

**TECHNICAL REPORT**

**Connecticut Alternate Assessment  
English Language Arts, Grades 3–8, 11  
Mathematics, Grades 3–8, 11**

**Test Administrations  
March 27–June 2, 2023**

*Submitted to:*

Connecticut State Department of Education (CSDE)

*Submitted by:*

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

*January 2024*

## TABLE OF CONTENTS

<b>1.</b>	<b>Introduction.....</b>	<b>1</b>
<b>2.</b>	<b>2023 Administration and Item Re-Evaluation .....</b>	<b>2</b>
2.1	TESTING WINDOW .....	2
2.2	TEST FORMS .....	2
2.3	TEST MODE.....	2
2.4	TEST ATTEMPTEDNESS .....	2
2.5	ITEM RE-EVALUATION.....	3
<b>3.</b>	<b>2023 State Summary .....</b>	<b>6</b>
3.1	STUDENT PARTICIPATION .....	6
3.2	SCORING METHOD REVIEW .....	8
3.3	SCORE SUMMARY .....	10
3.4	PERCENTAGE OF STUDENTS BY PERFORMANCE LEVEL .....	10
<b>4.</b>	<b>Reporting .....</b>	<b>11</b>
4.1	ONLINE REPORTING SYSTEM .....	11
4.2	PAPER REPORT.....	13
<b>5.</b>	<b>Reliability and Validity.....</b>	<b>17</b>
5.1	RELIABILITY .....	17
<b>6.</b>	<b>Quality Control .....</b>	<b>36</b>
6.1	QUALITY CONTROL IN TEST CONFIGURATION .....	36
6.2	PLATFORM REVIEW.....	36
6.3	USER ACCEPTANCE TESTING AND FINAL REVIEW .....	37
6.4	QUALITY ASSURANCE IN ONLINE DATA .....	37
6.5	QUALITY CONTROL ON SCORING .....	37
6.6	QUALITY ASSURANCE IN REPORTING .....	38
<b>7.</b>	<b>Reference .....</b>	<b>40</b>

**List of Tables**

Table 1. Number of Operational Items by Standards..... 1  
Table 2. Participation by Grade and Gender..... 7  
Table 3. Participation by Grade and Ethnicity ..... 8  
Table 4. Slope and Intercept ..... 9  
Table 5. Scale Score Cut Points ..... 9  
Table 6. Scale Score Summary ..... 10  
Table 7. Percentage of Students by Performance Level ..... 11  
Table 8. Types of Online Score Reports by Level of Aggregation..... 12  
Table 9. Types of Subgroups ..... 13  
Table 10. Cronbach’s Alpha ..... 18  
Table 11. Marginal Reliability and Marginal Standard Error of Measurement ..... 19  
Table 12. Classification Accuracy and Consistency ..... 27  
Table 13. Eigenvalues of The First Three Components..... 29  
Table 14. Observed Correlation Matrix Among Standards (ELA) ..... 30  
Table 15. Disattenuated Correlation Matrix Among Standards (ELA) ..... 31  
Table 16. Observed Correlation Matrix Among Standards (Mathematics) ..... 32  
Table 17. Disattenuated Correlation Matrix Among Standards (Mathematics)..... 33  
Table 18. Q3 Distribution ..... 36

**List of Figures**

Figure 1. Family Report Mock-Up ..... 14  
Figure 2. Family Report Mock-Up for Early Stop Students ..... 16  
Figure 3. Sample Test Information Function ..... 20  
Figure 4. CSEM by Test ..... 21  
Figure 5. Scree Plots ..... 27

# 1. Introduction

This report introduces the Connecticut Alternate Assessment (CTAA) used during the 2023 administration, summarizes test administration and performance results, and details the evaluation of the assessment quality.

Funded through a General Supervision Enhancement Grant (GSEG) from the U.S. Department of Education Office of Special Education Programs (OSEP), the National Center and State Collaborative (NCSC), a collaborative of 24 states and five organizations (National Center on Educational Outcomes [NCEO] at the University of Minnesota, National Center for the Improvement of Educational Assessment [Center for Assessment], University of North Carolina at Charlotte, University of Kentucky, and edCount, LLC), developed the multi-state comprehensive alternate assessment for students with significant cognitive disabilities to complement the work of the Race to the Top Common State Assessment Program (RTTA). As a member of this multi-state grant project, the Connecticut State Department of Education (CSDE) has adopted the NCSC English language arts (ELA) and mathematics tests since the spring 2015 administration. The American Institutes for Research (AIR) started administering the CTAA for the state in 2016. The CTAA tests are administered to students in grades 3–8 and 11.

The CTAA is the NCSC alternate assessment and is based on alternate achievement standards (AA-AAS). The 2023 CTAA assessment included

- assessments in mathematics and ELA for students in grades 3–8 and 11;
- approximately 29–41 operational items for each subject, mostly selected response;
- online assessments with paper-pencil tests as accommodations; and
- approximately 1.5–2 hours for each assessment (mathematics and ELA).

*Table 1. Number of Operational Items by Standards*

		ELA						
Standards		Grade	Grade	Grade	Grade	Grade	Grade	Grade
		3	4	5	6	7	8	11
Reading	Informational Text	10	9	9	10	8	11	9
	Literature	9	10	11	10	10	10	7
	Foundational Skills	10	10					
Language		2	3	3	4	4	4	4
Writing		10	8	8	8	10	10	9
<b>Total</b>		41	40	31	32	32	35	29
		Mathematics						
Standards		Grade	Grade	Grade	Grade	Grade	Grade	Grade
		3	4	5	6	7	8	11
Number & Operations in Base Ten		8	4	15				
	Numbers and Operations–Fractions	8	9	6				

<b>Number &amp; Quantity</b>							7
<b>Operations &amp; Algebraic Thinking</b>	12	12	4				
<b>Measurement &amp; Data</b>	8	8	7				
<b>The Number System</b>				12	8	3	
<b>Expressions &amp; Equations</b>				7	4	8	
<b>Statistics &amp; Probability</b>				4	3	8	6
<b>Ratios &amp; Proportional Relationships</b>				12	12		
<b>Functions</b>						8	
<b>Geometry</b>	4	4	4	4	8	8	2
<b>Algebra &amp; Functions</b>							16
<b>Total</b>	40	37	36	39	35	35	31

The information about test development, item alignment and system coherence, test administration, item calibration and analysis, field testing, item review, scoring and scaling, and standard setting can be found in the 2015 NCSC technical report located at [http://www.ncscpartners.org/Media/Default/PDFs/Resources/NCSC15\\_NCSC\\_TechnicalManualNarrative.pdf](http://www.ncscpartners.org/Media/Default/PDFs/Resources/NCSC15_NCSC_TechnicalManualNarrative.pdf). This document summarizes the test results, reporting, reliability and validity of the test, and the quality control process for the 2023 administration.

## 2. 2023 Administration and Item Re-Evaluation

### 2.1 TESTING WINDOW

The 2023 testing window started on March 27, 2023 and ended on June 2, 2023.

### 2.2 TEST FORMS

As described in the 2015 NCSC technical report, four forms were developed for each grade and subject test. In 2023, one of the ELA forms was adopted for each ELA test. The mathematics forms were newly built in 2016, and those forms were administered again in 2023. The form summary and their comparisons with their respective test blueprints can be found in Appendix A.

### 2.3 TEST MODE

The 2023 tests were administered online with paper-pencil forms as an accommodation. For paper-pencil tests, test administrators (TAs) entered item responses through the online system.

### 2.4 TEST ATTEMPTEDNESS

If a student logs in to the online testing system and answers at least one item, the student is counted as having attempted or participated in the test.

For CTAA, an early stopping rule (ESR) is established. The rule allows students who have difficulties taking the tests to exit after the first four items. If a student does not respond to the first four items, the TA is required to contact the state to determine if the ESR should be considered for the student. If the student qualifies for the ESR, the TA will not resume the test. CSDE will inform AIR, and AIR will submit the test after the fourth item. Then, AIR will open a second test of the other subject for the student, submit no-response (NR) for the first four items, and submit the second test. For example, if a student did not respond to the ELA test and was approved as an ESR student, the he or she also did not take the mathematics test. The responses to the first four items in the mathematics test were set to NR, and the test was submitted by AIR. If the student did not qualify for the ESR, the TA had to resume the test and the student had to answer the rest of the items through the end of the test. If a student logs in to the online testing system and answers at least one item, the student is counted as having attempted or participated in the test.

## 2.5 ITEM RE-EVALUATION

From the 2015 administration, NCSC item analysis was based on students from all member states. To ensure that the items performed as expected for Connecticut students, the items were re-evaluated using Connecticut students only after the administration in 2016 and 2017. Items that did not perform well were dropped from scoring. This section summarizes the methods, criteria, and results of the evaluation. The statistics used in item evaluation in 2017 can be found in Appendix B.

### 2.5.1 Item Difficulty

Since the ELA and mathematics tests contain only selected-response items, we compute the proportion of number correct responses ( $p$ -value). Items that are either extremely difficult ( $< 0.2$ ) or extremely easy ( $> 0.9$ ) are flagged for review.

### 2.5.2 Classical Item Discrimination

The item discrimination index indicates the extent to which each item differentiates between those examinees who possess the skills being measured and those who do not. In general, the higher the value, the better the item is able to differentiate between high- and low-achieving students. The discrimination index for items is calculated as the correlation between the item score and the overall score excluding that item. Items are flagged if the point-biserial correlation is less than 0.25. The point-biserial correlation is computed as

$$r_{pb} = \frac{M_1 - M_0}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}} \sqrt{\frac{n_1 n_0}{n^2}}, \quad (1)$$

where

$x$  is the overall test score excluding the item under evaluation. So the denominator is the standard deviation of  $x$ ;

$M_1$  is the mean of  $x$  for records that have a response of 1 for the item;

$M_0$  is the mean of  $x$  for records that have a response of 0 for the item;  
 $n_1$  is the number of records for records that have a response of 1 for the item; and  
 $n_0$  is the number of records for records that have a response of 0 for the item.

### 2.5.3 Item Response Theory Model Fit

The two-parameter logistic (2PL) model, as shown below, is used in calibration for each individual item. IRTPRO was used in the analysis in 2017.

$$P_i(X_i = 1 | \theta_j) = \frac{\exp[Da_i(\theta_j - b_i)]}{1 + \exp[Da_i(\theta_j - b_i)]}, \quad (2)$$

where

$x_i$  indexes the raw score on item  $i$ ;

$\theta_j$  is the ability of student  $j$ ;

$a_i$  is the item discrimination for item  $i$ ;

$b_i$  is the item difficulty for item  $i$ ; and

$D$  is the normalizing constant 1.701.

Fit statistics are used for evaluating the goodness-of-fit of item response theory (IRT) item parameters to the actual performance of students. That is, item fit statistics indicate how well the scores obtained for a given item fit an expected distribution of scores under a particular IRT model.

To evaluate model fit, the Q1 statistic (Yen, 1981) was calculated for all core items. Q1 is a fit statistic that compares observed and expected item performance. Q1 is calculated as

$$Q_{1i} = \sum_{j=1}^J \frac{N_{ij}(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}, \quad (3)$$

where  $N_{ij}$  is the number of examinees in cell  $j$  for item  $i$ . As presented in Yen (1981), the trait interval of 10 is used. Students are sorted by their ability levels from the lowest to highest. They are then divided into 10 cells.  $O_{ij}$  and  $E_{ij}$  are the observed and predicted proportions of examinees in cell  $j$  for item  $i$ . The expected or predicted proportion is calculated as

$$E_{ij} = \frac{1}{N_{ij}} \sum_{aej}^{N_{ij}} P_i(\hat{\theta}_a), \quad (4)$$

where  $P_i(\hat{\theta}_a)$  is the item characteristic function for item  $i$  and examinee  $a$ . The summation is taken over examinees in cell  $j$ . The generalization of Q1, or Generalized Q1, for items with multiple response categories is

$$gen Q_{1i} = \sum_{j=1}^J \sum_{k=1}^{m_i} \frac{N_{ij}(O_{ikj} - E_{ikj})^2}{E_{ikj}}, \quad (5)$$

with

$$\mathbf{E}_{ikj} = \frac{1}{N_{ij}} \sum_{aej}^{N_{ij}} \mathbf{P}_{ik}(\hat{\boldsymbol{\theta}}_a). \quad (6)$$

Both the Q1 and Generalized Q1 results are transformed into the statistic ZQ1, and are compared to a criterion,  $ZQ_{crit}$ , to determine acceptable fit.

$$ZQ_1 = \frac{Q_1 - df}{\sqrt{2df}}, \quad (7)$$

where Q is either Q1 or Generalized Q1 and df is the degrees of freedom for the statistic. The degree of freedom is calculated as  $J * (K - 1) - m$  where  $J$  is the trait interval,  $K$  is the number of score categories, and  $m$  is the number of estimated item parameters in the IRT model. For example, 2PL items have  $df = 10 * (2 - 1) - 2 = 8$ . Poor fit is indicated where ZQ1 is greater than  $ZQ_{crit}$ .

