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

2022–2023 Test Administration

Submitted to:
ELPA21

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

April 2024

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 2022–2023 summative testing windows for the six states discussed in this report are shown in Table 1.1.

Table 1.1 2022–2023 ELPA21 Summative Testing Windows by State

State	ELPA21 Summative
Arkansas	3/6/2023–4/14/2023
Iowa	1/30/2023–3/24/2023
Louisiana	2/13/23–3/24/23
Nebraska	2/6/23–3/24/23
Ohio	1/30/23–3/24/23
West Virginia	2/7/23–3/24/23

1.2 Test Design

The 2022–2023 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 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 2022–2023 summative assessments (see details in Table 1.5). Table S7.1 in the pooled Appendix for the summative assessment shows the 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	29	29	24	24	25	26	29	32	34	38	23	26
Reading	23	23	30	30	30	36	27	30	29	33	38	40
Speaking	11	27	9	25	9	25	8	30	7	27	7	27
Writing	18	18	20	20	14	24	13	30	8	28	8	28
Total	81	97	83	99	78	111	77	122	78	126	76	121

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

	Grade/Grade Band											
	K		**1**		**2–3**		**4–5**		**6–8**		**9–12**	
Domain	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*	*Items*	*Score Points*
Listening	28	28	22	22	23	24	24	27	30	31	21	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
Speaking	46	10	15	13	16	13	113
Writing	0	0	2	5	4	0	11
Total	46	10	17	18	20	13	124

1.3 Test Administration Manual

1.3.1 Directions for Test Administration

For the 2022–2023 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 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 (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 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 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

Business Scoring Rules for the Summative Assessment

Business rules and instructions applicable to the 2022–2023 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 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 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. 2022–2023 Summary

The 2022–2023 student participation and performance statistics for each state and the pooled analysis for the summative assessment are presented in Sections 1–5 of the pooled Appendix for the summative assessment. The figures and tables included in Sections 1–7 are listed in the following paragraphs:

- 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 (including grade, gender, and ethnicity) and by other characteristics (e.g., migrant, special education, Title I, or Section 504 Plan status). The pooled analysis includes the summary by grade, gender, and ethnicity. Subgroups vary across the states, for example, the female subgroups vary from 43.2%–48.7% while male subgroups vary from 50.9%–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, mean, 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 2022–2023 Student Participation

Table 2.1 summarizes student participation in each state. There were 211,879 students in total who participated in the 2022–2023 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		Iowa		Louisiana		Nebraska		Ohio		West Virginia		Total		
	2021–22	2022–23	2021–22	2022–23	2021–22	2022-23	2021–22	2022-23	2021–22	2022-23	2021–22	2022-23	2021–22	2022-23	Diff
K	≥4,550	≥4,380	≥4,610	≥4,830	≥3,930	≥4,030	≥3,920	≥3,890	≥10,230	≥10,580	≥230	≥230	≥27,500	≥27,960	≥450
1	≥4,250	≥4,500	≥4,100	≥4,300	≥3,880	≥4,360	≥3,680	≥3,850	≥9,380	≥10,570	≥230	≥230	≥25,550	≥27,830	≥2,270
2	≥4,260	≥3,810	≥3,640	≥3,630	≥3,380	≥3,680	≥3,190	≥3,250	≥8,530	≥8,610	≥190	≥230	≥23,210	≥23,230	≥20
3	≥3,480	≥3,580	≥2,800	≥3,060	≥2,860	≥2,900	≥2,320	≥2,580	≥6,580	≥7,270	≥190	≥190	≥18,250	≥19,600	≥1,350
4	≥3,030	≥2,970	≥2,380	≥2,660	≥2,460	≥2,640	≥1,820	≥2,140	≥5,320	≥6,070	≥130	≥160	≥15,160	≥16,660	≥1,500
5	≥2,720	≥2,700	≥2,100	≥2,210	≥2,050	≥2,200	≥1,440	≥1,610	≥4,650	≥5,110	≥130	≥130	≥13,110	≥13,980	≥860
6	≥2,610	≥2,570	≥1,890	≥2,040	≥2,080	≥1,970	≥1,180	≥1,400	≥3,720	≥4,640	≥100	≥130	≥11,590	≥12,780	≥1,180
7	≥2,620	≥2,550	≥1,780	≥1,800	≥1,830	≥2,120	≥1,140	≥1,230	≥3,610	≥3,900	≥120	≥110	≥11,120	≥11,740	≥610
8	≥2,490	≥2,650	≥1,930	≥1,840	≥1,830	≥2,000	≥1,110	≥1,260	≥3,490	≥4,050	≥130	≥140	≥10,990	≥11,980	≥980
9	≥2,840	≥2,780	≥2,200	≥2,270	≥2,610	≥2,810	≥1,570	≥1,760	≥4,780	≥5,040	≥140	≥180	≥14,170	≥14,880	≥710
10	≥2,510	≥2,770	≥2,110	≥2,180	≥1,480	≥1,950	≥1,130	≥1,530	≥3,550	≥4,300	≥120	≥170	≥10,910	≥12,930	≥2,010
11	≥2,280	≥2,380	≥1,990	≥1,930	≥1,390	≥1,270	≥1,010	≥1,070	≥3,100	≥3,290	≥120	≥130	≥9,910	≥10,090	≥170
12	≥2,060	≥1,920	≥1,380	≥1,590	≥880	≥1,020	≥830	≥920	≥2,510	≥2,570	≥90	≥120	≥7,770	≥8,180	≥400
Total	≥39,760	≥39,620	≥32,960	≥34,400	≥30,710	≥33,010	≥24,390	≥26,540	≥69,500	≥76,070	≥1,980	≥2,210	≥199,310	≥211,870	≥12,560

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

Table S1.2 in Section 1 of the pooled Appendix for the summative assessment shows student participation by subgroups. For the pooled analysis, the number of students tested decreases as the grade level increases from K–8. There were more male students (50.8%–55.6%) than female students (44.0%–48.5%) tested. In each test, most students were Hispanic or Latino (56.7%–63.1%), followed by Asian students (8.4%–15.0%) and White students (7.0%–9.2%).

The results from 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 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 2022–2023 Student Scale Score and Performance-Level Summary

Table 2.2–Table 2.4 summarize student performance in the 2022–2023 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 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, Table 2.7 shows the percentage of students in each overall proficiency category in each grade. Sections 5 and 6 of the pooled Appendix for the summative assessment 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 9 and 10 in reading and kindergarten and grades 1 and 9 in writing. For reading and writing, students across all grades have higher percentages in levels 1 and 2 than in levels 4 and 5. In the listening domain, in kindergarten and grade 7 and above, the highest percentage of students had PL 3. In the speaking domain, in kindergarten and grades 5 and above, the highest percentage of students had PL3.

For the listening domain, in grades 1–8 and 11–12, more students are in levels 4 and 5 than in levels 1 and 2. For the speaking domain, more students are in levels 4 and 5 than in levels 1 and 2 in kindergarten, grades 2–6, 8, and 11–12.

The percentage of students in each proficiency category is summarized in Table 2.7 and Section 6 of the pooled Appendix for the summative assessment. Table 2.7 shows that most students (59.1%–74.4%) are in the Progressing category in all grades. The percentage of students who are Progressing decreases from kindergarten to grade 2, and the largest increase occurs from grade 10 to grade 11. The largest drop occurs from grade 8 to grade 9 and then increases to grade 12. The percentage of students in the Emerging category decreases from kindergarten to grade 3, then increases with fluctuations (slight decreases in grades 6 and 8) until grade 9, and thereafter decreases consistently until grade 12.

*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	≥27,910	233	547.5	745	77.1	≥27,810	247	548.7	740	73.6
1	≥27,800	233	547.6	711	76.1	≥27,690	235	536.7	759	81.5
2	≥23,200	221	527.9	707	68.4	≥23,100	225	513.4	733	71.7
3	≥19,590	221	549.6	734	73.8	≥19,480	225	544.2	765	79.0
4	≥16,630	216	509.9	720	71.3	≥16,520	227	510.4	734	69.3
5	≥13,960	257	523.5	716	75.4	≥13,860	258	528.8	744	73.8
6	≥12,750	222	507.9	737	69.3	≥12,660	239	511.3	747	61.9
7	≥11,720	222	515.6	760	75.3	≥11,660	239	521.8	767	67.1
8	≥11,950	262	530.1	758	82.6	≥11,870	288	538.4	760	74.5
9	≥14,830	249	508.5	766	78.7	≥14,790	257	510.9	769	71.4
10	≥12,870	249	526.3	729	78.2	≥12,830	257	526.4	741	73.9
11	≥10,040	275	548.9	787	74.8	≥10,020	282	546.1	787	73.9
12	≥8,140	262	550.9	740	70.8	≥8,120	265	547.7	752	71.5

