TECHNICAL REPORT

Connecticut Alternate Assessment

English Language Arts, Grades 3–8, 11
Mathematics, Grades 3–8, 11

Test Administrations
April 18 – June 10, 2016

Submitted to:
Connecticut State Department of Education (CSDE)

Submitted by:
American Institutes for Research
1000 Thomas Jefferson Street NW, Suite 200
Washington, DC 20007

May 9, 2017

NOTE: Some tables (including all of Appendices) were suppressed from this document due to data confidentiality and the privacy of student educational records.

TABLE OF CONTENTS

List of Tables

List of Figures

1. Introduction

This report introduces the Connecticut Alternate Assessment (CTAA) used during the 2016 administration, summarizes the administration and performance results, and details the evaluation of the assessment quality.

Funded through a General Supervision Enhancement Grant (GSEG) from the United States Education Department (USED) 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 multistate 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 multistate grant project, the Connecticut Department of Education (CSDE) adopted the NCSC English language arts and mathematics test in the spring 2016 administration. Students in grades 3–8 and 11 took the tests.

The CTAA is the NCSC alternate assessment and is based on alternate achievement standards (AA-AAS). The 2016 CTAA assessment included

- Assessments in mathematics and English language arts (ELA) for students in grades 3–8 and 11,
- Around 30–35 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)

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. This document summarizes the test results, reporting, psychometric qualities of test forms, and the quality control process for the 2016 administration.

2. 2016 Administration and Item Re-evaluation

2.1 Test Window

The 2016 test window started on April 18 and ended on June 10.

2.2 Test Forms

As described in the 2015 NCSC technical report, four forms were developed 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. The form summary and their comparisons with their respective test blueprints can be found in Appendix A.

2.3 TEST MODE

The 2016 tests were administered online with paper forms as accommodation. For paper 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. If a student has no response to the first four items, the teacher is directed to consult with the state. If the state approves, the student is directed to exit the test. Otherwise, the student is required to respond to all items until the end of the test.

For CTAA,, an early stopping rule (ESR) is established. That is, the rule allows students who have difficulties taking the tests to exit the tests after the first four items. If a student does not respond to the first four items, the teacher 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 then submit the second test. For example, if a student did not respond to the ELA test and was approved as an ESR student, the student 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 does not qualify for the ESR, the TA must resume the test and the student has to answer the rest of the items through the end of the test.

2.5 ITEM RE-EVALUATION

CTAA item analysis was based on students from all member states. To ensure that the items performed as expected for Connecticut students, after 2016 administration, the items were re-evaluated using Connecticut students only. 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 can be found in Appendix B.

2.5.1 Item Difficulty

Since the ELA and mathematics tests only contain 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

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

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 IRT Model Fit

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

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

where

X_i indexes the raw score on item *i*,
θ_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 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.

The Q_1statistic described by Yen (1981) is used for item fit. The standardized fit values, referred to as ZQ_1statistics, are compared over items (CTB Documents A and B, 1998). 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 Checking

In 2015, each form 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 2016, one of the four ELA forms was adopted for each ELA test. The mathematics forms were newly built. To build conversion tables for scoring, items were evaluted based on Connecticut students only. 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, the forms were made sure that they were statistically parallel to the corresponding 2015 forms. The evaluation will take the following steps:

1. Free calibration is based on the item responses from the Connecticut 2016 administrations.
    a. Student records with more than 10 valid scores are used in the calibration process.
    b. The items in the verbal and nonverbal forms in the ELA grades 3 and 4 test need to be combined in calibration.
2. Stocking-Lord is used to equate the 2016 item parameters to the 2015 scale.
    a. Only items with positive point-biserial are used in the equating process.
3. Plot TCCs using the 2015 parameters and the equated 2016 parameters. More attention is paid to forms with large TCC differences.
4. The unsigned area (UA) of the differences of item response curves (ICCs [Raju, 1990]) is computed. The item with the largest ICC difference is flagged for review.
5. The TCCs and UA are taken into account simultaneously to decide if items with a large UA would be dropped from scoring.

Specifically, the differences of TCCs is evaluated as

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

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

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

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

2.5.5 Procedure for Item Evaluation

Flagged items are examined individually. The combined effect of statistics discussed above is taken into account. During the examining period, the content of the flagged items is reviewed. The items that are determined to be used in scoring are documented in Appendix C. They are approved by CSDE.

3. 2016 State Summary

3.1 Scoring Method Review

The two-parameter logistic model is used in calibration. Based on it, conversion tables are constructed for scoring. 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 test characteristic curves 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 1. 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 1. Slope and Intercept

Content Area	Grade	Slope (A)	Intercept (B)
Mathematics	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
ELA	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 2.

Table 2. Scale Score Cut Points

Content Area	Grade	scale.Cut 1	scale.Cut 2	scale.Cut 3
Mathematics	3	1236	1240	1254
Mathematics	4	1233	1240	1251
Mathematics	5	1231	1240	1255
Mathematics	6	1234	1240	1249
Mathematics	7	1232	1240	1254
Mathematics	8	1234	1240	1249
Mathematics	11	1234	1240	1249
ELA	3	1234	1240	1251
ELA	4	1234	1240	1258
ELA	5	1232	1240	1256
ELA	6	1231	1240	1253
ELA	7	1236	1240	1255
ELA	8	1230	1240	1250
ELA	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 16. The SEM of the scale score is the SEM of the theta score times the slope.

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

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

3.2 STUDENT PARTICIPATION

This section describes the demographics of participating students in spring 2016. Table 3 and Table 4 present the student demographics for participating students by gender and ethnicity in each grade for each subject.

Table 3. Participation by Grade and Gender

ELA						
Grade	**Total**		**Female**		**Male**	
	N	**Pct**	**N**	**Pct**	**N**	**Pct**
3	590	100	212	36	378	64
4	598	100	198	33	400	67
5	617	100	206	33	411	67
6	611	100	196	32	415	68
7	571	100	173	30	398	70
8	585	100	188	32	397	68
11	508	100	184	36	324	64
Total	4080	100	1357	33	2723	67
Mathematics						
Grade	**Total**		**Female**		**Male**	
	N	**Pct**	**N**	**Pct**	**N**	**Pct**
3	584	100	211	36	373	64
4	593	100	196	33	397	67
5	610	100	202	33	408	67
6	605	100	194	32	411	68
7	565	100	172	30	393	70
8	582	100	187	32	395	68
11	501	100	180	36	321	64
Total	4040	100	1342	33	2698	67

Table 4. Participation Ethnicity

Table 4 was deleted due to data confidentiality and the privacy of student educational records.

Demographic characteristics of the student population are relatively consistent across grades.

Approximately 30–36% of students are female in each grade and subject.

Among the participants, white students (39–50%) and Hispanic students (22–33%) make up the majority of the assessed students. African American students make up 17–24%. Asian students make up 3–6% of the assessed students in each grade, and multiracial students make up about 1–3% of the assessed student population.

3.3 Score Summary

Table 5 presents the summary statistics of the scale score by grade for ELA and mathematics.

Table 5. Scale Score Summary

Subject	Grade	N	MEAN	MEDIAN	STD	MIN	MAX
ELA	3	590	1234	1236	21	1200	1290
ELA	4	598	1236	1238	19	1200	1290
ELA	5	617	1236	1237	19	1200	1290
ELA	6	611	1233	1233	19	1200	1290
ELA	7	571	1234	1236	20	1200	1290

ELA	8	585	1232	1234	19	1200	1290
ELA	11	508	1239	1239	21	1200	1290
Mathematics	3	584	1235	1238	21	1200	1290
Mathematics	4	593	1233	1236	19	1200	1290
Mathematics	5	610	1236	1238	20	1200	1290
Mathematics	6	605	1234	1235	19	1200	1290
Mathematics	7	565	1234	1236	20	1200	1290
Mathematics	8	582	1234	1238	19	1200	1290
Mathematics	11	501	1234	1237	18	1200	1290

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

3.4 Score Summary by Subgroups

The scale score summary by subgroups is listed in Appendix F.

