Colorado Alternate Assessment Program

Technical Report

Science

2023

Table of Contents

Chapter 1: Introduction 7
1.1. Testing Requirements 7
1.2. Intended Population 7
1.3. CoAlt Background 8
1.4. Purpose of CoAlt 9
1.5. Assessment Development Partners 9
1.5.1. Colorado Department of Education 9
1.5.2. Colorado Educator Community 10
1.5.3. Pearson 10
1.5.4. Colorado Technical Advisory Committee 10
Chapter 2: Test Design 11
2.1. Alternate Academic Achievement Standards 11
2.2. Item Types 12
2.3. Test Frameworks and Blueprints 13
2.4. Performance Levels 14
2.5. Cognitive Complexity 15
2.6. Test Composition 15
2.7. Timing of Tests 16
Chapter 3: Item Development 17
3.1. Item Banking System 17
3.2. Item Development Plan 17
3.3. Item Writing 17
3.4. Item Review 18
3.4.1. Internal Review 18
3.4.2. External Content and Bias Review 18
3.5. Data Review 19
Chapter 4: Test Construction 21
Chapter 5: Test Administration 22
5.1. Manuals 22
5.2. Test Materials 22
5.3. Administration Training 23
5.4. Practice Resources 23
5.5. Accessibility Features and Accommodations 24
5.6. Test Security 25
5.7. Test Monitoring 25
5.7.1. Training 26
5.7.2. Process 26
5.7.3. Participation 26
5.7.4. Results 27
Chapter 6: Scoring 28
6.1. SR Scoring 28
6.2. SPT Scoring 28
Chapter 7: Standard Setting 29

Chapter 8: Reporting 30
8.1. Description of Scores 30
8.2. Score Reports 30
Chapter 9: Test Results and Analysis 31
9.1. Student Participation 31
9.2. Performance Results 31
9.3. Classical Item Analysis 32
9.4. Subclaim Correlations 33
Chapter 10: Calibration, Equating, and Scaling 34
10.1. IRT Model 34
10.2. Data Preparation 35
10.3. Calibration 35
10.4. Equating 36
10.5. Scaling 36
Chapter 11: Reliability 38
11.1. Internal Consistency (Coefficient Alpha) 38
11.2. Standard Error of Measurement (SEM) 39
11.3. Conditional Standard Error of Measurement (CSEM) 40
11.4. Decision Consistency and Accuracy 40
Chapter 12: Validity 42
12.1. Evidence Based on Test Content 42
12.2. Evidence Based on Response Processes 43
12.3. Evidence Based on Internal Structure 43
12.4. Evidence Based on Relations to Other Variables 43
12.5. Evidence for Validity and Consequences of Testing 44
12.6. Fairness 44
References 45
Appendix A: CoAlt Eligibility Guidelines 47
Appendix B: Sample Student Performance Report 49
Appendix C: Scale Score Distributions 51
Appendix D: Scale Score Distribution Histograms 54
Appendix E: Performance Results by Demographic Subgroup 56
Appendix F: Classical Item-Level Statistics 58
Appendix G: IRT Item-Level Statistics 61
Appendix H: Test Characteristic Curves (TCCs) 64
Appendix I: Test Information Curves (TICs) and CSEM Curves 66
Appendix J: Test Administrator Survey Responses 69

List of Tables

Table 1.1. Schedule of Major Events 10
Table 2.1. 2023 CoAlt Science Test Blueprint—Grade 5 14
Table 2.2. 2023 CoAlt Science Test Blueprint—Grade 8 14
Table 2.3. 2023 CoAlt Science Test Blueprint—Grade 11 14
Table 2.4. Performance Levels and Policy Claims 15
Table 2.5. 2023 CoAlt Science Test Designs 15
Table 3.1. Item Statistical Flagging Criteria 19
Table 3.2. Data Review Results 20
Table 5.1. Test Materials 23
Table 5.2. Number of Participating Schools in Test Monitoring 26
Table 5.3. Number of Participating Students (Observations) in Test Monitoring 26
Table 5.4. Test Monitoring Percent Agreement Rates between Transcribers 27
Table 6.1. SPT Scoring Rubric 28
Table 7.1. Performance Level Cut Scores 29
Table 9.1. Student Participation N-Count Demographic Distribution 31
Table 9.2. Scale Score Performance Summary and Performance Level Distributions 32
Table 9.3. Summary Statistics for Points Earned by Subclaim 32
Table 9.4. Correlations Between Subclaims 33
Table 11.1. Coefficient Alpha 39
Table 11.2. SEM 39
Table 11.3. Kappa Values 41
Table 11.4. Decision Consistency and Accuracy Estimates 41
Table 11.5. Accuracy of Cut Scores 41
Table 11.6. Consistency of Cut Scores 41
Table 12.1. Correlation Between Test Validity Questions and Student Scores 44
Table C.1. Scale Score Distribution—Science Grade 5 51
Table C.2. Scale Score Distribution—Science Grade 8 52
Table C.3. Scale Score Distribution—Science Grade 11 53
Table E.1. Scale Score Summary Statistics by Demographic Subgroup—Grade 5 56
Table E.2. Scale Score Summary Statistics by Demographic Subgroup—Grade 8 56
Table E.3. Scale Score Summary Statistics by Demographic Subgroup—Grade 11 57
Table F.1. SR Item Classical Statistics—Science Grade 5 58
Table F.2. SPT Item Classical Statistics—Science Grade 5 58
Table F.3. SR Item Classical Statistics—Science Grade 8 59
Table F.4. SPT Item Classical Statistics—Science Grade 8 59
Table F.5. SR Item Classical Statistics—Science Grade 11 60
Table F.6. SPT Item Classical Statistics—Science Grade 11 60
Table G.1. Operational Item Parameter Estimates—Science Grade 5 61
Table G.2. Operational Item Parameter Estimates—Science Grade 8 62
Table G.3. Operational Item Parameter Estimates—Science Grade 11 63

List of Figures

Figure D.1. Scale Score Distribution Histogram—Grade 5 54
Figure D.2. Scale Score Distribution Histogram—Grade 8 54
Figure D.3. Scale Score Distribution Histogram—Grade 11 55
Figure H.1. TCC—Grade 5 64
Figure H.2. TCC—Grade 8 64
Figure H.3. TCC—Grade 11 65
Figure I.1. TIC—Grade 5 66
Figure I.2. TIC—Grade 8 66
Figure I.3. TIC—Grade 11 67
Figure I.4. CSEM Curve—Grade 5 67
Figure I.5. CSEM Curve—Grade 8 68
Figure I.6. CSEM Curve—Grade 11 68

Chapter 1: Introduction

The purpose of this technical report is to inform users and other interested parties about the development, content, administration, and technical characteristics of the Spring 2023 Colorado Alternate (CoAlt) Science assessment in Grades 5, 8, and 11 for students with the most significant cognitive disabilities. The report includes an overview and summary of the components of the program, including information regarding the planning and administration of the assessments and details regarding item development, test construction, administration procedures, scoring, reporting, reliability, and validity, as well as a statistical summary of the Spring 2023 operational items.

1.1. Testing Requirements

All public schools in Colorado are required by state law to administer a standards-based summative assessment each year in specified content areas and grade levels. Every student, regardless of ability or language background, must be provided with the opportunity to demonstrate their content knowledge through the state assessments. The Colorado Measures of Academic Success (CMAS) assessments are Colorado's end-of-year standards-based assessments designed to measure students' achievement of the grade-level Colorado Academic Standards (CAS).

The Individuals with Disabilities in Education Act of 2004 (IDEA) mandates that all students have access to the general curriculum and be included in each state's accountability system. The Every Student Succeeds Act of 2015 (ESSA) continues to specify that states must provide an alternate assessment when implementing statewide accountability systems to help ensure the inclusion of all students in a state's accountability system. To ensure the participation of all students with the most significant cognitive disabilities, Colorado developed the CoAlt Science assessments aligned to Colorado's alternate academic achievement standards known as the Extended Evidence Outcomes (EEOs) of the CAS.[1]

In 2015, Colorado passed legislation (C.R.S. §22-7-1013 (8) (a-c)) that allows for parents/guardians to excuse their child(ren) from testing.

1.2. Intended Population

The CoAlt assessments are designed for students with the most significant cognitive disabilities who have significant limitations in cognitive functioning and deficits in adaptive behavior. These students may also exhibit limitations in communication, methods of response, sustaining attention, and short-term memory. A very small number of students with the most significant cognitive disabilities who cannot participate in the CMAS assessment, even with accommodations, may take the CoAlt assessment. These students must be identified as having a significant cognitive disability, although Intellectual Disability does not have to be the student's primary disability label for IDEA eligibility.

[1] The CoAlt English Language Arts (ELA) and Mathematics assessments are administered by the Dynamic Learning Maps (DLM) consortium and are documented in a separate technical report located online at https://dynamiclearningmaps.org/publications. Social studies was not assessed in Spring 2023.

CoAlt participation is determined by a student's Individualized Education Program (IEP) team that decides whether the student meets the criteria in the alternate academic achievement standards and the Alternate Assessment Participation Guidelines Worksheet provided in Appendix A:.[2] The IEP team can decide that the CoAlt assessment is most appropriate if the student meets all the following participation criteria:

- The student has been evaluated and determined to be eligible to receive special education services and has an IEP.
- The student has documented evidence of a most significant cognitive disability.
- The student has a significant cognitive disability.
- The student receives daily instruction based on the alternate academic achievement standards.

1.3. CoAlt Background

The CoAlt assessments follow the direction of the Office of Standards and Instruction (SIS) and Exceptional Student Services Unit (ESSU) at the Colorado Department of Education (CDE). A key element in ESSA is that alternate assessments must be aligned with the content standards for the grade level in which the student is enrolled. The CAS for science were originally adopted in December 2009. On August 3, 2011, the State Board of Education adopted the EEOs of the CAS for students with the most significant cognitive disabilities who qualify for an alternate assessment. In partnership with Colorado educators and Pearson, CDE developed the CoAlt Science assessments to evaluate student mastery of the CAS in science for students with the most significant cognitive disabilities. For eligible students, these end-of-year assessments provide an indicator of student progress toward the EEOs of the CAS, known as the alternate academic achievement standards. The first operational administration of the CoAlt Science assessments occurred in Spring 2014 for Grades 5 and 8 and in Fall 2014 for Grade 11.

The Spring 2020 CoAlt administration was cancelled due to the COVID-19 pandemic. In 2021, Colorado received a partial waiver of the federal assessment requirements from the U.S. Department of Education (USED) due to COVID-19 conditions in Colorado. With the exception of students with a parent/guardian excusal, only students in Grades 8 and 11 took the CoAlt Science assessment. The Grades 4 and 7 social studies assessments were also not administered.

In 2022, newly revised CAS were implemented for mathematics, ELA, and science. In 2008, Colorado passed Senate Bill 212 (also known as CAP4K) that required the State Board of Education to adopt content standards that prepare students for the 21st century workforce and for active citizenship upon receiving a high school diploma. It also required a revision to the CAS by July 1, 2018, and every six years thereafter. As such, the 2009/2010 CAS were reviewed and revised, resulting in the 2020 CAS. While minimal changes were made to the mathematics and ELA CAS, the science CAS underwent a substantial update to keep up with the shift to the Next Generation Science Standards (NGSS; NGSS Lead States, 2013). After the CAS were adopted, a committee of both special and content educators convened to adapt the Evidence Outcomes (EOs) from the 2020 CAS to the EEOs to which the CoAlt is aligned.

[2] The participation guideline worksheet is also available online at http://www.cde.state.co.us/cdesped/accommodationsmanual_participationguidelinesworksheet.

Schools were asked to complete full instructional implementation of the new three-dimensional science standards by 2021–2022, with item development for the new CoAlt Science assessment beginning in Spring 2021. Colorado students saw items aligned to the 2020 CAS for the first time in Spring 2022. The new assessment was administered to all tested students, which made it possible to test enough new content to allow for a robust item bank and to obtain a sufficient sample of students to conduct field test analyses. Standard setting was conducted in Fall 2022 so that full results with scale scores and performance levels could be reported for the Spring 2023 administration.

While the Spring 2022 CoAlt Science assessment reported percentile ranks only, the Spring 2023 assessment reported scale scores and performance levels. Social studies was not administered in Spring 2022 or Spring 2023. Also beginning with the Spring 2022 administration, scannable answer documents are used for score entry, and online score entry through PearsonAccess^{next} is no longer used.

1.4. Purpose of CoAlt

The goals of the Colorado Assessment System, including the CoAlt assessments, are to measure and support student progress toward the content standards; provide students, parents/guardians, and other stakeholders with information regarding student achievement; and gauge the quality and efficiency of educational programs in public schools. For CoAlt in particular, the primary purpose of the assessment program is to determine the level at which Colorado students with significant cognitive disabilities meet the EEOs of the CAS. CoAlt also promotes improved instruction toward grade-level expectations, growth over time toward independent performance, and high expectations toward achievement in the content areas. CoAlt results may be used in many ways, including to

- inform instruction in the classroom;
- inform district and school leaders about potential programming and instruction priorities;
- provide the community with information on how well the state's education system is meeting the goals of helping every student attain academic proficiency in accordance with Colorado's alternate standards;
- provide aggregated data for the state's accountability system; and
- allow students to demonstrate their mastery of skills and concepts in the EEOs.

1.5. Assessment Development Partners

Activities specific to the CoAlt Science assessments were conducted collaboratively by CDE, the Colorado educator community, and Pearson, the assessment contractor. Input and advice were also provided by the Colorado Technical Advisory Committee (TAC).

1.5.1. Colorado Department of Education

As the administrative arm of the State Board of Education, CDE is responsible for implementing state and federal education laws. CDE's Assessment Unit works closely with Colorado school districts, educators, community stakeholders, and assessment development partners to develop and administer the state assessments. CDE focuses on creating assessments that serve students, schools, districts, and the community while complying with state and federal legal requirements. CDE also works closely with Pearson on each facet of the assessment, with CDE serving as the ultimate approver of services and products provided.

1.5.2. Colorado Educator Community

Educator participation in the CoAlt development process is critical to ensuring that the assessments are aligned to the EEOs of the CAS, are appropriate for Colorado students with the most significant cognitive disabilities at the assessed grade level, and are free from potential bias and sensitivity issues. Throughout the test development process, educators provide input through participation in content and bias review, data review, and standard setting meetings. For each meeting, an effort is made to involve educators who are representative of the entire state of Colorado, familiar with this population of students, and experts in the content areas assessed. Table 1.1 presents a schedule of major events from the 2022–2023 testing cycle that includes meetings with educator participation.

Table 1.1. Schedule of Major Events

Event	Date(s)
CoAlt Science Standard Setting	October 25–26, 2022
District Assessment Coordinator (DAC) Administration Training	November 14–16, 2022
Item Writer Training	January 24–26, 2023
Spring 2023 Administration Window	April 10–28, 2023
CoAlt Content and Bias	July 18–21, 2023
Reports Released	July 6, 2023
Data Review	August 4, 2023

1.5.3. Pearson

As the primary contractor responsible for the end-to-end assessment cycle services and products, Pearson works closely with CDE throughout the CMAS and CoAlt Science assessment development and administration processes. This includes item and test development, forms creation, enrollment, packaging and distribution, test delivery, scoring, customer service, standard setting, score reporting, and psychometric services.

