TECHNICAL REPORT

Connecticut Alternate Science Assessment
Grades 5, 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

List of Tables

List of Figures

1. Introduction

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

The Individuals with Disabilities Education Act (IDEA) of 1997 established a legal requirement for all students to participate in statewide content-area assessments. The goal of this requirement was to ensure that every child—including special education students with the most significant cognitive disabilities—would have access to a rigorous curriculum, would receive effective instruction, and be subject to reasonable and high expectations of academic achievement. While students with the most significant cognitive disabilities do not always participate in the same grade-level academic classroom instruction as general education students, they are nevertheless expected to receive grade-level instruction with appropriate academic content and skills with simplifications in the breadth, depth, or complexity of the content standards.

The CTAS Assessment is an alternate assessment based on alternate achievement standards for students with significant cognitive disabilities. It has been developed to ensure that all students with significant cognitive disabilities can participate in an assessment that measures what they know and can do in relation to the Next Generation Science Standards (NGSS). The CTAS Assessment includes six performance tasks that are intended to be administered throughout the year. Teachers work with eligible students to rate student performance on the CTAS Core Extensions. Teachers administer various activities to the students and submit performance ratings into the Data Entry Interface (DEI). The CTAS Assessment must be administered to eligible students with significant cognitive disabilities in grades 5, 8, and 11. The grade 5 test consists of 44 items, and the grades 8 and 11 tests have 42 items. Table 1 displays the number of items in each strand.

Table 1. Number of Operational Items by Standards

Standards	Grade 5	Grade 8	Grade 11
Earth Science (ES)	18	18	16
Life Science (LS)	13	13	16
Physical Science (PS)	13	11	10
Total	44	42	42

1.1 Development and Design of the CTAS

Prior to beginning the design and development of the CTAS Assessment, the CSDE sought extensive formal and informal feedback from educators across the state of Connecticut on the science assessment format. This was done to ensure the format would be relevant and appropriate for students with the most significant cognitive disabilities and who were eligible for the alternate assessment. Based on that feedback, a number of guiding principles were established.

The CTAS Assessment should

- be meaningful and accessible to participating students;
- guide the science curriculum and instruction throughout the year by providing a coherent sequence of assessment activities;
- allow for administration of the assessment throughout the year;
- include an appropriate balance of the breadth and depth of NGSS Learning Progressions across grade bands;
- assess the three dimensions of the NGSS (i.e., science and engineering practices, disciplinary core ideas, and crosscutting concepts);
- incorporate scientific phenomena that students make sense of or use to solve a problem; and
- expect consistent demonstration of the performance expectations by students statewide.

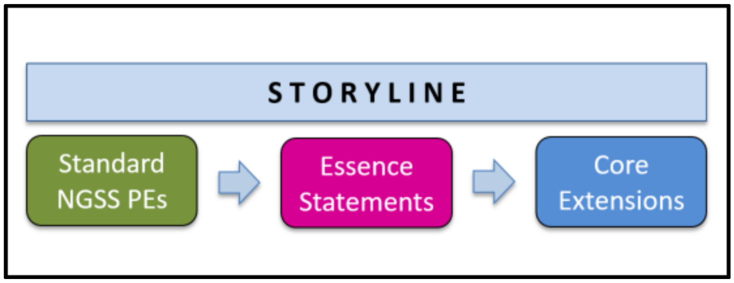
The guiding principles, basic format, and function of the CTAS Assessment were synthesized from feedback from a field of educators, which is comprised of the CTAS Committee and Connecticut educators with knowledge of the NGSS standards and/or experience with students with disabilities (particularly those with significant cognitive disabilities). This committee met several times to offer comprehensive guidance on test design and contributed to all phases of test development.

1.1.1 Design

In collaboration with the CSDE and the American Institutes for Research (AIR), the CTAS committee selected a variety of NGSS Standard Performance Expectations that were appropriate for students with significant cognitive disabilities in order to create derived Essence Statements. Essence Statements capture the most important elements of each standard and make them more accessible to participating students. The NGSS Standard Performance Expectations and CTAS Essence Statements were used to develop the assessment.

Each CTAS Essence Statement is associated with 2–4 Core Extensions. The extensions describe specific student performances and are connected to activities, which are to be administered to the student by the Trained Teacher Alternate Assessment (TEA). The Trained TEA then rates the student’s performance on a 0–2 scale. Additional details regarding rating/scoring procedures are included in the Student Score Worksheet. Figure 1 is a diagram of the primary components of the CTAS Assessments.

Figure 1. Primary Components of the Connecticut Alternate Science Assessment



The CTAS Assessment has been organized into six storylines in each assessed grade (i.e., grades 5, 8, and 11) with two storylines per content area: Earth Science (Storylines 1 and 2); Life Science (Storylines 3 and 4); and Physical Science (Storylines 5 and 6).