The standardized fit values, referred to as  $ZQ_1$  statistics, are compared over items. The parameters from the 2015 calibration by NCSC are used in the computation, since the 2015 parameters are used in scoring.

#### 2.5.4 Item Parameter Stability Check

Each form built in 2015 contained core items and non-core items. The core items were used in scoring. The non-core items were identified and dropped from scoring for the considerations of meeting blueprints and statistically parallel forms of each test.

In 2015, four forms were developed by the consortium for each grade and subject test. In 2016, one of the ELA forms was adopted for each ELA test. The mathematics forms were newly built in 2016, and those forms were administered again in 2016 and beyond. In 2017, items were evaluated based on Connecticut students only to build conversion tables for scoring. During the item evaluation, the core items that were used in scoring for ELA tests were evaluated. All items in mathematics forms were evaluated. At the end of the evaluation, it was verified that the forms were statistically parallel to the corresponding 2015 forms. The evaluation took the following steps.

1. Free calibration was based on the item responses from the Connecticut 2017 administrations.
  - a. Student records with more than 10 valid scores were used in the calibration process.
  - b. The items in the verbal and nonverbal forms in the ELA grades 3 and 4 test needed to be combined in calibration.
2. The Stocking-Lord method was used to equate the 2017 item parameters to the 2015 scale.
  - a. Only items with a positive point-biserial were used in the equating process.
3. Test characteristics curves (TCCs) were plotted using the 2015 parameters and the equated 2017 parameters. More attention was paid to forms with large TCC differences.
4. The unsigned area (UA) of the differences of item response curves (ICCs) was computed.
5. The TCCs and UA were taken into account simultaneously to decide whether items with a large UA would be dropped from scoring.

Specifically, the differences of TCCs was evaluated as



$$D_q = \sum_{i=1}^n (p_{y1}(\theta_q) - p_{y2}(\theta_q)), \quad (8)$$

where  $p_{y1}(\theta_q)$  is the 2PL model evaluated at quadrature point  $\theta_q$  using the parameters from 2015 calibration,  $p_{y2}(\theta_q)$  is the 2PL model evaluated at quadrature point  $\theta_q$  using the equated parameters, and  $n$  is the number of items.

The unsigned area is computed as below. In the item evaluation,  $UA \geq 2$  drew attention.

$$UA = \int_{-\infty}^{\infty} |p_{y1}(\theta_q) - p_{y2}(\theta_q)| d\theta \quad (9)$$

### 2.5.5 Procedure for Item Evaluation

Flagged items were examined individually. The combined effect of statistics discussed in the previous sections was taken into account. During the examining period, the content of the flagged items was reviewed. The items that were determined to be used in scoring in 2017 are documented in Appendix C. They were approved by CSDE. The items used for scoring in 2023 are the same as those in 2017.

## 3. 2023 State Summary

### 3.1 STUDENT PARTICIPATION

This section describes the demographics of participating students in spring 2023. Table 2 and Table 3 present the student demographics for participating students by gender and ethnicity in each grade for each subject.

Demographic characteristics of the student population are relatively consistent across grades. Approximately 29%–35% of students are female in each grade and subject.

Among the participants, white students (30%–42%) and Hispanic students (31%–38%) make up the majority of the assessed students. African American students make up 18%–22%. Asian students make up 4%–8% of the assessed students in each grade, and multiracial students make up about 3%–5% of the assessed student population.

Table 2. Participation by Grade and Gender

ELA							
Grade	Total	Female		Male		Missing	
	N	N	%	N	%	N	%
3	473	138	29	335	71	0	0
4	493	158	32	335	68	0	0
5	497	158	32	339	68	0	0
6	455	150	33	305	67	0	0
7	423	122	29	301	71	0	0
8	419	137	33	282	67	0	0
11	417	145	35	272	65	0	0
<b>Total</b>	3177	1008	32	2169	68	0	0
Mathematics							
Grade	Total	Female		Male		Missing	
	N	N	%	N	%	N	%
3	472	138	29	334	71	0	0
4	488	159	33	329	67	0	0
5	496	157	32	339	68	0	0
6	452	149	33	303	67	0	0
7	425	124	29	301	71	0	0
8	417	134	32	283	68	0	0
11	414	144	35	270	65	0	0
<b>Total</b>	3164	1005	32	2159	68	0	0

*Table 3. Participation by Grade and Ethnicity*

Note: Table has been deleted by the CSDE but is available on request.

### **3.2 SCORING METHOD REVIEW**

The two-parameter logistic model was used in calibration. Based on the 2PL model, conversion tables were constructed for scoring in 2016 and has been used in 2016 and beyond. The conversion tables are constructed by associating each raw score point on the  $y$ -axis with the corresponding theta point on the  $x$ -axis in the TCCs for each form.

The scale scores are computed as  $SS_G = A * \theta_G + B$ , where  $A$  is the slope and  $B$  is the intercept as listed in Table 4. The scale scores of CTAA tests range from 1200 to 1290. If the estimated scale score is less than 1200, the scale score is set to 1200; if the estimated scale score is greater than 1290, the scale score is set to 1290.

Table 4. Slope and Intercept

Content Area	Grade	Slope (A)	Intercept (B)
<b>Mathematics</b>	3	13.06	1243.67
	4	13.1	1239.87
	5	13.08	1241.41
	6	12.82	1241.25
	7	12.91	1243.24
	8	13.02	1242.36
	11	12.99	1242.48
<b>ELA</b>	3	11.72	1242.05
	4	12.06	1240.09
	5	12.42	1241.61
	6	12.35	1237.81
	7	12.3	1242.43
	8	12.61	1239.46
	11	11.49	1244.22

CTAA tests adopted four performance levels, Level 1 to Level 4, on the scale score range divided by three cut scores. The cut scores are listed in Table 5. The slopes and intercepts listed in Table 4 and the cut scores were set through a standard-setting meeting convened by NCSC on August 10–13, 2015. Details about the standard setting can be found in the [National Center and State Collaborative 2015 Operational Assessment Technical Manual](#).

Table 5. Scale Score Cut Points

Content Area	Grade	scale.Cut 1	scale.Cut 2	scale.Cut 3
<b>Mathematics</b>	3	1236	1240	1254
<b>Mathematics</b>	4	1233	1240	1251
<b>Mathematics</b>	5	1231	1240	1255
<b>Mathematics</b>	6	1234	1240	1249
<b>Mathematics</b>	7	1232	1240	1254
<b>Mathematics</b>	8	1234	1240	1249
<b>Mathematics</b>	11	1234	1240	1249
<b>ELA</b>	3	1234	1240	1251
<b>ELA</b>	4	1234	1240	1258
<b>ELA</b>	5	1232	1240	1256
<b>ELA</b>	6	1231	1240	1253
<b>ELA</b>	7	1236	1240	1255
<b>ELA</b>	8	1230	1240	1250
<b>ELA</b>	11	1236	1240	1255

Appendix D contains the conversion tables based on items listed in Appendix C. The conversion tables contain the raw score, theta score, adjusted theta score that is adjusted around the cuts, scale score, performance level, and the standard error of measurement (SEM) associated with each theta or scale score. The SEM of the theta score is the inverse of the square root of the test information function, as shown in equation 10. The SEM of the scale score is the SEM of the theta score times the slope.

$$\text{se}(\theta) = \frac{1}{\sqrt{-\left(\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta}\right)}}, \quad (10)$$

where  $\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta}$  is the second derivative of the log-likelihood with respect to  $\theta$ .

### 3.3 SCORE SUMMARY

Table 6 presents the summary statistics of the scale score by grade for ELA and mathematics. The mean scale score ranged from 1230 to 1238.

*Table 6. Scale Score Summary*

Subject	Grade	N	Mean	Median	STD	Min.	Max.
ELA	3	473	1233	1233	16	1200	1286
ELA	4	493	1232	1232	15	1200	1290
ELA	5	497	1233	1233	16	1200	1290
ELA	6	455	1230	1231	15	1200	1290
ELA	7	423	1234	1235	14	1200	1290
ELA	8	419	1231	1230	15	1200	1290
ELA	11	417	1238	1237	16	1200	1290
Mathematics	3	472	1235	1236	17	1200	1290
Mathematics	4	488	1230	1233	16	1200	1290
Mathematics	5	496	1235	1238	15	1200	1276
Mathematics	6	452	1232	1234	14	1200	1290
Mathematics	7	425	1234	1235	13	1200	1276
Mathematics	8	417	1236	1239	15	1200	1276
Mathematics	11	414	1236	1236	15	1200	1290

Appendix E lists the student scale score distribution by test. The reason that more students earned the score of 1200 is that most of those students answered only the first four items and exited early. Many of them were identified as ESR students. The scale score summary by subgroups is listed in Appendix F.

### 3.4 PERCENTAGE OF STUDENTS BY PERFORMANCE LEVEL

The percentages of students in each performance level are listed in Table 7. More than 30% of the students were in Level 1, except for mathematics grade 5. The percentages of students in each performance level are listed in Appendix G.

Table 7. Percentage of Students by Performance Level

Subject	Grade	Total	Level1 (%)	Level2 (%)	Level3 (%)	Level 4 (%)
ELA	3	473	51	17	19	13
ELA	4	493	58	12	27	4
ELA	5	497	42	33	20	5
ELA	6	455	49	29	14	8
ELA	7	423	57	17	20	6
ELA	8	419	48	29	13	9
ELA	11	417	41	22	29	8
Mathematics	3	472	44	21	28	7
Mathematics	4	488	49	20	23	7
Mathematics	5	496	23	42	31	4
Mathematics	6	452	47	31	15	6
Mathematics	7	425	36	38	21	5
Mathematics	8	417	32	22	30	16
Mathematics	11	414	34	28	27	11

## 4. Reporting

The CTAA test results were provided in two mediums: the Online Reporting System (ORS) and a printed family report to be sent home.

### 4.1 ONLINE REPORTING SYSTEM

The ORS generates a set of online score reports that includes reliable and valid information describing student performance for students, parents, educators, and other stakeholders. Because the score reports on student performance are updated in real time, authorized users (e.g., school principals, teachers) may view student performance on the tests and use the results to improve student learning. The ORS also provides participation information that helps monitor the progression of the test administration.

In addition, the ORS produces aggregate score reports for teachers, schools, districts, and states. To facilitate comparisons, each aggregate report contains the summary results for the selected aggregate unit, as well as all aggregate units above the selected aggregate. For example, if a school is selected, the summary results of the district to which the school belongs and the summary results of the state are also provided so that the school performance can be compared with district and state performance. If a teacher is selected, the summary results for the school, the district, and the state are also provided for comparison purposes. Table 8 lists the types of online reports and the levels at which they can be viewed (student, roster, teacher, school, state, and district).

#### 4.1.1 Types of Online Score Reports

The ORS is designed to help educators, students, and parents answer questions regarding how well students have performed in each subject area. The ORS is designed with great consideration for

stakeholders who are not technical measurement experts (e.g., teachers, parents, or students). It ensures that test results are easily readable. Simple language is used so that users can quickly understand assessment results and make valid inferences about student achievement. In addition, the ORS is designed to present student performance in a uniform format. For example, similar colors are used for groups of similar elements, such as achievement levels, throughout the design. This design strategy allows scorers to compare similar elements and to avoid comparing dissimilar elements.

Once authorized users log in to the ORS and select Score Reports, the online score reports are presented hierarchically. The ORS starts by presenting summaries on student performance by grade at a selected aggregate level. In order to view student performance for a specific aggregate unit, users can select the specific aggregate unit from a drop-down menu with a list of aggregate units (e.g., schools within a district, teachers within a school) to choose from. For more-detailed student assessment results for a school, teacher, or roster, users can select the grade on the online score reports.

Table 8 summarizes the types of online score reports available at the aggregate and individual student levels. Detailed information about the online score reports and instructions on how to navigate the online score reporting system can be found in the *Online Reporting System User Guide*, accessible using the help button in the ORS.

*Table 8. Types of Online Score Reports by Level of Aggregation*

LEVEL OF AGGREGATION	TYPES OF ONLINE SCORE REPORTS
<b>State</b> <b>District</b> <b>School</b> <b>Teacher</b> <b>Roster</b>	Number of students tested and percentage of students determined proficient (overall and by subgroup) Average scale scores (overall and by subgroup) Percentage of students at each performance level (overall and by subgroup) On-demand student roster report
<b>Student</b>	Scale scores and the standard errors of the scale scores Performance levels

### 4.1.2 Subgroup Report

The aggregate score reports at a selected aggregate level are provided. Users can see student assessment results by any subgroup. Table 9 presents the types of subgroups and subgroup categories provided in the ORS.