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

*Results include all records with valid scale scores. No special filter was used to exclude invalidated cases. 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	≥27,730	285	559.4	744	91.3	≥27,770	302	532.0	718	76.1
1	≥27,640	263	559.8	736	77.5	≥27,680	238	528.2	741	87.5
2	≥23,090	251	535.2	732	75.3	≥23,100	231	507.7	734	77.8
3	≥19,4804	251	558.0	751	80.4	≥19,480	231	539.6	764	83.0
4	≥16,560	235	532.8	729	75.7	≥16,540	222	503.0	718	74.9
5	≥13,880	250	540.7	737	79.3	≥13,860	254	520.3	740	78.2
6	≥12,690	260	532.0	748	77.8	≥12,670	235	501.5	731	75.3
7	≥11,670	260	534.5	732	82.3	≥11,660	235	511.4	768	80.5
8	≥11,880	288	543.5	759	87.3	≥11,870	283	525.2	766	86.8
9	≥14,7400	300	522.9	720	82.0	≥14,750	261	498.0	771	85.9
10	≥12,800	300	540.4	721	77.1	≥12,810	261	517.0	721	81.0
11	≥9,980	340	560.7	751	71.1	≥9,980	327	539.3	787	74.6
12	≥8,080	305	563.7	725	69.0	≥8,100	269	541.3	729	69.5

*Scores from domain tests marked as Exemption or Not Attempted are excluded.
*Results include all records with valid scale scores. No special filter was used to exclude invalidated cases. 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	≥27,940	3361	5474.8	6776	534.6	≥27,960	3160	5468.5	7023	592.1
1	≥27,810	3387	5451.5	6698	545.0	≥27,830	2967	5441.5	7032	621.6
2	≥23,220	3264	5297.8	6801	510.8	≥23,230	2934	5262.0	6905	564.9
3	≥19,600	3264	5487.5	6685	561.2	≥19,600	2934	5480.4	7150	613.9
4	≥16,650	3273	5223.8	6817	520.5	≥16,660	2877	5214.5	6869	564.4
5	≥13,970	3462	5346.0	6817	559.9	≥13,980	3134	5331.8	6922	597.1
6	≥12,770	3323	5209.0	6967	477.1	≥12,780	2993	5205.5	7008	549.1
7	≥11,740	3323	5277.2	6967	520.2	≥11,740	2993	5268.9	7103	590.9
8	≥11,970	3515	5397.5	6967	582.9	≥11,980	3352	5377.3	7150	644.2
9	≥14,870	3470	5223.1	7171	545.1	≥14,880	3220	5178.0	7050	616.9
10	≥12,910	3470	5343.5	7171	568.4	≥12,930	3220	5319.7	6859	598.5
11	≥10,080	3470	5498.9	7171	567.8	≥10,090	3479	5491.6	7313	563.9
12	≥8,170	3555	5511.8	7171	549.4	≥8,180	3282	5508.4	6935	532.5

*Results include all records with valid scale scores. No special filter was used to exclude invalidated cases. 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	≥27,910	15.7	14.7	48.4	10.1	11.1	≥27,810	16.3	17.5	36.4	13.7	16.2
1	≥27,800	9.8	6.4	28.7	24.7	30.3	≥27,690	26.8	17.3	27.8	12.6	15.4
2	≥23,200	6.1	4.9	25.0	31.5	32.5	≥23,100	24.3	16.4	29.1	15.1	15.1
3	≥19,590	5.9	5.1	25.3	34.9	28.7	≥19,480	29.9	15.6	30.2	14.6	9.7
4	≥16,630	8.6	6.3	20.2	40.8	24.1	≥16,520	23.2	14.2	31.9	18.7	12.1
5	≥13,960	10.6	7.9	12.7	45.5	23.4	≥13,860	23.0	15.3	38.0	15.5	8.2
6	≥12,750	10.1	7.4	24.0	35.9	22.6	≥12,660	23.4	17.9	38.4	13.1	7.3
7	≥11,720	15.8	13.0	36.4	20.6	14.2	≥11,660	31.7	24.6	33.2	7.0	3.5
8	≥11,950	15.8	11.4	33.4	22.8	16.6	≥11,870	29.9	22.5	38.6	5.9	3.1
9	≥14,830	28.8	13.1	32.9	15.8	9.5	≥14,790	42.8	21.1	30.6	3.7	1.9
10	≥12,870	21.5	13.1	32.4	18.4	14.6	≥12,830	35.5	20.7	34.6	5.4	3.8
11	≥10,040	12.5	11.5	32.7	21.2	22.0	≥10,020	25.6	20.3	39.4	8.5	6.1
12	≥8,140	9.7	11.9	35.5	21.5	21.3	≥8,120	23.6	21.9	40.4	8.3	5.8
Total	≥211,440	12.63	9.34	30.24	26.15	21.63	≥210,460	26.61	18.22	33.49	11.79	9.90

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

*Results include all records with valid scale scores. No special filter was used to exclude invalidated cases. 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	≥27,730	20.0	12.6	28.1	13.8	25.5	≥27,770	37.9	29.8	23.5	3.4	5.5
1	≥27,640	27.4	26.3	9.9	14.9	21.4	≥27,680	35.7	20.5	26.0	7.5	10.2
2	≥23,090	21.7	16.9	16.1	21.1	24.3	≥23,100	24.2	15.7	29.7	15.5	15.0
3	≥19,480	17.8	11.9	17.9	27.3	25.0	≥19,480	28.3	16.1	30.7	15.2	9.7
4	≥16,560	14.9	10.7	20.1	29.8	24.6	≥16,540	19.5	12.8	49.1	12.2	6.5
5	≥13,880	17.6	12.8	29.4	23.5	16.5	≥13,860	16.3	10.3	59.8	9.2	4.4
6	≥12,690	17.6	11.6	29.9	21.6	19.4	≥12,670	16.9	11.1	52.4	11.7	7.9
7	≥11,670	20.5	14.7	33.2	16.9	14.7	≥11,660	26.7	19.1	42.3	7.3	4.7
8	≥11,880	19.7	12.5	34.1	16.6	17.1	≥11,870	26.7	18.4	43.4	6.8	4.7
9	≥14,740	29.9	17.0	33.3	10.5	9.4	≥14,750	38.5	20.0	35.6	3.7	2.2
10	≥12,800	22.2	16.9	34.2	13.3	13.3	≥12,810	31.6	19.9	39.0	5.8	3.8
11	≥9,980	13.8	14.1	36.5	17.2	18.5	≥9,980	21.2	20.3	43.6	8.9	6.0
12	≥8,080	11.2	15.1	36.5	18.2	19.0	≥8,100	19.1	22.2	44.8	8.4	5.6
Total	≥210,270	20.46	15.46	25.09	18.77	20.22	≥210,320	27.97	18.74	37.06	8.97	7.30

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

*Results include all records with valid scale scores. No special filter was used to exclude invalidated cases. 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	≥27,960	21.6	72.1	6.3
1	≥27,830	15.0	70.6	14.4
2	≥23,230	10.8	66.4	22.8
3	≥19,600	10.8	70.2	19.0
4	≥16,660	13.8	70.0	16.2
5	≥13,980	16.6	72.6	10.9
6	≥12,780	16.2	71.5	12.3
7	≥11,740	24.6	69.0	6.3
8	≥11,980	23.8	70.3	5.9
9	≥14,880	38.3	59.1	2.6
10	≥12,930	31.1	63.8	5.1
11	≥10,090	20.6	70.8	8.6
12	≥8,180	17.6	74.4	8.0
Total	≥211,870	19.1	69.3	11.6

2.3 2022–2023 Testing Time for Online Summative Tests

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

Chapter 3. Reliability

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

- Cronbach's alpha;
- marginal standard error of measurement (MSEM);
- marginal reliability;
- conditional standard error of measurement (CSEM);
- classification accuracy (CA) and classification consistency (CC); and
- inter-rater analysis.

The methods used in the computation of test reliability are described in Part I, 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 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 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 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 or grade band.

3.1 Internal Consistency

Due to small examinee count (see Section 1 of the pooled Appendix for the summative assessment), scores earned by students who took braille and paper-pencil tests were excluded from the analysis. Table 3.1 shows the values of Cronbach's alpha for the pooled sample (across states) based on the items in each domain test, arranged by grade level. Values range from 0.81 to 0.96. 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 pooled Appendix for the summative assessment.