Each storyline includes the NGSS Standard Performance Expectations, the derived CTAS Essence Statement, and the corresponding Core Extensions, which are directly aligned to the activities in the performance tasks. Each activity provides a coherent sequence of instruction for the Trained TEA on how to assess student performance associated with each Core Extension. These activities ask students to make sense of real-world phenomena and/or engage with an engineering design problem. Table 2 includes an overview of the each of the six storylines and associated performance tasks by content area.

Table 2. Storylines and Performance Tasks Overview

Content Area	Storyline Number	Storyline and Performance Task	Grade-Level Performance Task (PT)
Earth Science	1	Earth Systems	Grade 5
			Grade 8
			Grade 11
	2	Natural Resources	Grade 5
			Grade 8
			Grade 11
Life Science	3	Living Organisms	Grade 5
			Grade 8
			Grade 11
	4	Healthy Ecosystems	Grade 5
			Grade 8
			Grade 11
Physical Science	5	Forces and Motion	Grade 5
			Grade 8
			Grade 11
	6	Using Energy Every Day	Grade 5
			Grade 8
			Grade 11

2. 2023 Administration

2.1 Testing Window

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

2.2 Test Forms

In 2023, one test form was administered for each grade. Each form contains six performance tasks. Each performance task follows a storyline and guiding questions to engage students in making sense of the scientific phenomena or thinking about an engineering design problem. Each performance task contains a list of activities supporting the storyline. For the grade 5 form, the test consists of 44 activities/items, and grades 8 and 11 tests have 42 activities/items. Table 1 displays the number of items in each strand.

2.3 Test Mode

Test administrators (TAs) entered ratings to the activities into the online data entry 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 the Connecticut Alternate Science (CTAS) Assessment, an early stopping rule (ESR) is established. This rule allows students who have difficulties taking the assessments to exit the tests after attempting the first activity in the first performance task. If a student does not respond to the first item in the first performance task, 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. If the student does not qualify for the ESR, the TA must resume the assessment and the student will have to answer the rest of the items through the end of the assessment.

2.4.1 Item Difficulty

Since the assessment contains only selected-response items, AIR computes 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. Table 3 presents the summary of the *p*-values. The average *p*-value was 0.49 for grades 5, 0.50 for grade 8, and 0.53 for grade 11. There were no items with *p*-values below 0.2 or above 0.9.

Table 3. Summary of Item Difficulty

Grade	Item Count	Minimum	Maximum	Mean	Standard Deviation
5	44	0.29	0.74	0.49	0.12
8	42	0.31	0.79	0.50	0.10
11	42	0.35	0.74	0.53	0.09

2.4.2 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:

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

where

- x is the overall test score, excluding the item under evaluation;
- 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.

Table 4 displays the summary of the point-biserial correlation. All items in all three grades had point-biserial values above 0.60.

Table 4. Summary of Item Discrimination

Grade	Item Count	Minimum	Maximum	Mean	Standard Deviation
5	44	0.60	0.92	0.79	0.07
8	42	0.69	0.89	0.79	0.05
11	42	0.67	0.88	0.77	0.06

3. 2023 State Data Summary

3.1 Student Participation

This section describes the demographics of participating students in spring 2023. Table 5 and Table 6 present the student demographics for participating students by gender and ethnicity in each grade.

Demographic characteristics of the student population were relatively consistent across grades. Approximately 32–34% of students in each grade were female.

Among the participants, white students (32–42%) and Hispanic students (31–38%) made up the majority of the assessed students. African American students made up 19–21%, Asian students made up 4–7%, and multiracial students made up about 3–4% of the assessed students.

Table 5. Participation by Grade and Gender

Grade	Total		Female		Male		Missing	
	N	%	N	%	N	%	N	%
5	480	100	152	31.7	328	68.3	-	-
8	391	100	128	32.7	263	67.3	-	-
11	433	100	146	33.7	287	66.3	-	-
Total	1304	100	426	32.7	878	67.3	-	-

Table 6. Participation by Ethnicity

Grade	Total		American Indian or Alaskan Native		Asian		Black or African American		Hispanic or Latino		Pacific Islander		Two or more races		White	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
5	480	100	1	0.2	34	7.1	93	19.4	180	37.5			21	4.4	151	31.5
8	391	100	1	0.3	15	3.8	82	21	139	35.5			14	3.6	140	35.8
11	433	100	4	0.9	18	4.2	80	18.5	136	31.4	1	0.2	12	2.8	182	42
Total	1304	100	6	0.5	67	5.1	255	19.6	455	34.9	1	0.1	47	3.6	473	36.3

3.2 Scoring

Student responses to CTAS Assessment items are coded according to the rating scale in Table 7. An item has a rating of 0, 1, or 2. No missing response is allowed.

Table 7. CTAS Item Scoring Rubric

2	1	0
Mastered/Independent	**Developing/Supported**	**Does Not Demonstrate**
The student demonstrates understanding independently without scaffolding.	The student demonstrates limited understanding typically requiring additional support through scaffolding.	The student does not demonstrate understanding.

After the spring 2023 administration, a standard-setting workshop was held to determine cut scores for three performance standards for each test. Based on the standard setting results, the raw scores for the CTAS Assessment are mapped into four performance levels:

1. Does Not Meet
2. Approaching
3. Meets
4. Exceeds

The process and detailed results of standard setting are described in the CTAS Assessment standard setting technical report. Table 8 lists cut scores for each test.