Table 9. Types of Subgroups

Breakdown by Category	Displayed Category
<b>Ethnicity</b>	Hispanic or Latino Ethnicity
	American Indian or Alaskan Native
	Asian
	Black or African American
	White
	Native Hawaiian or Other Pacific Islander
	Two or More Races
<b>Gender</b>	Male
	Female
<b>IDEA Indicator</b>	Special Education
	Unknown
<b>Limited English Proficiency Status</b>	Yes
	Unknown
<b>Enrolled Grade</b>	Grade 03
	Grade 04
	Grade 05
	Grade 06
	Grade 07
	Grade 08
	Grade 11


## 4.2 PAPER REPORT

Paper reports for the CTAA were also printed and shipped to the district at the end of the test administration.



Figure 1 shows the mock-up of the family report for students who finished the tests. Figure 2 shows the mock-up for students who stopped early.

Figure 1. Family Report Mock-Up



CONNECTICUT STATE  
DEPARTMENT OF EDUCATION  
**CSDE**

Student Name: **Jane Doe**  
 Grade: **5**  
 Date of Birth: **05/20/2012**  
 SASID: **1234567801**

School: **Demo Elementary School**  
 District: **Demo District**  
 Test Year: **2023**

Spring 2023 Connecticut Alternate Assessment Results

**Dear Parents and Guardians:**

This report shows your child’s scale score and performance level for the 2023 Connecticut Alternate Assessment (CTAA) in English language arts (ELA) and mathematics.

The CTAA content, developed by a group of states and national organizations, is Connecticut’s online alternate assessment for ELA and mathematics for Grades 3–8 and 11. The CTAA assesses students with significant cognitive disabilities and measures content that is derived from Connecticut’s academic standards. The test contains many built-in supports that allow students to take the test using materials they are most familiar with and to communicate what they know and can do as independently as possible. The entire test is designed to be read aloud to the student. In addition, the following built-in supports are provided:

- reduced passage length for the ELA reading passages;
- pictures and other graphics to help students understand what they read (or what is being read to them);
- models for students to use during the ELA and mathematics tests; and
- common geometric shapes and smaller numbers on the mathematics tests.

In order to support independent communication to the greatest extent possible, the CTAA is designed to work with different communication modes and systems. Please discuss the specific ways your child participated on these assessments with your child’s teacher.

The scale score and performance level summarize your child’s achievement based on Connecticut’s academic standards. Descriptors explain the knowledge and skills children at this level generally demonstrate.

You can find more information and resources for helping your child by talking to your child’s teacher and by accessing <https://ct.portal.cambiumast.com/alternate-assessment.html>.



Student Name: **Jane Doe**  
 Grade: **5**  
 Date of Birth: **05/20/2012**  
 SASID: **1234567801**

School: **Demo Elementary School**  
 District: **Demo District**  
 Test Year: **2023**

### Overall Results

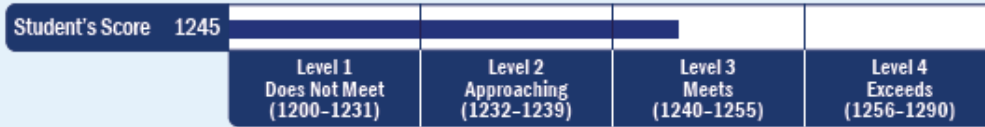
Jane scored at Level 3 on the CTAA English language arts test and scored at Level 2 on the mathematics test.

ELA			✓	
Mathematics		✓		
	Level 1	Level 2	Level 3	Level 4

### ELA Results

Jane's Total Scale Score= 1245

(Scale Score Range 1200-1290)



Your child's performance level is **Level 3: Meets the Achievement Standard**

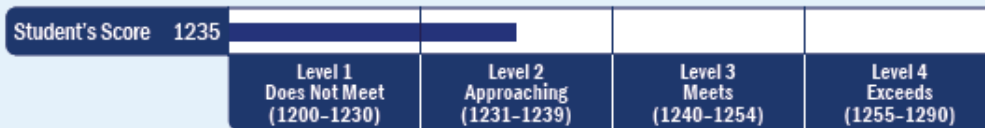
Children performing at this level use built-in supports to show what they know and can do. A child is generally able to: use literary texts with clear to implied ideas and varied sentences to compare characters, settings, and events, summarize a text, answer questions about what the text says, and use context to define multiple meaning words; use informational texts with clear to implied ideas and varied sentences to identify the main idea and supporting details, use details to support an author's point, compare and contrast information and events in different texts, and use context to define multiple meaning words; develop an explanatory text that is organized for a specific text structure and supported with relevant information; and develop a story by identifying beginning, middle, and end.

A student's test scores can vary if tests are taken several times. If Jane were tested again on ELA, the new scale score would probably fall between 1233 and 1257.

### Mathematics Results

Jane's Total Scale Score= 1235

(Scale Score Range 1200-1290)




Your child's performance level is **Level 2: Approaching the Achievement Standard**

Children performing at this level use built-in supports to show what they know and can do. A child is generally able to: solve simple problems with decimals using mathematical language and symbolic representations (e.g., <, >, =); identify place values; round decimal numbers; identify the effects of addition and multiplication; identify a representation of addition of fractions; and convert standard measurements.

A student's test scores can vary if tests are taken several times. If Jane were tested again on mathematics, the new scale score would probably fall between 1223 and 1247.

Figure 2. Family Report Mock-Up for Early Stop Students



Student Name: **Jolyne Doe**

Grade: **5**

Date of Birth: **05/20/2012**

SASID: **1234567803**

School: **Demo Elementary School**

District: **Demo District**

Test Year: **2023**

**Overall Results**

Jolyne did not complete the CTAA English language arts and math tests. No scores were provided.

ELA				
Mathematics				
	<b>Level 1</b>	<b>Level 2</b>	<b>Level 3</b>	<b>Level 4</b>

**ELA Results** Jolyne's Total Scale Score = NA (Scale Score Range 1200-1290)

Your child's ELA test was not completed because your child did not show a consistent observable mode of communication during the test. Your child's teacher followed Connecticut State Department of Education's guidance for the Early Stopping Rule. Please contact your child's teacher to discuss current progress.

Student's Score	NA				
		<b>Level 1 Does Not Meet (1200-1231)</b>	<b>Level 2 Approaching (1232-1239)</b>	<b>Level 3 Meets (1240-1255)</b>	<b>Level 4 Exceeds (1256-1290)</b>

Your child did not receive a score because the assessment was not completed. Your child's teacher may use other measures to capture your child's strengths and abilities.

**Mathematics Results** Jolyne's Total Scale Score = NA (Scale Score Range 1200-1290)

Your child's Math test was not completed because your child did not show a consistent observable mode of communication during the test. Your child's teacher followed Connecticut State Department of Education's guidance for the Early Stopping Rule. Please contact your child's teacher to discuss current progress.

Student's Score	NA				
		<b>Level 1 Does Not Meet (1200-1230)</b>	<b>Level 2 Approaching (1231-1239)</b>	<b>Level 3 Meets (1240-1254)</b>	<b>Level 4 Exceeds (1255-1290)</b>

Your child did not receive a score because the assessment was not completed. Your child's teacher may use other measures to capture your child's strengths and abilities.

## 5. Reliability and Validity

With the implementation of the CTAA tests, both reliability evidence and validity evidence are necessary to support appropriate inferences about students' achievement from the CTAA scores. This section provides empirical evidence about the reliability and validity of the 2022–2023 CTAA, given its intended uses.

Cronbach's alpha, marginal reliability, marginal standard error of measurement (MSEM), conditional standard error of measurement (CSEM), classification accuracy and consistency, internal consistency, and dimensionality are examined for each test.

### 5.1 RELIABILITY

#### 5.1.1 Internal Consistency

Reliability refers to consistency in test scores. Reliability can be defined as the degree to which individuals' deviation scores remain relatively consistent over repeated administrations of the same test or alternate test forms (Crocker & Algina, 1986). For example, if a person takes the same or parallel tests repeatedly, he or she should receive consistent results. The reliability coefficient refers to the ratio of true score variance to observed score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}. \quad (11)$$

There are various approaches for estimating the reliability of scores. Among these approaches, the internal consistency method is employed when it is not possible to conduct repeated test administrations. Whereas other methods often compute the correlation between two separate tests, this method considers each item within a test to be a one-item test. There are several other statistical methods based on this idea: coefficient *alpha* (Cronbach, 1951), Kuder-Richardson Formula 20 (Kuder & Richardson, 1937), Kuder-Richardson Formula 21 (Kuder & Richardson, 1937), stratified coefficient *alpha* (Qualls, 1995), and Feldt-Raju coefficient (Feldt & Qualls, 1996; Feldt & Brennan, 1989). In this report, Cronbach's alpha was computed for each test to assess the internal consistency of items.

Cronbach's alpha indicates how well the items within the test are related. For fixed-form tests, internal consistency can be estimated by Cronbach's coefficient alpha. Alpha coefficients range from 0 to 1. The closer an alpha is to 1, the more reliable the test is. An alpha of 0.8 or above is considered acceptable for tests of modest length.

Cronbach's coefficient alpha was computed as

$$\alpha = \frac{n}{n-1} \left[ 1 - \frac{\sum_{i=1}^n \sigma_i^2}{\sigma_x^2} \right], \quad (12)$$

where  $n$  is the sample size and  $\sigma_i^2$  is the raw score variance for item  $i$ .  $\sigma_x^2$  is the variance of the total raw scores.

The Cronbach's alpha coefficients are summarized in Table 10. The computation of Cronbach's alpha requires the full response matrix; therefore, the sample sizes are smaller. Mathematics

grades 5 has the lowest alpha coefficient, 0.65. According to Nunnally (1978), 0.7 is the minimum acceptable alpha coefficient. Therefore, all tests meet the minimum acceptable requirement except for Mathematics grades 5 and grade 7.

Table 10. Cronbach's Alpha

Subject	Grade	Sample Size	Number Items	Alpha
ELA	3	337	41	0.87
ELA	4	353	40	0.87
ELA	5	381	31	0.78
ELA	6	334	32	0.81
ELA	7	304	32	0.8
ELA	8	304	35	0.79
ELA	11	302	29	0.83
Mathematics	3	389	40	0.84
Mathematics	4	401	37	0.8
Mathematics	5	415	36	0.65
Mathematics	6	366	39	0.73
Mathematics	7	335	35	0.67
Mathematics	8	335	35	0.77
Mathematics	11	337	31	0.79

### 5.1.2 Marginal Reliability

Marginal reliability (Sireci, Thissen, & Wainer, 1991) is a measure of the overall reliability of the test based on the average conditional standard errors, estimated at different points on the achievement scale, for all students.

Within the IRT framework, measurement error varies across the range of ability. The amount of precision is indicated by the test information at any given point of a distribution. The inverse of the test information function represents the SEM, which is equal to the inverse square root of information. The larger the measurement error, the less test information is being provided. The amount of test information provided is at its maximum for students toward the center of the distribution, as opposed to students with more-extreme scores. Conversely, measurement error is minimal for the part of the underlying scale that is at the middle of the test distribution and greater on scaled values farther away from the middle.

Specifically, marginal reliability is based on the average CSEM estimated at different points on the achievement scale. The true score variance is the observed score variance minus the error variance. The marginal reliability ( $\bar{\rho}$ ) is computed as

$$\bar{\rho} = \left( \frac{\sigma_{true}^2}{\sigma_{obs}^2} \right) = \left( \frac{\sigma_{obs}^2 - \bar{\sigma}_{err}^2}{\sigma_{obs}^2} \right) \quad (13)$$

$$\bar{\sigma}_{err}^2 = \int \sigma_{err}^2 p(\theta) d\theta = \frac{\sum \sigma_{err}^2}{N}$$

where  $\sigma_{true}^2$  is true score variance,  $\sigma_{obs}^2$  is the observed score variance,  $\bar{\sigma}_{err}^2$  is the error variance,  $\sigma_{err}^2$  is the square of the CSEM at the ability estimate of each student, and  $N$  is the number of students. The maximum marginal reliability index is 1. A greater index indicates a greater precision of scores.

Another way to examine score reliability is MSEM computed as the square root of  $\bar{\sigma}_{err}^2$ . A smaller MSEM indicates a greater accuracy of scores. The marginal reliability  $\bar{\rho}$  and the test MSEM behave oppositely. The higher the  $\bar{\rho}$ , the lower the MSEM, and vice versa.