Table 3.1 Cronbach's Alpha by Domain and Grade

Grade	Listening	Reading	Speaking	Writing	Overall
K	.86	.81	.91	.89	.94
1	.86	.84	.84	.93	.95
2	.84	.84	.84	.87	.94
3	.86	.86	.86	.88	.95
4	.86	.85	.86	.90	.95
5	.87	.87	.88	.90	.95
6	.87	.82	.88	.91	.94
7	.88	.84	.89	.91	.95
8	.90	.87	.90	.92	.96
9	.87	.88	.93	.91	.96
10	.87	.89	.91	.90	.96
11	.86	.90	.90	.88	.95
12	.84	.89	.89	.85	.94

3.2 Marginal Standard Error of Measurement

Another way to examine score reliability is with the 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 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 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 of this technical report for more information. The results show that the listening tests for grades 1–3 have the lowest reliabilities, followed by the speaking tests. The reliabilities for the speaking domain from grades 4–12 are lower than the other domains. All the reliability indexes are above 0.8, except for the listening test in grade 1 and the comprehension test in grades K–3. 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	≥27,670	.86	.84	.91	.89	.80	.83
1	≥27,580	.78	.90	.82	.90	.72	.85
2	≥23,020	.81	.91	.85	.92	.77	.88
3	≥19,410	.81	.91	.86	.92	.79	.89
4	≥16,470	.87	.91	.86	.92	.83	.89
5	≥13,820	.87	.91	.87	.92	.84	.90
6	≥12,620	.89	.89	.88	.91	.84	.89
7	≥11,600	.90	.90	.89	.92	.86	.90
8	≥11,800	.91	.91	.90	.93	.87	.91
9	≥14,660	.92	.93	.92	.93	.90	.91
10	≥12,730	.92	.93	.90	.92	.90	.91
11	≥9,930	.90	.92	.89	.91	.89	.89
12	≥8,050	.89	.92	.88	.89	.88	.88

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

3.4 Classification Accuracy and Consistency

Table 3.3 shows the overall CA and CC in each domain. The 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 3.4 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 0.78. In listening and speaking, both indexes for cut score 3 and/or cut score 4 are relatively low in all grades, which indicates a lack of difficult items.

The CA and CC results for overall proficiency categories are summarized in Table 3.6 and Figure S11.3 in the pooled Appendix for the summative assessment. All CAs and CCs are above

0.84 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 3.4 Classification Accuracy for Each Cut Score by Grade and Domain**

Grade	Listening				Reading				Speaking				Writing			
	Cut 1	Cut 2	Cut 3	Cut 4	Cut 1	Cut 2	Cut 3	Cut 4	Cut 1	Cut 2	Cut 3	Cut 4	Cut 1	Cut 2	Cut 3	Cut 4
K	.95	.92	.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 3.5 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	.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 2022–2023 summative assessment, consistency of handscoring was evaluated for a total of 72 items (11 items in kindergarten, 9 items in grade 1, and 13 items in each of the other four grade bands). Handscored items on paper-pencil and braille forms were not included in the results due to the small sample size.

Table 3.7 contains the summary of kappa coefficients for each summative assessment in the pooled analysis. The description about kappa coefficients can be found in Chapter 4.5 of Part I of this technical report. The table shows that 55.2%–93.8% 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 0.641 to 0.909. In 2021–2022, the weighted kappa coefficients ranged from 0.649 to 0.925. The inter-rater consistencies are also assessed by item and are summarized in Section 12 of the pooled Appendix for the summative assessment. 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. There are two speaking items with exact agreement rate lower than 60%: one item in grade band 4–5 (58.1%) and another in grade band 1 (55.2%), which may be due to the higher score points (e.g., score point=5).

Table 3.7 Summary of Kappa Coefficients by Grade Band

Grade/Grade Band	Number of Items	Weighted Kappa		% Exact Agreement		% within 1 Agreement		% Not within 1 Agreement	
		Min	Max	Min	Max	Min	Max	Min	Max
K	11	.734	.859	66.0	92.8	96.4	99.7	0.3	3.6
1	9	.641	.873	55.2	93.8	96.0	99.2	0.8	4.0
2–3	13	.715	.882	61.7	90.1	97.0	99.3	0.7	3.0
4–5	13	.679	.892	58.1	86.5	94.5	99.4	0.6	5.5
6–8	13	.776	.909	62.7	87.8	97.2	99.3	0.7	2.8
9–12	13	.749	.905	61.8	81.1	95.2	99.1	0.9	4.8

Chapter 4. Validity

In this chapter, validity for the ELPA21 summative assessment is measured by examining the internal structure of the items, evidence based on consequences of testing, and the evidence related to fairness, which are mainly the third and fifth source of evidence for validity mentioned in Part I. 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 item difficulties on the theta metric. Evidence based on consequences of testing is assessed by measuring correlations between screener and summative and student progress from screener to summative. Fairness is assessed using a differential item functioning (DIF) procedure.

The analysis results for each state and the pooled analysis 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.

- 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 analysis results for each state and the pooled analysis are summarized in the following sections of 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 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 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 6.2 in this 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 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' ability estimates ($\hat{\boldsymbol{\theta}}_\iota$) 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 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 item difficulties match student abilities well.

4.3 Relationship between Summative and Screener Tests

Students who took the ELPA21 screener and were classified as English learners (ELs) (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 assess 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 Tests

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

These correlations show predictive validity between the two ELPA21 tests because they were given to the same students at different times.

4.3.2 Student Progress from Screener to Summative

Student progress from the time they took screener assessments to the time they took summative assessments 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 of the analyses.

4.4 Summary of Classical Item Difficulty and Item Discrimination

This section contains the summary of classical statistics for the spring 2022–2023 operational forms. The operational data file used for this analysis was the 100% (all schools) student data file. Cambium Assessment, Inc. (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 summary statistics for each domain and grade band (see details in Table 4.1). Each state’s summary is presented in state’s Appendix B4.1–B4.6.

Table 4.1 Operational Summary of Classical Item Difficulty and Item Discrimination Indices by Grade Band (Six States Combined)

Grade Band	Domain	N-Count	Item Difficulty		Item Discrimination	
			Mean	SD	Mean	SD
K	Listening	≥27,570	0.80	0.37	0.52	0.11
	Speaking	≥27,250	0.68	0.86	0.68	0.06
	Reading	≥27,430	0.79	0.37	0.50	0.11
	Writing	≥27,420	0.52	0.46	0.57	0.11
1	Listening	≥27,630	0.87	0.32	0.52	0.09
	Speaking	≥27,410	0.80	0.79	0.61	0.04
	Reading	≥27,440	0.64	0.43	0.45	0.16
	Writing	≥27,530	0.72	0.43	0.69	0.10
2–3	Listening	≥42,570	0.85	0.35	0.52	0.11
	Speaking	≥42,230	0.76	0.82	0.64	0.03
	Reading	≥42,290	0.67	0.48	0.52	0.15
	Writing	≥42,290	0.62	0.62	0.71	0.11
4–5	Listening	≥30,460	0.79	0.40	0.54	0.12
	Speaking	≥30,200	0.71	1.00	0.63	0.03
	Reading	≥30,280	0.58	0.50	0.50	0.15
	Writing	≥30,220	0.71	0.75	0.72	0.06
6–8	Listening	≥36,240	0.81	0.38	0.60	0.15
	Speaking	≥35,750	0.67	1.10	0.71	0.03
	Reading	≥36,060	0.55	0.50	0.45	0.17
	Writing	≥35,870	0.62	1.05	0.75	0.09
9–12	Listening	≥45,680	0.62	1.05	0.75	0.09
	Speaking	≥44,700	0.67	1.25	0.75	0.03
	Reading	≥45,550	0.50	0.49	0.48	0.17
	Writing	≥44,930	0.57	1.04	0.73	0.11

Note. These are the raw score mean regardless of points possible.

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 assessments of each state appear in Section 15 of the state's Appendix. It is noted that the mock-up for score reports is not included in the pooled Appendix for the pooled analysis.

Chapter 6. Classical Item and Test Analysis Results

6.1 Item Analysis Results

Cambium Assessment, Inc. (CAI) employs classical item analysis procedures to ensure that items function as intended with respect to the underlying scales. The operational summary statistics are based on Classical Test Theory (CTT) and include information such as the item difficulty and the discrimination summary statistics for each domain and grade band (see details in Table 4.1).

Item-level statistics for the 2022–2023 field-test items are presented in Tables A.1–A.6 by grade band in the Appendix A in Part II summative report. In Tables A.1–A.6, with the exception of a few high *p*-values and low item-total correlation values, all items fell well within the preset level of acceptance, both in terms of the *p*-value and point-biserial.

6.2 Domain Intercorrelations

Domain intercorrelations based on the scale scores of the four domains (speaking, listening, reading, and writing) were calculated using Pearson correlations to investigate the answers to these questions. Table 6.1 shows the intercorrelation of the four domains by grade band.

In Table 6.1, correlations between domains in terms of scale scores are presented for each grade band. In kindergarten (KG), for example, the correlations range from 0.59–0.95; for listening, the correlations are between 0.72–0.95 with other domains, and speaking shows lower correlations with other domains.