3.5 Percentage of Students by Performance Level

The percentages of students in each performance level are listed in Table 6.

Table 6. Percentage of Students by Performance Level

subject	grade	Total	Percent_level_1	Percent_level_2	Percent_level_3	Percent_level_4
ELA	3	590	44	15	22	19
ELA	4	598	41	10	36	12
ELA	5	617	34	24	30	12
ELA	6	611	41	25	20	15
ELA	7	571	46	12	29	13
ELA	8	585	41	29	13	17
ELA	11	508	36	17	32	15
Mathematics	3	584	39	15	32	15
Mathematics	4	593	42	16	25	16
Mathematics	5	610	25	32	28	16
Mathematics	6	605	39	25	20	16
Mathematics	7	565	33	30	26	11
Mathematics	8	582	38	20	22	19
Mathematics	11	501	35	26	27	13

3.6 Percentage of Students by Performance Level By Subgroup

The percentages of students in each performance level are listed in Appendix G.

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 to monitor the progression of 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 7.1 lists the types of online reports and the levels at which they can be viewed (student, roster, teacher, school, 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, students, et al.). 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 or teachers within a school) to choose from. For more detailed student assessment results for a school, a teacher, or a roster, users can select the grade on the online score reports.

Table 7.1 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 7.1 Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
State **District** **School** **Teacher** **Roster**	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
Student	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 7.2 presents the types of subgroups and subgroup categories provided in the ORS.

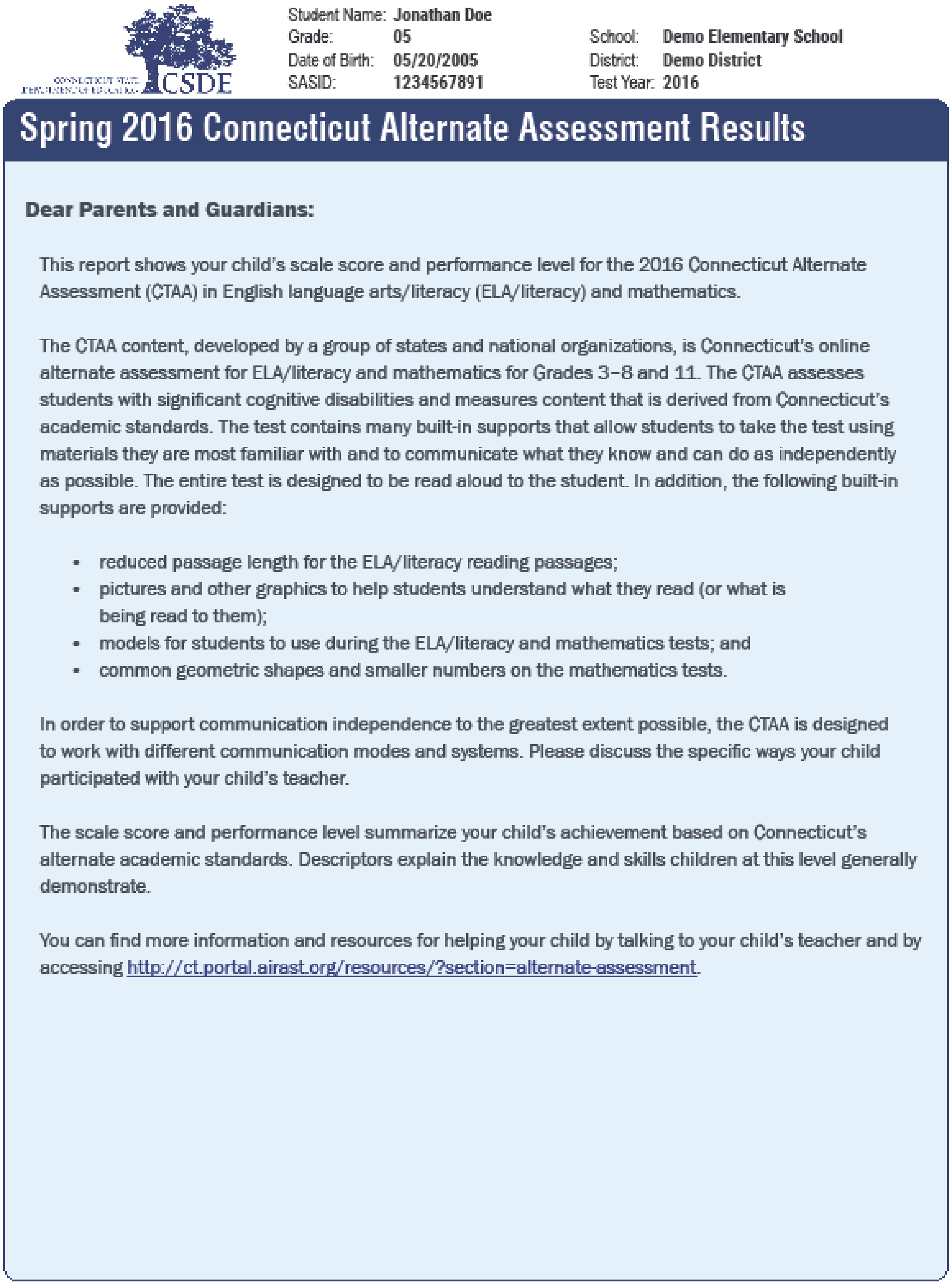
Table 7.2. Types of Subgroups

Breakdown by Category	Displayed Category
Ethnicity	Hispanic or Latino Ethnicity
	American Indian or Alaska Native
	Asian
	Black or African American
	White
	Native Hawaiian or Other Pacific Islander
	Multi-racial
Gender	Male
	Female
IDEA Indicator	Special Education
	Unknown
Limited English Proficiency Status	Yes
	Unknown
Enrolled Grade	Grade 03
	Grade 04
	Grade 05
	Grade 06
	Grade 07
	Grade 08
	Grade 09
	Grade 10
	Grade 11

4.2 Paper Report

Paper Reports for the CTAA were also printed and shipped to the district at the end of the 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. The text related to the Early Stopping Rule is circled.

Figure 1. Family Report Mock–Up



CSDE

Student Name:	**Jonathan Doe**		
Grade:	**05**	School:	**Demo Elementary School**
Date of Birth:	**05/20/2005**	District:	**Demo District**
SASID:	**1234567891**	Test Year:	**2016**

Spring 2016 Connecticut Alternate Assessment Results

Dear Parents and Guardians:

This report shows your child's scale score and performance level for the 2016 Connecticut Alternate Assessment (CTAA) in English language arts/literacy (ELA/literacy) and mathematics.