Table 8. Performance Level Cut Points for CTAS

Grade	Does Not Meet	Approaching	Meets	Exceeds
5	0–31	32–56	57–64	65–88
8	0–25	26–56	57–63	64–84
11	0–31	32–56	57–64	65–84

3.3 Score Summary

Table 9 presents the summary statistics of the raw score by grade. The mean raw score ranged from 36.6 to 38.8. Each item is worth two score points.

Table 9. Raw Score Summary

Grade	N	N of Items	Mean	Median	STD	Min	Max
5	480	44	37.3	40	26	0	88
8	391	42	36.6	40	25.2	0	84
11	433	42	38.8	42	24.9	0	83

Table 10 shows the summary statistics of the raw score by each performance task. In grade 5, the average score is between 4.1 and 8.4. The average score ranged from 3.6 to 9.3 in grade 8, and from 3.9 to 8.2 in grade 11.

Table 10. Raw Score Summary by Performance Task

Grade	Performance Task	Max Possible Score Points	MEAN	MEDIAN	STD	MIN	MAX
5	1	18	8.4	9	5.5	0	18
	2	18	6.8	7	5.4	0	18
	3	10	4.1	4	3.1	0	10
	4	16	6.3	7	4.8	0	16

Grade	Performance Task	Max Possible Score Points	MEAN	MEDIAN	STD	MIN	MAX
	5	14	5.7	5	4.9	0	14
	6	12	5.9	7	4.1	0	12
8	1	16	7.4	8	5.4	0	16
	2	20	9.3	10	6.1	0	20
	3	10	4.5	5	3.3	0	10
	4	16	6.5	7	4.8	0	16
	5	10	3.6	4	3.1	0	10
	6	12	5.3	6	4.2	0	12
11	1	14	7	7	4.4	0	14
	2	18	8	8	5.8	0	18
	3	14	7.1	8	4.6	0	14
	4	18	8.2	9	5.6	0	18
	5	10	3.9	4	3.1	0	10
	6	10	4.6	5	3.3	0	10

Appendix B presents the raw score distribution, and Appendix C lists the raw score summary by subgroups.

3.4 Percentage of Students by Performance Level

The percentage of students in each performance level is listed in Table 11. About a third of the students were in Level 1. The percentage of students in each performance level by subgroup is listed in Appendix D.

Table 11. Percentage of Students by Performance Level

Grade	Total N	Level 1 (%)	Level 2 (%)	Level 3 (%)	Level 4 (%)
5	480	40.2	33.1	6.7	20
8	391	34	38.9	11.3	15.9
11	433	38.1	34.2	9	18.7

4. Reporting

The CTAS Assessment results were provided in two mediums: (1) the Online Reporting System (ORS), and (2) 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, and districts. 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 the school performance can be compared with district and state performance. If a teacher is selected, the summary results for the school, district, and state are also provided for comparison purposes. Table 12 lists the types of online reports and the levels at which they can be viewed (i.e., 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 (e.g., teachers, parents, students) who are not technical measurement experts. 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.

The online score reports are presented hierarchically once authorized users log in to the ORS and select “Score Reports.” 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 12 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 ORS can be found in the *Online Reporting System User Guide*, accessible using the “Help” button in the ORS.

Table 12. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
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 to users. Users can see student assessment results by any subgroup. Table presents the types of subgroups and subgroup categories provided in the ORS.