1.5.4. Colorado Technical Advisory Committee

The Colorado TAC is comprised of psychometric, assessment, and special populations experts tasked with providing high-level consulting and expert advice regarding validity and reliability issues. Topics for which the TAC has provided input include the blueprint design, scaling and equating, scoring, reporting, alignment study feedback, peer review, and standard setting. The TAC included the following members during the 2023 assessment cycle:

- Dr. Elliot Asp, Senior Partner, The Colorado Education Initiative
- Dr. Jonathan Dings, Executive Director of Student Assessment and Program Evaluation, Boulder Valley School District
- Dr. Michael Kolen, Psychometric Consultant
- Dr. Suzanne Lane, Professor, University of Pittsburgh
- Dr. Martha Thurlow, Director, National Center on Educational Outcomes

Chapter 2: Test Design

The CoAlt Science assessment was designed to provide this unique population of students with the opportunity to demonstrate their knowledge of the EEOs. The assessment includes paper-based test books used by the test administrator to administer test items to the students. Each assessment is administered one-on-one and can be split over as many sessions/days as appropriate for the student. The test books are designed to sit on the table, allowing the test administrator to read the item to the student while the student views the answer options. The test books include scripted text for the test administrator to follow as they read the item stems and answer options to the student. There is flexibility for presentation and response based on the student's mode of communication, but the script and order in which the answer choices are presented to the student must remain the same.

2.1. Alternate Academic Achievement Standards

The EEOs are alternate academic achievement standards aligned to the grade-level 2020 CAS in science but reduced in depth, breadth, and complexity. They can be found online at http://www.cde.state.co.us/CoExtendedEO/StateStandards.

The standards are considered three-dimensional in that they incorporate Disciplinary Core Ideas (DCIs), Science and Engineering Practices (SEPs), and Crosscutting Concepts (CCCs). The DCIs encompass the content that occurs at each grade and provides the background knowledge for students to develop sense-making around phenomena in the three standards of Physical Science, Life Science, and Earth and Space Science:

- Physical Science: Students know and understand common properties, forms, and changes in matter and energy.
    - PS1: Matter and its interactions
    - PS2: Motion and stability: Forces and interactions
    - PS3: Energy
    - PS4: Waves and their applications in technologies for information transfer

- Life Science: Students know and understand the characteristics and structure of living things, the processes of life, and how living things interact with each other and their environment.
    - LS1: From molecules to organisms: Structures and processes
    - LS2: Ecosystems: Interactions, energy, and dynamics
    - LS3: Heredity: Inheritance and variation of traits
    - LS4: Biological evolution: Unity and diversity

- Earth and Space Science: Students know and understand the processes and interactions of Earth's systems and the structure and dynamics of Earth and other objects in space.
    - ESS1: Earth's place in the universe
    - ESS2: Earth's systems
    - ESS3: Earth and human activity

The SEPs describe how scientists investigate and build models and theories of the natural world or how engineers design and build systems. They reflect science and engineering as they are practiced and experienced. There are eight practices:

1. Asking questions (for science) and defining problems (for engineering)
2. Developing and using models
3. Planning and carrying out investigations
4. Analyzing and interpreting data
5. Using mathematics and computational thinking
6. Constructing explanations (for science) and designing solutions (for engineering)
7. Engaging in argument from evidence
8. Obtaining, evaluating, and communicating information

CCCs cross boundaries between science disciplines and provide an organizational framework to connect knowledge from various disciplines into a coherent and scientifically based view of the world. They build bridges between science and other disciplines and connect the DCIs and SEPs throughout the fields of science and engineering. There are seven CCCs:

1. Patterns
2. Cause and Effect
3. Scale, Proportion, and Quantity
4. Systems and System Models
5. Energy and Matter
6. Structure and Function
7. Stability and Change

The most substantial revision from the 2009 EEOs is the addition of a one-to one correspondence to each EO, thereby increasing the rigor for students with the most significant cognitive disabilities. Prior iterations of the EEOs had only 1–4 outcomes for each standard. SEPs and CCCs are incorporated into the EEOs, though not all EEOs are three-dimensional. SEPs and CCCs are also tested within the items.

The CoAlt Science assessment is administered in Grades 5, 8, and 11. Consistent with the standards, the Grade 5 assessment assesses the grade-level standards. Because the science standards are articulated by grade band at the middle school and high school levels rather than grade levels, the Grade 8 CoAlt Science assessment assesses all middle school science standards, and the Grade 11 assessment assesses all high school science standards.

2.2. Item Types

The CoAlt Science assessment includes 1-point selected response (SR) and 3-point supported performance task (SPT) item types.[3] The test administrator records student responses to the SR items and their scores on the SPT items on a scannable answer document included with the task manipulatives set provided for each test, which is then returned to Pearson for scoring.[4]

[3] Sample CoAlt Science items are available online at https://coassessments.com/practice-resources/.
[4] An example of the answer document is provided in the *CoAlt Test Administrator Manual* available online at https://coassessments.com/manuals/.

SR items are scaffolded items presented in a three-item cluster set (Part A, Part B, and Part C items) that correspond to the same phenomenon-based stimulus but are unrelated to each other. The stimulus provides the phenomenon that students reference to answer each item, and the art is repeated on the student-facing page with each item. The items are organized as follows:

- The first item in the set (Part A) has three picture answer options and is one-dimensional, testing only the DCI from the EEO. These items do not require sensemaking (i.e., the items are DOK Level 1, meaning they are just recall and do not require the student to figure something out).
- The Part B item has three picture answer options and is two-dimensional, requiring sensemaking and testing the DCI and either the SEP or CCC.
- The Part C item has four answer options that are primarily picture-based (and rarely text-based). It is three-dimensional and requires sensemaking.

SPT items consist of three related prompts (i.e., address the same EEO) that students respond to by placing a set of option cards in designated boxes within a chart or graphic. Students may manipulate the option cards independently or indicate the desired placement through their preferred mode of expressive communication, such as verbal directions, pointing, or eye gaze. Test administrators score the student's performance on each prompt using a 1-point scoring rubric that is built into the item (1 if the student responds correctly, 0 if the student responds incorrectly, NR if the student does not respond). The points for the three prompts are added together to provide one score for the SPT item. This item type reveals a different level of understanding of specific concepts and skills than those demonstrated through SR items alone. These items are all three-dimensional, are phenomenon based, and require sensemaking.

SPT items are classified as "Give a card" or "Find a card." For "Give a card" items, the test administrator gives the student a card to place in a table or other graphic organizer. The tasks have three answer cards, all of which are used. For "Find a card" items, the test administrator asks the student to search for a card of four provided cards in response to an item and place that card in a table or other graphic organizer. These tasks have four answer cards, one of which is not used.

2.3. Test Frameworks and Blueprints

The CoAlt assessment frameworks were developed to better identify the content standards that may be assessed on the CoAlt Science assessments. The frameworks were designed to assist educators, test developers, policymakers, and the public by clearly defining the elements of the EEOs that are suitable for state testing. The CoAlt frameworks were first used with the Spring 2023 assessments and can be found online at http://www.cde.state.co.us/assessment/newassess-coaltsss.

The test blueprints take the frameworks a step further by specifying the number of test items by content standard, grade-level expectation (GLE), EEO, and item type. The specificity of the test blueprints ensures that the assessments cover the breadth of the content indicated by the CAS within the associated grade or grade band. CDE and Pearson collaboratively developed the CoAlt Science test blueprints based on the CMAS blueprints. While the complete blueprints are used internally, Table 2.1, Table 2.2, and Table 2.3 present the high-level CoAlt Science blueprints that summarize the number of items and percentage of score points on each test.

Table 2.1. 2023 CoAlt Science Test Blueprint—Grade 5

Subclaim	#Item Sets	Total #Items	#1-Point SR Items	#3-Point SPT Items	Total #Points	% of Total Points
Physical Science	4	13 (14)	12 (13)	1	15-16	36-38%
Physical Science/Life Science	2	7	6	1	9	21%
Earth and Space Science	5	16 (15)	15 (14)	1	17-18	40-43%
Total	11	36	33	3	42	100%

Note. SR = selected-response, SPT = supported performance task. One Physical Science EEO will always be clustered with the Life Science EEOs. Physical Science 1.3 possibly in a cluster with ESS1 or ESS2.

Table 2.2. 2023 CoAlt Science Test Blueprint—Grade 8

Subclaim	#Item Sets	Total #Items	#1-Point SR Items	#3-Point SPT Items	Total #Points	% of Total Points
Physical Science	5	16	15	1	18	38%
Life Science	4	13	12	1	15	31%
Earth and Space Science	4	13	12	1	15	31%
Total	13	42	39	3	48	100%

Note. SR = selected-response, SPT = supported performance task.

Table 2.3. 2023 CoAlt Science Test Blueprint—Grade 11

Subclaim	#Item Sets	Total #Items	#1-Point SR Items	#3-Point SPT Items	Total #Points	% of Total Points
Physical Science	5	16 (17)	15 (16)	1	18-19	38-40%
Life Science	4	13	12	1	15	31%
Earth and Space Science	4	13 (12)	12 (11)	1	14-15	29-31%
Total	13	42	39	3	48	100%

Note. SR = selected-response, SPT = supported performance task. Possible Physical Science and Earth and Space Science crossover cluster.

2.4. Performance Levels

Student performance on the CoAlt Science assessment is categorized into four performance levels (*Emerging, Approaching Target, At Target,* and *Advanced*). The performance levels are based on the overall scale score, and cut scores divide the score scale for a grade and content area into the performance levels (see Chapter 7 for more information on the cut scores). Students in the *At Target* and *Advanced* levels are considered ready for continuing study in the content area.

The performance levels are accompanied by performance level descriptors (PLDs) that articulate what a student should know and be able to do in a particular performance level (e.g., the set of statements describing what it means for a Grade 8 student to reach *At Target* in science. The CoAlt Science assessment uses two types of PLDs: (1) policy PLDs (also known as policy claims) that provide a general idea of what is expected of a student at each level regardless of their grade level, as shown in Table 2.4, and (2) grade-level PLDs that provide detailed descriptions of performance levels by grade level, available online at https://www.cde.state.co.us/assessment/newassess-coaltsss and included on the Individual Student Performance Report and in the *CMAS and CoAlt Interpretive Guide to Assessment Reports*.

Table 2.4. Performance Levels and Policy Claims

Performance Level	*Emerging*	*Approaching Target*	*At Target*	*Advanced*
Policy Claim	Students performing at this level demonstrate an initial understanding of concepts and skills represented by the EEOs of the CAS. They will need extensive academic supports to engage successfully in further studies in the content area.	Students performing at this level demonstrate a limited understanding of concepts and skills represented by the EEOs of the CAS. They will likely need moderate academic supports to engage successfully in further studies in the content area.	Students performing at this level demonstrate a foundational understanding of concepts and skills represented by the EEOs of the CAS. They are academically prepared to engage in further studies in the content area with appropriate supports.	Students performing at this level demonstrate a solid understanding of the concepts and skills represented by the EEOs of the CAS. They are academically well prepared to engage in further studies in the content area with appropriate supports.
Scale Score	150–224	225–249	250–varies*	varies*–350

*varies by grade

2.5. Cognitive Complexity

All CoAlt Science items are assigned a Depth of Knowledge (DOK) level that indicates the cognitive complexity of the item. DOK refers to the level of rigor or sophistication of the task in an item designed to reflect the complexity of the CAS. To ensure that the assessments include a deep pool of items that span a full range of cognitive levels and skills, each item was evaluated and tagged with one of the following DOK levels: Level 1: Recall, Level 2: Skill & Concepts, and Level 3: Strategic Thinking. DOK Level 4: Extended Thinking items are not included because the tests do not contain any extended-response items.

2.6. Test Composition

The Spring 2023 test forms included a set of operational items held constant across all forms and a set of embedded field test items differing from form to form. Only the operational items were included in students' final scores. Table 2.5 presents the number of items on each test form, including the number of operational vs. embedded field test items and the total number of score points possible.

Table 2.5. 2023 CoAlt Science Test Designs

Grade	#Test Forms	Total #OP + FT Items	#1-Point OP SR Items	#3-Point OP SPT Items	#1-Point FT SR Items	#3-Point FT SPT Items	Total #OP Points
5	2	50	33	3	12	2	42
8	2	56	39	3	12	2	48
11	2	54	39	3	9	3	48

Note. OP = operational, FT = field test, SR = selected-response, SPT = supported performance task

2.7. Timing of Tests

The CoAlt Science assessments are untimed, and testing may extend over multiple days for a student. The assessment may be stopped or restarted at any time, but once an item is presented to the student, the item should be completed before stopping the assessment. The amount of time it takes the student to complete the assessment is recorded by the test administrator on the answer document after testing is complete.

Chapter 3: Item Development

The CoAlt Science item development follows the same process as the CMAS Science assessment to the extent possible, although it is modified to reflect the unique characteristics of the assessment program such as the item types and needs of the population of students who take alternate assessments. CDE relies greatly on input from Colorado educators—both general and special educators—and alternate assessment specialists to ensure that the CoAlt Science assessment is equitable for students and accurately measure the content standards.

The item development process is a tiered, inter-related process that begins with the development of the test blueprints for each grade level, followed by developing the item development plans (IDPs) to forecast the targeted number of items needed to create a robust item bank that is refreshed over time. Once written, all newly developed items go through multiple rounds of review, including contractor, CDE, and Colorado educator content, bias, and data reviews. While the Spring 2023 CoAlt Science item writing was conducted internally at Pearson, all items were reviewed by Colorado educators.

3.1. Item Banking System

Pearson's proprietary software, ABBI (Assessment Banking and Building solutions for Interoperable assessments), is used to support the test development processes from initial content authoring through the review cycles. ABBI is the authoritative source for all content, data, and functionality for all CoAlt system components. It serves as the repository where the item bank is housed, item revisions are catalogued, and items and item metadata are uploaded and revised by assessment specialists. Items can be moved into various statuses, each representing a step in the item development process. The items and associated stimuli are tracked, and revisions are recorded from creation through retirement in a secure environment.

Custom development reports can be generated out of ABBI, which allows users to generate Excel reports that capture metadata (e.g., unique item number, task type, cognitive complexity, associated stimulus, item status, item statistics, and comments) useful for analyzing the item bank. ABBI is the source of reference for how and when changes to the item and the metadata have been implemented.

3.2. Item Development Plan

An IDP is created at the beginning of each item development cycle to determine the number of items needed to construct the assessment based on the test blueprint requirements, informing item development targets that address item shortages. The grade-level IDPs delineate the target number of items per content standard/reporting category, GLE, and EEO and help to forecast the number of items needed to create a robust operational item bank that will be refreshed over time. To accomplish this, the item bank is analyzed and gaps are identified.

