620	≥10	0.41	0.59				0.59	0.55	0.66
1	Reading	7099	1	≥3600	≥10	0.61	0.39				0.39	0.46	0.56
1	Reading	7100	1	≥3500	≥10	0.25	0.75				0.75	0.48	0.65
1	Reading	7101	1	≥3610	≥10	0.59	0.41				0.41	0.36	0.44
1	Reading	7172	1	≥3680	≥20	0.74	0.26				0.26	0.31	0.41
1	Reading	7257	1	≥3680	≥20	0.67	0.33				0.33	0.10	0.12
1	Reading	7258	1	≥3680	≥20	0.74	0.26				0.26	-0.01	-0.02
1	Reading	7259	1	≥3690	≥20	0.19	0.81				0.81	0.04	0.07
1	Reading	7307	1	≥3480	≥20	0.69	0.31				0.31	0.39	0.48
1	Reading	7308	1	≥3470	≥20	0.76	0.24				0.24	0.17	0.22
1	Writing	7059	1	≥2960	≥10	0.24	0.76				0.76	0.53	0.71
1	Writing	7060	1	≥2900	≥10	0.23	0.77				0.77	0.56	0.76

Grade	Domain	Item #	Max. Points	N- count	Omit N-count	Proportion at Each Point/Option					*p*-value	Point Biserial	Biserial/ Polyserial
						0	1/A	2/B	3/C	4/D			
1	Writing	7087	1	≥2910	<10	0.24	0.76				0.76	0.29	0.39
1	Writing	7088	1	≥2910	<10	0.54	0.46				0.46	0.34	0.41
1	Writing	7089	1	≥2920	<10	0.68	0.32				0.32	0.10	0.13
1	Writing	7090	1	≥2860	≥10	0.28	0.72				0.72	0.42	0.55
1	Writing	7091	1	≥2990	≥10	0.78	0.22				0.22	0.26	0.36
1	Writing	7092	1	≥2750	≥10	0.68	0.32				0.32	0.31	0.39
1	Writing	7171	1	≥2830	<10	0.59	0.41				0.41	0.66	0.78
1	Writing	7253	1	≥2840	<10	0.40	0.60				0.60	0.71	0.85
1	Writing	7254	1	≥2960	≥10	0.63	0.37				0.37	0.66	0.81
1	Writing	7255	1	≥2890	≥10	0.78	0.22				0.22	0.53	0.71
1	Writing	7256	1	≥3010	≥10	0.71	0.29				0.29	0.57	0.75
1	Writing	7263	1	≥2890	≥10	0.47	0.53				0.53	0.66	0.78
1	Writing	7309	1	≥2890	<10	0.48	0.52				0.52	0.65	0.77
2-3	Listening	7199	1	≥18450	≥60	0.04	0.96				0.96	0.24	0.61
2-3	Listening	7202	1	≥18540	≥60	0.04	0.96				0.96	0.29	0.76
2-3	Listening	7315	1	≥18330	≥50	0.31	0.69				0.69	0.21	0.28
2-3	Listening	7317	1	≥18530	≥50	0.31	0.69				0.69	0.31	0.41
2-3	Reading	7061	1	≥7940	<10	0.33	0.67				0.67	0.48	0.61
2-3	Reading	7106	1	≥7710	≥10	0.35	0.65				0.65	0.59	0.72
2-3	Reading	7107	1	≥7830	<10	0.27	0.73				0.73	0.52	0.69
2-3	Reading	7108	1	≥7940	<10	0.61	0.39				0.39	0.39	0.48
2-3	Reading	7109	1	≥7770	≥10	0.57	0.43				0.43	0.44	0.53
2-3	Reading	7110	1	≥7860	<10	0.18	0.82				0.82	0.54	0.79
2-3	Reading	7111	1	≥7900	≥10	0.32	0.68				0.68	0.56	0.70
2-3	Reading	7170	1	≥7730	≥10	0.25	0.75				0.75	0.60	0.81
2-3	Reading	7260	1	≥7320	≥10	0.56	0.44				0.44	0.52	0.61
2-3	Reading	7265	1	≥7270	≥50	0.67	0.33				0.33	0.29	0.37
2-3	Reading	7266	1	≥7300	≥30	0.45	0.55				0.55	0.60	0.72