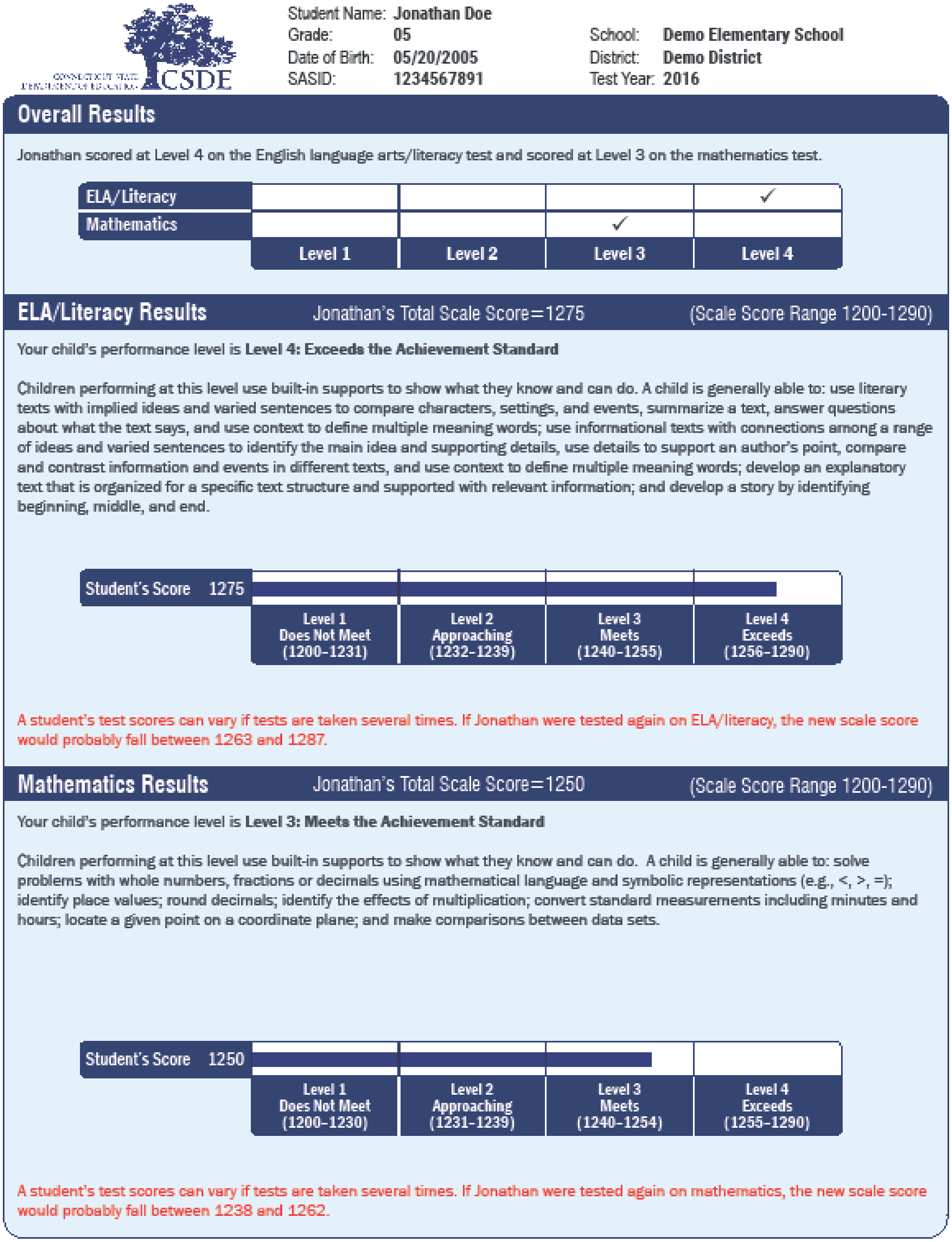
The CTAA content, developed by a group of states and national organizations, is Connecticut's online alternate assessment for ELA/literacy and mathematics for Grades 3–8 and 11. The CTAA assesses students with significant cognitive disabilities and measures content that is derived from Connecticut's 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/literacy 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/literacy and mathematics tests; and
- common geometric shapes and smaller numbers on the mathematics tests.

In order to support communication independence 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 with your child's teacher.

The scale score and performance level summarize your child's achievement based on Connecticut's alternate 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 http://ct.portal.airast.org/resources/?section=alternate-assessment.



CONNECTICUT STATE DEPARTMENT OF EDUCATION CSDE

Student Name:	Jonathan Doe		
Grade:	05	School:	Demo Elementary School
Date of Birth:	05/20/2005	District:	Demo District
SASID:	1234567891	Test Year:	2016

Overall Results

Jonathan scored at Level 4 on the English language arts/literacy test and scored at Level 3 on the mathematics test.

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

ELA/Literacy Results

Jonathan's Total Scale Score=1275 (Scale Score Range 1200-1290)

Your child's performance level is **Level 4: Exceeds 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 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 connections among a range of 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.

Student's Score 1275

Level 1 Does Not Meet (1200–1231)	Level 2 Approaching (1232–1239)	Level 3 Meets (1240–1255)	Level 4 Exceeds (1256–1290)

A student's test scores can vary if tests are taken several times. If Jonathan were tested again on ELA/literacy, the new scale score would probably fall between 1263 and 1287.

Mathematics Results

Jonathan's Total Scale Score=1250 (Scale Score Range 1200-1290)

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

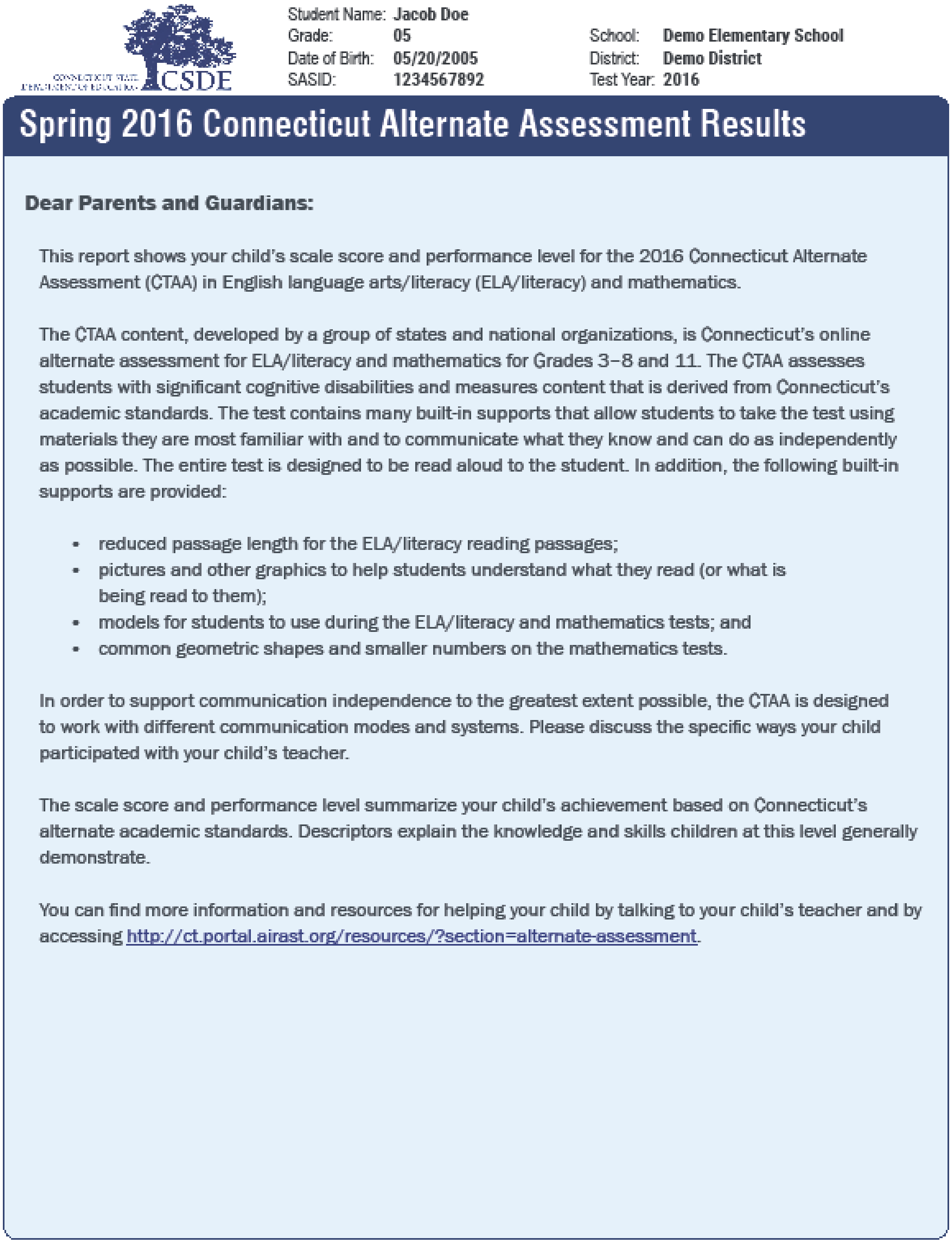
Children performing at this level use built-in supports to show what they know and can do. A child is generally able to: solve problems with whole numbers, fractions or decimals using mathematical language and symbolic representations (e.g., <, >, =); identify place values; round decimals; identify the effects of multiplication; convert standard measurements including minutes and hours; locate a given point on a coordinate plane; and make comparisons between data sets.