The 2023 results of marginal reliability, MSEM, and standard deviation (STD) of scale scores by test are listed in Table 11. It shows that, except for the ELA grade 11 test, the marginal reliability estimates exceed 0.80. The form MSEMs are about one third of the STD of scale scores. The results suggest that the test scores are mostly precisely estimated. The SEM is within a reasonable range. The results further indicate that the forms are statistically reliable in measuring student abilities.

For the ELA grade 11 test, in the conversion table in Appendix D, the CSEM at the maximum raw score point 25 is 43.7, which is significantly higher than the CSEMs for other score points in this form and others. It indicates that the test needed more difficult items for high-ability students. Eleven students earned a raw score of 25. Removing the 11 students, the marginal reliability, STD, MSEM, and the MSEM/STD become 0.81, 10.33, 4.54, and 0.44, respectively. The CSEM curve is steeper when scale scores go to both ends, which also indicates that more items are needed to better cover the scale score range. In addition, the conversion table shows that there are only 25 score points in this form. A shorter test will lower test reliability.

Table 11. Marginal Reliability and Marginal Standard Error of Measurement

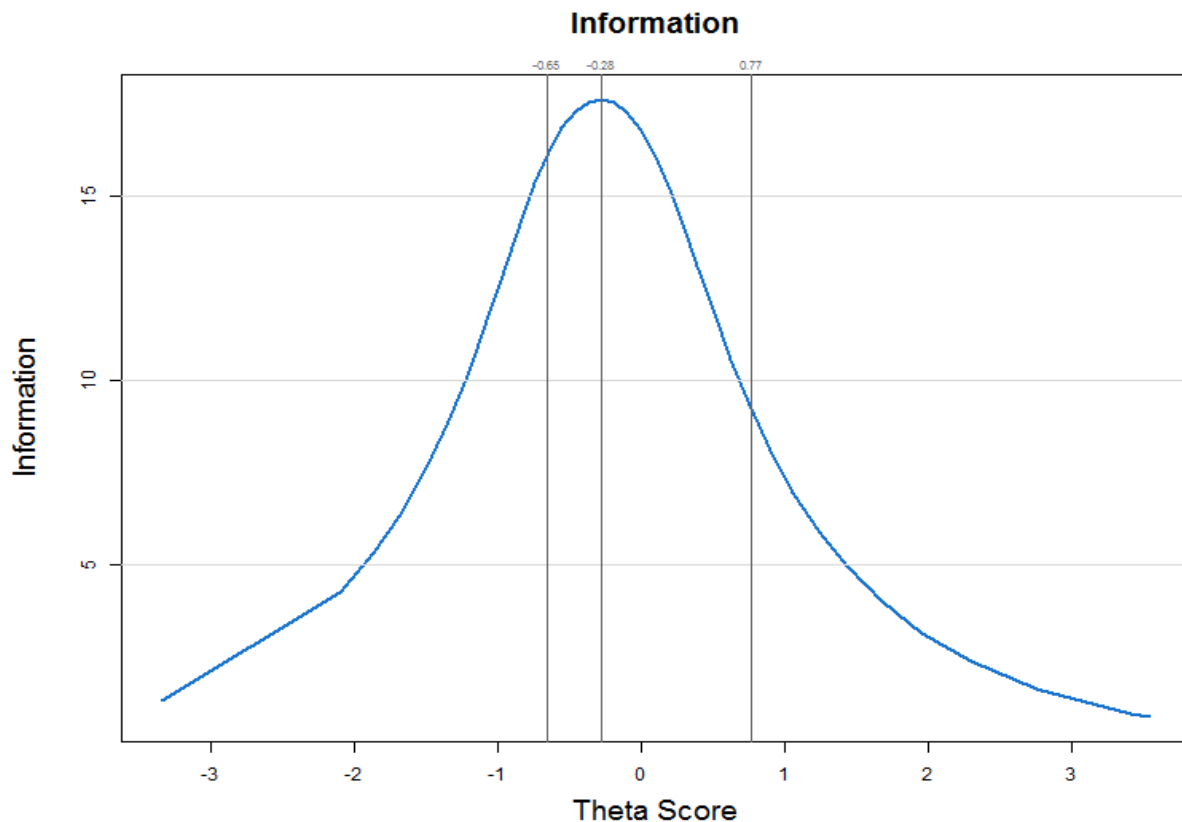
Subject	Grade	Sample Size	Marginal Reliability	STD	MSEM	MSEM/STD
ELA	3	438	0.87	13.94	5.04	0.36
ELA	4	432	0.86	12.61	4.73	0.38
ELA	5	458	0.84	13.88	5.55	0.40
ELA	6	421	0.85	14.24	5.54	0.39
ELA	7	380	0.84	13.11	5.22	0.40
ELA	8	381	0.85	14.12	5.42	0.38
ELA	11	393	0.53	14.47	9.88	0.68
Mathematics	3	437	0.89	13.92	4.72	0.34
Mathematics	4	429	0.84	13.91	5.64	0.41
Mathematics	5	457	0.81	12.34	5.37	0.44
Mathematics	6	418	0.86	12.62	4.77	0.38
Mathematics	7	381	0.81	11.66	5.07	0.43
Mathematics	8	380	0.85	13.75	5.35	0.39
Mathematics	11	391	0.83	13.37	5.45	0.41

### 5.1.3 Conditional Standard Error of Measurement

Within the IRT framework, measurement error varies across the range of ability as a result of the test information function (TIF). The TIF describes the amount of information provided by the test at each score point along the ability continuum. The inverse of the TIF is characterized as the conditional measurement error at each score point. For instance, if the measurement error is large, then less information is being provided by the assessment at the specific ability level.

Figure 3 displays a sample TIF from the CTAA mathematics grade 3 test. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most-precise scores in this range. The curve is lower at the tails, which indicates that the test provides less information about examinees at the tails relative to the center. The vertical lines are samples of the performance cuts.

Figure 3. Sample Test Information Function



The standard error for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

$$se(\theta_i) = \frac{1}{\sqrt{TIF(\theta_i)}} \quad (14)$$

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the standard errors are more useful for score interpretation. For this reason, standard error plots are



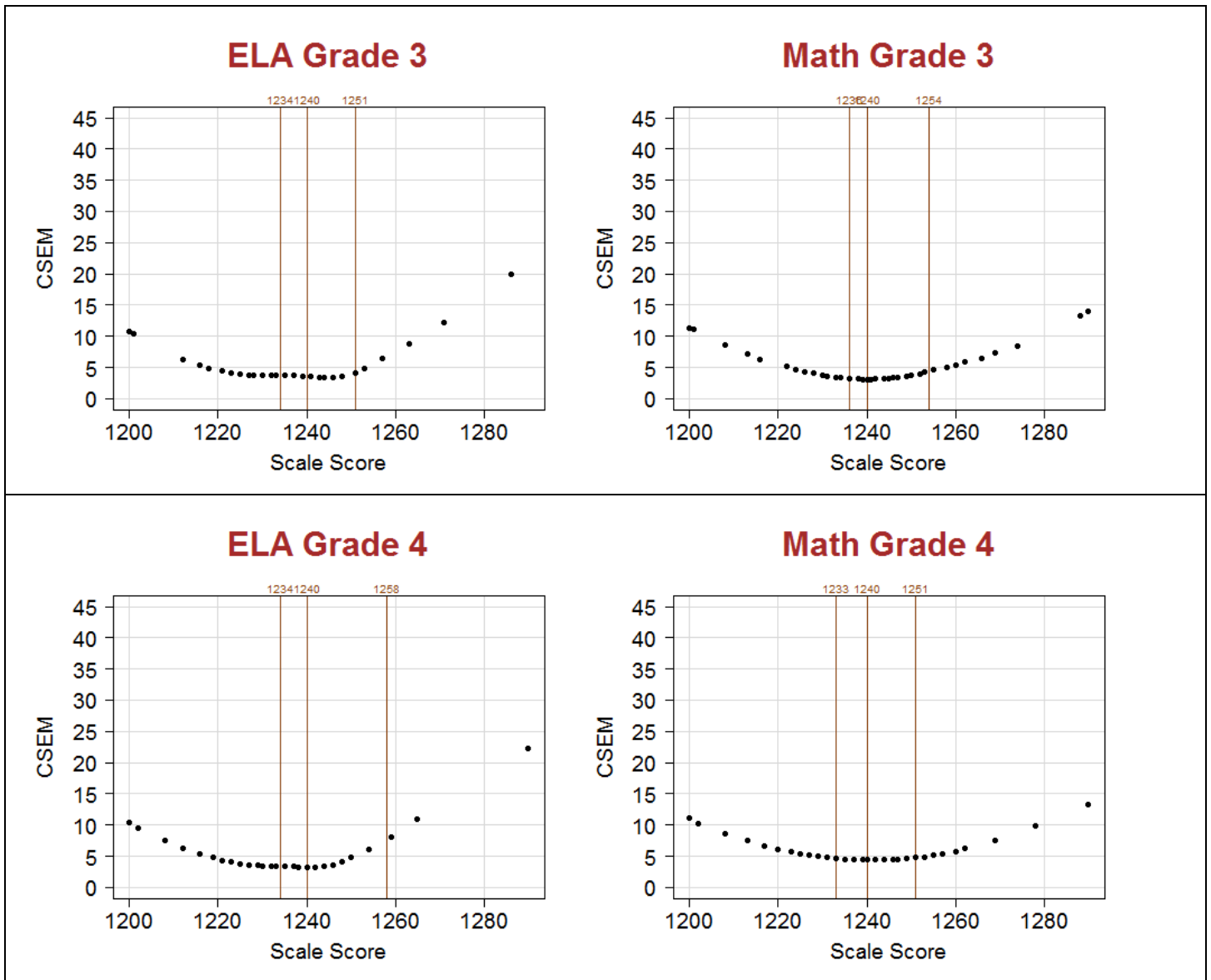
presented in Figure 4, instead of the TIFs. These plots are based on the scaled scores reported in 2023. Vertical lines represent the three performance category cut scores.

As described in Section 3.2, Scoring Method Review, the CSEM is computed as

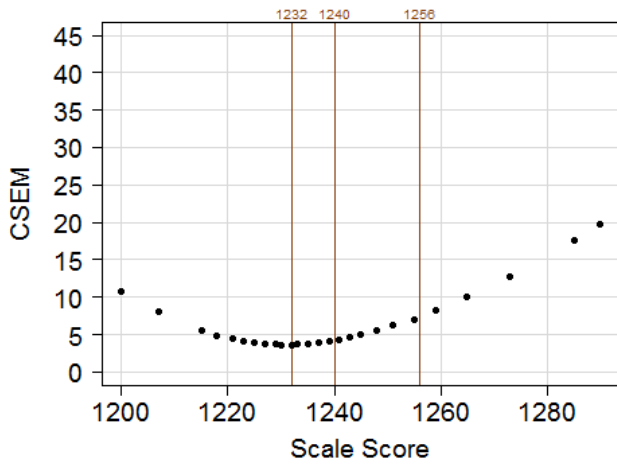
$$se(\theta) = \frac{1}{\sqrt{-\left(\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta}\right)}}, \quad (15)$$

where  $\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta}$  is the second derivative of the log-likelihood with respect to  $\theta$ .