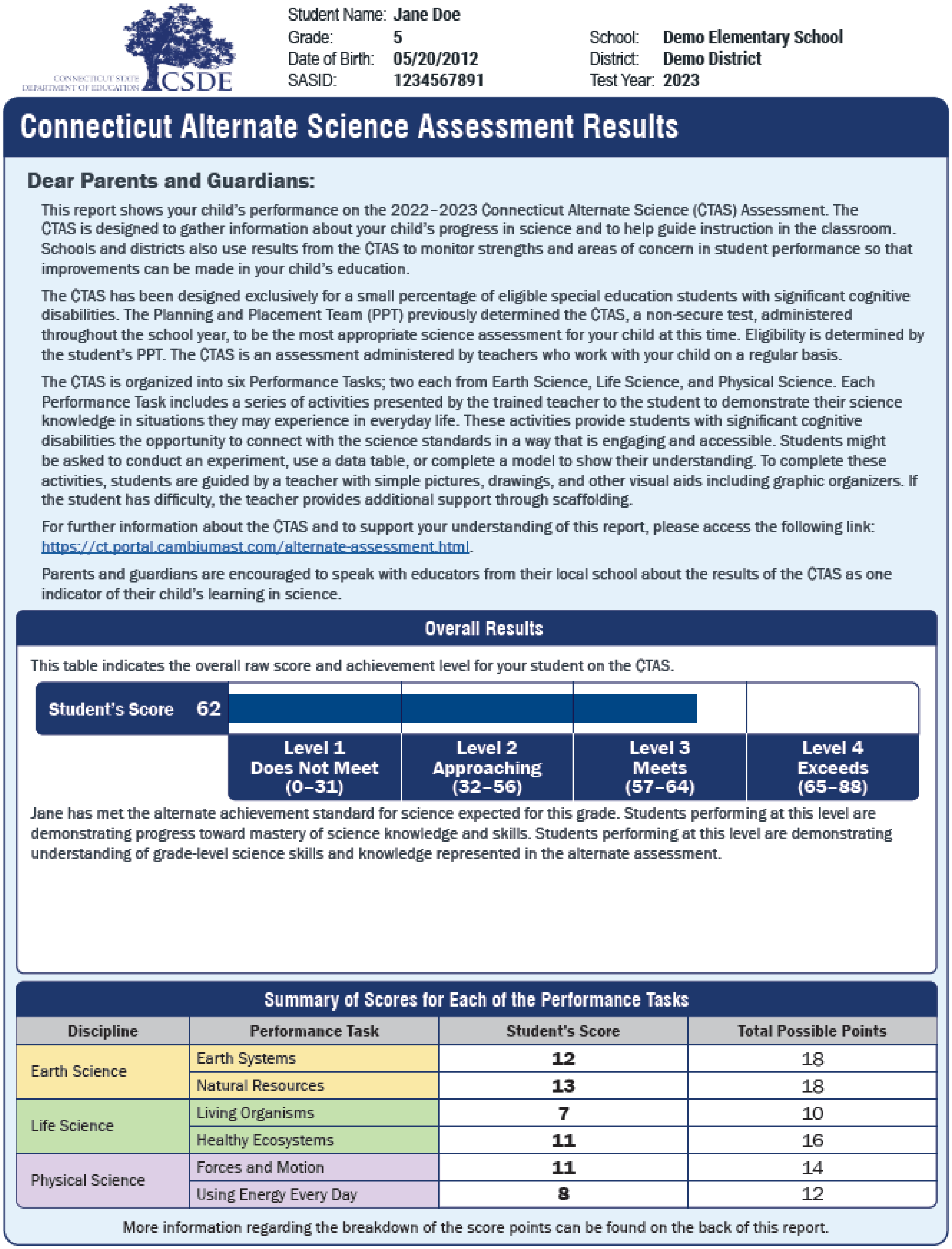
Table 13. Types of Subgroups

Breakdown by Category	Displayed Category
Ethnicity	Hispanic or Latino
	American Indian or Alaska Native
	Asian
	Black or African American
	White
	Native Hawaiian or Other Pacific Islander
	Two or More Races
Gender	Male
	Female
IDEA (Individuals with Disabilities Education Act) Indicator	Special Education
	Unknown
Limited English Proficiency Status	Yes
	Unknown
Enrolled Grade	Grade 5
	Grade 8
	Grade 11

4.2 Paper Report

Paper reports for the CTAS Assessment were also printed and shipped to the district at the end of the administration. Figure 2 shows the mock-up of the family report for students who finished the tests.

Figure 2. Family Report Mock-Up



Student Name: **Jane Doe**
Grade: **5**
Date of Birth: **05/20/2012**
SASID: **1234567891**

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

Connecticut Alternate Science Assessment Results

Dear Parents and Guardians:

This report shows your child's performance on the 2022–2023 Connecticut Alternate Science (CTAS) Assessment. The CTAS is designed to gather information about your child's progress in science and to help guide instruction in the classroom. Schools and districts also use results from the CTAS to monitor strengths and areas of concern in student performance so that improvements can be made in your child's education.

The CTAS has been designed exclusively for a small percentage of eligible special education students with significant cognitive disabilities. The Planning and Placement Team (PPT) previously determined the CTAS, a non-secure test, administered throughout the school year, to be the most appropriate science assessment for your child at this time. Eligibility is determined by the student's PPT. The CTAS is an assessment administered by teachers who work with your child on a regular basis.

The CTAS is organized into six Performance Tasks; two each from Earth Science, Life Science, and Physical Science. Each Performance Task includes a series of activities presented by the trained teacher to the student to demonstrate their science knowledge in situations they may experience in everyday life. These activities provide students with significant cognitive disabilities the opportunity to connect with the science standards in a way that is engaging and accessible. Students might be asked to conduct an experiment, use a data table, or complete a model to show their understanding. To complete these activities, students are guided by a teacher with simple pictures, drawings, and other visual aids including graphic organizers. If the student has difficulty, the teacher provides additional support through scaffolding.

For further information about the CTAS and to support your understanding of this report, please access the following link: https://ct.portal.cambiumast.com/alternate-assessment.html.

Parents and guardians are encouraged to speak with educators from their local school about the results of the CTAS as one indicator of their child's learning in science.

Overall Results

This table indicates the overall raw score and achievement level for your student on the CTAS.