3.3. Item Writing

After the test blueprints and IDPs were developed, the internal item writing process began at Pearson. SR and SPT items for each assessment were written to measure concepts and skills found in the EEOs. Item writers used various guides and resources developed during specifications development, including the content standards, item specifications, and item writing guidelines.

3.4. Item Review

3.4.1. Internal Review

After the CoAlt items were written and entered into ABBI, they underwent a content review at Pearson to evaluate the standard and knowledge-and-skill match, quality of the items, adherence to the universal design principles, cognitive demand, item relevance to the purpose of the test, readability, and appropriateness of graphics. Additional fact-checking was also conducted to ensure the accuracy of item content.

Pearson's editorial team checked items for clarity, correctness of language, appropriateness of language for the grade level, adherence to style guidelines, and conformity with acceptable item writing practices. Editors with content expertise in science also reviewed the items, adding a valuable layer of content validation and fact-checking. Alternate assessment specialists with expertise in the areas of special education and students with disabilities reviewed all items to ensure that they were appropriate for students with significant cognitive disabilities. Pearson also performed a universal design review to

- assess item accessibility irrespective of diversity of background, cultural tradition, and viewpoints;
- evaluate changing roles and attitudes toward various groups;
- review the role of language in setting and changing attitudes toward various groups;
- appraise contributions of diverse groups (including ethnic and minority groups, individuals with disabilities, and women) to the history and culture of the United States and the achievements of individuals within these groups; and
- edit for inappropriate language usage or stereotyping regarding sex, race, culture, ethnicity, class, or geographic region.

These reviews were conducted to ensure that all students would have an equal opportunity to demonstrate achievement regardless of their gender, ethnic background, religion, socio-economic status, or geographic region. Items that were accepted based on the Pearson reviews were re-classified in ABBI as ready for CDE review. CDE then reviewed the items, checking to make sure the content was accurate, the EEO alignment was appropriate, the language was appropriate for the grade level and student population, and the graphics were clear and relevant to the item. Items accepted based on the CDE review were re-classified in ABBI as accepted.

3.4.2. External Content and Bias Review

Items that passed the internal review were included in external content and bias review. Educators reviewed the items for content and bias concerns, evaluating whether they were properly aligned to the content standards and identifying any potential bias in the items while considering the unique needs of students with significant cognitive disabilities. These reviews included content-specific general educators, special educators, and teachers of students who are culturally and linguistically diverse. Items that were accepted based on the educator committee recommendation were re-classified in ABBI as ready for field testing.

3.5. Data Review

After item development was complete, selected items were placed on the operational assessments in embedded field test positions. The goal of field testing is to allow for the evaluation of the quality of the newly developed items through a review of item performance data to determine their inclusion in the operational item pool. To accomplish this, psychometricians performed statistical analyses on the field tested items to evaluate their quality.

Table 3.1 presents the statistical flags applied to the field tested items. Classical statistics included item means (*p*-values), item-total correlations/point biserials, and distribution of responses across answer options or score points, depending on the item type. Differential item functioning (DIF) analyses were conducted on various subgroups (gender, ethnicity, free and reduced lunch, and multilingual learner [ML]) using Mantel–Haenszel Delta DIF statistics (Dorans & Holland, 1992). Classification rules derived from National Assessment of Educational Progress (NAEP) guidelines (Allen et al., 1999) were used to classify items as having either negligible, moderate, or significant DIF. Items were then flagged based on the criteria in Table 3.1, and the flagged items were taken to a data review meeting where a committee of educators reviews the flagged items and their statistics along with student performance data.

Table 3.1. Item Statistical Flagging Criteria

Statistic	Criterion	Possible Indication
P-value	< 0.1 or > 0.9	Very difficult or easy item
Item-total correlation	< 0.15	Poorly discriminating item
Distractor item-total correlation (SR only)	> 0.0	Possible miskey*
Score point percentage (multi-point items only)**	<1% or >50%	Very few students or many students got a certain score
Differential item functioning (DIF)***	B, C	Item could be biased toward a certain student demographic group

*Possible miskey because the key should have a positive item-total correlation
**If a multi-point item has less than 1% for a score point or more than 50% zeros, the item is flagged.
***B DIF indicates moderate DIF, whereas C DIF indicates significant DIF.

During the data review meeting, educators were trained to interpret the statistical information and judge the appropriateness of the flagged items. The committee members used the data as a tool to direct them toward potential flaws in an item and discuss whether there were construct-irrelevant reasons for a data flag. A data flag, by itself, is not the sole reason an item is rejected. Committee members were instructed that their final judgments about the appropriateness or fairness of an item for any individual and subgroup encompassed by the data flag should be based on their expertise with their content area and experience as Colorado educators.

Committee members reviewed each item and recommended whether to accept or reject it. An accepted item indicated that the educators, through their varying expertise, determined that there is not a construct-irrelevant reason for the data flag within the item, whereas a rejected item indicated that the educators determined there was a construct-irrelevant reason for the data flag. Construct-irrelevant reasons for data flags could include issues such as language that is above grade-level or content that is biased against a particular group. In contrast, construct-relevant explanations could be difficult content that is part of the standards or distractors that reflect a very common misunderstanding of the concept covered by the item, which would not be a reason to reject the item.

Following the data review meeting, CDE reviewed the committee's recommendations and made final decisions. All accepted items were moved into "Ready for Operational" status. Rejected items were reclassified as "Do Not Use" or "Revise and Re-field Test" to eliminate them from use on an operational test. These items may be modified and field tested again on future test forms. Table 3.2 presents the results of the data review based on Spring 2023 data (i.e., the number of field tested items that were either accepted, accepted for revision and re-field test, or rejected as a result of the data review).

Table 3.2. Data Review Results

Grade	#Accepted	#Accepted for Revision and Re-field Test	#Rejected
5	7	0	6
8	3	0	6
11	6	0	2

Chapter 4: Test Construction

When building operational test forms, Pearson assessment specialists select a set of operational items in accordance with the test blueprint and test construction specifications. Items selected for use operationally must meet the test blueprint and should include a variety of topics and contexts with specified psychometric targets. The following guidelines are used during test form construction:

- Adherence to the test blueprints
- Efficient and deliberate use of varied content representative of the knowledge and skills in the content standards
- Balance of gender, ethnicity, geographic regions, and relevant demographic factors
- Thorough review of each item to verify that the content is up-to-date and relevant
- Review of the full form, including embedded field test items, for instances of clueing and/or content overlap

After the initial operational items are selected, the test form is reviewed by two Pearson assessment specialists who each verify that the test form meets the test blueprint and test construction specifications (i.e., the required number of items, EEO coverage, and item types). The psychometrician then verifies that the test form falls within the psychometric and test blueprint parameters and identifies the anchor item set within each operational test form. (See Chapter 10 for information about anchor items.) Once the test form is vetted internally, it is presented to CDE for review. If needed, CDE and Pearson assessment specialists and psychometricians collaborate to finalize the test form.

After the operational test form is approved, field test items are selected from the items in ABBI that are coded as ready for field testing. The assessment specialists assemble field test item sets so they comprise the appropriate distribution of standards, item types, topic coverage, and key distributions. They also review item replacement for future years to ensure appropriate item rotation. Items chosen are embedded on the operational test form in a designated location. The specific responsibilities for Pearson and CDE during test construction are outlined below:

- Pearson responsibilities:
  - generate a test construction schedule
  - select and sequence a proposed set of operational items
  - select and sequence a proposed set of anchor items
  - select and sequence a proposed set of field test items
  - conduct content and psychometric reviews of each proposed set of items
  - construct a test map that provides content and psychometric information for each item
  - manage the CDE review process
  - provide the CDE with copies of proposed items and the associated test map
  - revise the proposed item set based on CDE comments
  - document edits/comments provided by CDE

- CDE responsibilities:
  - review and approve item selection based on content and psychometric properties
  - review and approve the test form for layout, item sequencing, and avoidance of cueing

Chapter 5: Test Administration

The CoAlt Science assessments are paper-based assessments administered one-on-one by a test administrator who records student responses on a scannable answer document. Prior to the administration, training of Colorado districts, schools, and teachers was a high priority because the assessments involve specifically developed materials, administration requirements, and score entry steps. CoAlt Science assessment administration and training procedures were standardized to ensure that students would receive comparable test results while allowing flexibility to accommodate the unique needs of students in this population. Test administration procedures were communicated to the appropriate individuals via manuals and virtual and recorded trainings.

The District Assessment Coordinator (DAC) is responsible for establishing the administration schedule and ensuring that every student taking a CoAlt Science assessment is assessed within the state assessment window. Districts may use the entire state testing window for administration of this assessment, but it is expected that students taking the CoAlt Science assessment will test during the same testing window as their peers taking the CMAS assessments. It is important that scheduling of the assessment is based on the individual needs of the student.

5.1. Manuals

The following manuals are available online at https://coassessments.com/manuals/ to support the CoAlt Science administration:

- The *CoAlt Science Test Administrator Manual* provides instructions for administering the CoAlt Science assessments, including scoring procedures, as well as the before, during, and after testing tasks for the Test Administrator. Test administration policies and procedures, including scoring information, are to be followed as written so that all testing conditions are uniform statewide, ensuring that every student in Colorado receives the same standard directions and scoring during the test administration.
- The *CMAS and CoAlt Procedures Manual* provides instructions for the coordination of the CoAlt Science assessments. Instructions include the protocols that all school staff are to follow related to test security, test administration, and providing accommodations. The manual also includes the tasks to be completed by DACs, School Assessment Coordinators (SACs), and District Technology Coordinators (DTCs) before, during, and after the test administration.
- The *PearsonAccess^{next} Online User Guide* provides guidance for DACs, SACs, DTCs, and student enrollment personnel who use PearsonAccess^{next}, the website used for student registration, test setup, administration preparation, and assessment and data management.

5.2. Test Materials

Table 5.1 presents the paper-based test materials used by the test administrator during the administration of the CoAlt Science assessment, distributed by the SAC, as provided in the *CMAS and CoAlt Procedures Manual*. For the SR items, the student marks/points/indicates their response in the test book, and the test administrator marks the student answer on the answer document. SPT items have cutout cards that the student places/indicates placement of in the correct box in the test book. The test administrator scores the student response and marks the student's score in the answer document.

Table 5.1. Test Materials

Test Material	Description
CoAlt *Test Administrator Manual*	Provides information necessary for the administration and scoring of the CoAlt science assessment for use by the test administrator. The manual contains the SPT Score Flow Chart for scoring SPT items.
Test Books	The test administrator uses the CoAlt test book to read the administration script from the Test Administrator page while the student response pages face the student.
Task Manipulatives	Students use task manipulatives to respond to the SPT items. Prior to testing, test administrators must prepare the task manipulatives by cutting them apart.
Answer Document	The test administrators use the answer document to record student responses during testing. After testing, answer documents are returned to Pearson for scoring
Secure Return Envelope	Transport test materials between the testing environment and the central storage area in an unsealed secure return envelope. Task manipulatives should be stored and returned in the envelope. (Note: Test books will not fit in the envelopes.)

5.3. Administration Training

Administration training is intended to make sure all individuals involved in the CoAlt Science assessment activities at the school and district levels are prepared to follow administration processes and procedures with fidelity, as well as support adherence to security procedures. Fidelity to standardized test administration processes and procedures helps to ensure the comparability of resulting scores and accurate interpretation of results.

Thorough trainings were conducted by CDE for DACs and district-based special education staff across Colorado. The virtual trainings contained information regarding proper procedures for administration. Training sessions covered CoAlt Science assessment eligibility requirements, the test design, accommodations, distribution of materials, test security, and PearsonAccess^{next} tasks necessary to set up and administer the assessment and access test results. The trainings were posted on the CDE website at http://www.cde.state.co.us/assessment/trainings-archive. Administration training materials such as web-based modules, slide decks, and manuals were also available on the CDE website for training SACs. After CDE trained DACs and special education staff, these individuals trained SACs and any other individuals within the district who planned to participate in the CoAlt Science assessment administration.

Pearson customer service center staff were also trained to answer questions thoroughly and knowledgeably about the administration, and to escalate inquiries as necessary. A knowledge base of commonly asked questions was created to ensure accurate and consistent responses to school and district personnel. The knowledge base was created by the CDE and Pearson based on information covered in the training materials and manuals. Revisions and additions were made to the knowledge base as needed. CDE met with Pearson daily during the administration window to review questions from districts and ensure that appropriate answers were provided. Policy questions received by the Pearson customer service center were referred to CDE.

5.4. Practice Resources

Colorado Practice Resources (CPRs) are available online at https://coassessments.com/practice-resources/ to help students become familiar with the SR and SPT item types on the CoAlt Science assessments. Each grade has multiple SR clusters and SPT samples. As the assessment system progresses, the CPRs will be updated to reflect the current assessment.

5.5. Accessibility Features and Accommodations

The CoAlt Science assessments were developed to be accessible for students with the most significant cognitive disabilities. Accessibility was considered from the beginning of the test development process and is inherent within the CoAlt Science assessments and administration procedures. For example, CoAlt Science assessments are read aloud to students and all students who take CoAlt Science assessments are assessed individually. The assessments can also be administered over several days for students who need more time due to limitations in behavioral control, stamina, or communication. Even though the assessments are designed to be accessible, students with disabilities taking the assessments may still require changes to the assessment procedures, or accommodations, to accurately demonstrate their knowledge and skills of the content. This also includes students classified as ML) who need language supports to demonstrate their knowledge of the content.

In addition to incorporating accessibility into the assessment, accommodations are also available to students who need additional changes to the test administration to access the assessment. Accommodations provide a student with an opportunity to engage with the assessment while not affecting the reliability or validity of the assessment. Accommodations can be adjustments to the test presentation, materials, environment, or response mode of the student and are based on student need. Accommodations should not provide an unfair advantage to any student. Providing an accommodation for the sole purpose of increasing test scores is not ethical and CDE provides extensive training on how to implement accommodations. Accommodations must be documented in the student's IEP and used regularly during classroom instruction and assessments prior to the assessment window to ensure the student can successfully use the accommodation.

Although accommodations are used for classroom instruction and assessments, some may not be appropriate for use on statewide assessments. As a result, it is important that educators become familiar with the state assessment policies about the appropriate use of accommodations and that districts have a plan in place to ensure and monitor the appropriate use of accommodations. Accommodations for the CoAlt Science assessments could include the following:

- Assistive technology
- Eye gaze
- Modified picture symbols (enlarged pictures and/or pictures of real objects)
- Objects (three-dimensional or representational objects)
- Sign language
- Translation into student's native language
- Other
- None

5.6. Test Security

Test security procedures are put in place to enhance the likelihood that security is maintained before, during, and after assessment administration. For example, materials used during the administration of the assessment are to be kept in locked storage locations when not under the direct supervision of Pearson or approved assessment coordinators and administrators. All district and school personnel involved in the CoAlt Science test administration are required to participate in annual local training. DACs and district special education staff are responsible for overseeing training for the district, including verifying that the SACs are trained. SACs are responsible for ensuring that all individuals involved in handling test materials at the school level are trained and subsequently act in accordance with all security requirements.