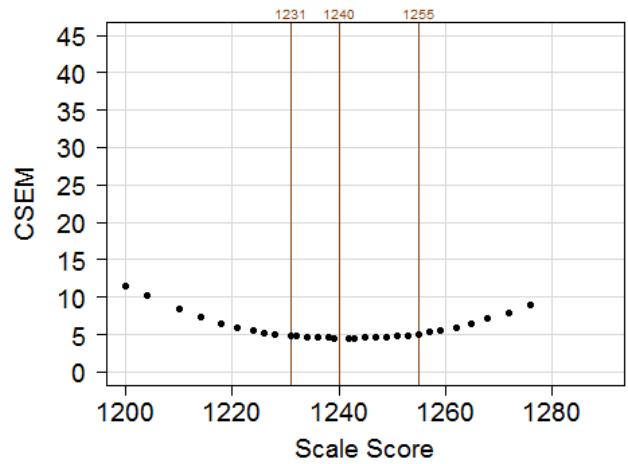
Figure 4. CSEM by Test



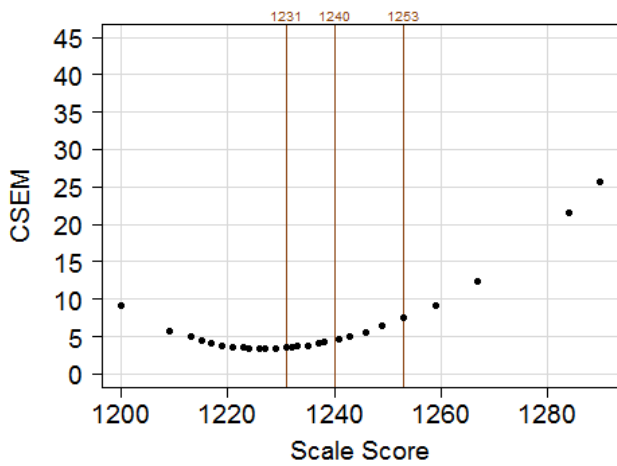
### ELA Grade 5



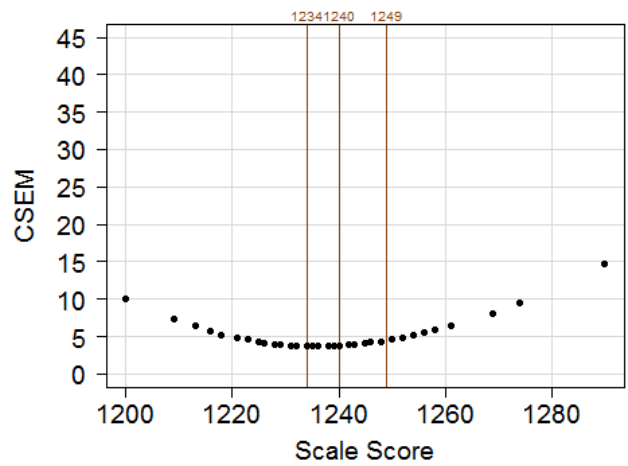
### Math Grade 5



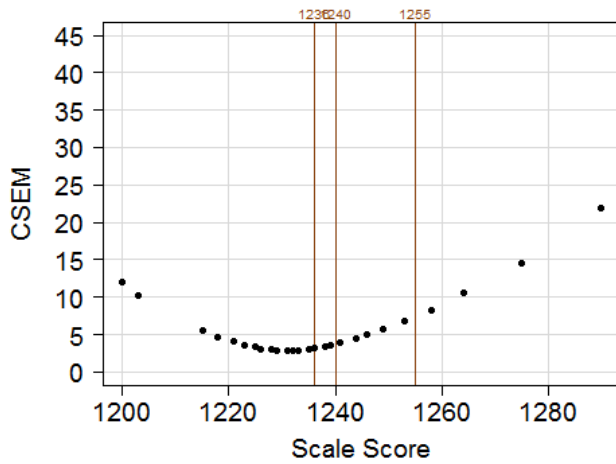
### ELA Grade 6



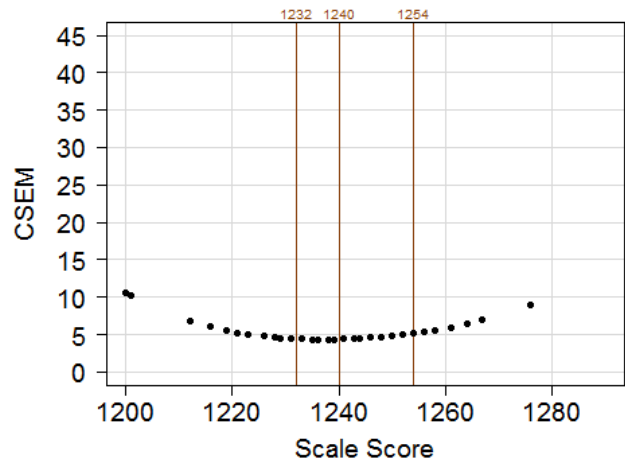
### Math Grade 6



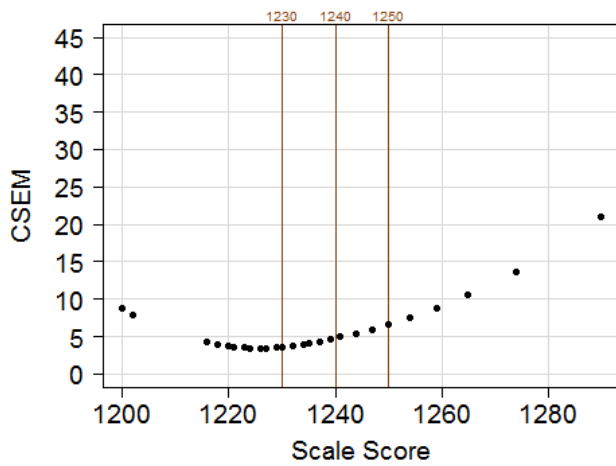
### ELA Grade 7



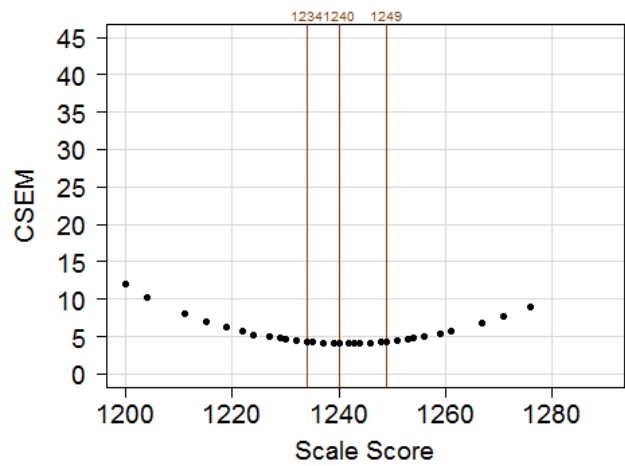
### Math Grade 7

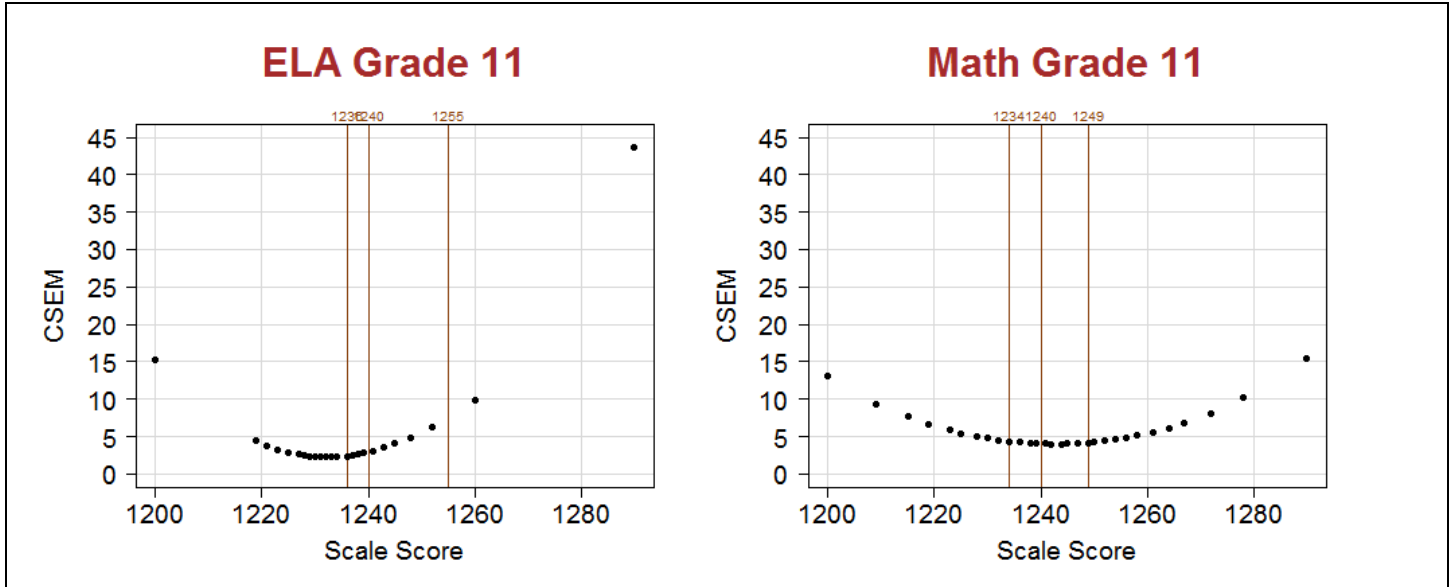


### ELA Grade 8



### Math Grade 8





Generally, the relationship between CSEM and scale score is U-shaped, with larger CSEMs towards the ends of the scale and smaller CSEMs towards the middle. That is because there are more items with medium difficulties in each test, which leads to greater measurement information and, therefore, lower SEM in the middle range.

Compared with other tests, the CSEMs for the ELA grade 11 test increased more rapidly when scale scores go to both ends on the  $x$ -axis, which lead to lower reliability of the test. The reason is that the CSEMs for the extreme scores are higher, and the test is shorter, with only 25 score points, as shown in the conversion table, and fewer items at the middle range.

#### 5.1.4 Classification Accuracy and Consistency

When student performance is reported in terms of achievement levels, the reliability of achievement classification is evaluated in terms of the probabilities of consistent classification of students as specified in Standard 2.16 in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Both classification accuracy and consistency are examined.

Classification accuracy refers to the degree to which a student's true score and observed score would fall within the same performance level (Rudner, 2001). Classification consistency refers to the degree to which examinees are classified into the same performance level assuming the test is administered twice independently (Lee, Hanson, & Brennan, 2002) or, in other words, the percentages of students who are consistently classified in the same performance levels on two equivalent test forms. In reality, the true ability is unknown, and students do not take an alternate, equivalent form; therefore, classification accuracy and consistency are estimated based on students' item scores and the item parameters, and the assumed underlying latent ability distribution. For the CTAA tests, the classification accuracy and classification consistency are examined at each performance level using the Rudner classification index (Rudner, 2005).

For the  $i$ th student, the student's estimated ability is  $\hat{\theta}_i$  with an SEM of  $se(\hat{\theta}_i)$ , and the estimated ability is distributed, as  $\hat{\theta}_i \sim N(\theta_i, se(\hat{\theta}_i))$ , assuming a normal distribution, where  $\theta_i$  is the unknown true ability of the  $i$ th student. The probability of the true score at achievement level 1 based on the cut scores  $c_{l-1}$  and  $c_l$  is estimated as

$$p_{il} = p(c_{l-1} \leq \theta_i < c_l) = p\left(\frac{c_{l-1} - \hat{\theta}_i}{se(\hat{\theta}_i)} \leq \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} < \frac{c_l - \hat{\theta}_i}{se(\hat{\theta}_i)}\right) = p\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)} \leq \frac{\hat{\theta}_i - \theta_i}{se(\hat{\theta}_i)} < \frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) = \Phi\left(\frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) - \Phi\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)}\right). \quad (16)$$

For level 1,  $c_0 = -\infty$ , and for level  $L$ ,  $c_L = \infty$ .

### Classification Accuracy

Using  $p_{il}$ , we can construct an  $L \times L$  table as

$$\begin{pmatrix} n_{a11} & \cdots & n_{a1L} \\ \vdots & \vdots & \vdots \\ n_{aL1} & \cdots & n_{aLL} \end{pmatrix},$$

where  $n_{alm} = \sum_{p_{l_i=l}} p_{im}$ ,  $p_{l_i}$  is the  $i$ th student's achievement level. In the above table, the row represents the observed level and the column represents the expected level.

Based on the above table, the classification accuracy (CA) for the cut  $c_l$  ( $l = 1, \dots, L - 1$ ) is estimated by

$$CA_{c_l} = \frac{\sum_{k,m=1}^l n_{akm} + \sum_{k,m=l+1}^L n_{akm}}{N}, \quad (17)$$

where  $N$  is the total number of students.

The overall classification accuracy is computed as

$$CA = \frac{\sum_{i=1}^L n_{aii}}{N}. \quad (18)$$

### Classification Consistency

Using  $p_{il}$ , similar to accuracy, we can construct another  $L \times L$  table by assuming that the test is administered twice independently to the same student group; hence we have

$$\begin{pmatrix} n_{c11} & \cdots & n_{c1L} \\ \vdots & \vdots & \vdots \\ n_{cL1} & \cdots & n_{cLL} \end{pmatrix},$$

where  $n_{clm} = \sum_{i=1}^N p_{il} p_{im}$ .