Student's Score 62

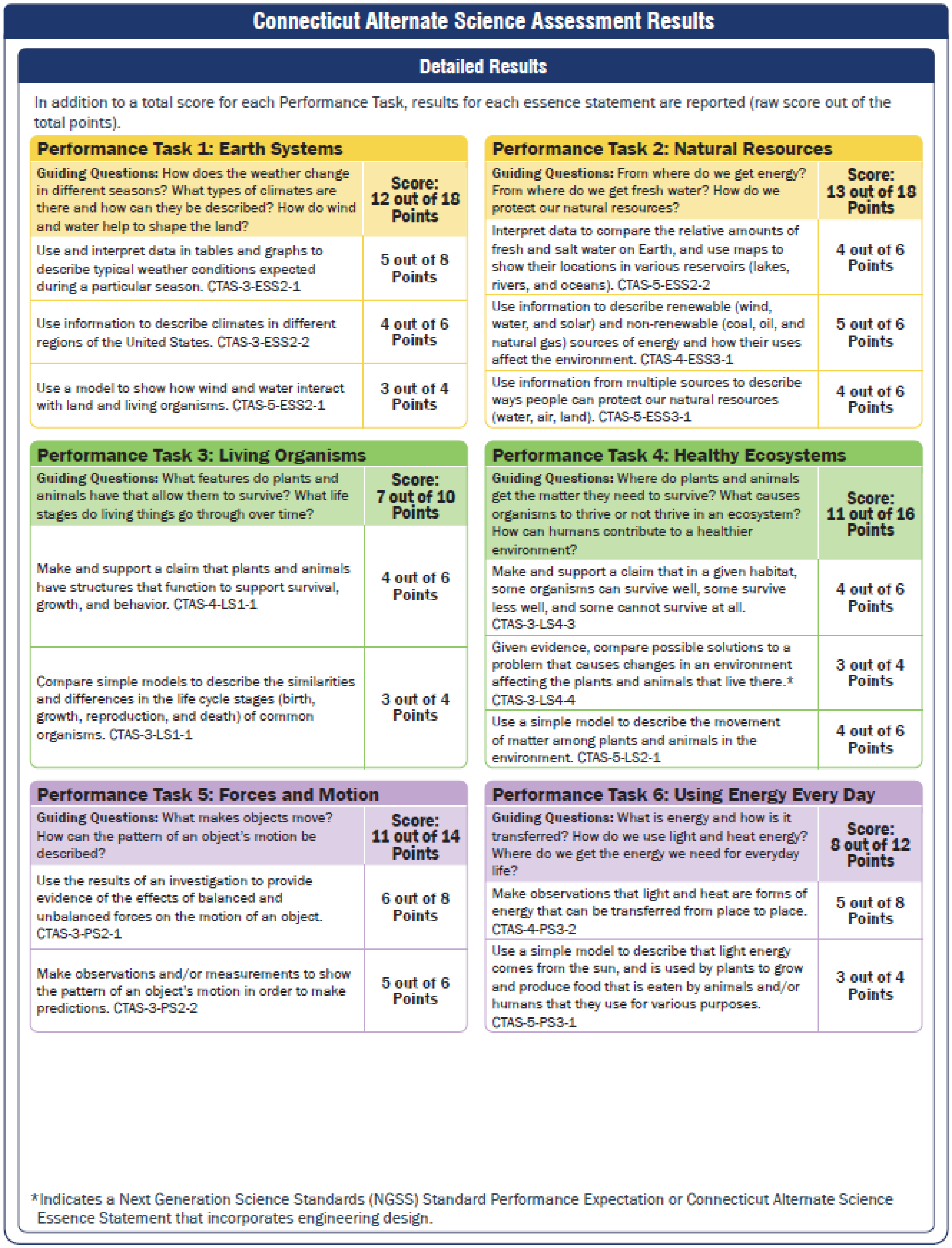
Level 1 Does Not Meet (0–31)	Level 2 Approaching (32–56)	Level 3 Meets (57–64)	Level 4 Exceeds (65–88)

Jane has met the alternate achievement standard for science expected for this grade. Students performing at this level are demonstrating progress toward mastery of science knowledge and skills. Students performing at this level are demonstrating understanding of grade-level science skills and knowledge represented in the alternate assessment.

Summary of Scores for Each of the Performance Tasks

Discipline	Performance Task	Student's Score	Total Possible Points
Earth Science	Earth Systems	**12**	18
	Natural Resources	**13**	18
Life Science	Living Organisms	**7**	10
	Healthy Ecosystems	**11**	16
Physical Science	Forces and Motion	**11**	14
	Using Energy Every Day	**8**	12

More information regarding the breakdown of the score points can be found on the back of this report.



Connecticut Alternate Science Assessment Results

Detailed Results

In addition to a total score for each Performance Task, results for each essence statement are reported (raw score out of the total points).