Student's Score 1250

Level 1 Does Not Meet (1200–1230)	Level 2 Approaching (1231–1239)	Level 3 Meets (1240–1254)	Level 4 Exceeds (1255–1290)

A student's test scores can vary if tests are taken several times. If Jonathan were tested again on mathematics, the new scale score would probably fall between 1238 and 1262.

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



Student Name:	Jacob Doe		
Grade:	05	School:	Demo Elementary School
Date of Birth:	05/20/2005	District:	Demo District
SASID:	1234567892	Test Year:	2016

Spring 2016 Connecticut Alternate Assessment Results

Dear Parents and Guardians:

This report shows your child's scale score and performance level for the 2016 Connecticut Alternate Assessment (CTAA) in English language arts/literacy (ELA/literacy) and mathematics.

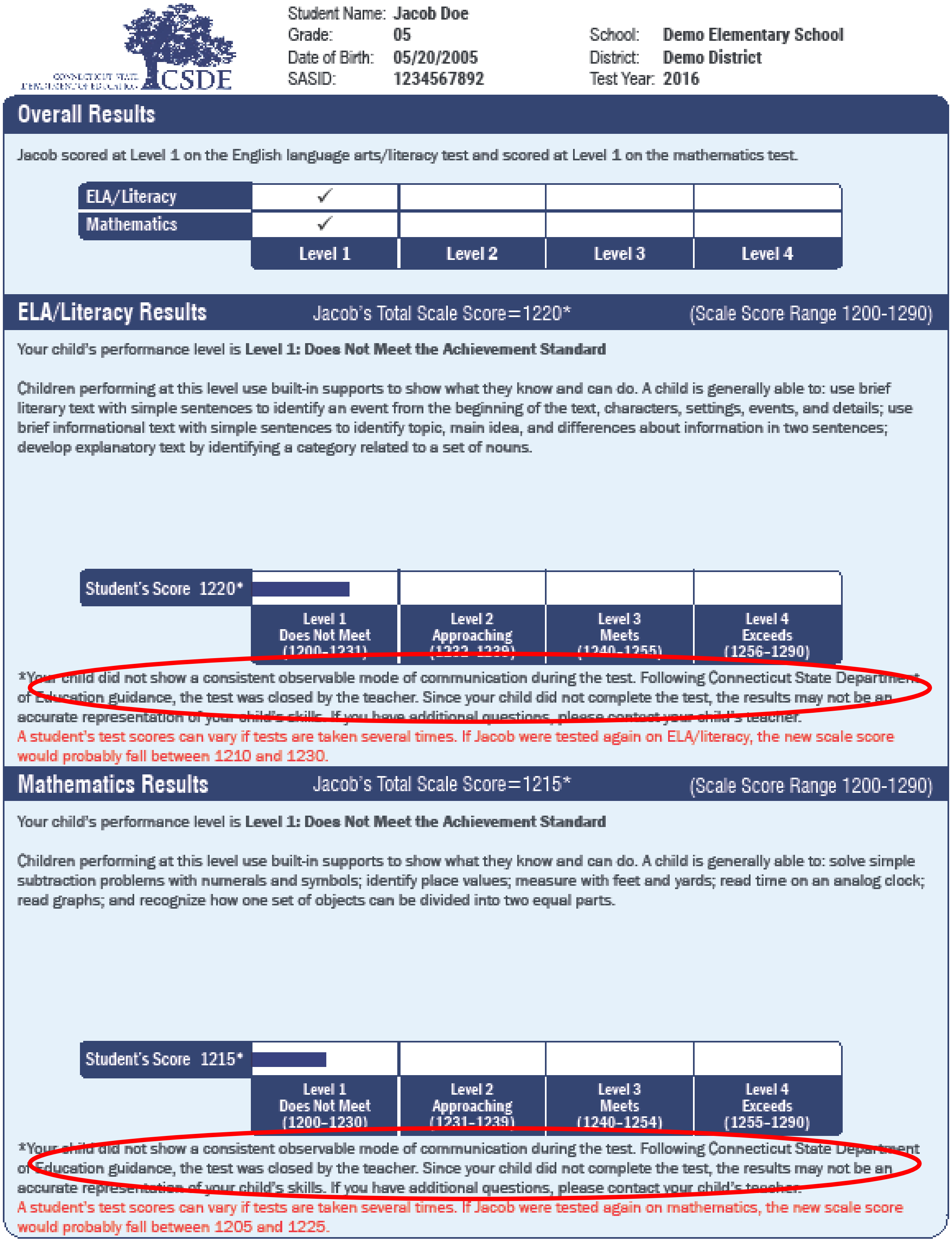
The CTAA content, developed by a group of states and national organizations, is Connecticut's online alternate assessment for ELA/literacy and mathematics for Grades 3–8 and 11. The CTAA assesses students with significant cognitive disabilities and measures content that is derived from Connecticut's 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/literacy 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/literacy and mathematics tests; and
- common geometric shapes and smaller numbers on the mathematics tests.

In order to support communication independence 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 with your child's teacher.

The scale score and performance level summarize your child's achievement based on Connecticut's alternate 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 http://ct.portal.airast.org/resources/?section=alternate-assessment.



CONNECTICUT STATE DEPARTMENT OF EDUCATION CSDE

Student Name: Jacob Doe
Grade: 05
Date of Birth: 05/20/2005
SASID: 1234567892

School: Demo Elementary School
District: Demo District
Test Year: 2016

Overall Results

Jacob scored at Level 1 on the English language arts/literacy test and scored at Level 1 on the mathematics test.

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

ELA/Literacy Results

Jacob's Total Scale Score=1220* (Scale Score Range 1200-1290)

Your child's performance level is **Level 1: Does Not Meet 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 brief literary text with simple sentences to identify an event from the beginning of the text, characters, settings, events, and details; use brief informational text with simple sentences to identify topic, main idea, and differences about information in two sentences; develop explanatory text by identifying a category related to a set of nouns.

Student's Score 1220*	▬			
	Level 1 Does Not Meet (1200–1231)	Level 2 Approaching (1232–1239)	Level 3 Meets (1240–1255)	Level 4 Exceeds (1256–1290)

*Your child did not show a consistent observable mode of communication during the test. Following Connecticut State Department of Education guidance, the test was closed by the teacher. Since your child did not complete the test, the results may not be an accurate representation of your child's skills. If you have additional questions, please contact your child's teacher.

A student's test scores can vary if tests are taken several times. If Jacob were tested again on ELA/literacy, the new scale score would probably fall between 1210 and 1230.

Mathematics Results

Jacob's Total Scale Score=1215* (Scale Score Range 1200-1290)

Your child's performance level is **Level 1: Does Not Meet 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 subtraction problems with numerals and symbols; identify place values; measure with feet and yards; read time on an analog clock; read graphs; and recognize how one set of objects can be divided into two equal parts.

Student's Score 1215*	▬			
	Level 1 Does Not Meet (1200–1230)	Level 2 Approaching (1231–1239)	Level 3 Meets (1240–1254)	Level 4 Exceeds (1255–1290)

*Your child did not show a consistent observable mode of communication during the test. Following Connecticut State Department of Education guidance, the test was closed by the teacher. Since your child did not complete the test, the results may not be an accurate representation of your child's skills. If you have additional questions, please contact your child's teacher.

A student's test scores can vary if tests are taken several times. If Jacob were tested again on mathematics, the new scale score would probably fall between 1205 and 1225.

5. Technical Quality

Technical quality of the operational forms is discussed in this section. 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 and Marginal Standard Error of Measurement

Test reliability is assessed by marginal reliability and Cronbach's alpha. Marginal reliability (Sireci, Thissen, & Wainer, 1991) assesses the precision of scoring. Cronbach's alpha assesses the internal consistency of items.