Based on the previously mentioned table, the classification consistency (CC) for the cut  $c_l$  ( $l = 1, \dots, L - 1$ ) is estimated by

$$CC_{c_l} = \frac{\sum_{k,m=1}^l n_{ckm} + \sum_{k,m=l+1}^L n_{ckm}}{N}. \quad (19)$$

The overall classification consistency is computed as

$$CC = \frac{\sum_{i=1}^L n_{cti}}{N}. \quad (20)$$

Besides the overall CA and CC for each test, CA and CC analyses were also conducted for each cut point. The early stopped students were excluded from the analysis. The result is shown in Table 12. The overall classification accuracy of the test ranges from 0.75 to 0.82 for ELA, and from 0.69 to 0.75 for mathematics. The overall cut accuracy rates are much higher, denoting that the degree to which we can reliably differentiate students between adjacent performance levels is typically above or close to 0.9. The overall classification consistency values are from 0.66 to 0.75 for ELA, and from 0.59 to 0.69 for mathematics. The classification consistency values for each cut are near or above 0.8.

In all performance levels, classification accuracy is slightly higher than classification consistency. Classification consistency rates can be lower than classification accuracy because the consistency is based on two tests with measurement errors, while the accuracy is based on one test with a measurement error and the true score. The accuracy and consistency rates for each performance level are higher for the levels with smaller standard error.

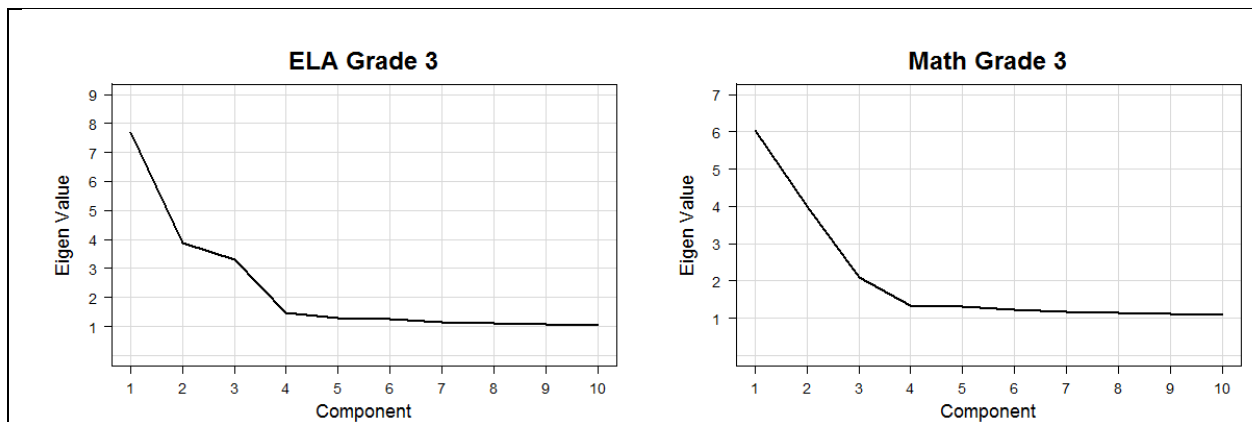
Table 12. Classification Accuracy and Consistency

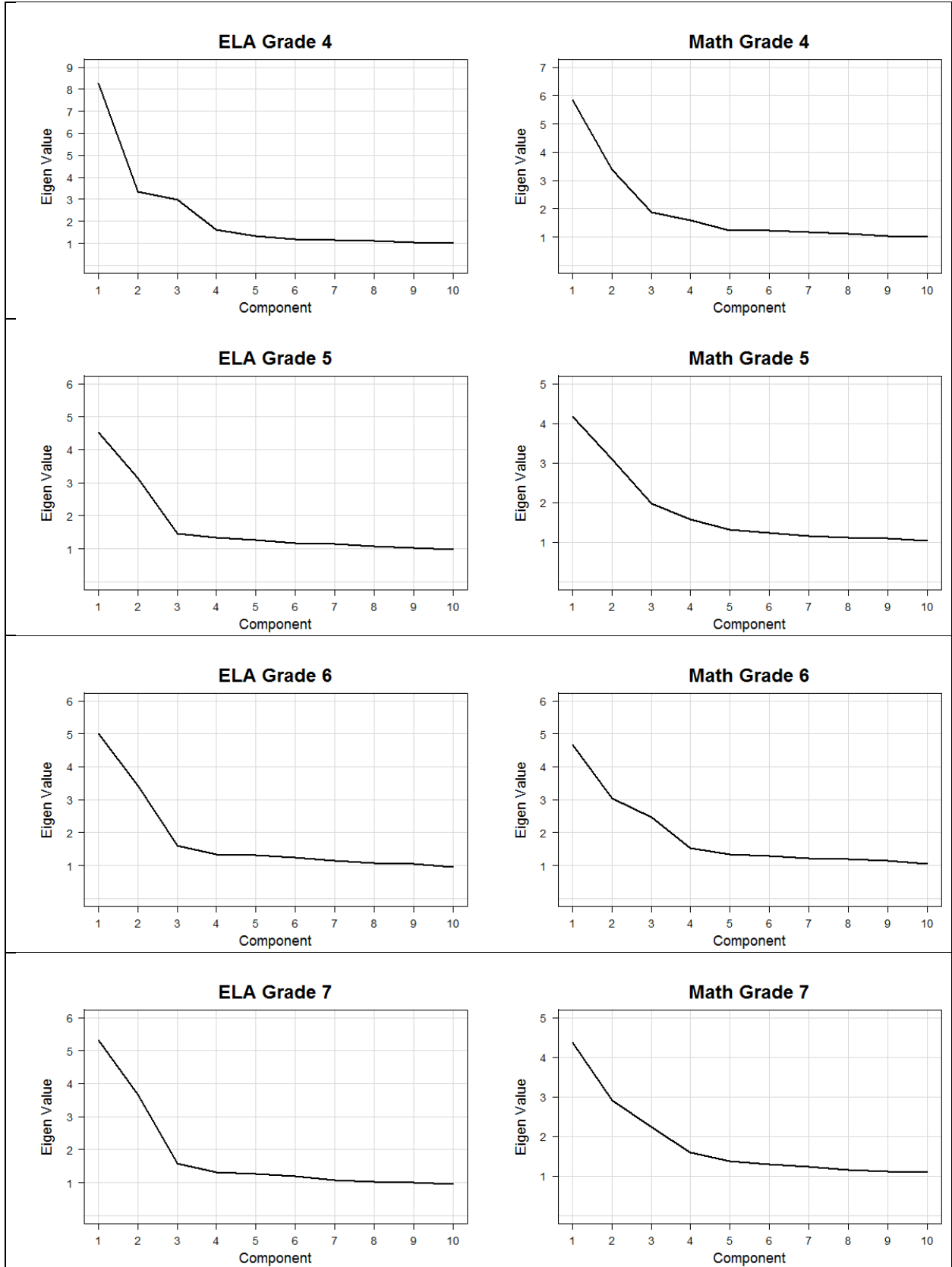
Subject	Grade	Count	CA. Overall	CC. Overall	CA.Cut1	CA.Cut2	CA.Cut3	CC.Cut1	CC.Cut2	CC.Cut3
ELA	3	438	0.77	0.69	0.89	0.93	0.95	0.84	0.89	0.93
ELA	4	432	0.82	0.75	0.91	0.92	0.98	0.87	0.89	0.97
ELA	5	458	0.75	0.66	0.87	0.90	0.97	0.82	0.86	0.96
ELA	6	421	0.78	0.70	0.89	0.93	0.96	0.85	0.89	0.95
ELA	7	380	0.78	0.72	0.89	0.92	0.96	0.85	0.88	0.94
ELA	8	381	0.76	0.68	0.88	0.92	0.95	0.84	0.89	0.93
ELA	11	393	0.77	0.69	0.90	0.91	0.95	0.85	0.86	0.93
Math	3	437	0.75	0.69	0.87	0.91	0.97	0.83	0.87	0.96
Math	4	429	0.74	0.66	0.89	0.89	0.96	0.84	0.85	0.94
Math	5	457	0.71	0.61	0.89	0.84	0.97	0.85	0.78	0.96
Math	6	418	0.72	0.63	0.84	0.90	0.97	0.79	0.86	0.95
Math	7	381	0.69	0.59	0.84	0.87	0.97	0.78	0.82	0.96
Math	8	380	0.69	0.60	0.89	0.88	0.91	0.85	0.83	0.88
Math	11	391	0.69	0.61	0.87	0.86	0.95	0.82	0.81	0.93

### 5.1.5 Principal Component Analysis

The test dimensionality is investigated using principal component analysis (PCA) with an orthogonal rotation method (Jolliffe, 2002; Cook, Kallen, & Amtmann, 2009). The results are presented in the scree plots in Figure 5. The graphs show that the first three components explain the majority of the variation. Table 13 shows the eigenvalues of the first three components. The PCA results does not necessarily suggest that the forms measure one dominant construct. To further investigate the dimensionality, analyses using correlations among the standards and Q3 statistics for local independence are performed, and the results are presented in the following sections.

Figure 5. Scree Plots







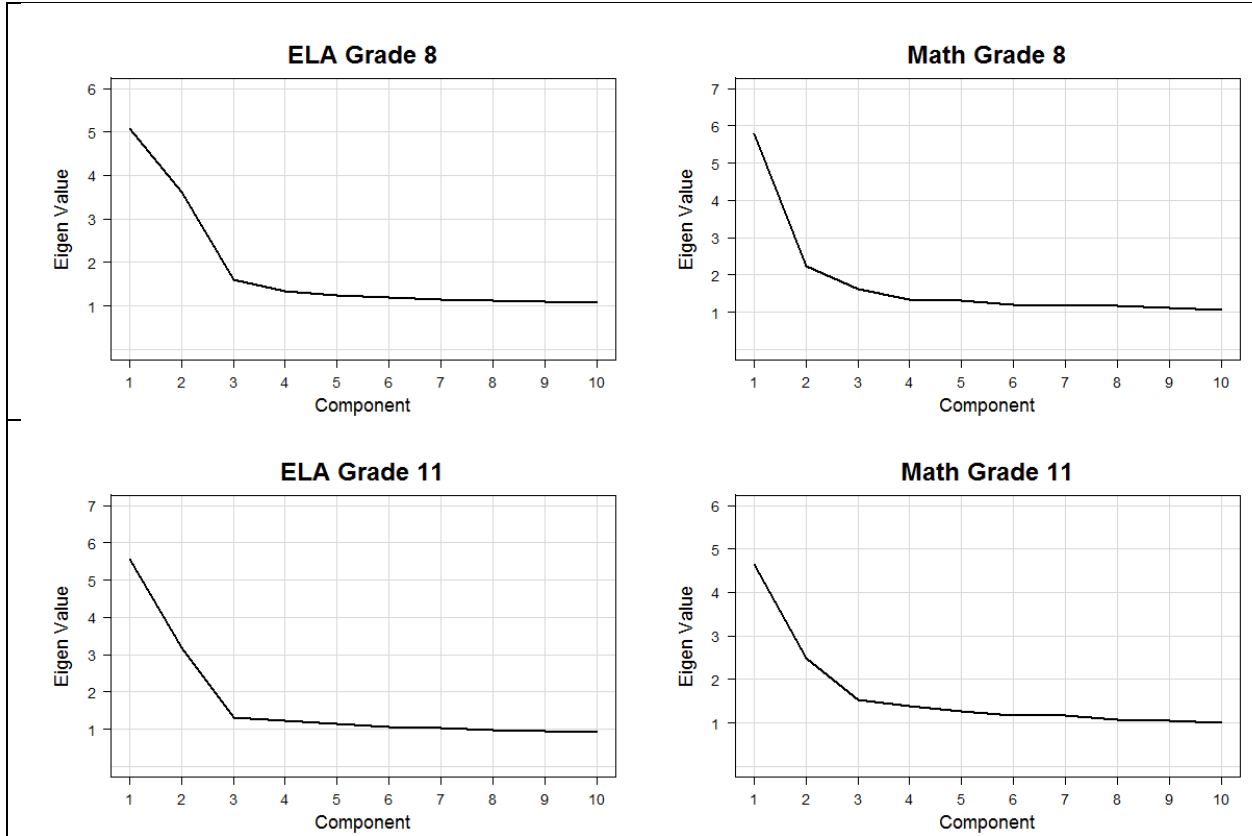


Table 13. Eigenvalues of The First Three Components

ELA							
Component	3	4	5	6	7	8	11
1	7.6823	8.2707	4.5299	5.0072	5.3140	5.0698	5.5655
2	3.8657	3.3413	3.1248	3.4190	3.6662	3.6107	3.1863
3	3.3082	2.9778	1.4660	1.6066	1.5789	1.6019	1.3072
Mathematics							
Component	3	4	5	6	7	8	11
1	6.0240	5.8491	4.1798	4.6668	4.3732	5.7750	4.6544
2	4.0073	3.3783	3.0918	3.0358	2.9165	2.2406	2.4819
3	2.1073	1.8716	1.9729	2.4680	2.2246	1.6282	1.5302

### Correlations Among Strand Scores

This section explores the internal structure of the assessment using the scores provided at the strand level. The relationship of the subscores is just one indicator of the test dimensionality.