Performance Task 1: Earth Systems	
Guiding Questions: How does the weather change in different seasons? What types of climates are there and how can they be described? How do wind and water help to shape the land?	**Score: 12 out of 18 Points**
Use and interpret data in tables and graphs to describe typical weather conditions expected during a particular season. CTAS-3-ESS2-1	5 out of 8 Points
Use information to describe climates in different regions of the United States. CTAS-3-ESS2-2	4 out of 6 Points
Use a model to show how wind and water interact with land and living organisms. CTAS-5-ESS2-1	3 out of 4 Points

Performance Task 2: Natural Resources	
Guiding Questions: From where do we get energy? From where do we get fresh water? How do we protect our natural resources?	**Score: 13 out of 18 Points**
Interpret data to compare the relative amounts of fresh and salt water on Earth, and use maps to show their locations in various reservoirs (lakes, rivers, and oceans). CTAS-5-ESS2-2	4 out of 6 Points
Use information to describe renewable (wind, water, and solar) and non-renewable (coal, oil, and natural gas) sources of energy and how their uses affect the environment. CTAS-4-ESS3-1	5 out of 6 Points
Use information from multiple sources to describe ways people can protect our natural resources (water, air, land). CTAS-5-ESS3-1	4 out of 6 Points

Performance Task 3: Living Organisms	
Guiding Questions: What features do plants and animals have that allow them to survive? What life stages do living things go through over time?	**Score: 7 out of 10 Points**
Make and support a claim that plants and animals have structures that function to support survival, growth, and behavior. CTAS-4-LS1-1	4 out of 6 Points
Compare simple models to describe the similarities and differences in the life cycle stages (birth, growth, reproduction, and death) of common organisms. CTAS-3-LS1-1	3 out of 4 Points

Performance Task 4: Healthy Ecosystems	
Guiding Questions: Where do plants and animals get the matter they need to survive? What causes organisms to thrive or not thrive in an ecosystem? How can humans contribute to a healthier environment?	**Score: 11 out of 16 Points**
Make and support a claim that in a given habitat, some organisms can survive well, some survive less well, and some cannot survive at all. CTAS-3-LS4-3	4 out of 6 Points
Given evidence, compare possible solutions to a problem that causes changes in an environment affecting the plants and animals that live there.* CTAS-3-LS4-4	3 out of 4 Points
Use a simple model to describe the movement of matter among plants and animals in the environment. CTAS-5-LS2-1	4 out of 6 Points

Performance Task 5: Forces and Motion	
Guiding Questions: What makes objects move? How can the pattern of an object's motion be described?	**Score: 11 out of 14 Points**
Use the results of an investigation to provide evidence of the effects of balanced and unbalanced forces on the motion of an object. CTAS-3-PS2-1	6 out of 8 Points
Make observations and/or measurements to show the pattern of an object's motion in order to make predictions. CTAS-3-PS2-2	5 out of 6 Points

Performance Task 6: Using Energy Every Day	
Guiding Questions: What is energy and how is it transferred? How do we use light and heat energy? Where do we get the energy we need for everyday life?	**Score: 8 out of 12 Points**
Make observations that light and heat are forms of energy that can be transferred from place to place. CTAS-4-PS3-2	5 out of 8 Points
Use a simple model to describe that light energy comes from the sun, and is used by plants to grow and produce food that is eaten by animals and/or humans that they use for various purposes. CTAS-5-PS3-1	3 out of 4 Points

*Indicates a Next Generation Science Standards (NGSS) Standard Performance Expectation or Connecticut Alternate Science Essence Statement that incorporates engineering design.

5. Reliability and Validity

With the implementation of the Connecticut Alternate Science (CTAS) Assessments, both reliability evidence and validity evidence are necessary to support appropriate inferences of student's achievement from the CTAS Assessment scores. This section provides empirical evidence about the reliability and validity of the 2022–2023 CTAS Assessment, given its intended uses.

Cronbach's alpha, 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:

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

There are various approaches for estimating the reliability of scores. Among the various approaches for estimating the reliability of scores, 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

$$\propto = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_x^2}\right] \quad (3),$$

where n is the sample size, and σ_i^2 is the raw score variance for item i. σ_x^2 is the variance of the total raw scores.

The Cronbach's alpha coefficients are summarized in Table 14. The data files for reliability analyses excludes students with the early stopping rule (ESR) flag. In addition, the computation of Cronbach's alpha requires the full response matrix; therefore, the sample sizes are smaller. Grades 5 and 11 have the alpha coefficient of 0.97, and Grades 8 has the alpha coefficient of 0.98.

Table 14. Cronbach's Alpha

Grade	Sample Size	Number Items	Alpha
5	400	44	0.97
8	319	42	0.98
11	361	42	0.97

5.1.2 Standard Error of Measurement

Another way to view reliability is to consider its relationship with the Standard Errors of Measurement (SEM)—the smaller the standard error, the higher the precision of the test scores. For example, the Classical Test Theory (CTT) assumes that an observed score (X) of each individual can be expressed as a true score (T) plus some error (E), $X = T + E$. The variance of X can be shown to be the sum of two orthogonal variance components:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \quad (4).$$

Returning to the definition of reliability as the ratio of true score variance to observed score variance, the following applies:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \quad (5).$$