A chain of custody plan for materials is required to be written and implemented to ensure that materials are securely distributed from DACs to SACs to Test Administrators and securely returned from Test Administrators to SACs and then to DACs. SACs are required to distribute materials to and collect materials from the Test Administrators each day of testing and to securely store and deliver materials to DACs after testing is completed in accordance with the instructions in the *CMAS and CoAlt Procedures Manual*.

All individuals involved in the test administration are required to sign a security agreement prior to handling test materials, which requires them to follow all procedures set forth in the aforementioned manuals and prevents them from divulging the contents of the assessment, copying any part of the assessment, reviewing test items with the students, allowing students to remove test materials from the testing room, or interfering with the independent work of any student taking the assessment.

PearsonAccess^{next} used during the administration includes permissions-based user role access to all information within the system, including accessing student information, setting up student tests, and accessing reports. Access to this information is tightly controlled before, during and after test administration, requiring a login ID and password to enter the system.

After all testing is completed at a school, used and unused materials are required to be securely stored and returned to the DAC by the district deadline for shipment to Pearson. DACs are required to report any missing test materials or test irregularities and to complete the appropriate documentation.

5.7. Test Monitoring

During the Spring 2023 administration, six assessment specialists were selected by Pearson and approved by CDE to serve as test monitors who were sent out to a small sample of schools to observe the administration of the CoAlt Science assessments. The assessment specialists were familiar with administering alternate assessments, including CoAlt Science, and with the population of students who take alternate assessments. The test monitor's task was to record several metrics during their observations, including adherence to administration procedures, security measures, and score entry. The observations were scheduled to mitigate any impact on the classroom and will be used to evaluate the training procedures and manuals for the following year.

5.7.1. Training

Prior to monitoring the test administrations, the test monitors participated in training developed by CDE and Pearson via teleconference to ensure that they would be consistent in their methods. The training facilitator reviewed the test monitors' region assignments, the purpose of test monitoring, the test monitor materials, and the expectations of the test monitors. The test monitor materials included materials such as a security agreement form, the CoAlt Procedures Manual, the CoAlt Test Administrator Manual, a Test Monitor Checklist, and an answer document.

5.7.2. Process

Test monitors used a Test Monitor Checklist containing questions related to the test administration and test security to indicate how well test administrators were adhering to the test administration procedures and security measures. The test monitors also transcribed student responses from their observations onto a CoAlt answer document that would later be used to evaluate score entry. Once all observations were completed, the Test Monitor Checklists and transcriptions were returned to Pearson for analysis. Response frequencies were generated for the Test Monitor Checklist questions to evaluate how well test administrators were following the test administration procedures and security measures. To evaluate score entry, the test monitor's student responses were compared to the test administrator's student responses to determine the amount of agreement between the set of responses.

5.7.3. Participation

Pearson and CDE worked together to recruit schools to participate in test monitoring. Schools were selected so that the sample of observed students would be representative of the geographic regions of the state and diverse in terms of gender and ethnicity. As shown in Table 5.2, the participating school districts represented seven of the eight geographic regions of the state. As shown in Table 5.3 that presents the number of observations conducted (counted once for each student) compared to the total CoAlt Science student population, 17 observations were conducted for Grades 5 and 8, and 11 observations were conducted for Grade 11.

Table 5.2. Number of Participating Schools in Test Monitoring

Geographic Region	Grade 5	Grade 8	Grade 11
Metro	3	1	2
North Central	4	3	3
Northeast	–	–	1
Northwest	–	–	–
Pikes Peak	4	5	2
Southeast	–	1	–
Southwest	1	–	–
West Central	–	2	1

Table 5.3. Number of Participating Students (Observations) in Test Monitoring

	Population			Sample		
Grade	N	Male	Female	N	Male	Female
5	386	62%	38%	17	76%	24%
8	465	61%	39%	17	53%	47%
11	400	57%	43%	11	64%	36%

5.7.4. Results

In general, most test monitors indicated that the testing environment was appropriate. Test administrators seemed comfortable with the students, were well prepared for administering the test, provided the accommodations needed for the students, and administered the assessment at an appropriate pace. The test monitors also noted that the testing rooms had adequate space and were free of visible materials that could aid with the test items. However, the test monitors noted some challenges. For example, the test administrator did not always follow the standardized script provided in the test books in some instances, and there were sometimes interruptions/distractions in the testing room. CDE noted the issues and will use the test monitor feedback as part of test administrator training sessions in the next year.

Test monitors also transcribed student responses as part of their observations. To evaluate the transcription, the student responses transcribed by the test monitor and the test administrator were compared to determine perfect agreement (i.e., when the test monitor and test administrator assign the same response to the same item). Test monitors could not always observe the student taking all the test items, such as when students were tested across multiple days, which led to instances where test monitor scores were missing. When this occurred, only the items with responses from both the test monitor and the test administrator were included in the analysis. Three students were also excluded from the analysis because their score entry data could not be matched between the test monitor and the test administrator. As shown in Table 5.4 that presents the resulting agreement results, the perfect agreement rates indicate high levels of agreement between the sets of transcribed student responses.

Table 5.4. Test Monitoring Percent Agreement Rates between Transcribers

Grade	Perfect Agreement	Non-Perfect Agreement
5	97%	3%
8	94%	6%
11	98%	2%

Chapter 6: Scoring

The test administrator marks a student's responses to the 1-point SR items (A, B, C, D, or NR when there is no response from the student) and indicates their assigned scores for the 3-point SPT items (0, 1, or NR) in the scannable answer document that is then returned to Pearson. The 1-point SR items are then machine-scored, whereas each of the three prompts in an SPT item had already been scored by the test administrator using the built-in rubric to evaluate student performance.

6.1. SR Scoring

The SR items are key-based multiple-choice items. Initial scoring expectations are developed during item development and are included in the item review process. The scoring rules and correct responses are included in the items' XML coding. Prior to scoring, key checks are completed for all SR items to verify that the machine is correctly identifying correct and incorrect responses. If there is a discrepancy in the scoring, content experts review the item and adjustments are made as needed. During testing, actual distribution of scores is compared to expected distribution. Further evaluation is completed if a discrepancy is identified.

6.2. SPT Scoring

SPT items consist of three related items called prompts. Students are required to manipulate option cards by placing them in designated areas on a diagram or chart to respond to each prompt. Student performance on each prompt is scored using a 1-point rubric by the test administrator during the administration, as shown in Table 6.1. To administer the item, the test administrator has the student response page and option cards ready for the student to engage with the item. The test administrator then presents the scripted text for the first prompt. Scores are assigned by the test administrator based on the following scenarios:

- If the student responds correctly, they receive 1 point.
- If the student responds incorrectly, they receive 0 points.
- If the student does not provide a response to the prompt, they receive an NR, or no response, that represents 0 points.

Table 6.1. SPT Scoring Rubric

Score Point	Requirement
1	Student responds correctly
0	Student responds incorrectly
NR	Student does not respond

Note. NR = no response, which represents 0 points. This rubric is used for each of the three prompts within each SPT item.

If an incorrect response is given or the student does not respond, the test administrator places the correct option card in the response box and tells the student the correct answer. After the first prompt is completed, the test administrator completes the same steps for the remaining two prompts. For scoring and reporting purposes, the points for the three prompts are added together to provide one score for the SPT item that can range from 0–3 points.

Chapter 7: Standard Setting

To support the interpretation of student results, student performance on the CoAlt Science assessment is described in terms of performance levels as presented in Table 2.4. Standard setting is the process of translating those policy-driven performance standards into scores on the assessment. The purpose of a standard setting study is to determine the boundaries—or cut scores—along the score scale that differentiate student performance among performance levels (e.g., Cizek et al., 2004; Kane, 1994).

Standard setting for the new CoAlt Science assessment aligned to the EEOs of the 2020 CAS took place from October 25–26, 2022, with Colorado educators using a modified version of the Item Descriptor (ID) Matching method (Ferrara et al., 2012), as detailed in the *CoAlt Science 2022 Standard Setting Report* (Pearson, 2024). Three grade-level panels were convened, with a total of 35 educators participating across all panels. The recommendations from the standard setting panels were then presented to CDE and ultimately the Colorado State Board of Education for consideration and final approval on December 14, 2022.

Table 7.1 presents the resulting scale score cut scores for each grade that will be used to report student results on the CoAlt Science assessments starting in Spring 2023.

Table 7.1. Performance Level Cut Scores

Grade	*Emerging*	*Approaching Target*	*At Target*	*Advanced*
5	150–224	225–249	250–272	273–350
8	150–224	225–249	250–276	277–350
11	150–224	225–249	250–276	277–350

Chapter 8: Reporting

8.1. Description of Scores

The CoAlt Science reports provide information on student performance in terms of scale scores, performance levels, and percent earned scores. A scale score is a conversion of a student's total test score (i.e., the total number of points earned on a test) to a scale that is common to all test forms for that assessment. Scale scores are particularly useful for comparing test scores over time and creating comparable scores when a test has multiple forms. Students taking the CoAlt Science assessment receive an overall test scale score that ranges from 150 to 350, as shown in Table 7.1. In addition to the overall test scale score, an indicator of the range of scale scores a student would likely receive if the assessment was taken multiple times is also provided.

Performance levels and their PLDs are reported at the overall assessment level. Students are classified into performance levels based on their scale score and the cut scores obtained from standard setting. CoAlt Science has four performance levels: *Emerging*, *Approaching Target*, *At Target*, and *Advanced*. Students in the top two performance levels indicate that with the appropriate supports, the student is prepared for further study in the content area.

To prevent incorrect interpretations and provide a metric that is more generally understood, student performance is also reported as the percentage of points earned (i.e., the number of points a student earned out of the total number of points possible) for the content standards (Physical Science, Life Science, and Earth and Space Science) and the SEPs. Unlike scale scores, the percent of points earned scores cannot be compared across years because individual items change from year to year and are not constructed to be comparable in difficulty.

8.2. Score Reports

Two types of score reports are provided to communicate student performance on the CoAlt Science assessments: (1) the student-level Student Performance Report and (2) the aggregate reports. The Student Performance Report provides information about the performance of a particular student. The student's scale score, associated performance level, and percent of points earned are displayed on a one-page report, along with comparative information related to state performance. PLDs are also provided. Student Performance Reports are printed and shipped to districts for distribution to students and parents. Electronic reports are available in PearsonAccess^{next}.

Two types of aggregate reports are produced for schools and districts: Performance Level Summaries and Content Standards Rosters. These reports are produced at the school, district, and state levels and provide summary information for a given school or district. State, district, and school reports are provided electronically through PearsonAccess^{next}. Access to the reports is limited to authorized users.

Appendix B presents a sample Student Performance Report, and examples of each type of aggregate report and a detailed explanation are provided in the *CMAS and CoAlt Interpretive Guide to Assessment Reports*. For a detailed explanation of the information provided in all reports, refer to the *CMAS and CoAlt Interpretive Guide to Assessment Reports* located online at https://www.cde.state.co.us/assessment/cmas_coalt_interpretiveguide_2023.

Chapter 9: Test Results and Analysis

9.1. Student Participation

Table 9.1 presents a breakdown of the number of students who took the Spring 2023 CoAlt Science assessment by various demographic characteristics. All forms were administered in paper format. Approximately 1,251 students across grades took the assessment in Spring 2023.

Table 9.1. Student Participation N-Count Demographic Distribution

Subgroup	Grade 5	Grade 8	Grade 11
Total	386	465	400
No IEP	*	*	*
IEP	386	464	400
No Accommodation	*	*	*
Accommodation	386	465	400
Am. Indian/Alaska Native	*	*	*
Asian	17	18	*
Black	35	38	33
Hispanic	163	207	190
White	149	178	138
Hawaiian/Pacific Islander	*	*	*
Two or More Races	18	19	25
Missing	*	*	*
No Economic Disadvantage	163	216	196
Economic Disadvantage	223	249	204
Female	147	181	174
Male	239	284	226
Language Proficiency NA	300	366	315
Language Proficiency NEP	61	45	37
Language Proficiency LEP	*	*	*
Language Proficiency FEP	20	45	43
Not Migrant	385	465	399
Migrant	*	*	*

*n-count less than 16

9.2. Performance Results

Table 9.2 presents the scale score performance summary and performance level distributions (i.e., the percentage of students classified into each performance level), and Table 9.3 presents the summary statistics for points earned by subclaim. Appendix C presents the cumulative scale score distributions by grade, Appendix D displays the same information in graphical form, and Appendix E presents the summary statistics for the overall scale scores by demographic subgroup.

Table 9.2. Scale Score Performance Summary and Performance Level Distributions

Grade	N	Mean	SD	Median	%*Emerging*	%*Approaching Target*	%*At Target*	%*Advanced*
5	386	238	32.0	236	33.4	31.4	20.7	14.5
8	465	234	31.1	232	38.9	28.6	23.7	8.8
11	400	235	38.0	238	35.0	32.0	21.8	11.3

Table 9.3. Summary Statistics for Points Earned by Subclaim

Subclaim	Grade	Mean	SD	Min.	Max.	Average % Correct
Physical Science	5	7.3	3.1	0	15	45.94
	8	8.4	3.5	0	18	46.44
	11	7.9	4.1	0	18	44.13
Life Science	5	4.9	2.1	0	9	54.70
	8	8.6	3.6	0	15	57.12
	11	7.2	3.4	0	15	47.72
Earth and Space Science	5	6.5	3.2	0	15	37.95
	8	6.8	3.2	0	14	45.06
	11	5.6	2.6	0	14	37.64

Note. Life Science is Physical Science/Life Science in Grade 5.

9.3. Classical Item Analysis

Appendix F presents the item-level classical statistics for the Spring 2023 CoAlt Science assessments, including the omit rate, *p*-value, item-total correlation, and the percentage of students earning each score point (SPT items only).

Item difficulty is measured by the *p*-value bounded by 0.0 and 1.0 that indicates how easy or hard an item is. The *p*-value for 1-point items is the proportion of students who answered an item correctly and is calculated by dividing the number of students who got the item correct by the total number of students who answered it. For multiple-point items, the *p*-value is the average item score (i.e., the sum of student scores on an item divided by the total number of students who responded to the item) that is then put on a 0 to 1 scale by dividing the average item score by the maximum number of points for the item. A high *p*-value indicates that an item is easy (high proportion of students answered it correctly), whereas a low *p*-value indicates that an item is difficult. Easy and hard items are both necessary to include on an assessment to balance the test difficulty.

Item discrimination is represented by the item-total correlation (also known as the point-biserial correlation), is bounded by -1.0 and 1.0, and indicates how well an item discriminates, or distinguishes, between low-performing and high-performing students. The item-total correlation is based on the relationship between student performance on a specific item and performance on the entire test based on their test score. Students who do well on a test are expected to do well on a given item, and students who do not do well on a test are expected to not do well on a given item. This means that for a highly discriminating item, students who get the item correct will have a higher average test score than students who get the item incorrect. An item with a high positive item-total correlation discriminates between low-performing and high-performing students better than an item with an item-total correlation near zero. A negative item-total correlation indicates that low-performing students did better on that item than high-performing students.