Specifically, marginal reliability is based on the average conditional standard error of measurement 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

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

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

where σ_{true}^2 is true score variance, σ_{obs}^2 is the observed score variance, $\bar{\sigma}_{err}^2$ is the error variance, and σ_{err}^2 is the square of the conditional standard error of measurement 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.

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

$$\propto = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_x^2}\right]$$

where *n* is the sample size, σ_i^2 is the raw score variance for item i. σ_x^2 is the variance of the total raw 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 2017 results of marginal reliability, MSEM, and standard deviation of scale scores (STD) by test are listed in Table 8. It shows that except for the ELA grade 11 test, the marginal reliability estimates exceed 0.87. The form MSEMs are about one third of the standard deviations (STD) of scale scores. The results suggest that the test scores are mostly precisely estimated. The standard error of measurement 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 conditional SEM at the maximum raw score point 25 is 43.7, which is significantly higher than the conditional SEMs for other score points in this form and others. It indicates that the test needed harder items for high-ability students. Thirty-one students earned raw score 25. Removing the 31 students, the marginal reliability, STD, MSEM, and the MSEM/STD become 0.82, 17.08, 7.19, and 0.42, respectively. Besides, 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 8: Marginal Reliability and Marginal Standard Error of Measurement

Subject	Grade	Sample Size	Marginal Reliability	STD	MSEM	MSEM/STD
ELA	3	590	0.89	21.48	7.22	0.34
ELA	4	598	0.88	19.40	6.72	0.35
ELA	5	617	0.87	19.48	6.92	0.36
ELA	6	611	0.87	19.03	6.79	0.36
ELA	7	571	0.87	19.70	7.22	0.37
ELA	8	585	0.88	19.35	6.58	0.34
ELA	11	508	0.63	21.07	12.86	0.61
Mathematics	3	584	0.91	21.49	6.54	0.30
Mathematics	4	593	0.88	19.33	6.60	0.34
Mathematics	5	610	0.89	19.70	6.56	0.33
Mathematics	6	605	0.90	18.88	5.92	0.31
Mathematics	7	565	0.90	19.73	6.39	0.32
Mathematics	8	582	0.87	19.11	6.90	0.36
Mathematics	11	501	0.87	17.98	6.44	0.36

The Cronbach’s alpha coefficients are summarized in Table 9. Except mathematics grade 5 and grade 11 tests, no other test has alpha coefficients below 0.80, which indicates that the items within each test are reasonably consistent in measuring the construct that the test is designed to measure. The computation of Cronbach’s alpha requires the full response matrix; therefore, the sample sizes are smaller.

Table 9: Cronbach’s Alpha

Subject	Grade	Sample Size	Number Items	Alpha
ELA	3	340	42	0.89
ELA	4	386	41	0.88
ELA	5	368	32	0.81
ELA	6	383	33	0.85
ELA	7	360	33	0.83
ELA	8	371	35	0.83

Subject	Grade	Sample Size	Number Items	Alpha
ELA	11	349	32	0.84
Mathematics	3	358	40	0.87
Mathematics	4	405	40	0.83
Mathematics	5	417	40	0.77
Mathematics	6	419	40	0.84
Mathematics	7	353	40	0.80
Mathematics	8	380	40	0.81
Mathematics	11	340	40	0.76

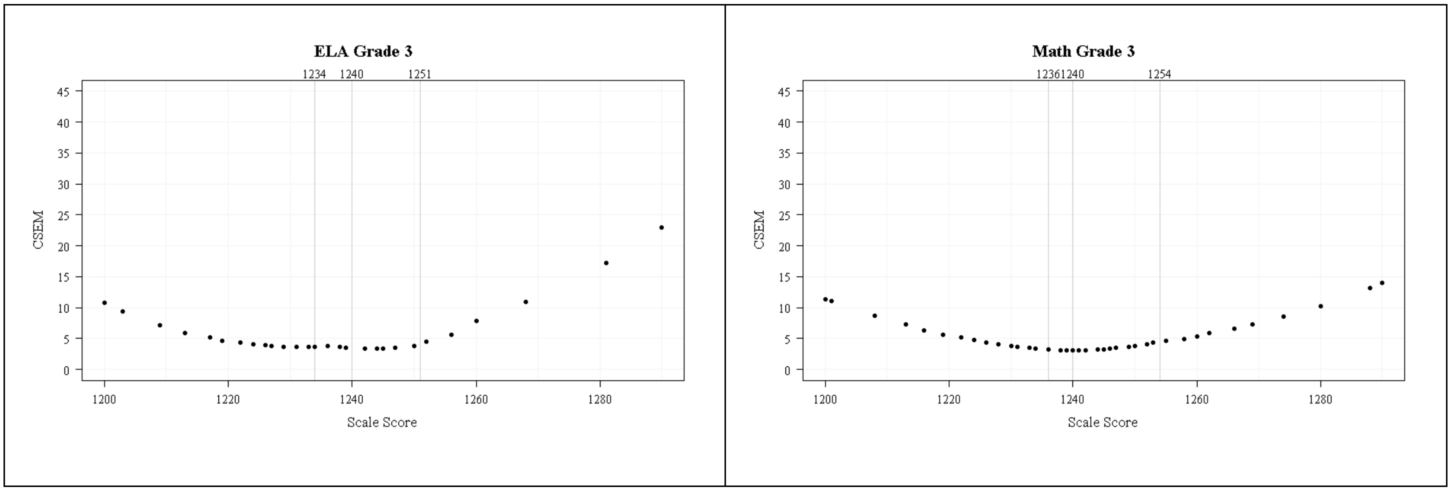
5.2 Conditional Standard Error of Measurement

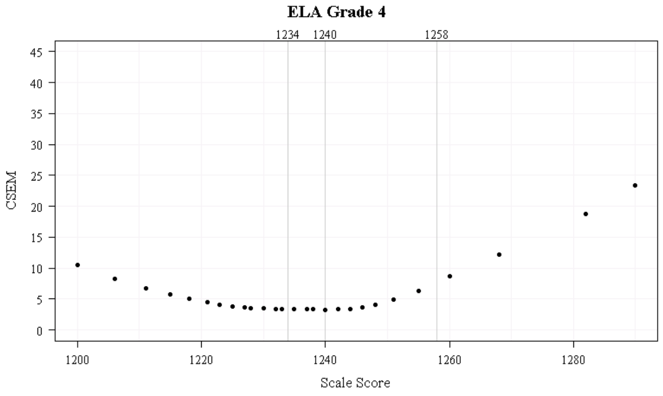
As described in Section 3.1 Scoring Method Review, the conditional SEM is computed as

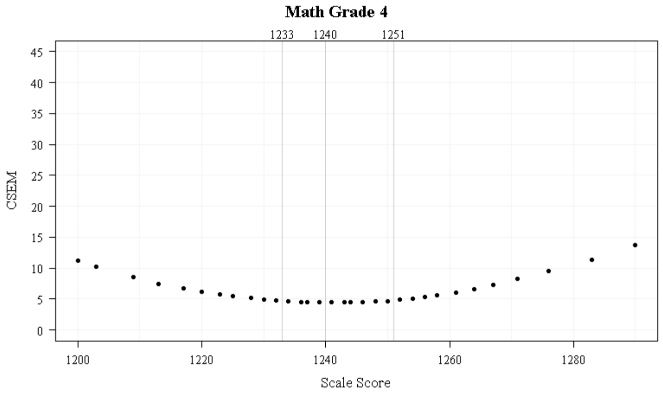
$$\mathrm{se}(\theta) = \frac{1}{\sqrt{-\left(\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta}\right)}}$$

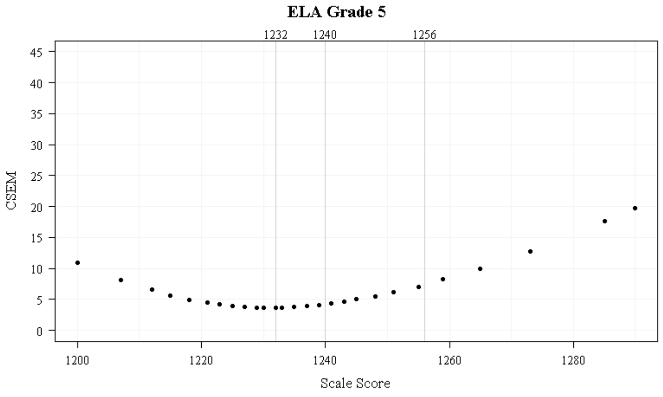
where $\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta}$ is the second derivative of the log-likelihood with respect to θ.