As the fraction of error variance to observed score variance tends to zero, the reliability then tends to 1. The SEM of the CTT, which assumes a homoscedastic error, is derived from the classical notion expressed earlier as $\sigma_X\sqrt{1 - \rho_{XX'}}$, where σ_X is the standard deviation of the scaled score and $\rho_{XX'}$ is a reliability coefficient. Based on the definition of reliability, this formula can be derived:

$$\rho_{XX'} = 1 - \frac{\sigma_E^2}{\sigma_X^2},$$

$$\frac{\sigma_E^2}{\sigma_X^2} = 1 - \rho_{XX'},$$

$$\sigma_E^2 = \sigma_X^2(1 - \rho_{XX'}),$$

$$\sigma_E = \sigma_X\sqrt{(1 - \rho_{XX'})} \quad (6).$$

Table 15 presents the SEM of each test. The SEM can be interpreted with the confidence interval. For example, if a grade 5 student obtains a score of 40, there are two out of three chances (68%) that the student's true score would fall between 40–3.76 and 40+3.76.

Table 15. Standard Error of Measurement

Grade	Reliability	SD of Observed Score	SEM
5	0.97	23.75	3.76
8	0.98	22.71	3.58
11	0.97	22.27	3.70

5.1.3 Classification Accuracy and Consistency

Students are placed into one of four performance levels given their raw score. As described above, the cut scores for student classification into the different performance levels were determined after the CTAS Assessment standard-setting process.

Classification accuracy refers to the degree to which a student's true score and observed score would fall within the same performance level. Classification consistency refers to the degree to which examinees are classified into the same performance level, assuming the test is administered twice independently—that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test forms. In reality, however, the true ability is unknown, and students do not take an alternate, equivalent form.

The Livingston and Lewis (1995) method was used to compute classification accuracy and consistency. For classification consistency, the observed score distribution and the observed score distribution for a parallel form predicted from the beta-binomial model were compared. For classification accuracy, the observed score distribution and the true score distribution predicted from the beta-binomial model were compared. The distribution of true scores is estimated by fitting a four-parameter beta distribution. The parameters are estimated from the observed distribution.

Table 16 and Table 17 display classification accuracy and consistency, respectively. Overall, classification accuracy falls between 0.85 and 0.87, which suggests 85–87% of the students estimated to have a true score status are correctly classified into that category by their observed scores. The false positive rate is expressed as the proportion of individuals who scored above the cut score based on their observed score, but their true score would otherwise have classified them as below the cut score. The false negative rate is expressed as the proportion of individuals who scored below the cut score based on their observed score, but otherwise would have been classified as above the cut score based on their true scores. The false positive rate is 5–8%, and the false negative rate is 8–9%.

The range of classification consistency is from 0.76 to 0.82. Kappa values are between 0.66 and 0.72. 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.

Table 16. Classification Accuracy

Grade	Accuracy	False Positive	False Negative
5	0.85	0.08	0.09
8	0.87	0.05	0.08
11	0.85	0.06	0.08