Grade	Domain	Item #	Max. Points	N- count	Omit N-count	Proportion at Each Point/Option					*p*-value	Point Biserial	Biserial/ Polyserial
						0	1/A	2/B	3/C	4/D			
2-3	Reading	7267	1	≥7290	≥40	0.62	0.38				0.38	0.31	0.38
2-3	Reading	7318	1	≥7480	≥30	0.69	0.31				0.31	0.30	0.37
2-3	Reading	7319	1	≥7490	≥20	0.60	0.40				0.40	0.29	0.36
2-3	Reading	7320	1	≥7490	≥20	0.62	0.38				0.38	0.25	0.31
2-3	Reading	7321	1	≥7490	≥20	0.66	0.34				0.34	0.45	0.55
2-3	Writing	7261	1	≥14740	≥20	0.72	0.28				0.28	0.60	0.77
2-3	Writing	7310	1	≥14850	≥20	0.59	0.41				0.41	0.63	0.76
2-3	Writing	7311	1	≥14890	≥20	0.64	0.36				0.36	0.67	0.82
2-3	Writing	7312	1	≥15010	≥20	0.57	0.43				0.43	0.68	0.80
2-3	Writing	7314	1	≥14800	≥40	0.58	0.42				0.42	0.70	0.84
4-5	Listening	7340	1	≥53810	≥60	0.45	0.55				0.55	0.43	0.51
4-5	Listening	7341	1	≥53800	≥70	0.19	0.81				0.81	0.55	0.72
4-5	Listening	7342	1	≥53810	≥60	0.55	0.45				0.45	0.15	0.18
4-5	Reading	7322	1	≥11090	≥10	0.53	0.47				0.47	0.39	0.47
4-5	Reading	7323	1	≥11090	≥10	0.63	0.37				0.37	0.33	0.41
4-5	Reading	7324	1	≥11080	≥10	0.61	0.39				0.39	0.29	0.36
4-5	Reading	7325	2	≥11090	≥10	0.17	0.42	0.41			0.62	0.53	0.58
4-5	Reading	7326	1	≥11060	≥40	0.42	0.58				0.58	0.47	0.56
4-5	Reading	7327	1	≥11100	≥20	0.64	0.36				0.36	0.15	0.20
4-5	Reading	7328	1	≥11090	≥30	0.44	0.56				0.56	0.54	0.64
4-5	Reading	7329	1	≥11090	≥20	0.73	0.27				0.27	0.06	0.08
4-5	Reading	7330	1	≥11090	≥20	0.59	0.41				0.41	0.39	0.48
4-5	Reading	7334	1	≥11470	≥10	0.79	0.21				0.21	0.18	0.26
4-5	Reading	7335	1	≥11460	≥20	0.76	0.24				0.24	0.17	0.23
4-5	Reading	7336	2	≥11460	≥20	0.37	0.53	0.11			0.37	0.29	0.33
4-5	Reading	7337	1	≥11460	≥20	0.41	0.59				0.59	0.43	0.52
4-5	Reading	7338	1	≥11460	≥20	0.64	0.36				0.36	0.18	0.23
4-5	Reading	7339	1	≥11450	≥20	0.52	0.48				0.48	0.35	0.43

Grade	Domain	Item #	Max. Points	N- count	Omit N-count	Proportion at Each Point/Option					*p*-value	Point Biserial	Biserial/ Polyserial
						0	1/A	2/B	3/C	4/D			
4-5	Reading	7343	1	≥11060	<10	0.28	0.72				0.72	0.53	0.66
4-5	Reading	7344	1	≥11060	≥10	0.30	0.70				0.70	0.54	0.66
4-5	Reading	7345	1	≥11050	≥10	0.34	0.66				0.66	0.51	0.62
4-5	Reading	7346	1	≥11060	≥10	0.37	0.63				0.63	0.53	0.63
4-5	Reading	7347	2	≥11130	<10	0.13	0.46	0.41			0.64	0.53	0.58
4-5	Reading	7348	1	≥11120	≥10	0.53	0.47				0.47	0.45	0.54
4-5	Reading	7349	2	≥11120	≥10	0.22	0.55	0.22			0.50	0.34	0.38
4-5	Reading	7350	1	≥11120	≥10	0.24	0.76				0.76	0.52	0.67
4-5	Writing	7214	1	≥5960	<10	0.11	0.89				0.89	0.43	0.66
4-5	Writing	7215	1	≥5870	<10	0.26	0.74				0.74	0.46	0.58
4-5	Writing	7216	1	≥5830	<10	0.57	0.43				0.43	0.33	0.41
4-5	Writing	7217	1	≥5990	<10	0.65	0.35				0.35	0.15	0.19
4-5	Writing	7218	1	≥5990	<10	0.59	0.41				0.41	0.28	0.35
4-5	Writing	7262	1	≥5820	<10	0.60	0.40				0.40	0.49	0.62
4-5	Writing	7276	1	≥6030	<10	0.84	0.16				0.16	0.35	0.57
4-5	Writing	7286	1	≥5720	<10	0.28	0.72				0.72	0.71	0.83
4-5	Writing	7332	1	≥5950	<10	0.54	0.46				0.46	0.64	0.78
4-5	Writing	7333	1	≥5850	<10	0.34	0.66				0.66	0.67	0.77
6-8	Listening	7353	1	≥32430	≥40	0.58	0.42				0.42	0.44	0.54
6-8	Listening	7354	1	≥32280	≥180	0.72	0.28				0.28	0.55	0.73
6-8	Listening	7357	1	≥32410	≥50	0.51	0.49				0.49	0.21	0.26
6-8	Listening	7359	1	≥32420	≥50	0.57	0.43				0.43	0.18	0.23
6-8	Listening	7366	2	≥32840	≥50	0.32	0.50	0.18			0.43	0.22	0.25
6-8	Listening	7368	1	≥32830	≥50	0.37	0.63				0.63	0.54	0.66
6-8	Listening	7370	1	≥32820	≥70	0.43	0.57				0.57	0.39	0.48
6-8	Listening	7372	1	≥32840	≥50	0.67	0.33				0.33	0.21	0.27
6-8	Reading	2542	1	≥13240	≥10	0.51	0.49				0.49	0.36	0.43
6-8	Reading	7169	1	≥14100	≥10	0.37	0.63				0.63	0.55	0.65