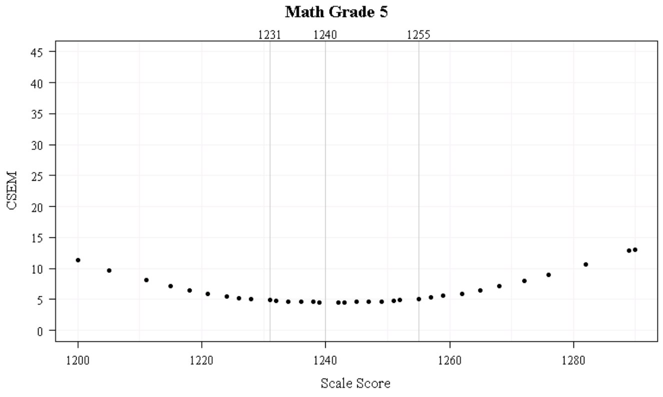
Figure 3. CSEM by Test

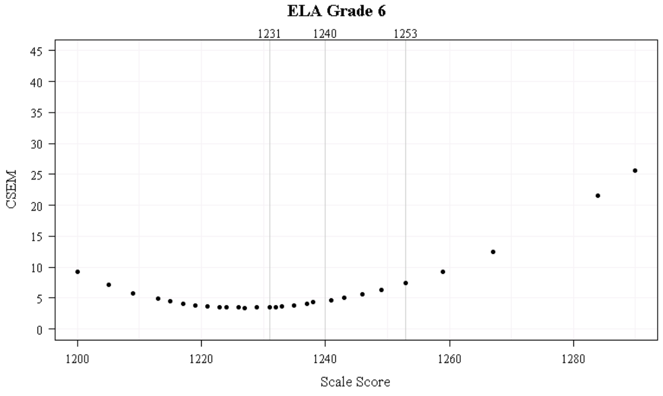


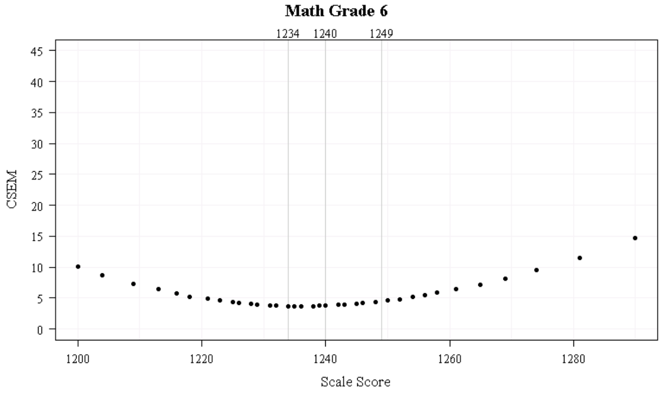

ELA Grade 4
1234
1240
1258
CSEM
0
5
10
15
20
25
30
35
40
45
1200
1220
1240
1260
1280
Scale Score

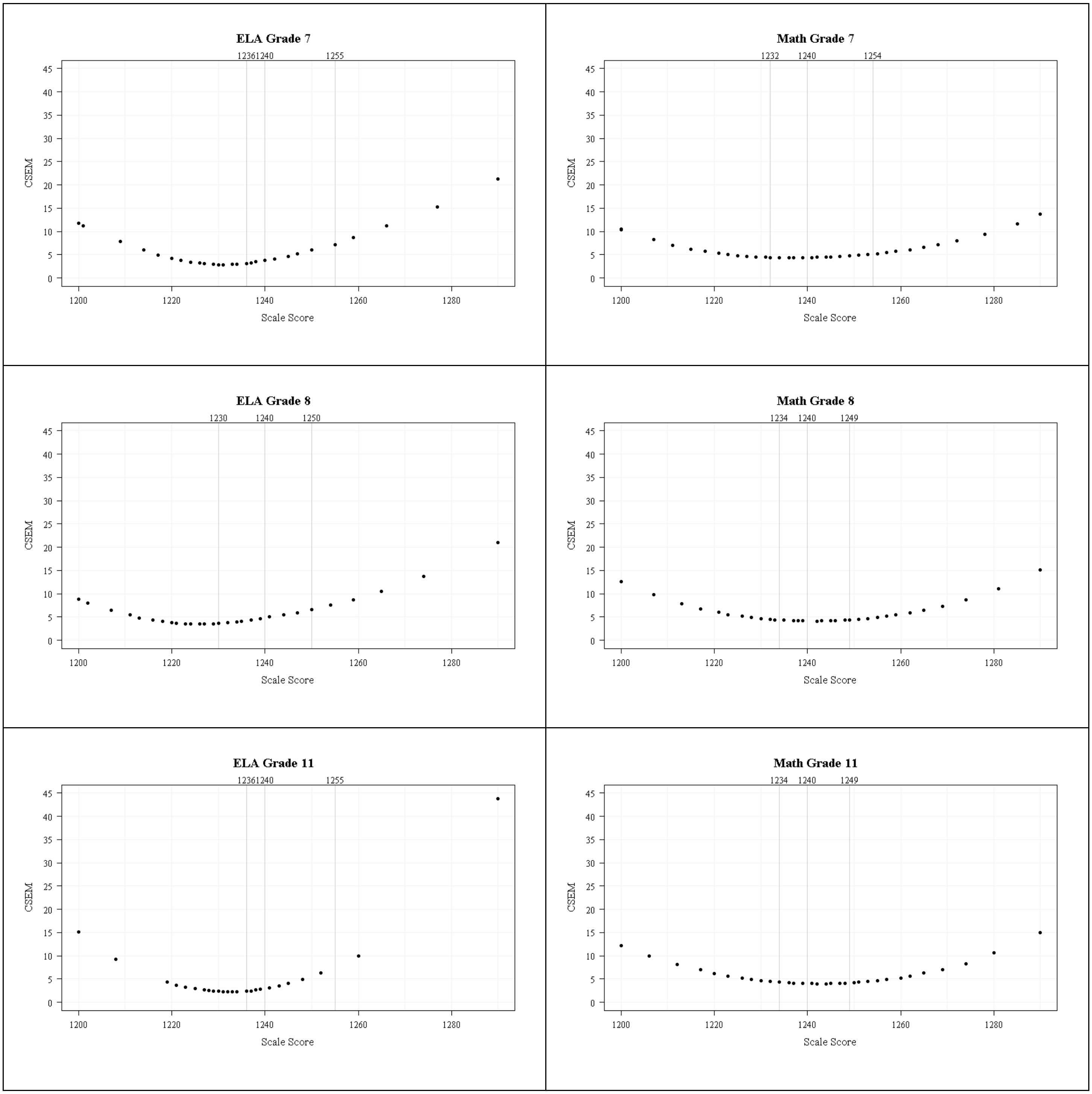

Math Grade 4
1233
1240
1251
CSEM
0
5
10
15
20
25
30
35
40
45
1200
1220
1240
1260
1280
Scale Score


ELA Grade 5
1232
1240
1256
CSEM
0
5
10
15
20
25
30
35
40
45
1200
1220
1240
1260
1280
Scale Score


Math Grade 5
1231
1240
1255
CSEM
0
5
10
15
20
25
30
35
40
45
1200
1220
1240
1260
1280
Scale Score


ELA Grade 6
1231
1240
1253
CSEM
0
5
10
15
20
25
30
35
40
45
1200
1220
1240
1260
1280
Scale Score


Math Grade 6
1234
1240
1249
CSEM
0
5
10
15
20
25
30
35
40
45
1200
1220
1240
1260
1280
Scale Score



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

5.3 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 analysis investigates how precisely students are classified into each performance level. It refers to the agreement between the observed classifications and the classifications that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. Classification consistency investigates how consistently students are classified into the same performance level across two independent administrations of alternate but equivalent forms. For the CTAA tests, the classification accuracy and classification consistency are examined at each performance level using the Rudner classification index (Rudner, 2005).