Table 6.1 Intercorrelation between the Domain Scale Scores by Grade Band (Six States Only)

Grade Level	Domain	Listening	Reading	Speaking	Writing
KG	Listening	1			
	Reading	0.95	1		
	Speaking	0.80	0.78	1	
	Writing	0.72	0.71	0.59	1
G1	Listening	1			
	Reading	0.80	1		
	Speaking	0.79	0.72	1	
	Writing	0.80	0.94	0.74	1
G2–3	Listening	1			
	Reading	0.86	1		
	Speaking	0.82	0.78	1	
	Writing	0.86	0.96	0.81	1
G4–5	Listening	1			
	Reading	0.89	1		
	Speaking	0.82	0.77	1	
	Writing	0.91	0.93	0.83	1
G6–8	Listening	1			
	Reading	0.91	1		
	Speaking	0.83	0.77	1	
	Writing	0.92	0.89	0.85	1
G9–12	Listening	1			
	Reading	0.94	1		
	Speaking	0.83	0.77	1	
	Writing	0.94	0.90	0.83	1

6.3 Differential Item Functioning (DIF) Results

DIF analysis only included online tests. Paper tests were not included due to low sample size. Table 6.2 provides sample sizes used for the DIF analysis groups. Due to a small sample size in some ethnic subgroups, all seven states' data were combined for the DIF analysis. Table 6.3–Table 6.8 provide a summary of the number of moderate (B) and large (C) DIF items by grade band and domains based on the combined seven states' data. Large DIF items were found for kindergarten, grade 1, grades 4–5, grades 6–8, and grades 9–12 listening; kindergarten and grades 2–3 writing; and kindergarten, grade 1, grades 2–3, and grades 4–5 reading. The special education (SPED)/Individualized Education Program (IEP)/Section 504 Plan group had the highest number of DIF items, followed by the Asian, African, female, Hispanic, and White categories. Results from a sample size less than 200 needed to be interpreted with caution.

Table 6.2 DIF Sample Sizes for DIF Groups

		K	1	2–3	4–5	6–8	9–12	Overall
Gender	Female	≥1,060	≥810	≥1,450	≥780	≥1,820	≥2,5570	≥8,490
	Male	≥1,120	≥850	≥1,590	≥920	≥2,270	≥3,220	≥9,980
African American vs. Non-African American	African American	≥230	≥170	≥320	≥180	≥430	≥650	≥2,000
	Non-African American	≥1,970	≥1,500	≥2,730	≥1,540	≥3,670	≥5,150	≥16,580
White vs. Non-White	White	≥190	≥140	≥250	≥130	≥300	≥400	≥1,430
	Non-White	≥2,010	≥1,530	≥2,810	≥1,590	≥3,800	≥5,400	≥17,150
Hispanic vs. Non-Hispanic	Hispanic	≥1,260	≥970	≥1,800	≥1,020	≥2,480	≥3,660	≥11,210
	Non-Hispanic	≥940	≥700	≥1,260	≥690	≥1,620	≥2,140	≥7,370
Asian vs. Non-Asian	Asian	≥330	≥240	≥370	≥180	≥350	≥480	≥1,960
	Non-Asian	≥1,870	1,430	≥2,680	≥1,540	≥3,750	≥5,320	≥16,620
SPED, IEP, or Section 504 Plan vs. Non-SPED, IEP, or Section 504 Plan	SPED, IEP, or Section 504 Plan	≥120	≥120	≥340	≥290	≥740	≥750	≥2,390
	Non-SPED, IEP, or Section 504 Plan	≥1,860	≥1,390	≥2,510	≥1,340	≥3,160	≥4,830	≥15,100
Overall		≥2,100	≥1,670	≥3,060	≥1,720	≥4,110	≥5,800	≥18,480

Note. DIF results with N < 200 should be interpreted with caution.

Table 6.3 2022–2023 Machine-Scored Field-Test Results of DIF Analyses (Female vs. Male)

Grade Band	Domain	Number of Items			
		All Items	DIF Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	27	2	1	1
	Reading	22	0	0	0
	Writing	20	0	0	0
1	Listening	24	1	1	0
	Reading	20	0	0	0
	Writing	27	1	1	0
2–3	Listening	21	0	0	0
	Reading	15	1	1	0
	Writing	12	0	0	0
4–5	Listening	22	2	1	1
	Reading	27	2	1	1
	Writing	11	1	1	0
6–8	Listening	15	3	1	2
	Reading	28	2	2	0
	Writing	10	0	0	0
9–12	Listening	21	1	0	1
	Reading	19	0	0	0
	Writing	4	0	0	0

Table 6.4 2022–2023 Machine-Scored Field-Test Results of DIF Analyses (Black vs. Non-Black)

Grade Band	Domain	Number of Items			
		All Items	DIF Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	27	0	0	0
	Reading	22	1	1	0
	Writing	20	3	3	0
1	Listening	24	0	0	0
	Reading	20	2	1	1
	Writing	27	0	0	0
2–3	Listening	21	1	1	0
	Reading	15	1	0	1
	Writing	12	0	0	0
4–5	Listening	22	1	1	0
	Reading	27	1	1	0
	Writing	11	0	0	0
6–8	Listening	15	3	2	1
	Reading	28	0	0	0
	Writing	10	1	1	0
9–12	Listening	21	1	0	1
	Reading	19	1	1	0
	Writing	4	0	0	0

Table 6.5 2022–2023 Machine-Scored Field-Test Results of DIF Analyses (White vs. Non-White)

Grade Band	Domain	Number of Items			
		All Items	DIF Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	27	1	1	0
	Reading	22	1	0	0
	Writing	20	0	0	0
1	Listening	24	1	1	0
	Reading	20	0	0	0
	Writing	27	2	2	0
2–3	Listening	21	1	1	0
	Reading	15	0	0	0
	Writing	12	0	0	0
4–5	Listening	22	2	2	0
	Reading	27	2	1	1
	Writing	11	0	0	0
6–8	Listening	15	1	0	1
	Reading	28	1	1	0
	Writing	10	0	0	0
9–12	Listening	21	1	1	0
	Reading	19	0	0	0
	Writing	4	0	0	0

Table 6.6 2022–2023 Machine-Scored Field-Test Results of DIF Analyses (Hispanic vs. Non-Hispanic)

Grade Band	Domain	Number of Items			
		All Items	DIF Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	27	2	2	0
	Reading	22	0	0	0
	Writing	20	1	1	0
1	Listening	24	1	1	0
	Reading	20	0	0	0
	Writing	27	2	2	0
2–3	Listening	21	0	0	0
	Reading	15	0	0	0
	Writing	12	1	1	0
4–5	Listening	22	1	1	0
	Reading	27	2	2	0
	Writing	11	0	0	0
6–8	Listening	15	3	3	0
	Reading	28	0	0	0
	Writing	10	0	0	0
9–12	Listening	21	0	0	0
	Reading	19	0	0	0
	Writing	4	0	0	0

Table 6.7 2022–2023 Machine-Scored Field-Test Results of DIF Analyses (Asian vs. Non-Asian)

Grade Band	Domain	Number of Items			
		All Items	DIF Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	27	1	1	0
	Reading	22	0	0	0
	Writing	20	2	2	0
1	Listening	24	1	1	0
	Reading	20	3	2	1
	Writing	27	1	1	0
2–3	Listening	21	1	1	0
	Reading	15	2	2	0
	Writing	12	2	1	1
4–5	Listening	22	3	1	2
	Reading	27	2	1	1
	Writing	11	2	2	0
6–8	Listening	15	3	2	1
	Reading	28	0	0	0
	Writing	10	0	0	0
9–12	Listening	21	3	2	1
	Reading	19	0	0	0
	Writing	4	0	0	0

Table 6.8 2022–2023 Machine-Scored Field-Test Results of DIF Analyses (SPED, IEP, or Section 504 Plan vs. Non-SPED, IEP, or Section 504 Plan)

Grade Band	Domain	Number of Items			
		All Items	DIF Items	Moderate (B) DIF Items	Large (C) DIF Items
K	Listening	27	3	1	2
	Reading	22	3	2	1
	Writing	20	1	0	1
1	Listening	24	4	2	2
	Reading	20	1	1	0
	Writing	27	0	0	0
2–3	Listening	21	4	4	0
	Reading	15	1	1	0
	Writing	12	0	0	0
4–5	Listening	22	6	3	3
	Reading	27	6	5	1
	Writing	11	1	1	0
6–8	Listening	15	4	2	2
	Reading	28	2	2	0
	Writing	10	0	0	0
9–12	Listening	21	3	1	2
	Reading	19	0	0	0
	Writing	4	0	0	0

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.