Grade	Domain	Item #	Max. Points	N- count	Omit N-count	Proportion at Each Point/Option					*p*-value	Point Biserial	Biserial/ Polyserial
						0	1/A	2/B	3/C	4/D			
6-8	Reading	7207	1	≥13200	≥10	0.65	0.35				0.35	0.36	0.45
6-8	Reading	7208	1	≥13200	≥10	0.65	0.35				0.35	0.12	0.16
6-8	Reading	7209	1	≥13190	≥20	0.42	0.58				0.58	0.25	0.31
6-8	Reading	7210	1	≥13210	≥10	0.46	0.54				0.54	0.35	0.43
6-8	Reading	7211	1	≥13200	≥20	0.70	0.30				0.30	0.24	0.31
6-8	Reading	7238	1	≥14090	≥20	0.36	0.64				0.64	0.38	0.48
6-8	Reading	7239	1	≥14080	≥20	0.50	0.50				0.50	0.32	0.39
6-8	Reading	7240	1	≥14090	≥10	0.50	0.50				0.50	0.37	0.44
6-8	Reading	7241	1	≥14090	≥20	0.41	0.59				0.59	0.34	0.42
6-8	Reading	7242	1	≥14080	≥20	0.67	0.33				0.33	0.39	0.47
6-8	Reading	7268	1	≥13740	≥10	0.47	0.53				0.53	0.43	0.52
6-8	Reading	7269	1	≥13730	≥20	0.31	0.69				0.69	0.43	0.53
6-8	Reading	7270	1	≥13730	≥20	0.70	0.30				0.30	0.20	0.25
6-8	Reading	7271	1	≥13730	≥20	0.31	0.69				0.69	0.46	0.57
6-8	Reading	7272	1	≥13730	≥20	0.56	0.44				0.44	0.44	0.53
6-8	Reading	7273	1	≥13730	≥20	0.56	0.44				0.44	0.44	0.53
6-8	Reading	7274	1	≥13230	≥10	0.55	0.45				0.45	0.49	0.58
6-8	Writing	7225	3	≥10000	<10	0.19	0.31	0.30	0.20		0.50	0.59	0.61
6-8	Writing	7228	3	≥9940	≥10	0.12	0.21	0.28	0.38		0.64	0.67	0.70
6-8	Writing	7229	3	≥9970	<10	0.22	0.33	0.32	0.13		0.46	0.56	0.58
6-8	Writing	7230	3	≥9900	<10	0.16	0.31	0.30	0.23		0.53	0.57	0.59
6-8	Writing	7231	3	≥9800	<10	0.16	0.25	0.28	0.31		0.58	0.67	0.70
6-8	Writing	7232	3	≥9850	<10	0.18	0.26	0.28	0.27		0.55	0.66	0.69
6-8	Writing	7235	3	≥9810	<10	0.14	0.29	0.33	0.24		0.56	0.56	0.58
9-12	Listening	2926	1	≥25790	≥230	0.26	0.74				0.74	0.44	0.57
9-12	Listening	7367	2	≥25500	≥90	0.08	0.42	0.50			0.71	0.52	0.59
9-12	Listening	7369	1	≥25500	≥90	0.35	0.65				0.65	0.54	0.65
9-12	Listening	7371	2	≥25510	≥80	0.06	0.27	0.67			0.81	0.56	0.67