9.4. Subclaim Correlations

The CoAlt Science assessments have three subclaim scores: Physical Science, Life Science, and Earth and Space Science. One way to assess the internal structure of a test is through the evaluation of correlations among subclaim subscores, as presented in Table 9.4. There is evidence of unidimensionality if the components within a content area are strongly related to each other. The intercorrelations between the subclaims were between 0.563 and 0.703 which indicates a moderate to strong relationship between the subclaims. The correlations between Physical Science and Life Science tended to be higher than the correlations of those subclaims with Earth and Space Science. Correlations between the subclaims and the total test ranged from 0.563 to 0.860.

Table 9.4. Correlations Between Subclaims

Grade	Subclaim	Life Science	Earth and Space Science	Total Test
5	Physical Science	0.582	0.576	0.789
	Life Science	–	0.563	0.741
	Earth and Space Science	–	–	0.809
8	Physical Science	0.703	0.690	0.849
	Life Science	–	0.656	0.840
	Earth and Space Science	–	–	0.840
11	Physical Science	0.681	0.622	0.860
	Life Science	–	0.616	0.832
	Earth and Space Science	–	–	0.782

Note. Life Science is Physical Science/Life Science in Grade 5.

Chapter 10: Calibration, Equating, and Scaling

The item response theory (IRT) Rasch partial credit model (RPCM) was used to develop, calibrate, equate, and scale the CoAlt Science assessments and to maintain and build the item bank. All test analyses including calibrations, scaling, and item model fit were accomplished within the IRT framework. The Spring 2023 operational administration determined the base scale for the CoAlt Science assessments. In the following years, equating will be used to place the new test forms on this newly developed operational scale. All steps in the calibration, equating, and scaling processes were repeated for each CoAlt Science assessment and were independently replicated by at least two members of the Pearson psychometric team to ensure accuracy.

Calibration is the process of estimating the parameters (such as item difficulty) for each item on an assessment so that all items are placed on a common scale. To maintain the same performance standards across different administrations of a particular test, it is necessary for each administration of the test to be of comparable difficulty. It is not fair to compare students to a common standard if the overall difficulty of the forms changes from year to year. Maintaining test form difficulty across administrations is achieved through equating. Equating adjusts for differences in overall test difficulty of test forms so that the scores resulting from two different administrations can be considered interchangeable.

Equating and scaling typically occur in sequence. First, equating is used to adjust for differences in test difficulty so resulting estimates of student proficiency (i.e., equated raw scores, theta estimates) are on a common metric. The equated estimates of proficiency are then converted to scale scores for reporting purposes.

10.1. IRT Model

For each grade-level assessment, the RPCM was used to place the CoAlt Science items and student proficiency on the same scale. The model is an extension of the Rasch one-parameter IRT model attributed to Georg Rasch (1966), as extended by Wright and Stone (1979), Masters (1982), and Wright and Masters (1982). The RPCM was selected because of its flexibility in accommodating various item types, including the 1-point SR and multi-point SPT items. The RPCM maintains a one-to-one relationship between scale scores and raw scores, meaning each raw score is associated with a unique scale score. It is the underlying Rasch scale that allows for comparisons of student performance across years and facilitates the maintenance of equivalent performance standards across years.

The RPCM is a mathematical measurement model with a single item parameter relating a student's performance on a given item involving *m*+1 score categories. The probability of student *n* scoring *x* on *m* steps of item *i* is a function of the student's proficiency level, θ_n (also referred to as "ability"), and the step difficulties, δ_{ij}, of the *m* steps in item *i* as follows:

$$P_{xni} = \frac{exp \sum_{j=0}^{x} (\theta_n - \delta_{ij})}{\sum_{k=0}^{m_i} exp \sum_{j=0}^{k} (\theta_n - \delta_{ij})}, x = 0, 1, \ldots m_i$$

10.2. Data Preparation

Prior to any analyses, several steps were completed in preparation.

- The data file containing student responses was verified and exclusion rules were applied.
- Traditional item analyses of all items were conducted prior to calibration.
- Complete data matrices (CDMs) and incomplete data matrices (IDMs) were created for calibrations.

A traditional item analysis of all operational and embedded field test items was conducted prior to calibration. The purpose of this analysis was to obtain classical statistics used to evaluate item performance. The following statistics were calculated:

- Item sample size
- Response distribution
- Item mean score
- Item-total correlation

10.3. Calibration

To obtain the Rasch item parameter estimates for the new Spring 2023 assessments, the RPCM was applied to the operational and embedded field test items. Winsteps (Linacre, 2021) was used for all grade-level calibrations. The calibration of the operational and embedded field test items for each assessment occurred in several steps. First, the operational items were calibrated creating the base scale for the assessment. Next, the embedded field test items were calibrated with the operational items using fixed common item parameter calibration. With this calibration method, the embedded field test items are calibrated with the operational item parameters fixed at their previously estimated values to place the embedded field test items on the same scale as the operational items. Several different calibrations were done to obtain item parameter estimates for the operational and embedded field test items:

- Operational Items
    - Used Winsteps control files and CDM to obtain operational item parameter estimates.
        - Obtained operational Rasch item difficulty values, step deviation values, and item fit values.
        - Created a test characteristic curve (TCC) using the new operational item parameter estimates. Appendix H presents plots of the TCCs and each cut score for a given grade is indicated with a red vertical line.
- Embedded Field Test Items
    - Used Winsteps control files and IDM to scale the embedded field test item parameter estimates to the operational scale by fixing the item parameter estimates of the operational items.
        - Obtained embedded field test Rasch item difficulty values, step deviation values, and item fit values.

After the item parameter estimates were obtained for the Spring 2023 operational items for each grade-level assessment, student proficiencies were estimated by conducting an anchored calibration of the operational items' item parameter estimates. Estimates were obtained via the joint maximum likelihood method (JMLE) applied within the Winsteps software program.

Student proficiency estimates are generated only for students who meet the attemptedness criteria. To be classified as attempted, a student must respond to at least nine items in section one of the test. The nine items can be operational or field test items.

10.4. Equating

Equating is used to place new forms onto the operational base scale. Equating of the operational test forms involves adjusting for differences in the difficulty of forms, both within and across assessment administrations, to ensure that students taking one form of a test are neither advantaged nor disadvantaged when compared to students taking a different form. Each time a new form is constructed, equating is used to allow scores on the new form to be comparable to scores on the previous form. If the IRT models fit the data and the model assumptions are met, calibration of test items places both items and students on a scale that is independent of any sample of students up to a linear transformation. Equating is used to determine and apply a scale transformation that allows for meaningful comparisons of student performance across different forms or administrations of the test.

After calibrating the 2023 operational items to set the base scale for each assessment, a subset of the new 2023 item parameters was used as an anchor set to equate the 2022 item parameters and cut scores to the 2023 base scale. The newly equated cut scores were then used to generate scale scores and performance levels for the 2023 assessments. The operational 2023 administration was chosen to be the base scale due to concerns about the validity of the 2022 administration since it was the first time students had seen items aligned to the new content standards.

The Spring 2022 item parameters and cut scores were equated to the 2023 base scale using a fixed common items approach so that all items from the bank would be available for future test construction. The subset of 2023 operational items used for equating are called anchor items. The anchor items are a set of common items already equated to the base scale and are also included on the 2022 test forms. This set of items represents the test blueprint in terms of content and item types. To obtain equated Rasch parameter estimates for the Spring 2022 item parameters and cut scores, anchor item parameter estimates were fixed to their previous item parameter estimates already on the base scale before calibrating the remaining non-anchor items. This method places the non-anchor items on the same scale as the anchor items.

10.5. Scaling

Student proficiencies were then transformed to scale scores ranging from 150 to 350 using the cut scores determined from standard setting. The CoAlt Science scale scores represent linear transformations of the student proficiencies (θ). The transformation is made by first multiplying any given θ by a slope (a) and then adding an intercept (b). The following linear transformation was used to convert student proficiency estimates into scaled scores (SS):

$$SS = (a * \theta) + b$$

The *a* and *b* values are referred to as scaling constants. These scaling constants will be applied each year to the Rasch proficiency estimates for that year's set of operational items. After the scale scores were obtained, the lowest observable scale score (LOSS) and the highest observable scale score (HOSS) for the performance levels were applied. The LOSS and HOSS for the performance levels were set to 150 and 350, respectively.

Chapter 11: Reliability

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) refer to reliability as the "consistency of scores across replications of a testing procedure" (p. 33). A reliable test produces stable scores; very similar score distributions would result if the test were administered repeatedly under similar conditions to the same students without memory or fatigue affecting the scores. The level of reliability/precision of scores has implications for validity. In other words, scores must be consistent and precise enough to be useful for intended purposes. If scores are to be meaningful, tests should produce stable scores if the same group of students were to take the same test repeatedly without any fatigue or memory of the test. The range of certainty around the score should also be small enough to support educational decisions. Reliability for the CoAlt Science assessment is evaluated with the following analyses:

- Internal consistency (coefficient alpha)
- Standard error of measurement (SEM)
- Conditional standard error of measurement (CSEM)
- Decision consistent and accuracy

11.1. Internal Consistency (Coefficient Alpha)

Within the framework of classical test theory, an observed test score is defined as the sum of a student's true score and error ($X = T + E$, where X = the observed score, T = the true score, and E = error). A true score is considered the student's true standing on the measure, while the error score reflects a random error component. Thus, error is the discrepancy between a student's observed and true score. Internal consistency is typically measured via correlations among the items on an assessment and provides an indication of how much the items measure the same general construct. High reliability of test scores implies that the test items within a subclaim are measuring a single construct, which is a necessary condition for validity when the intention is to measure a single construct.

The reliability coefficient of a measure is the proportion of variance in observed scores accounted for by the variance in true scores. The coefficient can be interpreted as the degree to which scores remain consistent over parallel forms of an assessment (Ferguson & Takane, 1989; Crocker & Algina, 1986). In the internal consistency method used to estimate reliability for the CoAlt Science assessments, a single form is administered to the same group of students to determine whether students respond consistently across the items within a test. A basic estimate of internal consistency reliability is Cronbach's coefficient alpha statistic (Cronbach, 1951). Coefficient alpha is equivalent to the average split-half correlation based on all possible divisions of a test into two halves. Coefficient alpha can be used on any combination of dichotomous and polytomous test items and is computed as follows:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} S_j^2}{S_X^2}\right)$$

where n is the number of items, S_j^2 is the variance of students' scores on item j, and S_X^2 is the variance of the total-test scores.

Coefficient alpha ranges from 0.0 to 1.0, where higher values indicate a greater proportion of observed score variance. Two factors affect estimates of internal consistency: test length and homogeneity of items. The longer the test, the more observed score variance is likely to be true score variance. The more similar the items, the more likely students will respond consistently across items within the test.

Coefficient alpha estimates for CoAlt Science are provided for the overall test and by subclaim, as shown in Table 11.1. The coefficient alpha for the total group across the science assessments ranged from 0.80 to 0.87. Given the differences in length, it is expected that the coefficient alpha for the overall test will be higher than that of the subscales. Appendix E presents the coefficient alphas by demographic subgroup.

Table 11.1. Coefficient Alpha

Grade	Physical Science	Life* Science	Earth and Space Science	Total Test
5	0.64	0.57	0.59	0.80
8	0.69	0.74	0.59	0.86
11	0.75	0.71	0.64	0.87

*For Grade 5, the subclaim is Physical Science/Life Science.

11.2. Standard Error of Measurement (SEM)

The SEM is another measure of reliability. This statistic uses the standard deviation of test scores along with a reliability coefficient (e.g., coefficient alpha) to estimate the number of score points that a student's test score would be expected to vary if the student was tested multiple times with equivalent forms of the assessment. It is calculated as follows:

$$SEM = s_x \sqrt{1-\rho_{XX'}}$$

where s_x is the standard deviation of test scores, and $\rho_{XX'}$ is the reliability coefficient.

There is an inverse relationship between the reliability coefficient and SEM: the higher the reliability, the lower the SEM. Table 11.2 presents the SEM results by subclaim for the CoAlt Science assessment. The SEM values for the total group ranged from 3.06 to 3.44.

Table 11.2. SEM

Grade	Physical Science	Life* Science	Earth and Space Science	Total Test
5	1.86	1.20	2.03	3.06
8	2.10	1.57	2.12	3.35
11	2.15	1.58	2.08	3.44

*For Grade 5, the subclaim is Physical Science/Life Science.

11.3. Conditional Standard Error of Measurement (CSEM)

While the SEM provides an estimate of precision for an assessment, the CSEM considers how measurement error likely varies across the scale score. In other words, the CSEM provides a measurement error estimate at each score point on an assessment, so the CSEM estimate could be used to indicate what the most likely range of scores would be for students receiving that score if they tested multiple times. The CSEM is defined as the standard deviation of observed scores given a particular true score and is estimated within the IRT framework as the inverse of the test information function. Appendix I presents plots of test information curves (TICs) and CSEM curves across the score scale range.

Because there is typically more information about students with scores in the middle of the score distribution where scores are most frequent, the CSEM is usually smallest, and thus the scores are most reliable, in the middle of the score distribution. An IRT method for estimating score-level CSEM is used because test- and item-level difficulties for CoAlt Science were calibrated using the Rasch measurement model. By using CSEMs that are specific to each scale score, a more precise error band can be placed around each student's observed score. During test construction, CSEMs are reviewed to ensure that they are minimized around the performance level cut scores.

11.4. Decision Consistency and Accuracy

The CoAlt Science scales are divided into four performance levels: *Emerging*, *Approaching Target*, *At Target*, and *Advanced*. Based on a student's scale score, the student is classified into one of the four performance levels. The consistency and accuracy of these performance level classifications is another important aspect of reliability to examine.

The consistency of a decision refers to the extent to which the same classification would result if a student were to take two parallel forms of the same assessment. However, since test-retest data are not available, psychometric models can be used to estimate the decision consistency based on test scores from a single administration. The accuracy of a decision refers to the agreement between a student's observed score classification and a student's true score classification, if a student's true score could be known.

Procedures developed by Livingston and Lewis (1995) were used to estimate the consistency and accuracy of performance level classifications. For the overall test, consistency and accuracy estimates, along with PChance (i.e., the probability of a consistent classification due to chance) and Cohen's Kappa (κ) coefficient (Cohen, 1960), are calculated as follows:

$$K = \frac{P - P_c}{1 - P_c}$$

where P is the probability of consistent classification, and P_c is the probability of consistent classification by chance (Lee et al., 2000).

Table 11.3 presents the kappa interpretations. Table 11.4 presents the decision consistency and accuracy results, and Table 11.5 and Table 11.6 present the consistency and accuracy estimates at each cut score.