Appendix A

Table A.1 to Table A.6 present the classical item statistics of 2023-2024 field-test items in each grade band. There are two major item formats: multiple-choice (MC) items and QTI (Question and Test Interoperability) items which are either machine-scored technology-enhanced items or hand-scored 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 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 Field test items' Classical Test Theory (CTT) summary statistics: Grade K (Combined Seven States' Data)

Domain	Item #	Item Format	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			
Listening	6598	QTI	1	≥1650	≥10	0.27	0.73				0.73	0.39	0.51
	6599	QTI	1	≥1620	≥10	0.34	0.66				0.66	0.44	0.55
	6600	QTI	1	≥1620	≥10	0.18	0.82				0.82	0.40	0.56
	6601	QTI	1	≥1560	≥10	0.09	0.91				0.91	0.38	0.63
	6602	QTI	1	≥1590	≥10	0.09	0.91				0.91	0.37	0.62
	6603	QTI	1	≥1620	<10	0.04	0.96				0.96	0.26	0.57
	6604	QTI	1	≥1680	<10	0.28	0.72				0.72	0.45	0.57
	6605	QTI	1	≥1680	≥10	0.04	0.96				0.96	0.23	0.52
	6606	QTI	1	≥1590	<10	0.11	0.89				0.89	0.39	0.62
	6607	QTI	1	≥1600	<10	0.09	0.91				0.91	0.33	0.58
	6608	QTI	1	≥1650	≥10	0.45	0.55				0.55	0.20	0.25
	6609	QTI	1	≥1620	≥10	0.17	0.83				0.83	0.36	0.52
	6842	QTI	1	≥1570	≥10	0.14	0.86				0.86	0.44	0.65
	6843	QTI	1	≥1570	≥10	0.50	0.50				0.50	0.15	0.19
	6844	QTI	1	≥1570	<10	0.02	0.98				0.98	0.18	0.50
	6845	QTI	1	≥1710	≥10	0.05	0.95				0.95	0.38	0.80
	6846	QTI	1	≥1710	≥10	0.21	0.79				0.79	0.29	0.40
	6847	MC	1	≥1700	≥20		0.25	0.10	**0.66**		0.66	0.37	0.47
	6610	MC	1	≥1610	≥10		0.29	**0.54**	0.18		0.54	0.22	0.28
	6611	QTI	1	≥1620	≥10	0.33	0.67				0.67	0.52	0.65

Domain	Item #	Item Format	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			
	6612	MC	1	≥1620	≥10		0.39	**0.43**	0.18		0.43	0.25	0.31
	6613	QTI	1	≥1570	≥10	0.39	0.61				0.61	0.30	0.38
	6614	MC	1	≥1560	≥10		**0.55**	0.23	0.21		0.55	0.26	0.32
	6615	MC	1	≥1570	≥10		0.24	0.16	**0.60**		0.60	0.30	0.38
	6833	QTI	1	≥1660	≥10	0.57	0.43				0.43	0.36	0.43
	6834	QTI	1	≥1660	≥10	0.26	0.74				0.74	0.45	0.57
	6835	MC	1	≥1650	≥10		0.27	0.26	**0.46**		0.46	0.25	0.31
Reading	6904	QTI	1	≥3510	≥20	0.32	0.68				0.68	0.37	0.47
	6905	MC	1	≥3490	≥40		0.11	**0.83**	0.06		0.83	0.49	0.70
	6906	MC	1	≥3490	≥40		0.26	0.18	**0.56**		0.56	0.27	0.34
	6907	MC	1	≥3510	≥30		0.15	0.12	**0.73**		0.73	0.36	0.46
	6908	MC	1	≥3500	≥40		**0.63**	0.17	0.20		0.63	0.37	0.46
	6909	MC	1	≥3510	≥30		0.38	0.19	**0.43**		0.43	0.11	**0.14**
	6901	QTI	1	≥3510	≥30	0.24	0.76				0.76	0.43	0.57
	6902	MC	1	≥3500	≥50		**0.54**	0.13	0.33		0.54	0.25	0.31
	6903	MC	1	≥3500	≥50		0.15	0.30	**0.55**		0.55	0.26	0.32
	6616	QTI	1	≥3320	≥20	0.09	0.91				0.91	0.41	0.71
	6617	QTI	1	≥3270	≥10	0.06	0.94				0.94	0.36	0.68
	6618	QTI	1	≥3380	≥20	0.05	0.95				0.95	0.26	0.55
	6619	QTI	1	≥3480	≥40	0.08	0.92				0.92	0.31	0.55
	6620	QTI	1	≥3480	≥40	0.12	0.88				0.88	0.36	0.57
	6621	QTI	1	≥3480	≥40	0.04	0.96				0.96	0.21	0.47
	6622	QTI	1	≥3480	≥30	0.03	0.97				0.97	0.15	0.35
	6623	MC	1	≥3470	≥40		**0.74**	0.07	0.19		0.74	0.24	0.33
	6624	QTI	1	≥3450	≥30	0.06	0.94				0.94	0.25	0.51
	6625	QTI	1	≥3460	≥30	0.04	0.96				0.96	0.23	0.50
	6626	QTI	1	≥3460	≥20	0.02	0.98				0.98	0.16	0.43
	6627	QTI	1	≥3450	≥30	0.04	0.96				0.96	0.23	0.52
	6628	MC	1	≥3460	≥30		0.20	0.11	**0.69**		0.69	0.37	0.47
Writing	6929	QTI	1	≥1590	≥20	0.13	0.87				0.87	0.28	0.45
	6930	QTI	1	≥1590	≥10	0.23	0.77				0.77	0.27	0.38
	6932	QTI	1	≥1590	≥10	0.36	0.64				0.64	0.48	0.60
	6933	QTI	1	≥1590	≥10	0.44	0.56				0.56	0.35	0.44
	7025	QTI	1	≥1520	≥10	0.35	0.65				0.65	0.27	0.35
	7026	QTI	1	≥1520	≥10	0.43	0.57				0.57	0.41	0.51
	6629	QTI	1	≥1620	≥10	0.24	0.76				0.76	0.17	**0.23**
	6630	QTI	1	≥1630	≥10	0.11	0.89				0.89	0.28	0.47
	6631	QTI	1	≥1590	≥10	0.25	0.75				0.75	0.48	0.64
	6632	QTI	1	≥1630	≥10	0.45	0.55				0.55	0.40	0.49

Domain	Item #	Item Format	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			
	6633	QTI	1	≥1610	≥10	0.29	0.71				0.71	0.10	**0.14**
	6634	QTI	1	≥1650	≥20	0.34	0.66				0.66	0.46	0.58
	6635	QTI	1	≥1580	≥10	0.67	0.33				0.33	0.34	0.43
	6636	QTI	1	≥1630	≥10	0.13	0.87				0.87	0.28	0.45
	6637	QTI	1	≥1600	≥10	0.50	0.50				0.50	0.42	0.51
	6638	QTI	1	≥1610	≥10	0.42	0.58				0.58	0.41	0.51
	6639	QTI	1	≥1660	≥10	0.36	0.64				0.64	0.50	0.62
	6640	QTI	1	≥1560	≥10	0.42	0.58				0.58	0.51	0.63
	6641	QTI	1	≥1660	≥10	0.35	0.65				0.65	0.49	0.61
	6642	QTI	1	≥1600	≥10	0.77	0.23				0.23	0.31	0.42

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

Table A.2 Field test items' Classical Test Theory (CTT) summary statistics: Grade 1 (Combined Seven States' Data)

Domain	Item #	Item Format	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			
Listening	6865	QTI	1	≥2170	<10	0.35	0.65				0.65	0.34	0.43
	6866	QTI	1	≥2230	<10	0.12	0.88				0.88	0.32	0.49
	6867	MC	1	≥2220	<10		0.11	**0.69**	0.19		0.69	0.32	0.41
	6868	QTI	1	≥2110	≥10	0.26	0.74				0.74	0.27	0.35
	6663	QTI	1	≥2210	<10	0.08	0.92				0.92	0.29	0.48
	6664	QTI	1	≥2210	<10	0.24	0.76				0.76	0.36	0.47
	6665	MC	1	≥2210	<10		0.18	0.08	**0.74**		0.74	0.39	0.50
	6869	QTI	1	≥2120	<10	0.03	0.97				0.97	0.24	0.60
	6870	QTI	1	≥2120	<10	0.04	0.96				0.96	0.21	0.47
	6871	MC	1	≥2120	<10		0.34	**0.52**	0.14		0.52	0.25	0.31
	6877	QTI	1	≥2090	<10	0.02	0.98				0.98	0.25	0.75
	6878	QTI	1	≥2080	<10	0.56	0.44				0.44	0.20	0.25
	6890	MC	1	≥2080	<10		0.09	**0.76**	0.15		0.76	0.42	0.55
	6891	MC	1	≥2080	<10		**0.87**	0.04	0.09		0.87	0.36	0.55
	6848	QTI	1	≥1880	<10	0.26	0.74				0.74	0.36	0.46
	6849	QTI	1	≥1880	<10	0.27	0.73				0.73	0.34	0.43
	6850	QTI	1	≥1880	<10	0.11	0.89				0.89	0.39	0.59
	6851	QTI	1	≥1880	<10	0.45	0.55				0.55	0.28	0.35
	6852	QTI	1	≥1870	<10	0.17	0.83				0.83	0.23	0.34
	6666	QTI	1	≥2140	<10	0.05	0.95				0.95	0.34	0.71
	6667	QTI	1	≥2110	<10	0.11	0.89				0.89	0.39	0.61
	6668	QTI	1	≥2190	<10	0.02	0.98				0.98	0.23	0.66
	6669	QTI	1	≥2050	<10	0.07	0.93				0.93	0.25	0.45
	6670	QTI	1	≥2110	<10	0.02	0.98				0.98	0.20	0.59
Reading	6671	QTI	1	≥1760	≥10	0.12	0.88				0.88	0.26	0.40
	6672	MC	1	≥1770	≥10		0.40	**0.46**	0.14		0.46	0.28	0.34
	6673	MC	1	≥1760	≥20		**0.61**	0.25	0.14		0.61	0.32	0.41
	6674	QTI	1	≥1760	≥10	0.48	0.52				0.52	0.22	0.28
	6675	QTI	1	≥1750	≥10	0.38	0.62				0.62	0.37	0.47
	6676	MC	1	≥1760	≥10		0.34	**0.56**	0.10		0.56	0.38	0.47
	6677	MC	1	≥1740	<10		0.40	0.28	**0.32**		0.32	0.35	0.44
	6678	QTI	1	≥1730	<10	0.46	0.54				0.54	0.42	0.51
	6934	QTI	1	≥1810	<10	0.22	0.78				0.78	0.22	0.30
	6935	QTI	1	≥1810	<10	0.68	0.32				0.32	0.54	0.69
	6679	MC	1	≥1890	≥10		**0.72**	0.14	0.14		0.72	0.40	0.52
	6680	MC	1	≥1790	≥10		0.35	0.21	**0.45**		0.45	0.32	0.40
	6681	MC	1	≥1880	≥10		0.27	0.14	**0.59**		0.59	0.47	0.58