Grade	Domain	Item #	Max. Points	N- count	Omit N-count	Proportion at Each Point/Option					*p*-value	Point Biserial	Biserial/ Polyserial
						0	1/A	2/B	3/C	4/D			
9-12	Listening	7373	1	≥25540	≥60	0.35	0.65				0.65	0.50	0.61
9-12	Listening	7374	1	≥25280	≥180	0.55	0.45				0.45	0.54	0.64
9-12	Listening	7375	1	≥25270	≥190	0.54	0.46				0.46	0.40	0.48
9-12	Listening	7376	2	≥25320	≥150	0.12	0.41	0.48			0.68	0.57	0.64
9-12	Listening	7377	1	≥25390	≥70	0.26	0.74				0.74	0.51	0.66
9-12	Reading	7220	1	≥13010	≥10	0.48	0.52				0.52	0.41	0.49
9-12	Reading	7221	1	≥13000	≥30	0.38	0.62				0.62	0.60	0.71
9-12	Reading	7222	1	≥12980	≥50	0.71	0.29				0.29	0.42	0.55
9-12	Reading	7223	1	≥12990	≥30	0.42	0.58				0.58	0.47	0.57
9-12	Reading	7224	1	≥12990	≥30	0.32	0.68				0.68	0.31	0.39
9-12	Reading	7227	1	≥13000	≥20	0.80	0.20				0.20	0.23	0.33
9-12	Reading	7243	1	≥13000	≥20	0.36	0.64				0.64	0.31	0.40
9-12	Reading	7244	1	≥12990	≥20	0.50	0.50				0.50	0.23	0.29
9-12	Reading	7245	1	≥12990	≥20	0.67	0.33				0.33	0.32	0.40
9-12	Reading	7246	2	≥13000	≥20	0.19	0.41	0.40			0.61	0.55	0.59
9-12	Reading	7247	1	≥12990	≥20	0.65	0.35				0.35	0.52	0.65
9-12	Reading	7248	1	≥12990	≥30	0.55	0.45				0.45	0.52	0.62
9-12	Reading	7277	1	≥13720	≥10	0.35	0.65				0.65	0.30	0.38
9-12	Reading	7278	1	≥13700	≥30	0.56	0.44				0.44	0.24	0.29
9-12	Reading	7279	1	≥13700	≥30	0.26	0.74				0.74	0.53	0.69
9-12	Reading	7280	1	≥13430	≥10	0.48	0.52				0.52	0.30	0.37
9-12	Reading	7281	1	≥13410	≥20	0.36	0.64				0.64	0.55	0.66
9-12	Reading	7282	1	≥13220	≥210	0.64	0.36				0.36	0.54	0.64
9-12	Reading	7283	1	≥13320	≥110	0.58	0.42				0.42	0.27	0.33
9-12	Reading	7356	1	≥13040	≥20	0.43	0.57				0.57	0.43	0.52
9-12	Reading	7358	1	≥13040	≥20	0.36	0.64				0.64	0.55	0.65
9-12	Reading	7360	1	≥13040	≥20	0.31	0.69				0.69	0.54	0.68
9-12	Reading	7362	1	≥13030	≥20	0.54	0.46				0.46	0.40	0.48

Grade	Domain	Item #	Max. Points	N- count	Omit N-count	Proportion at Each Point/Option					*p*-value	Point Biserial	Biserial/ Polyserial
						0	1/A	2/B	3/C	4/D			
9-12	Reading	7363	1	≥13040	≥10	0.31	0.69				0.69	0.55	0.68
9-12	Reading	7364	2	≥13040	≥20	0.09	0.52	0.39			0.65	0.52	0.59
9-12	Writing	7226	3	≥15870	≥20	0.25	0.39	0.28	0.09		0.40	0.45	0.48
9-12	Writing	7233	3	≥15680	≥20	0.24	0.40	0.26	0.10		0.41	0.37	0.38
9-12	Writing	7234	3	≥15740	≥20	0.11	0.24	0.31	0.34		0.63	0.62	0.65
9-12	Writing	7236	3	≥15610	≥10	0.33	0.33	0.21	0.12		0.37	0.47	0.49
9-12	Writing	7237	3	≥15660	≥10	0.18	0.30	0.30	0.21		0.52	0.61	0.64

Note. All field-tested items are included in the table, and some may not be added to the operational bank.