Table 11.3. Kappa Values

Value of Kappa	Strength of Agreement
< 0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very Good

Table 11.4. Decision Consistency and Accuracy Estimates

Grade	Accuracy	Consistency	PChance	Kappa
5	0.65	0.55	0.27	0.39
8	0.73	0.64	0.30	0.49
11	0.72	0.62	0.28	0.47

Table 11.5. Accuracy of Cut Scores

Grade	*Approaching Target* Cut	*At Target* Cut	*Advanced* Cut
5	0.88	0.86	0.90
8	0.90	0.89	0.94
11	0.90	0.89	0.93

Table 11.6. Consistency of Cut Scores

Grade	*Approaching Target* Cut	*At Target* Cut	*Advanced* Cut
5	0.83	0.80	0.86
8	0.85	0.85	0.92
11	0.86	0.84	0.90

Chapter 12: Validity

"Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014). As such, it is not the CoAlt Science assessments that are validated but rather the interpretations of the scores. The purpose of the CoAlt Science assessment is to provide information about a student's level of mastery of the EEOs of the CAS. In support of this, this technical report has described processes that were implemented throughout the CoAlt Science assessment cycle with validity and fairness considerations in mind. This chapter describes the various sources of validity evidence as outlined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), often referencing other chapters and sections of this report. As the CoAlt Science assessments mature, validity evidence supporting the assessments' interpretations will continue to be collected and documented.

12.1. Evidence Based on Test Content

It is important to examine the extent to which the items on an assessment measure the intended construct. The CoAlt Science assessments intend to measure the EEOs of the CAS, and steps are put in place throughout the development process with a focus on this goal, as outlined in Chapters 2 and 3 of this report. For example, an item goes through numerous reviews to confirm that it adequately aligns to the EEO that it is intended to measure. Statistical bias analyses (i.e., DIF analyses) were also conducted on the items to identify any items that may be measuring a dimension unrelated to the intended construct. The test blueprints were carefully developed with specificity at multiple levels to most optimally measure the EEOs.

An independent alignment study was conducted by the Human Resources Research Organization (HumRRO) in 2023 to provide further evidence to support the claim that the content of the CoAlt Science test items matches the intended content as specified in the EEOs (Revivo et al., 2023). For the study, three panels (one per grade) of Colorado educators were convened to review the alignment between the CoAlt Science items and the EEOs. Every effort was made to recruit panels consisting of teachers reflecting the various demographic subgroups and regions across Colorado. HumRRO applied alignment criteria drawn from the principles of Achieve (2018), Webb (1997, 1999, 2002) and Links for Academic Learning (Flowers et al., 2007). This procedure required the panelists to (a) provide Depth of Knowledge (DOK) ratings for each item, (b) indicate the EEO best aligned to each item, (c) indicate if each item aligned to an SEP or CCC, and (d) indicate if each item was amenable to supports and accommodations.

Overall, the results of the study provide validity evidence to support the claim that the content of the CoAlt Science test items matches the intended content as specified in the EEOs and test blueprint. The panelists' ratings strongly support that the assessment is composed of multidimensional items that reflect a range of the 2020 CAS, although the study also found that the item DOK levels may not reflect the intended distributions found in the blueprint. Finally, items tended to be rated as accessible to a wide range of student groups and amenable to accommodations. The results of the alignment study have been considered during the item development process for subsequent administrations.

12.2. Evidence Based on Response Processes

Evidence based on response processes pertains to the cognitive aspect behind how students respond to items and the processes by which judges or observers evaluate student performance. As part of the test administration, test administrators were asked a set of questions about students' instruction, their communication modes, and their item responses. These results, presented in Appendix J, help support the validity of the students' responses on the assessment.

One of the test validity questions asked teachers if they believe that student responses accurately reflect their understanding of the material. This question provides evidence as to whether teachers believe that students are using their knowledge of the content when responding to the items. The results from this question indicate that most teachers believe that students are using their content knowledge to answer test items, although these results need to be considered in conjunction with the other data related to the number of hours of instruction in the content area, teacher's familiarity with the content and the student, and the characteristics of the student population.

The test validity question regarding students' receptive and expressive communication methods provides evidence to support the test design and the types of accommodations provided on the assessment. The results from this question indicate that most students use oral administration or picture communication to receive information, and they use these same methods when responding to others.

12.3. Evidence Based on Internal Structure

The internal structure of an assessment pertains to the degree to which the items on an assessment measure one underlying construct. When assessments are designed to measure one underlying construct, the internal components of the assessments should exhibit a high degree of homogeneity that can be measured in terms of the internal consistency estimates of reliability. As a result, the internal consistency for the CoAlt Science assessments is evaluated using reliability coefficients as provided in Section 11.1.

12.4. Evidence Based on Relations to Other Variables

Evidence was collected showing the correlation between student scores and variables related to the student. Student test scores were correlated with test administrators' responses in Appendix J for several test validity questions to determine the strength of relationship between the variables. Table 12.1 presents the correlation coefficients between the student scores and these variables, providing validity evidence based on relations to other variables. The test validity questions are variables related to the student (e.g., how familiar are you with this student? How many hours per week does this student spend in instruction on this content area? Approximately how much instructional time for this content area is in the general education classroom?).

As shown in Table 12.1, the correlations between student scores and the familiarity of the test administrator with the student are small and indicate no meaningful relationship between the variables. The correlations between student scores and the instructional hours and instructional time variables are low positive correlations which indicate a relationship between student scores and the instructional hours and instructional time variables. The strength of these relationships will be reviewed for future administrations as Test Administrators and students have more opportunity to engage with the CAS in the classroom setting.

Table 12.1. Correlation Between Test Validity Questions and Student Scores

	Familiarity with the Student		Hours Per Week in Instruction on the Content Area		How Much Instructional Time in the Content Area Is in the General Education Classroom	
Grade	N	Correlation	N	Correlation	N	Correlation
5	378	0.04	378	0.12	380	0.26
8	450	-0.02	449	0.24	450	0.29
11	379	-0.02	375	0.12	368	0.27

12.5. Evidence for Validity and Consequences of Testing

As the CAS become more fully integrated into the classroom, and with additional administrations of the CoAlt Science assessment, it is intended that information around the consequences of the assessment will be collected. Some of the intended consequences include the appropriate use of the assessment for students with the most significant cognitive disabilities, the inclusion of those students in the state assessment system, and the effective instruction of students with the most significant cognitive disabilities in the EEOs of the CAS. Longitudinal comparisons can begin with the Spring 2024 administration.

12.6. Fairness

Fairness is an important aspect of validity, as it is critical that an assessment provide accurate measurements for **all** students. To that end, the following fairness considerations were woven into the development and administration of the CoAlt Science assessments:

- Sample items that provide the opportunity for teachers and students to become familiar with the test design and scoring of the assessments before experiencing the items on an operational test (Section 5.4)
- Universal design principles that are adhered to during the test development process with the goal of avoiding construct-irrelevant aspects of the assessment that could impact student performance (Chapter 3)
- DIF analyses to identify any items that appear to be unfairly favoring one subgroup over another. All items which show DIF are reviewed by educators for potential bias in the item. (Chapter 3)
- Accessibility tools and accommodations to allow students to fully demonstrate their content knowledge without being hindered by non-construct related elements in addition to being developed to be accessible for students with significant cognitive disabilities (Chapters 2 and 3, Section 5.5)

References

Achieve, Inc. (2018). *Criteria for procuring and evaluating high-quality and aligned summative science assessments*. https://www.nextgenscience.org/sites/default/files/Criteria03202018.pdf

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report (NCES 1999–452)*. National Center for Education Statistics, US Department of Education.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. AERA.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–50.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 35–66). Erlbaum.

Ferguson, G. A., & Takane, Y. (1989). *Statistical analysis in psychology and education* (6th ed.). McGraw-Hill.

Ferrara, S., Perie, M., & Johnson, E. (2008). Matching the judgmental task with standard setting panelist expertise: The item-descriptor (ID) matching method. *Journal of Applied Testing Technology*, *9*(1), 1–22.

Flowers, C., Wakeman, S., Browder, D. M., & Karvonen, M. (2007). Links for Academic Learning (LAL): A conceptual model for investigating alignment of alternate assessments based on alternate achievement standards. *Educational Measurement: Issues and Practice, 28*, 25–37.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425–461.

Linacre, J. M. (2021). Winsteps® (version 4.8.1.0) [computer program]. Winsteps.com.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

National Research Council (NRC). (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press. https://doi.org/10.17226/13165

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academic Press. https://www.nextgenscience.org/search-standards

Pearson. (2024). *Colorado Alternate (CoAlt) Science 2022 standard setting report*. Report developed under contract for the Colorado Department of Education (CDE).

Rasch, G. (1966). An individualistic approach to item analysis. In P. Lazarfeld & N. W. Henry (Eds.), *Readings in Mathematical Social Science* (pp. 89–107). Science Research Associates.

Revivo, R. Z., Dickinson, E. R., & Borawski, E. A. (2023, October). *Colorado Alternate (CoAlt) Science alignment study report*. Human Resources Research Organization (HumRRO).

Stout, W. F. (1990). A new item response theory modelling approach and applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293–325.

Webb, N. L. (1997). *Research Monograph No. 6: Criteria for alignment of expectations and assessments in science and science education*. Council of Chief State Schools Officers.

Webb, N. L. (1999). *Research Monograph No. 18: Alignment of science and science standards and assessments in four states*. National Institute for Science Education and Council of Chief State School Officers. (ERIC Document Reproduction Service No. ED440852).

Webb, N. L. (2005). *Webb alignment tool: Training manual*. Wisconsin Center for Education Research.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Wright, B.D., & Stone, M.H. (1979). *Best test design*. MESA Press.

Appendix A: CoAlt Eligibility Guidelines

Alternate Academic Achievement Standards and Alternate Assessment Participation Guidelines Worksheet

* For further clarification of terms used in this worksheet, please refer to the companion document *Participation Guidelines: Alternate Academic Achievement Standards for Instruction and Alternate Assessment*

Criterion #1: The student has been evaluated and determined to be eligible to receive special education services and has an IEP.	Response:
☐ Has the student been determined to be a student with a disability eligible to receive special education services under the Individuals with Disabilities Education Act (IDEA)?	☐ **No. Stop here.** The student must meet Special Education *Determination of Eligibility* criteria in one or more disability categories defined in ECEA Rules http://www.cde.state.co.us/cdesped/IEP_Forms.asp
☐ Is a current Individualized Education Program (IEP) in place or being developed for the student?	☐ **Yes.** If both elements can be affirmed, **continue to Criterion #2.**
Criterion #2: The student has documented evidence of a cognitive disability.	**Response:**
☐ During the process of determining eligibility for a student to receive special education services, did the IEP Team review a body of evidence that supports the existence of a cognitive disability?	☐ **No. Stop here.** The student must have documented evidence of the existence of a cognitive disability, regardless of the special education disability category. ☐ **Yes.** Empirical evidence of a cognitive disability is documented in the IEP. **Continue to Criterion #3.**
Criterion #3: The student has a significant cognitive disability.	**Response Options:**
☐ The student's demonstrated cognitive functioning and adaptive behavior in the home, school, and community environments are **significantly below age expectations, even with program modifications, adaptations and accommodations** and ☐ the School Psychologist (or other personnel trained in administering psychometric evaluation) presents evidence that the student's cognitive and adaptive functioning is consistent with that of a student with a **significant cognitive disability***.	☐ **Yes.** Both elements affirm that the student's evaluated performance falls within range of the most **significant cognitive disability**. The student (a) requires extensive, repeated individualized instruction and support that is not of a temporary or transient nature and (b) uses substantially adapted and modified materials and individualized methods of accessing information in alternative ways to acquire, maintain, generalize, demonstrate and transfer academic and functional skills necessary for application in school, work, home and community environments. Daily modified instruction is linked to the enrolled grade level **Colorado Academic Standards Extended Evidence Outcomes** (EEOs). For students receiving instruction on alternate standards and taking alternate assessment, the IEP must contain measurable annual goals **and** objectives for content areas. **Continue to 4B to select alternate standards-based instruction and appropriate alternate assessment.**
	☐ The documented evidence supports the existence of a significant cognitive disability. However the IEP Team determines that with appropriate adaptations (supports and accommodations), the student will receive daily instruction based on the Colorado Academic Standards enrolled **grade-level expectations**. (The student then does not qualify for instruction on alternate academic achievement standards or to take alternate assessment based on alternate academic achievement standards.) **Continue to 4A to select Grade-level standards-based instruction and appropriate grade-level assessment.**
Empirical evidence includes, but is not limited to, formal testing results, multi-disciplinary team evaluations, and other evaluative data.	☐ **Yes.** Although the documented evidence supporting the existence of a significant cognitive disability does not fall into the lower ranges, the IEP Team has considered the impact and severity of the disability along with other related factors in order to determine that the student qualifies to receive modified daily instruction based on the Colorado Academic Standards **Extended Evidence Outcomes** (alternate academic achievement standards) and participate in alternate assessment based on alternate academic achievement standards. **Continue to 4B to select Alternate standards-based instruction and appropriate alternate assessment.**

For questions related to this optional worksheet and companion guidance, please contact:
Gina Herrera Herrera_g@cde.state.co.us Exceptional Student Services Unit Colorado Dept. of Education
Rev. 9/15

Alternate Academic Achievement Standards and Alternate Assessment Participation Guidelines Worksheet

Tested Content Areas	4A Instruction and Assessment based on **Grade-Level Academic Achievement Standards** (Grade-level Expectations / Evidence Outcomes)	4B Instruction based on **Extended Evidence Outcomes (EEOs)** and ***Alternate Assessment** based on Alternate Academic Achievement Standards (AA-AAS)
CMAS: Reading/ Writing (ELA) Math Social Studies Science	☐ Grade-level classroom/ district assessments ☐ with accommodation ☐ without accommodation ☐ **State** Summative Assessment ☐ with accommodations allowed for use on state assessment ☐ without accommodation ☐ Unique Request- pending approval by CDE Assessment Unit	☐ Alternate classroom/ district assessments based on alternate standards ☐ Alternate **State** Summative Assessments (Gr. 3-9 and 11) Note: With the passage of IDEA in 1997 and its reauthorization in 2004, it is required that both state and districts provide an alternate assessment for students who cannot participate in general state and district assessments.
Other	☐ ACCESS for ELLs (K-12) ☐ with allowable accommodations ☐ Grade 10 Preparatory Exam ☐ Grade 11 College Entrance Exam	☐ Alternate ACCESS for ELLs (Gr. 1-12) ☐ 10th Grade DLM Alternate Assessment ☐ 11th Grade DLM Alternate Assessment
Dual Assessment	*Dual assessment is NOT an option beginning with the 2014-15 school year. If a student meets the guidelines to receive instruction on alternate standards and take alternate assessment based upon those alternate standards, then ALL tested content areas or other state-mandated assessments required for the student's enrolled grade level, will be ALTERNATE assessments.	