Domain	Item #	Item Format	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			
	6682	MC	1	≥1840	<10		0.26	**0.61**	0.13		0.61	0.44	0.54
	6683	MC	1	≥1890	<10		**0.78**	0.12	0.10		0.78	0.45	0.61
	6684	MC	1	≥1790	<10		0.07	0.06	**0.87**		0.87	0.46	0.72
	6685	MC	1	≥1830	<10		0.18	0.12	**0.70**		0.70	0.55	0.69
	6686	MC	1	≥1900	<10		0.26	**0.58**	0.16		0.58	0.41	0.50
	6687	MC	1	≥1840	<10		0.12	**0.74**	0.14		0.74	0.51	0.67
	6688	MC	1	≥1890	<10		0.13	**0.74**	0.13		0.74	0.51	0.67
Writing	6970	QTI	1	≥1220	<10	0.17	0.83				0.83	0.43	0.62
	6971	QTI	1	≥1220	<10	0.16	0.84				0.84	0.50	0.71
	6972	QTI	1	≥1190	<10	0.15	0.85				0.85	0.45	0.67
	6973	QTI	1	≥1190	<10	0.14	0.86				0.86	0.43	0.66
	7001	QTI	1	≥1240	<10	0.15	0.85				0.85	0.41	0.63
	7002	QTI	1	≥1240	<10	0.14	0.86				0.86	0.50	0.75
	7003	QTI	1	≥1250	<10	0.21	0.79				0.79	0.43	0.60
	7004	QTI	1	≥1250	<10	0.24	0.76				0.76	0.42	0.56
	6689	QTI	1	≥1120		0.46	0.54				0.54	0.66	0.79
	6690	QTI	1	≥1280	<10	0.38	0.62				0.62	0.68	0.84
	6691	QTI	1	≥1170	<10	0.46	0.54				0.54	0.64	0.77
	6692	QTI	1	≥1220	<10	0.35	0.65				0.65	0.65	0.79
	6693	QTI	1	≥1130	<10	0.36	0.64				0.64	0.61	0.74
	6694	QTI	1	≥1180	<10	0.46	0.54				0.54	0.65	0.80
	6695	QTI	1	≥1150	<10	0.44	0.56				0.56	0.63	0.75
	6696	QTI	1	≥1190	<10	0.24	0.76				0.76	0.63	0.80
	6697	QTI	1	≥1130	<10	0.37	0.63				0.63	0.56	0.68
	6698	QTI	1	≥1170	<10	0.43	0.57				0.57	0.67	0.81
	6699	QTI	1	≥1180	<10	0.17	0.83				0.83	0.41	0.57
	6700	QTI	1	≥1230	<10	0.31	0.69				0.69	0.30	0.39
	6701	QTI	1	≥1250	<10	0.28	0.72				0.72	0.44	0.56
	6702	QTI	1	≥1130	<10	0.17	0.83				0.83	0.47	0.67
	6703	QTI	1	≥1180	<10	0.09	0.91				0.91	0.35	0.60
	6704	QTI	1	≥1190	<10	0.24	0.76				0.76	0.42	0.55
	6705	QTI	1	≥1240	<10	0.09	0.91				0.91	0.21	0.35
	7031	QTI	1	≥1200	<10	0.76	0.24				0.24	0.36	0.50
	7036	QTI	1	≥1200	<10	0.09	0.91				0.91	0.42	0.72

Note. P-values of the key (multiple-choice items) are in bold.

Table A.3 Field test items' Classical Test Theory (CTT) summary statistics: Grades 2–3 (Combined Seven States' Data)

Domain	**Item #**	**Item Format**	**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**			
Listening	6706	QTI	1	≥2520	<10	0.13	0.87				0.87	0.35	0.48
	6707	QTI	1	≥2490	≥10	0.11	0.89				0.89	0.34	0.50
	6708	QTI	1	≥2450	<10	0.72	0.28				0.28	0.15	**0.21**
	6709	QTI	1	≥2520	≥10	0.05	0.95				0.95	0.29	0.50
	6710	QTI	1	≥2430	<10	0.07	0.93				0.93	0.40	0.65
	6712	QTI	1	≥2620	<10	0.09	0.91				0.91	0.33	0.49
	6713	QTI	1	≥2490	<10	0.85	0.15				0.15	0.01	**0.02**
	6714	QTI	1	≥2440	<10	0.05	0.95				0.95	0.37	0.67
	6715	QTI	1	≥2450	<10	0.03	0.97				0.97	0.33	0.72
	6716	QTI	1	≥2580	<10	0.27	0.73				0.73	0.44	0.54
	6717	QTI	1	≥2440	≥10	0.62	0.38				0.38	0.11	**0.14**
	6718	QTI	1	≥2580	<10	0.07	0.93				0.93	0.40	0.67
	6719	QTI	1	≥2580	<10	0.11	0.89				0.89	0.40	0.57
	6864	QTI	1	≥2580	<10	0.19	0.81				0.81	0.34	0.44
	6872	MC	1	≥2580	<10		**0.72**	0.24	0.04		0.72	0.07	**0.09**
	6873	MC	1	≥2580	<10		0.08	**0.80**	0.12		0.80	0.39	0.49
	6874	MC	1	≥2570	<10		0.06	0.09	**0.85**		0.85	0.43	0.58
	6875	MC	1	≥2570	<10		0.15	0.04	**0.81**		0.81	0.44	0.56
	6876	QTI	1	≥2420	<10	0.36	0.64				0.64	0.31	0.38
	6859	QTI	1	≥2520	<10	0.16	0.84				0.84	0.40	0.52
Reading	6860	QTI	2	≥3820	≥20	0.08	0.57	0.35			0.63	0.18	**0.21**
	6926	MC	1	≥3820	≥10		**0.70**	0.17	0.13		0.70	0.44	0.55
	6927	MC	1	≥3820	≥10		0.25	**0.43**	0.32		0.43	0.27	0.33
	6928	MC	1	≥3830	≥10		0.20	0.09	**0.71**		0.71	0.57	0.68
	6720	MC	1	≥3880	<10		**0.81**	0.10	0.09		0.81	0.46	0.61
	6721	MC	1	≥3850	<10		0.14	0.30	**0.56**		0.56	0.46	0.56
	6722	MC	1	≥3750	<10		**0.82**	0.09	0.10		0.82	0.40	0.53
	6723	MC	1	≥3800	<10		**0.92**	0.06	0.02		0.92	0.38	0.63
	6724	MC	1	≥3930	<10		0.08	0.08	**0.84**		0.84	0.57	0.77
	6726	MC	1	≥3840	<10		**0.85**	0.08	0.07		0.85	0.50	0.67
	6727	QTI	1	≥3840	≥20	0.06	0.94				0.94	0.42	0.75
	6728	QTI	1	≥3750	≥30	0.07	0.93				0.93	0.43	0.72
	6729	MC	1	≥3960	≥10		0.12	0.12	**0.76**		0.76	0.60	0.74
	6950	MC	1	≥3950	≥20		0.23	**0.67**	0.10		0.67	0.41	0.51
	6951	QTI	1	≥3950	≥20	0.19	0.81				0.81	0.49	0.64
Writing	6952	QTI	1	≥3070	<10	0.37	0.63				0.63	0.62	0.74
	6730	QTI	1	≥3050	<10	0.44	0.56				0.56	0.52	0.63

Domain	Item #	Item Format	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			
	6731	QTI	1	≥3060	<10	0.23	0.77				0.77	0.60	0.75
	6732	QTI	1	≥3060	<10	0.26	0.74				0.74	0.60	0.73
	6733	QTI	1	≥3000	<10	0.38	0.62				0.62	0.36	0.44
	6734	QTI	1	≥2990	<10	0.59	0.41				0.41	0.36	0.46
	6735	QTI	1	≥2960	<10	0.31	0.69				0.69	0.51	0.62
	6736	QTI	1	≥3010	<10	0.33	0.67				0.67	0.51	0.62
	6737	QTI	1	≥3010	<10	0.53	0.47				0.47	0.29	0.36
	6738	QTI	1	≥2980	<10	0.42	0.58				0.58	0.37	0.46
	6739	QTI	1	≥3040	<10	0.56	0.44				0.44	0.34	0.42
	6740	QTI	1	≥3040	<10	0.20	0.80				0.80	0.48	0.62