In ELA grades 3 and 4, there are five standards per grade: Reading-Literature, Reading-Informational Texts, Reading-Foundational Skills, Language, and Writing. Grades 5–8 and grade 11 have the same standards, with the exception of Reading-Foundational Skills. In mathematics, strand levels differ in each grade or course (see Table 1 for details).

Raw scores based on each standard were computed for this analysis even though these scores were not reported to students. It may not be reasonable to expect that the strand scores are completely orthogonal. This would suggest that there are no relationships among strand scores and would make justifying a unidimensional IRT model difficult. However, if the standards were perfectly correlated, one could justify a unidimensional model.

One pathway to explore the internal structure of the test is to look at observed correlations between the subscores. However, as each standard is measured with a small number of items, the standard errors of the observed scores within each standard are typically larger than the standard error of the total test score. Disattenuating for measurement error could offer some insight into the theoretical true score correlations. Both observed correlations and disattenuated correlations are provided in the following paragraph.

Table 14 through Table 17 present the observed and disattenuated correlation matrix of the strand raw scores for each subject area. In ELA, the correlations among the standards range from 0.15 to 0.61. Reading-Foundational Skills items exhibited lower correlations with other standards. For mathematics, the correlations were between 0.00 and 0.68. Operations and Algebraic Thinking in grade 5 showed lower correlations with other strands. Negative values are reported as 0.00\*\*.

In some instances, these correlations were lower than one might expect. However, as previously noted, the correlations were subject to a large amount of measurement error at the strand level, given the limited number of items from which the scores were derived. Consequently, over-interpretation of these correlations, as either high or low, should be made cautiously.

Disattenuated values greater than 1.00 are reported as 1.00\*\*. In ELA, the minimum was 0.22, and the average was 0.79. In mathematics, the minimum was 0.14, and the average was 0.68.

*Table 14. Observed Correlation Matrix Among Standards (ELA)*

GRADE	STANDARDS	NUMBER OF ITEMS	CAT1	CAT2	CAT3	CAT4	CAT5
3	Reading-Foundational Skills (Cat1)	10	1.00				
	Reading-Informational Text (Cat2)	10	0.32	1.00			
	Reading-Literature(Cat3)	9	0.15	0.58	1.00		
	Writing (Cat4)	10	0.32	0.61	0.53	1.00	
	Language(Cat5)*	2					
4	Reading-Foundational Skills (Cat1)	10	1.00				
	Reading-Informational Text (Cat2)	9	0.29	1.00			
	Reading-Literature(Cat3)	10	0.3	0.5	1.00		
	Writing (Cat4)	8	0.29	0.45	0.49	1.00	
	Language(Cat5)*	3					
5	Reading-Informational Text (Cat1)	9	1.00				
	Reading-Literature(Cat2)	11	0.48	1.00			
	Writing (Cat3)	8	0.47	0.55	1.00		
	Language(Cat4)*	3					

GRADE	STANDARDS	NUMBER OF ITEMS	CAT1	CAT2	CAT3	CAT4	CAT5
6	Reading-Informational Text (Cat1)	10	1.00				
	Reading-Literature(Cat2)	10	0.5	1.00			
	Writing (Cat3)	8	0.47	0.5	1.00		
	Language(Cat4)	4	0.44	0.55	0.37	1.00	
7	Reading-Informational Text (Cat1)	8	1.00				
	Reading-Literature(Cat2)	10	0.54	1.00			
	Writing (Cat3)	10	0.52	0.54	1.00		
	Language(Cat4)	4	0.53	0.54	0.52	1.00	
8	Reading-Informational Text (Cat1)	11	1.00				
	Reading-Literature(Cat2)	10	0.49	1.00			
	Writing (Cat3)	10	0.55	0.48	1.00		
	Language(Cat4)	4	0.34	0.5	0.38	1.00	
11	Reading-Informational Text (Cat1)	9	1.00				
	Reading-Literature(Cat2)	7	0.58	1.00			
	Writing (Cat3)	9	0.54	0.58	1.00		
	Language(Cat4)	4	0.57	0.59	0.54	1.00	

\* Correlations were not computed for the standard with the number of items < 4

Table 15. Disattenuated Correlation Matrix Among Standards (ELA)

GRADE	STANDARDS	NUMBER OF ITEMS	CAT1	CAT2	CAT3	CAT4	CAT5
3	Reading-Foundational Skills (Cat1)	10	1.00				
	Reading-Informational Text (Cat2)	10	0.43	1.00			
	Reading-Literature (Cat3)	9	0.22	1.00**	1.00		
	Writing (Cat4)	10	0.44	1.00**	0.99	1.00	
	Language(Cat5)*	2					
4	Reading-Foundational Skills (Cat1)	10	1.00				
	Reading-Informational Text (Cat2)	9	0.44	1.00			
	Reading-Literature(Cat3)	10	0.41	0.97	1.00		
	Writing (Cat4)	8	0.47	1.00**	1.00**	1.00	
	Language(Cat5)*	3					
5	Reading-Informational Text (Cat1)	9	1.00				
	Reading-Literature(Cat2)	11	0.97	1.00			
	Writing (Cat3)	8	0.95	1.00**	1.00		
	Language(Cat4)*	3					
6	Reading-Informational Text (Cat1)	10	1.00				
	Reading-Literature(Cat2)	10	0.85	1.00			
	Writing (Cat3)	8	1.00**	0.99	1.00		

GRADE	STANDARDS	NUMBER OF ITEMS	CAT1	CAT2	CAT3	CAT4	CAT5
	Language(Cat4)	4	0.89	0.97	0.86	1.00	
7	Reading-Informational Text (Cat1)	8	1.00				
	Reading-Literature(Cat2)	10	0.92	1.00			
	Writing (Cat3)	10	1.00**	1.00**	1.00		
	Language(Cat4)	4	1.00**	1.00**	1.00**	1.00	
8	Reading-Informational Text (Cat1)	11	1.00				
	Reading-Literature(Cat2)	10	0.93	1.00			
	Writing (Cat3)	10	1.00**	0.89	1.00		
	Language(Cat4)	4	0.71	0.99	0.77	1.00	
11	Reading-Informational Text (Cat1)	9	1.00				
	Reading-Literature(Cat2)	7	0.99	1.00			
	Writing (Cat3)	9	1.00**	0.96	1.00		
	Language(Cat4)	4	1.00**	1	1.00**	1.00	

\* Correlations were not computed for the standard with the number of items < 4;

\*\* Correlations were marked as \*\* for those values less than 0.00 or greater than 1.00.

Table 16. Observed Correlation Matrix Among Standards (Mathematics)

GRADE	STANDARDS	NUMBER OF ITEMS	CAT1	CAT2	CAT3	CAT4	CAT5
3	Geometry (Cat1)	4	1.00				
	Measurement & Data (Cat2)	8	0.4	1.00			
	Number and Operations Base Ten (Cat3)	8	0.28	0.48	1.00		
	Number and Operations Fractions (Cat4)	8	0.26	0.4	0.42	1.00	
	Operations and Algebraic Thinking (Cat5)	12	0.37	0.68	0.59	0.47	1.00
4	Geometry (Cat1)	4	1.00				
	Measurement & Data (Cat2)	8	0.39	1.00			
	Number and Operations Base Ten (Cat3)	4	0.25	0.43	1.00		
	Number and Operations Fractions (Cat4)	9	0.08	0.4	0.36	1.00	
	Operations and Algebraic Thinking (Cat5)	12	0.26	0.54	0.44	0.41	1.00
5	Geometry (Cat1)	4	1.00				
	Measurement & Data (Cat2)	7	0.2	1.00			
	Number and Operations Base Ten (Cat3)	15	0.28	0.3	1.00		
	Number and Operations Fractions (Cat4)	6	0.29	0.41	0.21	1.00	
	Operations and Algebraic Thinking (Cat5)	4	0.05	0.00**	0.21	0.00**	1.00
6	Geometry (Cat1)	4	1.00				
	The Number System (Cat2)	12	0.34	1.00			
	Ratio and Proportions (Cat3)	12	0.24	0.56	1.00		
	Statistics & Probability (Cat4)	4	0.27	0.41	0.36	1.00	

GRADE	STANDARDS	NUMBER OF ITEMS	CAT1	CAT2	CAT3	CAT4	CAT5
	Expressions & Equations (Cat5)	7	0	0.23	0.27	0.14	1.00
7	Geometry (Cat1)	8	1.00				
	The Number System (Cat2)	8	0.16	1.00			
	Ratio and Proportions (Cat3)	12	0.25	0.17	1.00		
	Expressions & Equations (Cat4)	4	0.23	0.34	0.2	1.00	
	Statistics & Probability (Cat5)*	3					
8	Expressions & Equations (Cat1)	8	1.00				
	Functions (Cat2)	8	0.59	1.00			
	Geometry (Cat3)	8	0.36	0.23	1.00		
	Statistics & Probability (Cat4)	8	0.5	0.51	0.28	1.00	
	The Number System (Cat5)*	3					
11	Algebra and Functions (Cat1)	16	1.00				
	Number & Quantity (Cat2)	7	0.48	1.00			
	Statistics & Probability (Cat3)	6	0.49	0.45	1.00		
	Geometry (Cat4)*	2					

\* Correlations were not computed for the standard with the number of items < 4;

\*\* Correlations were marked as \*\* for those values less than 0.00 or greater than 1.00.

Table 17. Disattenuated Correlation Matrix Among Standards (Mathematics)

GRADE	STANDARDS	NUMBER OF ITEMS	CAT1	CAT2	CAT3	CAT4	CAT5
3	Geometry (Cat1)	4	1.00				
	Measurement & Data (Cat2)	8	0.73	1.00			
	Number and Operations Base Ten (Cat3)	8	0.63	0.92	1.00		
	Number and Operations Fractions (Cat4)	8	0.66	0.85	1.00**	1.00	
	Operations and Algebraic Thinking (Cat5)	12	0.68	1.00**	1.00**	1.00**	1.00
4	Geometry (Cat1)	4	1.00				
	Measurement & Data (Cat2)	8	0.64	1.00			
	Number and Operations Base Ten (Cat3)	4	0.57	0.98	1.00		
	Number and Operations Fractions (Cat4)	9	0.14	0.74	0.91	1.00	
	Operations and Algebraic Thinking (Cat5)	12	0.44	0.93	1.00**	0.78	1.00
5	Geometry (Cat1)	4	1.00				
	Measurement & Data (Cat2)	7	0.53	1.00			
	Number and Operations Base Ten (Cat3)	15	0.69	0.72	1.00		
	Number and Operations Fractions (Cat4)	6	0.69	0.98	0.46	1.00	
	Operations and Algebraic Thinking (Cat5)	4	0.24	0.00**	1.00**	0.00**	1.00
6	Geometry (Cat1)	4	1.00				

GRADE	STANDARDS	NUMBER OF ITEMS	CAT1	CAT2	CAT3	CAT4	CAT5
	The Number System (Cat2)	12	0.79	1.00			
	Ratio and Proportions (Cat3)	12	0.56	1.00**	1.00		
	Statistics & Probability (Cat4)	4	1.00**	1.00**	1.00**	1.00	
	Expressions & Equations (Cat5)	7	0.00**	0.56	0.64	1.00**	1.00
7	Geometry (Cat1)	8	1.00				
	The Number System (Cat2)	8	0.38	1.00			
	Ratio and Proportions (Cat3)	12	0.49	0.42	1.00		
	Expressions & Equations (Cat4)	4	0.99	1.00**	0.95	1.00	
	Statistics & Probability (Cat5)*	3					
8	Expressions & Equations (Cat1)	8	1.00				
	Functions (Cat2)	8	1.00**	1.00			
	Geometry (Cat3)	8	0.95	0.56	1.00		
	Statistics & Probability (Cat4)	8	0.84	0.82	0.62	1.00	
	The Number System (Cat5)*	3					
11	Algebra and Functions (Cat1)	16	1.00				
	Number & Quantity (Cat2)	7	0.77	1.00			
	Statistics & Probability (Cat3)	6	1.00**	1.00**	1.00		
	Geometry (Cat4)*	2					

\* Correlations were not computed for the standard with the number of items < 4;

\*\* Correlations were marked as \*\* for those values less than 0.00 or greater than 1.00.