Exclusionary Factors:

The **IEP Team affirms**

- ☐ that annual assessment data was reviewed for each content area and
- ☐ the decision for participation in the Alternate Assessment is NOT based on:
  1. A disability category or label
  2. Poor attendance or extended absences
  3. Native language/social/cultural or economic difference
  4. Expected poor performance on the grade-level assessment
  5. Services student receives
  6. Educational environment or instructional setting
  7. Percent of time receiving special education
  8. English Language Learner (ELL) status
  9. low reading level/academic level
  10. Anticipated student's disruptive behavior
  11. Impact of student scores on accountability system
  12. Administrator decision
  13. Anticipated student's emotional duress

IEP Team Consensus: (Record decision on IEP Form)

☐ **Student meets participation guidelines as a student with a significant cognitive disability and will receive instruction based upon alternate academic achievement standards and participate in alternate assessment as indicated above.**

* For further clarification of terms used in this worksheet, please refer to the companion document ***Participation Guidelines: Alternate Academic Achievement Standards for Instruction and Alternate Assessment***

For questions related to this optional worksheet and companion guidance, please contact:
Gina Herrera Herrera_g@cde.state.co.us Exceptional Student Services Unit Colorado Dept. of Education
Rev. 9/15

Appendix B: Sample Student Performance Report

Confidential Student Performance Report

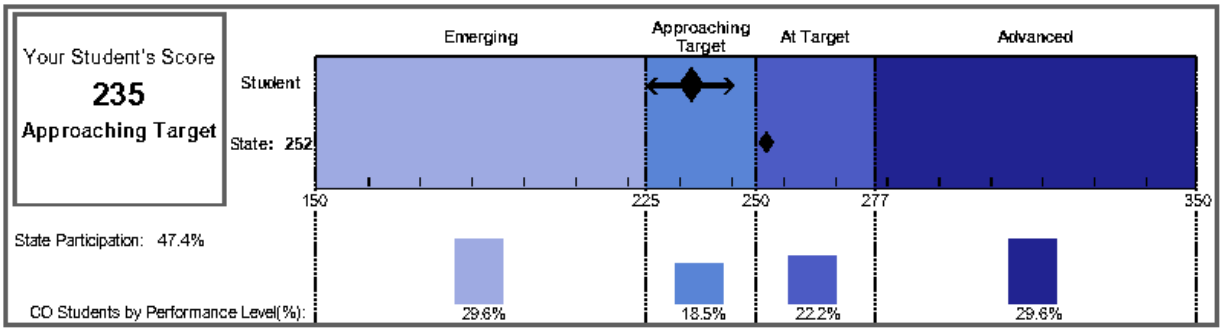
Colorado Alternate Assessment
Based on the 2020 Colorado Academic Standards
Student: **FIRSTNAME LASTNAME**
SASID: 9999999999 Birthdate: MM/DD/YYYY
School: MIDDLE SCHOOL (9999)
District: SAMPLE DISTRICT (9999)
Spring 2023

Science — Grade 8

This Colorado Alternate Assessment (CoAlt) report provides information about your student's understanding of the Extended Evidence Outcomes (EEOs) of Colorado's middle school science standards. Scan the QR code at the bottom of this page to see a video that will talk you through your student's report.

Your student's performance is shown as:
- A scale score: A numerical score based on Colorado's middle school science expectations
- A performance level: Your student's performance level is described at the bottom of this page

Consider as you review this report:
- Arrows around your student's diamond show where your student may have scored if the assessment was taken multiple times.
- Make state comparisons with caution if participation is low.
- Talk with your student's teacher about your student's progress in science.



Performance Level Description - Approaching Target

FIRSTNAME006 showed a limited understanding of the EEOs of Colorado's middle school science standards and will likely need moderate academic support to successfully engage in the next grade level. Students in the Approaching Target level typically:

- Identify that the amount of or the mass of atoms does not change in a chemical reaction.
- Identify simple molecules, such as water or oxygen gas.
- Identify a device that releases or absorbs heat energy by chemical processes and a device that either minimizes or maximizesheat energy transfer.
- Identify the relative amounts of kinetic and potential energy in a system.
- Identify that different materials can affect the reflection, absorption, or transmission of a light or sound wave.
- Identify how characteristic animal behaviors and specialized plant structures help the plants and animals survive, and identify examples of competitive, predatory, and mutually beneficial relationships between organisms.
- Identify an example of the cycling of matter and energy among living and nonliving parts of an ecosystem.
- Identify that variations of traits in populations increase some individuals' probability of surviving and reproducing and that natural selection works over many generations.
- Identify two locations of similar or different climates.
- Identify that regional climate is based on the region's landforms and latitude.
- Identify that Earth's resources are limited and unevenly distributed.
- Identify gravity as what keeps Earth and the Moon in their orbits and as what draws and holds together the matter making up Earth and the Moon.

8

To view a video report and the full version of the performance level descriptor, visit **https://coassessments.com/parentsandguardians/** or access the QR code.

Watch a video about this report!

Information about the Colorado Academic Standards measured by this assessment:
http://www.cde.state.co.us/coextendedeo.

Page 1 of 2

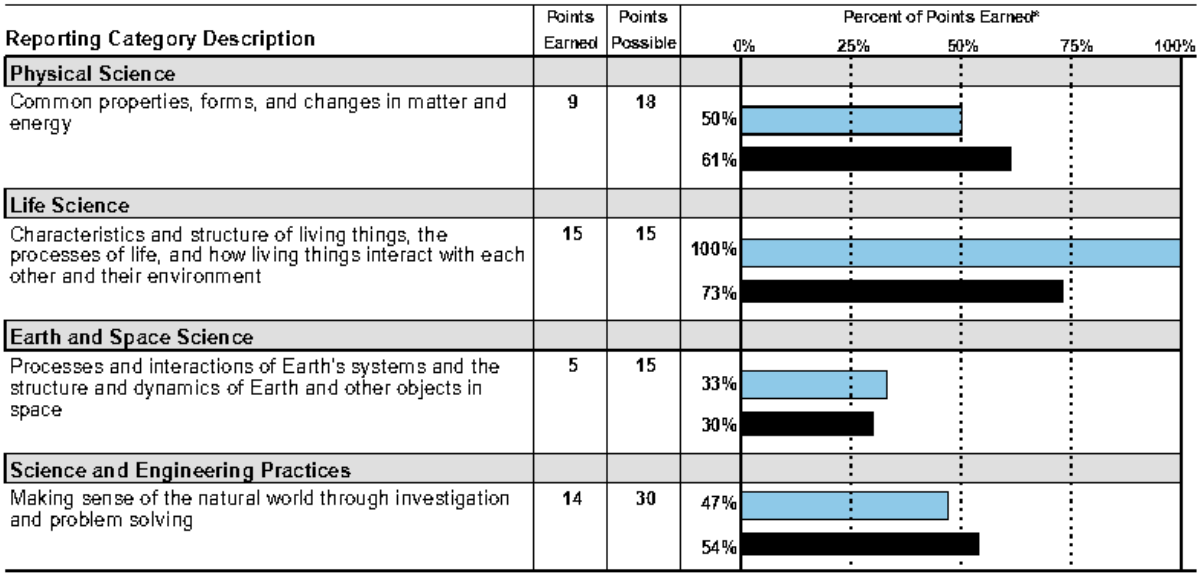
06082023-Z9999999-9999-9999- 0000000

FIRSTNAME LASTNAME

Content Standard Performance

Content Standard Performance

Reporting Category Description	Points Earned	Points Possible	Percent of Points Earned*
Physical Science			
Common properties, forms, and changes in matter and energy	9	18	50% 61%
Life Science			
Characteristics and structure of living things, the processes of life, and how living things interact with each other and their environment	15	15	100% 73%
Earth and Space Science			
Processes and interactions of Earth's systems and the structure and dynamics of Earth and other objects in space	5	15	33% 30%
Science and Engineering Practices			
Making sense of the natural world through investigation and problem solving	14	30	47% 54%



*The percent of points earned cannot be compared across years because individual items change from year to year. They also cannot be compared across Standards because the number of items and the difficulty of items may not be the same.

For information on the CoAlt assessment program, visit
http://www.cde.state.co.us/assessment.

Page 2 of 2

Appendix C: Scale Score Distributions

Table C.1. Scale Score Distribution—Science Grade 5

Scale Score	Freq.	%	Cum. Freq.	Cum. %
150	13	3.37	13	3.37
156	1	0.26	14	3.63
166	3	0.78	17	4.40
175	2	0.52	19	4.92
182	3	0.78	22	5.70
189	3	0.78	25	6.48
195	6	1.55	31	8.03
200	8	2.07	39	10.10
205	7	1.81	46	11.92
210	14	3.63	60	15.54
215	22	5.70	82	21.24
219	19	4.92	101	26.17
224	28	7.25	129	33.42
225	24	6.22	153	39.64
232	19	4.92	172	44.56
236	26	6.74	198	51.30
240	16	4.15	214	55.44
244	19	4.92	233	60.36
248	17	4.40	250	64.77
250	19	4.92	269	69.69
256	18	4.66	287	74.35
260	18	4.66	305	79.02
264	18	4.66	323	83.68
268	7	1.81	330	85.49
273	11	2.85	341	88.34
277	10	2.59	351	90.93
281	12	3.11	363	94.04
286	7	1.81	370	95.85
291	3	0.78	373	96.63
296	3	0.78	376	97.41
302	5	1.30	381	98.70
308	2	0.52	383	99.22
322	2	0.52	385	99.74
330	1	0.26	386	100.00

Table C.2. Scale Score Distribution—Science Grade 8

Scale Score	Freq.	%	Cum. Freq.	Cum. %
150	14	3.01	14	3.01
155	2	0.43	16	3.44
163	1	0.22	17	3.66
170	1	0.22	18	3.87
176	2	0.43	20	4.30
186	4	0.86	24	5.16
190	4	0.86	28	6.02
194	8	1.72	36	7.74
198	5	1.08	41	8.82
201	8	1.72	49	10.54
204	10	2.15	59	12.69
208	20	4.30	79	16.99
211	17	3.66	96	20.65
214	19	4.09	115	24.73
217	24	5.16	139	29.89
220	20	4.30	159	34.19
223	22	4.73	181	38.92
225	13	2.80	194	41.72
229	21	4.52	215	46.24
232	18	3.87	233	50.11
235	18	3.87	251	53.98
238	20	4.30	271	58.28
240	16	3.44	287	61.72
243	17	3.66	304	65.38
246	10	2.15	314	67.53
250	21	4.52	335	72.04
253	19	4.09	354	76.13
256	17	3.66	371	79.78
259	11	2.37	382	82.15
263	15	3.23	397	85.38
266	8	1.72	405	87.10
270	13	2.80	418	89.89
274	6	1.29	424	91.18
277	8	1.72	432	92.90
282	5	1.08	437	93.98
286	10	2.15	447	96.13
291	6	1.29	453	97.42
297	2	0.43	455	97.85
303	5	1.08	460	98.92
310	2	0.43	462	99.35
319	1	0.22	463	99.57
329	1	0.22	464	99.78
343	1	0.22	465	100.00

Table C.3. Scale Score Distribution—Science Grade 11

Scale Score	Freq.	%	Cum. Freq.	Cum. %
150	17	4.25	17	4.25
154	6	1.50	23	5.75
163	2	0.50	25	6.25
170	6	1.50	31	7.75
177	5	1.25	36	9.00
184	7	1.75	43	10.75
189	9	2.25	52	13.00
195	8	2.00	60	15.00
200	8	2.00	68	17.00
205	11	2.75	79	19.75
209	11	2.75	90	22.50
214	16	4.00	106	26.50
218	11	2.75	117	29.25
222	23	5.75	140	35.00
225	14	3.50	154	38.50
230	21	5.25	175	43.75
234	16	4.00	191	47.75
238	24	6.00	215	53.75
242	12	3.00	227	56.75
246	20	5.00	247	61.75
249	21	5.25	268	67.00
250	8	2.00	276	69.00
257	16	4.00	292	73.00
261	18	4.50	310	77.50
264	16	4.00	326	81.50
268	11	2.75	337	84.25
272	12	3.00	349	87.25
276	6	1.50	355	88.75
277	5	1.25	360	90.00
284	16	4.00	376	94.00
289	2	0.50	378	94.50
293	4	1.00	382	95.50
298	6	1.50	388	97.00
303	3	0.75	391	97.75
308	1	0.25	392	98.00
320	2	0.50	394	98.50
327	3	0.75	397	99.25
343	2	0.50	399	99.75
350	1	0.25	400	100.00