Table A.4 Field test items' Classical Test Theory (CTT) summary statistics: Grades 4–5 (Combined Seven States' Data)

Domain	Item #	Item Format	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			
Listening	6744	QTI	1	≥1540	<10	0.10	0.90				0.90	0.38	0.54
	6745	QTI	1	≥1490	<10	0.06	0.94				0.94	0.41	0.69
	6746	QTI	1	≥1530	<10	0.02	0.98				0.98	0.26	0.55
	6747	QTI	1	≥1530	<10	0.16	0.84				0.84	0.28	0.37
	6748	QTI	1	≥1550	<10	0.08	0.92				0.92	0.51	0.74
	6749	QTI	1	≥1480	<10	0.23	0.77				0.77	0.36	0.45
	6750	QTI	1	≥1520	<10	0.02	0.98				0.98	0.30	0.66
	6751	QTI	1	≥1520	≥10	0.07	0.93				0.93	0.46	0.71
	6752	QTI	1	≥1520	<10	0.02	0.98				0.98	0.29	0.68
	6754	QTI	1	≥1490	<10	0.11	0.89				0.89	0.59	0.79
	6755	QTI	1	≥1560	<10	0.56	0.44				0.44	0.23	0.29
	6756	QTI	1	≥1530	<10	0.48	0.52				0.52	0.45	0.56
	6757	QTI	1	≥1510	<10	0.06	0.94				0.94	0.48	0.75
	6832	QTI	1	≥1460	<10	0.59	0.41				0.41	0.08	**0.10**
	6839	QTI	1	≥1470	<10	0.35	0.65				0.65	0.41	0.50
	6840	QTI	1	≥1560	<10	0.44	0.56				0.56	0.44	0.53
	6841	QTI	1	≥1470	<10	0.22	0.78				0.78	0.59	0.71
	6993	QTI	1	≥1540	<10	0.32	0.68				0.68	0.34	0.42
	6836	MC	1	≥1590	<10		0.04	0.02	**0.91**	0.03	0.91	0.41	0.60
	6837	MC	1	≥1590	<10		0.03	0.02	0.03	**0.92**	0.92	0.46	0.67
	6838	MC	1	≥1590			0.11	0.05	**0.81**	0.02	0.81	0.51	0.62
	7037	QTI	1	≥1490	<10	0.28	0.72				0.72	0.34	0.43
Reading	6893	MC	1	≥1730	<10		0.16	**0.57**	0.18	0.09	0.57	0.41	0.50
	6894	MC	1	≥1730	<10		0.28	**0.47**	0.14	0.11	0.47	0.31	0.39
	6895	MC	1	≥1730	<10		0.18	0.17	**0.36**	0.29	0.36	0.15	**0.20**
	6896	QTI	1	≥1730	<10	0.40	0.60				0.60	0.36	0.45
	6758	MC	1	≥1700	<10		0.14	**0.78**	0.04	0.04	0.78	0.29	0.37
	6759	MC	1	≥1650			**0.87**	0.06	0.03	0.04	0.87	0.42	0.56
	6760	MC	1	≥1640	<10		0.19	0.06	0.12	**0.63**	0.63	0.47	0.57
	6761	MC	1	≥1690	<10		**0.89**	0.03	0.05	0.03	0.89	0.53	0.70
	6762	MC	1	≥1600	<10		0.04	0.04	**0.91**	0.02	0.91	0.46	0.67
	6763	MC	1	≥1650	<10		0.09	**0.77**	0.02	0.12	0.77	0.28	0.36
	6764	MC	1	≥1670			0.07	**0.87**	0.03	0.03	0.87	0.56	0.73
	6765	MC	1	≥1620	<10		**0.82**	0.05	0.05	0.07	0.82	0.50	0.62
	6766	MC	1	≥1640	<10		0.25	**0.46**	0.11	0.18	0.46	0.20	0.25
	6767	MC	1	≥1660	<10		**0.85**	0.05	0.07	0.03	0.85	0.57	0.70

Domain	Item #	Item Format	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**			
	6768	MC	1	≥1710	<10		0.04	0.18	**0.74**	0.04	0.74	0.43	0.52
	6769	MC	1	≥1640	<10		**0.90**	0.03	0.03	0.03	0.90	0.52	0.71
	6770	MC	1	≥1680	<10		0.09	0.25	**0.61**	0.05	0.61	0.27	0.34
	6771	MC	1	≥1670			0.06	0.05	0.03	**0.86**	0.86	0.62	0.78
	6897	QTI	1	≥1710	<10	0.63	0.37				0.37	0.14	**0.19**
	6898	QTI	1	≥1710	<10	0.38	0.62				0.62	0.43	0.52
	6899	MC	1	≥1710	<10		0.18	0.25	0.14	**0.44**	0.44	0.38	0.48
	6900	MC	1	≥1710	<10		0.15	0.19	0.12	**0.54**	0.54	0.41	0.51
	6914	MC	1	≥1820	<10		0.28	**0.41**	0.12	0.19	0.41	0.27	0.35
	6915	QTI	1	≥1820	<10	0.51	0.49				0.49	0.04	**0.05**
	6916	MC	1	≥1820	<10		0.34	0.17	**0.40**	0.09	0.40	0.29	0.38
	6917	QTI	1	≥1820	<10	0.28	0.72				0.72	0.52	0.62
	7038	MC	1	≥1680	<10		0.26	0.02	0.02	**0.69**	0.69	0.11	**0.14**
Writing	6772	QTI, QTI, QTI	3	≥2140	<10	0.10	0.17	0.29	0.45		0.70	0.64	0.66
	6773	QTI, QTI, QTI	3	≥2170	<10	0.10	0.17	0.26	0.46		0.69	0.65	0.67
	6774	QTI, QTI, QTI	3	≥2150	<10	0.10	0.24	0.47	0.19		0.59	0.51	0.53
	6775	QTI, QTI, QTI	3	≥2130	<10	0.20	0.24	0.29	0.28		0.55	0.63	0.65
	6776	QTI	1	≥2240	<10	0.27	0.73				0.73	0.61	0.72
	6777	QTI	1	≥2090		0.21	0.79				0.79	0.58	0.71
	6778	QTI	1	≥2120		0.47	0.53				0.53	0.34	0.42
	6779	QTI	1	≥2150	<10	0.79	0.21				0.21	0.35	0.57
	6780	QTI	1	≥2170		0.22	0.78				0.78	0.68	0.79
	6781	QTI	1	≥2120	<10	0.25	0.75				0.75	0.58	0.69
	7039	QTI, QTI, QTI	3	≥2200	<10	0.17	0.41	0.30	0.12		0.46	0.33	0.35

Table A.5 Field test items' Classical Test Theory (CTT) summary statistics: Grades 6–8 (Combined Seven States' Data)