In reality, the true ability is unknown and students do not take an alternate, equivalent form; therefore, the classification accuracy and consistency is estimated based on students' item scores and the CSEMs of the scores, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with a measurement error.

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

$$\begin{aligned} p_{il} = p(c_{l-1} \le \theta_i < c_l) &= p\left(\frac{c_{l-1} - \hat{\theta}_i}{se(\hat{\theta}_i)} \le \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} < \frac{c_l - \hat{\theta}_i}{se(\hat{\theta}_i)}\right) \\ &= p\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)} \le \frac{\hat{\theta}_i - \theta_i}{se(\hat{\theta}_i)} < \frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) = \Phi\left(\frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) - \Phi\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)}\right). \end{aligned}$$

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 $n_{alm} = \sum_{pl_i = l} p_{im}$, pl_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 $(l = 1, \cdots, L-1)$ is estimated by

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

where N is the total number of students.

The overall classification accuracy is computed as $\mathrm{CA} = \frac{\sum_{i=1}^{L} n_{aii}}{N}$

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 above table, the classification consistency (CC) for the cut c_l $(l = 1, \cdots, L-1)$ is estimated by

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

The overall classification consistency is computed as

$$\mathrm{CC} = \frac{\sum_{i=1}^{L} n_{cii}}{N}$$

Besides the overall CA and CC for each test, CA and CC analyses were also conducted for each cut point. The results show that the overall CA indices are all above 0.72, and the overall CC indices are all above 0.64. The CA indices at each cut point are all around or above 0.90, while the CC indices at each cut point are all around or above 0.85.

The result is shown in Table 10.

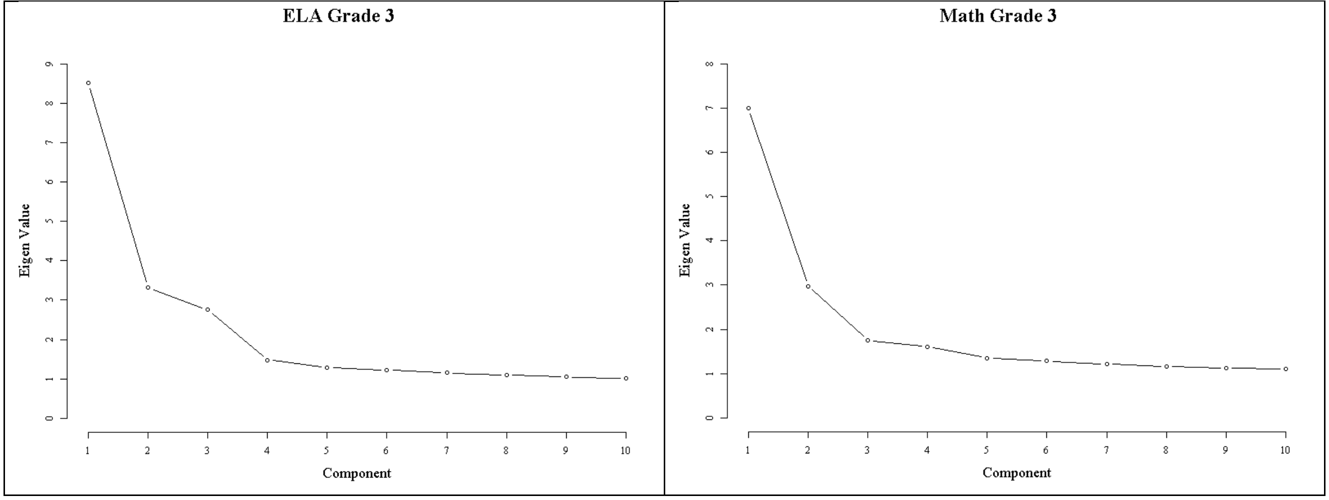
Table 10: Classification Accuracy and Consistency

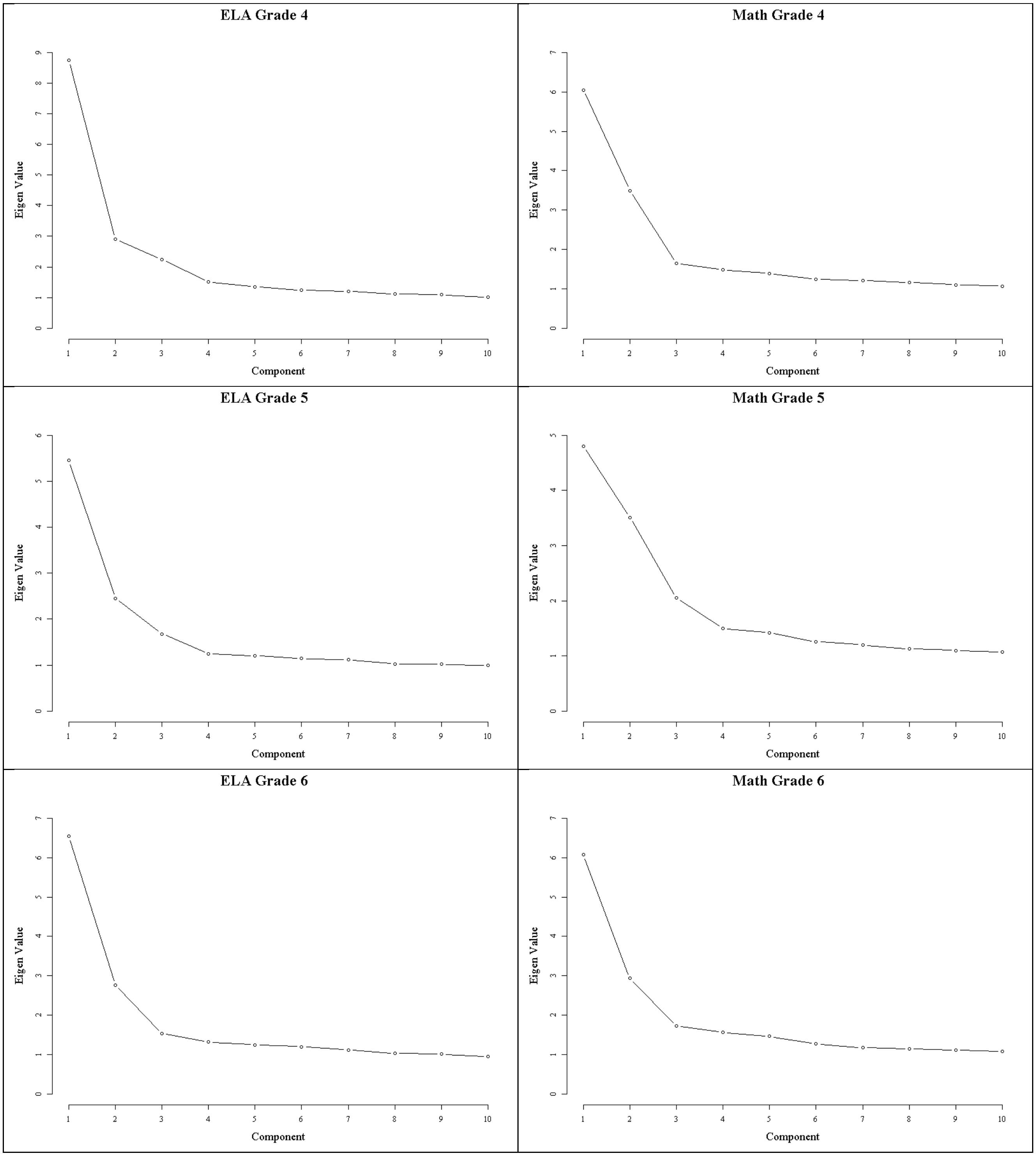
Subject	Grade	Count	Accuracy. Overall	Consistency. Overall	Accuracy. Cut1	Accuracy. Cut2	Accuracy. Cut3	Consistency. Cut1	Consistency. Cut2	Consistency. Cut 3
ELA	3	590	0.81	0.75	0.93	0.94	0.94	0.91	0.92	0.91
ELA	4	598	0.82	0.75	0.94	0.93	0.94	0.91	0.90	0.92
ELA	5	617	0.78	0.70	0.93	0.92	0.94	0.89	0.88	0.91
ELA	6	611	0.79	0.71	0.93	0.92	0.94	0.90	0.89	0.91
ELA	7	571	0.80	0.73	0.92	0.93	0.94	0.89	0.89	0.92
ELA	8	585	0.78	0.70	0.92	0.92	0.94	0.88	0.88	0.92
ELA	11	508	0.78	0.70	0.93	0.91	0.93	0.89	0.86	0.90
Math	3	584	0.82	0.76	0.93	0.93	0.96	0.90	0.89	0.94
Math	4	593	0.78	0.70	0.93	0.92	0.93	0.90	0.88	0.90
Math	5	610	0.76	0.68	0.93	0.89	0.94	0.89	0.85	0.92
Math	6	605	0.76	0.69	0.89	0.91	0.95	0.86	0.87	0.93
Math	7	565	0.76	0.67	0.89	0.90	0.96	0.86	0.86	0.94
Math	8	582	0.74	0.67	0.92	0.89	0.92	0.89	0.85	0.89
Math	11	501	0.72	0.64	0.90	0.87	0.94	0.86	0.83	0.91