### 5.1.6 Local Independence

The validity of the application of IRT depends greatly on meeting the underlying assumptions of the models. One such assumption is local independence, which means that for a given proficiency estimate, the (marginal) likelihood is maximized, assuming the probability of correct responses is the product of independent probabilities over all items (Chen & Thissen, 1997):

$$L(\theta) = \int \prod_{j=1}^K \Pr(x_j | \theta) f(\theta) d\theta. \quad (21)$$

When local independence is not met, there are issues of multidimensionality that are unaccounted for in the modeling of the data (Bejar, 1980). In fact, Lord (1980) noted that “local independence follows automatically from unidimensionality” (as cited in Bejar, 1980, p. 5). From a dimensionality perspective, there may be nuisance factors that are influencing relationships among certain items, after accounting for the intended construct of interest. These nuisance factors can be influenced by a number of testing features, such as speededness, fatigue, item chaining, and item or response formats (Yen, 1993).

Yen’s  $Q_3$  statistic (Yen, 1984) was used to measure local independence, which was derived from the correlation between the performances of two items. Simply, the  $Q_3$  statistic is the correlation among IRT residuals and is computed using the following equations:

$$\mathbf{d}_{ij} = \mathbf{u}_{ij} - \mathbf{T}_j(\hat{\theta}_i), \quad (22)$$

where  $u_{ij}$  is the item score of the  $i$ th examinee for item  $j$ .  $T_j(\hat{\theta}_i)$  is the estimated true score for item  $j$  of examinee  $i$ , which is defined as

$$\mathbf{T}_j(\hat{\theta}_i) = \sum_{k=1}^m y_{jk} \mathbf{P}_{jk}(\hat{\theta}_i), \quad (23)$$

where  $y_{jk}$  is the weight for response category  $k$ ,  $m$  is the number of response categories, and  $P_{jk}(\hat{\theta}_i)$  is the probability of response category  $k$  to item  $j$  by examinee  $i$  with the ability estimate  $\hat{\theta}_i$ .

The pairwise index of local dependence  $Q_3$  between item  $j$  and item  $j'$  is

$$Q_{3jj'} = r(\mathbf{d}_j, \mathbf{d}_{j'}), \quad (24)$$

where  $r$  refers to the Pearson product-moment correlation.

When there are  $n$  items,  $n(n-1)/2$ ,  $Q_3$  statistics will be produced. The  $Q_3$  values are expected to be small. Table 18 presents summaries of the distributions of  $Q_3$  statistics: minimum, 5th percentile, median, 95th percentile, and maximum values from each grade and subject. The results show that at least 90% of the items, between the 5th and 95th percentiles, for all grades and subjects except for grades 3 and 4 ELA, were smaller than a critical value of 0.2 for  $|Q_3|$  (Chen & Thissen, 1997). In general,  $Q_3$  values from ELA are higher than  $Q_3$  values from Math.

Table 18. Q3 Distribution

Q3 DISTRIBUTION						
	GRADE	MINIMUM	5TH PERCENTILE	MEDIAN	95TH PERCENTILE	MAXIMUM
ELA	3	-0.263	-0.190	-0.033	0.519	0.715
	4	-0.256	-0.183	-0.036	0.524	0.683
	5	-0.283	-0.196	-0.042	0.165	0.359
	6	-0.297	-0.205	-0.036	0.162	0.379
	7	-0.370	-0.247	-0.041	0.239	0.437
	8	-0.249	-0.187	-0.037	0.170	0.359
	11	-0.307	-0.216	-0.032	0.170	0.342
Math	3	-0.295	-0.185	-0.038	0.201	0.489
	4	-0.305	-0.207	-0.044	0.180	0.552
	5	-0.296	-0.191	-0.041	0.159	0.592
	6	-0.310	-0.182	-0.037	0.172	0.313
	7	-0.295	-0.205	-0.041	0.181	0.345
	8	-0.284	-0.185	-0.035	0.164	0.447
	11	-0.260	-0.168	-0.041	0.122	0.217

## 6. Quality Control

Thorough quality control has been integrated into every aspect of the CTAA test administration, scoring, and reporting. This section highlights the key procedures.

### 6.1 QUALITY CONTROL IN TEST CONFIGURATION

For online testing, the configuration files contain the complete information required for test administration and scoring, such as the test blueprint specification, slopes and intercepts for theta-to-scale score transformation, cut scores, and the item information (e.g., answer keys, item attributes, item parameters, passage information). The accuracy of the configuration file is checked and confirmed numerous times independently by multiple staff members before the testing window.

### 6.2 PLATFORM REVIEW

A platform is a combination of a hardware device and an operating system. Platform review is a process in which each item is checked to ensure that it is displayed appropriately on each tested platform. In recent years, the number of platforms has proliferated, and platform review now takes place on various platforms that are significantly different from one another.

AIR's test delivery system (TDS) supports a variety of item layouts. Each item goes through an extensive platform review on different operating systems, including Windows, Linux, and iOS, to ensure that the item looks consistent in all systems.

Platform review is conducted by a team. The team leader projects the item as it was web-approved in the Item Tracking System (ITS), and team members, each behind a different platform, look at the same item to see that it renders as expected.



### **6.3 USER ACCEPTANCE TESTING AND FINAL REVIEW**

Both internal and external user acceptance testing (UAT) was conducted before the testing window opened for the TDS and the ORS.

For the TDS, detailed protocols were developed, and reviewers were given detailed instructions to note or report issues related to system functionality, item display, or scoring. During the internal UAT, AIR created pseudo-tests that covered the entire range of possibilities of item responses and the complete set of scoring rules. The pseudo-tests were then manually entered into the TDS. When issues were found, AIR took immediate actions to solve them. When the TDS was updated, the related pseudo-cases could be re-entered into the system. The process was repeated until all issues were resolved. Pseudo-tests were also created for external UAT so that CSDE could conduct a hands-on review of the system prior to the opening of the testing window. CSDE approved the TDS before the system was opened for testing.

For the ORS, the same procedure was followed: both AIR and CSDE staff conducted internal and external UAT of the system to ensure that the system functioned as intended before opening to the public.

### **6.4 QUALITY ASSURANCE IN ONLINE DATA**

AIR's TDS has a real-time, quality-monitoring component built in. After a test is administered to a student, the TDS passes the resulting data to our quality assurance (QA) system. QA conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item, keys for multiple-choice items, score points in each item, and the total number of field-test items and operational items, and that the test record contains no data from items that have been invalidated.

Data pass directly from the Quality Monitoring System (QMS) to the Database of Record (DoR), which serves as the repository for all test information, and from which all test information for reporting is pulled. The Data Extract Generator (DEG) is the tool used to pull data from the DoR for delivery to CSDE. AIR staff ensure that data in the extract files match the DoR prior to delivery to CSDE.

### **6.5 QUALITY CONTROL ON SCORING**

AIR's scoring engine is used for operational scoring. Before operational scoring, AIR created mock-ups of student records that cover all scoring scenarios. The records were scored by both AIR's analysis team (responsible for the scoring engine) and AIR psychometricians, independently. They compared their results and solved discrepancies iteratively until a 100% match of scores was reached.

When the testing window closed, psychometricians scored the operational records and compared with the scores from the scoring engine again. All discrepancies were investigated and resolved before scores were released to the state and students.

## 6.6 QUALITY ASSURANCE IN REPORTING

Two types of score reports were produced for the CTAA tests: online reports and printed family reports.

### 6.6.1 Online Report Quality Assurance

Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DoR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the official record is stored. Only after scores have passed the QA checks and are uploaded to the DoR are they passed to the ORS, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the ORS until it passes all of the QA system's validation checks.

### 6.6.2 Paper Report Quality Assurance

#### **Statistical Programming**

The family reports contain custom programming and require rigorous QA processes to ensure their accuracy. All custom programming is guided by detailed and precise specifications. Upon approval of the specifications, analytic rules are programmed, and each program is extensively tested on test decks and real data from other programs. The final programs are reviewed by two senior statisticians and one senior programmer to ensure that they implement agreed-upon procedures. Custom programming is implemented independently by two statistical programming teams working from the specifications. Only when the output from both teams matches exactly are the scripts released for production. Quality control, however, does not stop there.

Much of the statistical processing is repeated, and AIR has implemented a structured software development process to ensure that the repeated tasks are implemented correctly and identically each time. AIR writes small programs called macros that take specified data as input and produce data sets containing derived variables as output. Approximately 30 such macros reside in AIR's library. Each macro is extensively tested and stored in a central development server. Once a macro is tested and stored, changes to the macro must be approved by the director of score reporting and the director of psychometrics, as well as by the project directors for affected projects.

Each change is followed by a complete retesting with the entire collection of scenarios on which the macro was originally tested. The main statistical program is mostly made up of calls to various macros, including macros that read in and verify the data and conversion tables and macros that do the many complex calculations. This program is developed and tested using artificial data generated to test both typical and extreme cases. In addition, the program goes through a rigorous code review by a senior statistician.

#### **Display Programming**

The paper report development process uses graphical programming, which takes place in a Xerox-developed programming language called Variable Data Intelligent PostScript Printware (VIPP) and allows almost infinite control of the visual appearance of the reports. After designers

at AIR create backgrounds, our VIPP programmers write code that indicates where to place all variable information (data, graphics, and text) in the reports. The VIPP code is tested using both artificial and real data. AIR's data generation utilities can read the output layout specifications and generate artificial data for direct input into the VIPP programs. This allows the testing of these programs to begin before the statistical programming is complete.

In later stages, artificial data are generated according to the input layout and run through the score reporting statistical programs, and the output is formatted as VIPP input. This enables AIR to test the entire system. Programmed output goes through multiple stages of review and revision by graphics editors and the score reporting team to ensure that design elements are accurately reproduced and that data are correctly displayed. Once we receive final data and VIPP programs, the AIR score reporting team reviews proofs that contain actual data based on our standard quality assurance documentation. In addition, we compare data independently calculated by AIR psychometricians with data on the reports. A large sample of reports is reviewed by several AIR staff members to make sure that all data are correctly placed on reports. This rigorous review is typically conducted over several days and takes place in a secure location at AIR. All reports containing actual data are stored in a locked storage area. Prior to printing the reports, AIR provides a live data file and individual student reports with sample districts for data file.

### **Sample Paper Report QC**

Before the final paper reports are generated, AIR's research assistants conduct a thorough comparison between the statistics on the paper report and the statistics generated from the DoR, the database that contains test results. If discrepancies are found, actions are taken until all discrepancies are resolved. The sample reports are sent to CSDE for approval. Upon CSDE's approval, the final student paper reports are produced and distributed.

## 7. Reference

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Bejar, I. I. (1980). *A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates*. *Journal of Educational Statistics*, 17, 282–296.
- Chen, W., & Thissen, D. (1997). *Local dependence indexes for item pairs using item response theory*. *Journal of Educational and Behavioral Statistics*, 22, 265–289.
- Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). *Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption*. *Quality of Life Research*, 18: 447–460.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Toronto: Holt, Rinehart and Winston.
- Cronbach, L. J. (1951). *Coefficient alpha and the internal structure of tests*. *Psychometrika*, 16: 297–334.
- Feldt, L. S., & Brennan, R. L. (1989). *Reliability*. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). Washington, DC: American Council on Education.
- Feldt, L. S., & Qualls, A. L. (1996). *Bias in coefficient alpha arising from heterogeneity*. *Applied Measurement in Education*, 9(3), 277–286.
- Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.), New York: Springer-Verlag.
- Kuder, G. F., & Richardson, M. W. (1937). *The Theory and Estimation of Test Reliability*. *Psychometrika*, 2: 151–160.
- Lee, W., Hanson, B. A., & Brennan, R. L. (2002). *Estimating consistency and accuracy indices for multiple classifications*. *Applied Psychological Measurement*, 26, 412–432.
- Lord, F. M. (1980). *Applications of item response theory practical testing problems*. Hillsdale, NJ: Erlbaum.
- Nunnally, J. C. (1978). *Psychometric Theory* (2d ed.). New York: McGraw-Hill.
- Qualls, L. A. (1995). *Estimating the reliability of a test containing multiple item formats*. *Applied Measurement in Education*, 8, 111–120.
- Rudner, L.M. (2001). *Computing the expected proportions of misclassified examinees*. *Practical Assessment, Research & Evaluation*, 7(14).
- Rudner, L.M. (2005). *Expected classification accuracy*. *Practical Assessment, Research & Evaluation*, 10(13).
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). *On the reliability of testlet-based tests*. *Journal of Educational Measurement*, 28(3), 234–247.
- Yen, W. M. (1981). *Using simulation results to choose a latent trait model*. *Applied Psychological Measurement*, 5, 245-262.

- Yen, W. M. (1984). *Effects of local item dependence on the fit and equating performance of the three-parameter logistic model*. *Applied Psychological Measurement*, 8, 125–145.
- Yen, W. M. (1993). *Scaling performance assessments: Strategies for managing local item dependence*. *Journal of Educational Measurement*, 30, 187–213.