Domain	Item #	Item Format	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			
Listening	6854	MC	1	≥4100	<10		0.04	0.07	**0.86**	0.02	0.86	0.42	0.58
	6855	MC	1	≥4100	<10		0.07	0.04	**0.87**	0.02	0.87	0.63	0.85
	6856	MC	1	≥4100	<10		0.02	0.25	0.03	**0.69**	0.69	0.62	0.73
	6857	MC	1	≥4110	<10		**0.82**	0.09	0.07	0.02	0.82	0.54	0.68
	6858	QTI	1	≥4090	≥30	0.47	0.53				0.53	0.53	0.65
	6861	QTI	1	≥4100	≥20	0.42	0.58				0.58	0.55	0.66
	6862	MC	1	≥4120	<10		0.07	0.07	0.17	**0.69**	0.69	0.48	0.58
	6863	MC	1	≥4120	<10		0.09	**0.83**	0.04	0.05	0.83	0.53	0.68
	6785	QTI	1	≥3990	≥20	0.06	0.94				0.94	0.36	0.62
	6786	QTI	1	≥3980	≥10	0.28	0.72				0.72	0.37	0.46
	6787	QTI	1	≥3990	≥20	0.11	0.89				0.89	0.32	0.47
	6788	QTI	1	≥3890	≥20	0.15	0.85				0.85	0.33	0.44
	6789	QTI	1	≥4010	≥10	0.13	0.87				0.87	0.44	0.61
	6790	QTI	1	≥4000	≥20	0.25	0.75				0.75	0.27	0.36
	6889	QTI	1	≥4050	≥40	0.41	0.59				0.59	0.51	0.62
Reading	6936	MC	1	≥4230	<10		0.20	**0.58**	0.16	0.05	0.58	0.45	0.54
	6937	MC	1	≥4230	<10		**0.53**	0.10	0.11	0.26	0.53	0.30	0.37
	6938	MC	1	≥4220	<10		0.08	0.32	**0.50**	0.11	0.50	0.37	0.46
	6939	MC	1	≥4220	<10		0.18	0.27	0.16	**0.39**	0.39	0.32	0.41
	6940	QTI	2	≥4220	<10	0.18	0.70	0.11			0.47	0.04	**0.05**
	6941	MC	1	≥4190	<10		0.14	0.15	**0.61**	0.11	0.61	0.37	0.46
	6942	MC	1	≥4190	<10		**0.53**	0.15	0.23	0.08	0.53	0.39	0.48
	6943	MC	1	≥4190	<10		0.10	0.32	0.19	**0.40**	0.40	0.34	0.44
	6944	MC	1	≥4190	<10		0.18	0.23	**0.45**	0.15	0.45	0.26	0.32
	6945	QTI	2	≥4190	<10	0.15	0.48	0.37			0.61	0.45	0.50
	6920	MC	1	≥4190	<10		0.30	0.20	**0.32**	0.19	0.32	0.14	**0.19**
	6921	MC	1	≥4190	<10		0.23	**0.39**	0.15	0.22	0.39	0.20	0.26
	6922	MC	1	≥4180	≥10		0.21	0.22	0.20	**0.37**	0.37	0.34	0.44
	6923	QTI	1	≥4120	≥60	0.66	0.34				0.34	0.34	0.45
	6924	QTI	2	≥4180	≥10	0.15	0.42	0.44			0.65	0.47	0.52
	6925	MC	1	≥4180	<10		0.15	0.26	**0.37**	0.22	0.37	0.13	**0.16**
	6791	MC	1	≥3900	<10		0.08	**0.71**	0.05	0.16	0.71	0.39	0.48
	6792	MC	1	≥3900	<10		0.03	0.08	**0.80**	0.09	0.80	0.44	0.56
	6793	MC	1	≥3810	<10		**0.75**	0.07	0.12	0.06	0.75	0.44	0.54
	6794	MC	1	≥3810	<10		0.10	0.23	**0.61**	0.06	0.61	0.46	0.55
	6795	MC	1	≥3990	<10		0.15	0.24	**0.25**	0.36	**0.25**	0.23	0.32
	6796	MC	1	≥3990	<10		0.14	0.09	**0.53**	0.24	0.53	0.45	0.54

Domain	**Item #**	**Item Format**	**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**			
	6797	MC	1	≥3900	<10		0.04	**0.83**	0.05	0.09	0.83	0.57	0.73
	6798	MC	1	≥3890	<10		0.04	0.09	0.08	**0.79**	0.79	0.59	0.72
	6799	MC	1	≥3940	<10		**0.59**	0.24	0.06	0.12	0.59	0.32	0.39
	6800	MC	1	≥3930	<10		0.40	0.15	0.08	**0.38**	0.38	0.28	0.36
	6801	MC	1	≥3780	≥10		0.04	0.03	0.08	**0.85**	0.85	0.51	0.68
	6802	MC	1	≥3780	≥10		0.19	0.23	**0.52**	0.06	0.52	0.48	0.58
Writing	6803	QTI, QTI, QTI	3	≥4430	<10	0.15	0.24	0.28	0.32		0.59	0.66	0.69
	6804	QTI, QTI, QTI	3	≥4430	<10	0.08	0.17	0.33	0.42		0.69	0.65	0.68
	6805	QTI, QTI, QTI	3	≥4620	<10	0.10	0.17	0.30	0.43		0.69	0.67	0.69
	6806	QTI, QTI, QTI	3	≥4480	<10	0.10	0.20	0.25	0.45		0.68	0.65	0.67
	6807	QTI, QTI, QTI	3	≥4370	<10	0.12	0.18	0.27	0.43		0.67	0.68	0.70
	6808	QTI, QTI, QTI	3	≥4390	<10	0.06	0.17	0.22	0.55		0.75	0.65	0.69

Table A.6 Field test items' Classical Test Theory (CTT) summary statistics: Grades 9–12 (Combined Seven States' Data)

Domain	Item #	Item Format	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			
Listening	6879	QTI	1	≥3120	≥20	0.66	0.34				0.34	0.27	0.35
	6880	MC	1	≥3140	<10		0.21	0.09	0.06	**0.64**	0.64	0.60	0.71
	6881	QTI	1	≥3120	≥20	0.38	0.62				0.62	0.58	0.69
	6882	MC	1	≥3140	<10		0.05	**0.81**	0.05	0.09	0.81	0.49	0.65
	6883	MC	1	≥2990	<10		0.06	**0.56**	0.08	0.31	0.56	0.38	0.46
	6884	MC	1	≥2990	<10		0.06	0.06	0.09	**0.79**	0.79	0.59	0.74
	6885	MC	1	≥2990	<10		0.04	0.04	**0.84**	0.08	0.84	0.50	0.70
	6886	MC	1	≥2990	<10		**0.75**	0.10	0.10	0.05	0.75	0.57	0.70
	6887	QTI	1	≥2960	≥10	0.41	0.59				0.59	0.42	0.52
	6888	QTI	1	≥3130	≥40	0.34	0.66				0.66	0.60	0.71
	7000	QTI	1	≥2970	≥20	0.36	0.64				0.64	0.58	0.69
	6814	QTI	1	≥3040	≥20	0.39	0.61				0.61	0.43	0.52
	6815	QTI	1	≥2960	≥20	0.10	0.90				0.90	0.41	0.65
	6816	QTI	1	≥3070	≥20	0.03	0.97				0.97	0.30	0.78
	6817	QTI	1	≥2960	≥10	0.19	0.81				0.81	0.42	0.56
	6818	QTI	1	≥3030	≥10	0.53	0.47				0.47	0.06	**0.08**
	6819	QTI	1	≥3020	≥20	0.07	0.93				0.93	0.43	0.78
	6294	QTI	1	≥3010	≥10	0.23	0.77				0.77	0.55	0.70
	7040	QTI	1	≥3040	≥30	0.03	0.97				0.97	0.26	0.60
	7041	QTI	1	≥3100	≥20	0.20	0.80				0.80	0.32	0.44
	7042	QTI	1	≥3080	<10	0.11	0.89				0.89	0.11	**0.18**
Reading	6958	MC	1	≥7730	≥10		0.13	0.18	**0.53**	0.16	0.53	0.43	0.52
	6959	QTI	1	≥7650	≥90	0.56	0.44				0.44	0.38	0.47
	6960	MC	1	≥7720	≥20		0.11	0.39	**0.31**	0.19	0.31	0.28	0.36
	6961	MC	1	≥7720	≥20		**0.31**	0.32	0.17	0.20	0.31	0.13	**0.17**
	6962	QTI	2	≥7720	≥20	0.30	0.56	0.14			0.42	0.24	0.27
	6963	MC	1	≥7720	≥20		0.22	0.20	0.22	**0.35**	0.35	0.35	0.44
	6820	MC	1	≥7640	≥10		0.08	0.15	0.10	**0.67**	0.67	0.55	0.65
	6821	MC	1	≥7640	≥10		0.27	**0.55**	0.07	0.11	0.55	0.47	0.56
	6822	MC	1	≥7430	≥10		0.10	0.15	0.09	**0.66**	0.66	0.56	0.66
	6823	MC	1	≥7430	≥10		0.12	**0.73**	0.10	0.05	0.73	0.50	0.63
	6824	MC	1	≥7400	≥10		0.11	0.11	**0.74**	0.04	0.74	0.56	0.69
	6825	MC	1	≥7400	≥20		0.11	0.09	**0.72**	0.08	0.72	0.63	0.76
	6826	MC	1	≥7620	≥10		0.11	**0.57**	0.26	0.06	0.57	0.45	0.54
	6827	QTI	1	≥7620	≥10	0.29	0.71				0.71	0.67	0.80
	6953	MC	1	≥7950	≥20		**0.47**	0.19	0.21	0.13	0.47	0.25	0.31
	6954	MC	1	≥7940	≥20		0.18	0.21	**0.43**	0.19	0.43	0.20	**0.25**

Domain	Item #	Item Format	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			
	6955	QTI	1	≥7950	≥20	0.43	0.57				0.57	0.38	0.46
	6956	QTI	1	≥7850	≥110	0.35	0.65				0.65	0.30	0.38
	6957	QTI	2	≥7940	≥20	0.31	0.46	0.23			0.46	0.45	0.49
Writing	6828	QTI, QTI, QTI	3	≥11280	≥10	0.15	0.35	0.36	0.14		0.50	0.46	0.48
	6829	QTI, QTI, QTI	3	≥11230	≥10	0.21	0.20	0.25	0.34		0.57	0.72	0.74
	6830	QTI, QTI, QTI	3	≥11420	≥20	0.14	0.24	0.26	0.36		0.61	0.69	0.71
	6831	QTI, QTI, QTI	3	≥11440	≥20	0.17	0.20	0.23	0.40		0.62	0.72	0.74