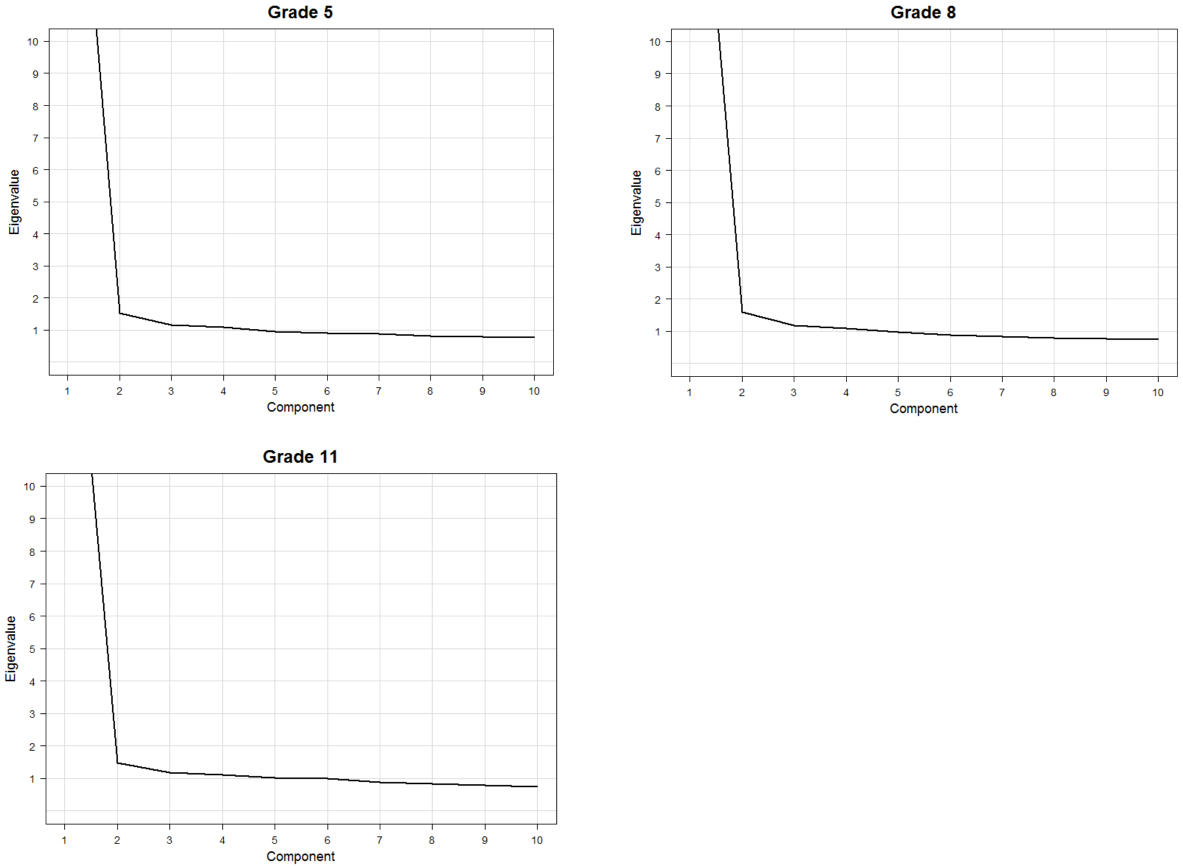
Table 17. Classification Consistency

Grade	Consistency	Kappa	Probability of Misclassification
5	0.80	0.69	0.20
8	0.82	0.75	0.18
11	0.79	0.68	0.21

5.1.4 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 3. The graphs show that the first component explains the majority of the variation. The PCA results suggest that the forms measure one dominant construct.

Figure 3. Scree Plots



5.2 EVIDENCE ON INTERNAL-EXTERNAL STRUCTURE

5.2.1 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.

Each grade has three strands: Earth Science (ES), Life Science (LS), and Physical Science (PS). 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. On the contrary, if the standards were perfectly correlated, we could justify a unidimensional model.

One pathway to explore the internal structure of the test is to explore 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 Table 18 and Table 19, respectively.

Table 18 and Table 19 present the observed and disattenuated correlation matrix of the strand raw scores. The correlations among the standards range from 0.85 to 0.91. Disattenuated correlations range from 0.91 to 0.98. Disattenuated correlations greater than 1.00 are reported as 1.00**. 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.

Table 18. Observed Correlation Matrix Among Standards

GRADE	STANDARDS	NUMBER OF ITEMS	ES	LS	PS
5	ES	18	1.00		
	LS	13	0.91	1.00	
	PS	13	0.9	0.9	1.00
8	ES	18	1.00		
	LS	13	0.91	1.00	
	PS	11	0.86	0.86	1.00
11	ES	16	1.00		
	LS	16	0.9	1.00	
	PS	10	0.85	0.85	1.00

Table 19. Disattenuated Correlation Matrix Among Standards

GRADE	STANDARDS	NUMBER OF ITEMS	ES	LS	PS
5	ES	18	1.00**		
	LS	13	0.98	1.00**	
	PS	13	0.96	0.97	1.00**
8	ES	18	1.00**		
	LS	13	0.97	1.00**	
	PS	11	0.91	0.92	1.00**
11	ES	16	1.00**		
	LS	16	0.96	1.00**	
	PS	10	0.92	0.93	1.00**

6. Quality Control

Thorough quality control has been integrated into every aspect of the Connecticut Alternate Science (CTAS) Assessment 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 specifications, 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 to 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.

The American Institutes for Research's (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 for TDS and the Online Reporting System (ORS) before the testing window was opened.

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 the Connecticut State Department of Education (CSDE) could conduct a hands-on review of the system prior to the opening of the testing window. The CSDE approved TDS before the system was opened for testing.

For the ORS, the same procedure is followed. Both AIR and CSDE staff conducted internal and external UAT of the system to ensure that the system functions 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 AIR's quality assurance (QA) system. The QA system 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 Monitor 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 CSDE. AIR staff ensure that data in the extract files match the DoR prior to delivery to the CSDE.

6.5 Quality Control on Scoring

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

When the testing window closes, psychometricians score the operational records and compare them with the scores from the scoring engine again. All discrepancies are investigated and resolved before scores are released to the state and students.

6.6 Quality Assurance in Reporting

Two types of score reports were produced for the CTAS Assessments: (1) online reports and (2) 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 specifications in our reporting specifications document. 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 independently implemented 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. The AIR team 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 made up mostly of calls to various macros, including macros that read in and verify the data, 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, AIR’s VIPP programmers write code that indicates where to place all variable information (i.e., 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 program testing 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, with the output 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 data are correctly displayed. Once AIR receives final data and VIPP programs, the AIR score reporting team reviews proofs that contain actual data based AIR’s standard quality assurance documentation. In addition, the AIR score reporting team compares 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 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.

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. If discrepancies are found, actions are taken until all discrepancies are resolved. The sample reports are sent to the CSDE for approval. Upon the CSDE’s approval, the final student paper reports are produced and distributed.

7. References

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 & Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrica*, *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., and M. W. Richardson. (1937). The Theory and Estimation of Test Reliability. *Psychometrica*, *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: Erlb.

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. (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.