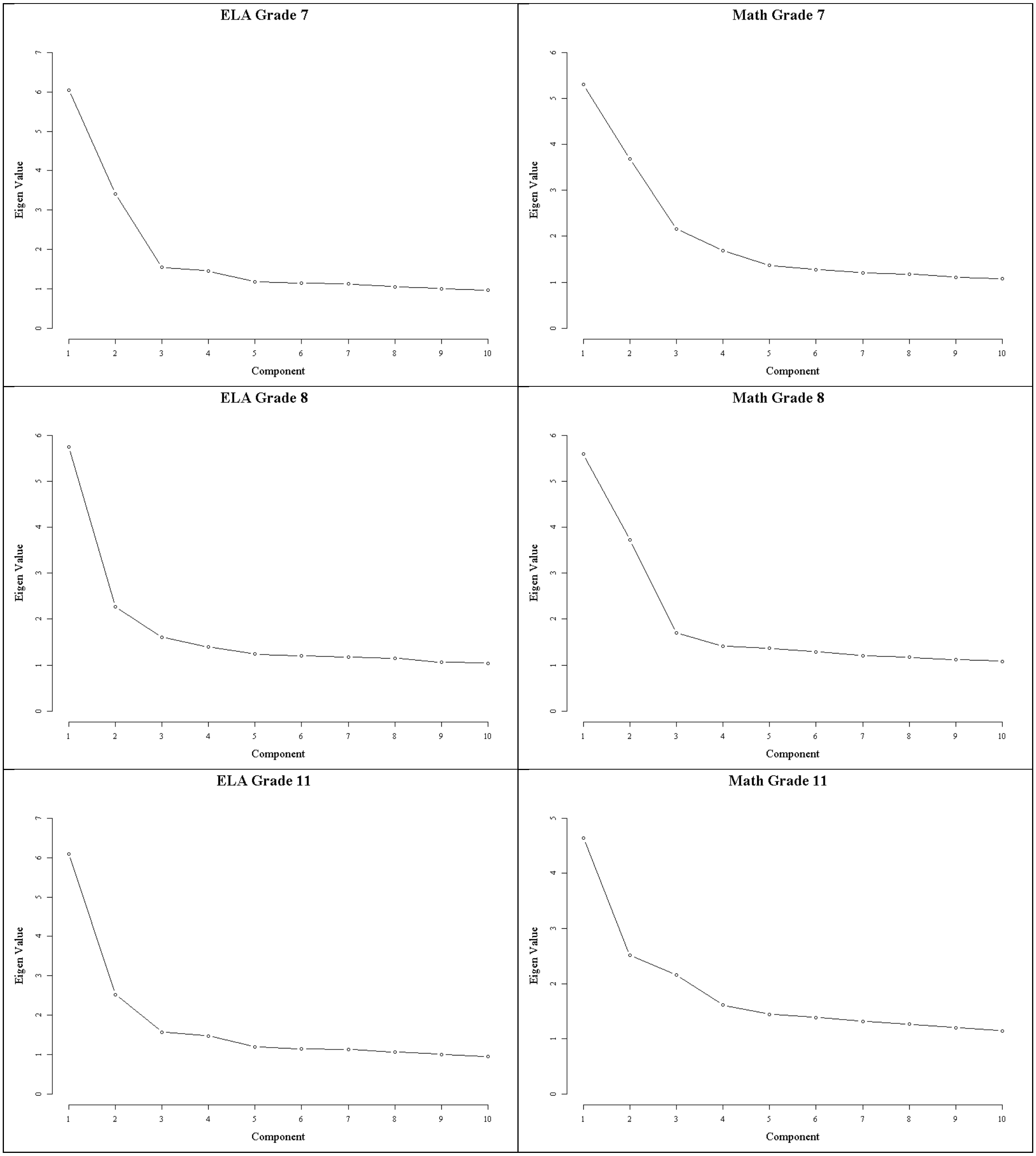
5.4 DIMENSIONALITY

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 Figure 4. The graphs show that the magnitude of the first eigenvalue is always much larger than the magnitude of the second and the following factors in all tests, which indicates that the forms measure one dominant construct.

Figure 4: Scree Plot




ELA Grade 4
Eigen Value
0
1
2
3
4
5
6
7
8
9
Component
1
2
3
4
5
6
7
8
9
10
Math Grade 4
Eigen Value
0
1
2
3
4
5
6
7
Component
1
2
3
4
5
6
7
8
9
10
ELA Grade 5
Eigen Value
0
1
2
3
4
5
6
Component
1
2
3
4
5
6
7
8
9
10
Math Grade 5
Eigen Value
0
1
2
3
4
5
Component
1
2
3
4
5
6
7
8
9
10
ELA Grade 6
Eigen Value
0
1
2
3
4
5
6
7
Component
1
2
3
4
5
6
7
8
9
10
Math Grade 6
Eigen Value
0
1
2
3
4
5
6
7
Component
1
2
3
4
5
6
7
8
9
10



ELA Grade 7

Math Grade 7

ELA Grade 8

Math Grade 8

ELA Grade 11

Math Grade 11

6. Quality Control

Thorough quality control has been integrated into every aspect of the CTAA test administration, scoring, and reporting. This chapter 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 (i.e., 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 prior 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 TDS and the ORS.

For 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 TDS. When issues were found, AIR took immediate actions to solve them. When 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 TDS before the system was opened for testing.

For the ORS, the same procedure is followed, both AIR and the Department staff conducted internal and external UAT of the system to ensure that the system function 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, 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 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 that is used to pull data from the DoR for delivery to the Department. AIR staff ensure that data in the extract files match the DoR prior to delivery to the Department.

6.5 Quality Control on Scoring

The scoring engine is used for operational scoring. Before operational scoring, AIR created mock-ups of student records that covers all scoring scenarios. The records are scored by both AIR's analysis team (responsible for the scoring engine) and AIR psychometricians independently. They compared their results and solve 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 quality assurance processes to ensure their accuracy. All custom programming is guided by detailed and precise. 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. We write 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 our 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 virtually 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 us 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 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 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.