Appendix D: Scale Score Distribution Histograms

Figure D.1. Scale Score Distribution Histogram—Grade 5

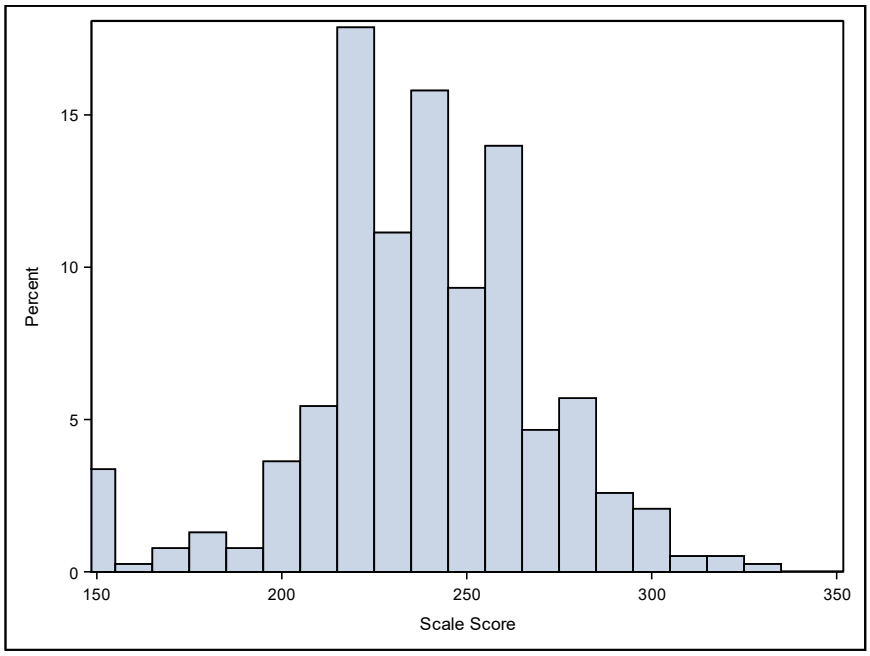


Figure D.2. Scale Score Distribution Histogram—Grade 8

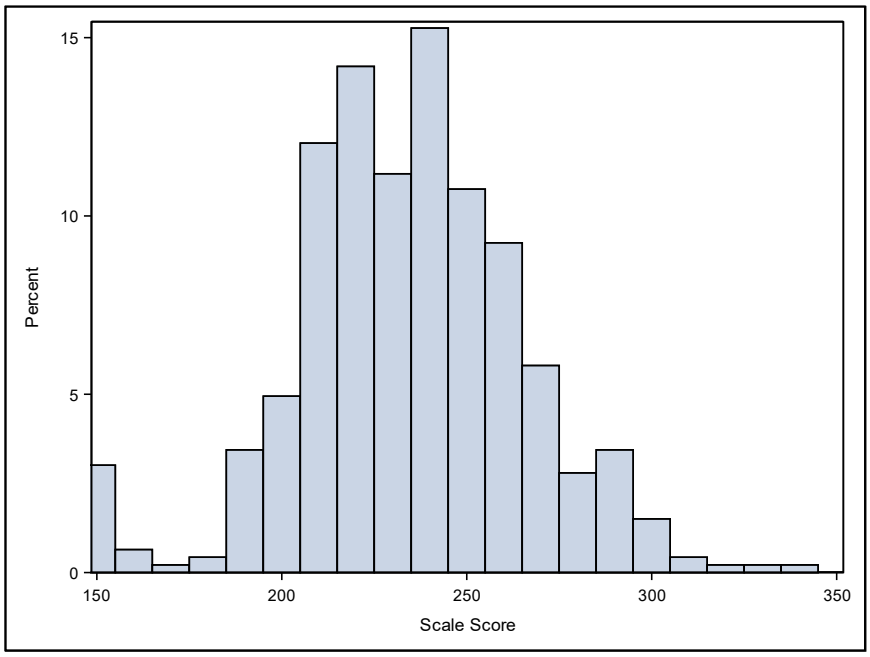
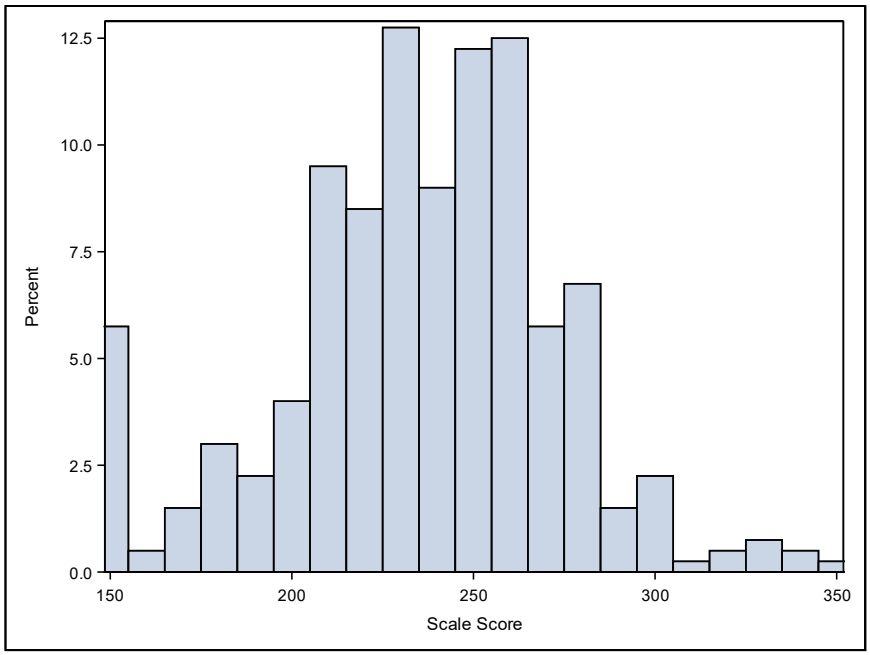


Figure D.3. Scale Score Distribution Histogram—Grade 11



Appendix E: Performance Results by Demographic Subgroup

Table E.1. Scale Score Summary Statistics by Demographic Subgroup—Grade 5

Subgroup	N	Mean	SD	Min.	Max.	Alpha
No IEP	*	*	*	*	*	*
IEP	386	237.53	31.99	150	330	0.80
No Accommodation	*	*	*	*	*	*
Accommodation	386	237.53	31.99	150	330	0.80
Am. Indian/Alaska Native	*	*	*	*	*	*
Asian	17	233.06	29.35	175	281	0.80
Black	35	233.00	30.80	156	330	0.77
Hispanic	163	236.31	30.55	150	322	0.78
White	149	239.64	33.88	150	322	0.83
Hawaiian/Pacific Islander	*	*	*	*	*	*
Two or More Races	18	236.94	30.95	150	281	0.81
Missing	*	*	*	*	*	*
No Economic Disadvantage	163	238.35	29.36	150	322	0.77
Economic Disadvantage	223	236.92	33.84	150	330	0.82
Female	147	232.58	31.24	150	302	0.79
Male	239	240.57	32.13	150	330	0.81
Language Proficiency NA	300	238.72	32.76	150	330	0.81
Language Proficiency NEP	61	229.82	30.71	150	302	0.81
Language Proficiency LEP	*	*	*	*	*	*
Language Proficiency FEP	20	244.75	23.86	210	286	0.68
Not Migrant	385	237.52	32.03	150	330	0.81
Migrant	*	*	*	*	*	*

*n-count less than 16

Table E.2. Scale Score Summary Statistics by Demographic Subgroup—Grade 8

Subgroup	N	Mean	SD	Min.	Max.	Alpha
No IEP	*	*	*	*	*	*
IEP	464	234.34	30.89	150	343	0.86
No Accommodation	*	*	*	*	*	*
Accommodation	465	234.16	31.11	150	343	0.86
Am. Indian/Alaska Native	*	*	*	*	*	*
Asian	18	234.56	22.49	194	277	0.75
Black	38	232.11	31.42	150	286	0.88
Hispanic	207	233.28	29.92	150	310	0.85
White	178	236.48	33.00	150	343	0.86
Hawaiian/Pacific Islander	*	*	*	*	*	*
Two or More Races	19	226.53	35.33	150	286	0.91
Missing	*	*	*	*	*	*
No Economic Disadvantage	216	232.03	30.96	150	343	0.85
Economic Disadvantage	249	236.01	31.17	150	319	0.86
Female	181	233.60	29.47	150	329	0.85
Male	284	234.51	32.15	150	343	0.86

Subgroup	N	Mean	SD	Min.	Max.	Alpha
Language Proficiency NA	366	235.62	31.78	150	343	0.86
Language Proficiency NEP	45	219.89	31.59	150	303	0.87
Language Proficiency LEP	*	*	*	*	*	*
Language Proficiency FEP	45	238.40	20.92	204	297	0.67
Not Migrant	465	234.16	31.11	150	343	0.86
Migrant	*	*	*	*	*	*

*n-count less than 16

Table E.3. Scale Score Summary Statistics by Demographic Subgroup—Grade 11

Subgroup	N	Mean	SD	Min.	Max.	Alpha
No IEP	*	*	*	*	*	*
IEP	400	235.17	37.96	150	350	0.87
No Accommodation	*	*	*	*	*	*
Accommodation	400	235.17	37.96	150	350	0.87
Am. Indian/Alaska Native	*	*	*	*	*	*
Asian	*	*	*	*	*	*
Black	33	230.21	32.46	170	293	0.84
Hispanic	190	234.92	38.07	150	343	0.88
White	138	235.78	38.05	150	350	0.88
Hawaiian/Pacific Islander	*	*	*	*	*	*
Two or More Races	25	238.60	44.81	150	327	0.90
Missing	*	*	*	*	*	*
No Economic Disadvantage	196	231.60	34.58	150	327	0.84
Economic Disadvantage	204	238.60	40.73	150	350	0.89
Female	174	233.55	36.40	150	327	0.87
Male	226	236.42	39.15	150	350	0.88
Language Proficiency NA	315	236.42	38.47	150	350	0.88
Language Proficiency NEP	37	220.54	42.78	150	343	0.91
Language Proficiency LEP	*	*	*	*	*	*
Language Proficiency FEP	43	238.67	28.06	177	308	0.77
Not Migrant	399	235.14	38.00	150	350	0.87
Migrant	*	*	*	*	*	*

*n-count less than 16

Appendix F: Classical Item-Level Statistics

Table F.1. SR Item Classical Statistics—Science Grade 5

Item	Omit %	*P*-value	Item–Total Correlation
1	0.83	0.58	0.53
2	0.83	0.37	0.18
3	0.83	0.23	0.25
4	1.10	0.49	0.39
5	1.10	0.30	0.28
6	1.10	0.21	0.18
7	0.83	0.47	0.51
8	0.83	0.63	0.32
9	0.83	0.14	0.20
10	0.83	0.32	0.33
11	0.83	0.45	0.33
12	0.83	0.37	0.47
13	0.83	0.42	0.35
14	0.83	0.41	0.53
15	0.83	0.25	0.40
16	0.00	0.50	0.37
17	0.00	0.46	0.52
18	0.00	0.46	0.55
19	0.00	0.60	0.49
20	0.00	0.64	0.35
21	0.00	0.37	0.05
22	0.00	0.47	0.40
23	0.00	0.47	0.25
24	0.28	0.32	0.26
25	0.28	0.39	0.43
26	0.28	0.41	0.41
27	0.28	0.33	0.40
28	0.28	0.47	0.49
29	0.28	0.37	0.45
30	0.28	0.45	0.23
31	0.00	0.43	0.16
32	0.00	0.61	0.24
33	0.00	0.34	0.30

Table F.2. SPT Item Classical Statistics—Science Grade 5

Item	Max. Points	Omit %	0%	1%	2%	3%	P-value	Item-Total Correlation
1	3	1.10	25.07	31.68	25.62	16.53	0.44	0.55
2	3	0.00	8.26	24.52	39.67	27.55	0.62	0.53
3	3	0.55	12.12	28.37	43.53	15.43	0.54	0.44

Table F.3. SR Item Classical Statistics—Science Grade 8

Item	Omit %	*P*-value	Item–Total Correlation
1	0.68	0.47	0.46
2	0.68	0.46	0.35
3	0.68	0.36	0.44
4	0.90	0.48	0.33
5	0.90	0.59	0.51
6	1.13	0.38	0.20
7	0.90	0.48	0.34
8	1.13	0.41	0.22
9	1.13	0.30	0.28
10	1.13	0.64	0.60
11	1.13	0.44	0.41
12	1.36	0.43	0.36
13	0.45	0.64	0.52
14	0.45	0.56	0.55
15	0.45	0.49	0.54
16	0.45	0.56	0.30
17	0.45	0.39	0.42
18	0.45	0.32	0.37
19	0.45	0.65	0.37
20	0.45	0.71	0.39
21	0.45	0.25	0.26
22	0.45	0.57	0.41
23	0.45	0.51	0.45
24	0.45	0.62	0.36
25	0.90	0.59	0.44
26	0.90	0.67	0.44
27	0.90	0.23	0.30
28	1.13	0.65	0.44
29	1.13	0.50	0.54
30	1.13	0.20	0.25
31	0.90	0.53	0.58
32	0.90	0.72	0.49
33	0.90	0.38	0.15
34	0.90	0.45	0.43
35	0.90	0.42	0.38
36	0.90	0.24	0.30
37	0.90	0.43	0.07
38	1.13	0.64	0.55
39	0.90	0.24	0.28

Table F.4. SPT Item Classical Statistics—Science Grade 8

Item	Max. Points	Omit %	0%	1%	2%	3%	P-value	Item-Total Correlation
1	3	1.58	25.11	33.26	23.98	16.06	0.43	0.56
2	3	0.90	9.73	12.67	10.86	65.84	0.77	0.68
3	3	0.90	20.81	27.15	28.28	22.85	0.51	0.45

Table F.5. SR Item Classical Statistics—Science Grade 11

Item	Omit %	*P*-value	Item–Total Correlation
1	0.80	0.39	0.38
2	0.80	0.40	0.31
3	0.80	0.41	0.33
4	0.80	0.73	0.53
5	0.80	0.25	0.12
6	0.80	0.27	0.24
7	1.07	0.62	0.47
8	1.07	0.40	0.25
9	1.34	0.14	0.21
10	1.07	0.56	0.51
11	1.07	0.30	0.18
12	1.07	0.38	0.42
13	0.00	0.43	0.46
14	0.00	0.51	0.45
15	0.00	0.58	0.43
16	0.00	0.62	0.56
17	0.00	0.44	0.42
18	0.00	0.28	0.16
19	0.00	0.56	0.37
20	0.00	0.32	0.31
21	0.00	0.54	0.57
22	0.00	0.77	0.51
23	0.00	0.16	0.17
24	0.00	0.38	0.39
25	0.80	0.59	0.60
26	0.80	0.47	0.38
27	0.80	0.39	0.38
28	0.54	0.29	0.25
29	0.54	0.45	0.46
30	0.54	0.47	0.49
31	0.80	0.51	0.49
32	0.80	0.44	0.48
33	0.80	0.43	0.44
34	0.80	0.58	0.40
35	0.80	0.32	0.34
36	0.80	0.32	0.42
37	0.27	0.56	0.47
38	0.27	0.46	0.51
39	0.80	0.71	0.47

Table F.6. SPT Item Classical Statistics—Science Grade 11

Item	Max. Points	Omit %	0%	1%	2%	3%	P-value	Item-Total Correlation
1	3	2.14	38.61	40.21	14.48	4.56	0.28	0.35
2	3	0.27	19.57	26.27	29.76	24.13	0.53	0.65
3	3	1.07	42.09	30.03	16.09	10.72	0.31	0.46

Appendix G: IRT Item-Level Statistics

Table G.1. Operational Item Parameter Estimates—Science Grade 5

Item	Item Type	Model	B	D1	D2	D3	D4	Infit	Outfit
1	SPT	Rasch	-0.133	0	-0.797	0.092	0.705	1.02	1.02
2	SPT	Rasch	-1.007	0	-1.357	0.050	1.307	1.02	1.02
3	SPT	Rasch	-0.515	0	-1.313	-0.284	1.597	1.16	1.21
4	SR	Rasch	-0.745	0	0	–	–	0.86	0.83
5	SR	Rasch	0.248	0	0	–	–	1.13	1.16
6	SR	Rasch	1.010	0	0	–	–	1.02	1.31
7	SR	Rasch	-0.308	0	0	–	–	0.97	1.07
8	SR	Rasch	0.613	0	0	–	–	1.04	1.02
9	SR	Rasch	1.117	0	0	–	–	1.09	1.14
10	SR	Rasch	-0.232	0	0	–	–	0.88	0.85
11	SR	Rasch	-0.974	0	0	–	–	1.04	1.02
12	SR	Rasch	1.636	0	0	–	–	1.06	1.08
13	SR	Rasch	0.482	0	0	–	–	1.00	1.06
14	SR	Rasch	-0.156	0	0	–	–	1.03	1.01
15	SR	Rasch	0.234	0	0	–	–	0.90	0.84
16	SR	Rasch	0.024	0	0	–	–	1.00	0.97
17	SR	Rasch	0.037	0	0	–	–	0.85	0.82
18	SR	Rasch	0.876	0	0	–	–	0.94	0.86
19	SR	Rasch	-0.346	0	0	–	–	0.99	0.97
20	SR	Rasch	-0.169	0	0	–	–	0.86	0.84
21	SR	Rasch	-0.169	0	0	–	–	0.84	0.81
22	SR	Rasch	-0.838	0	0	–	–	0.89	0.92
23	SR	Rasch	-1.015	0	0	–	–	1.01	0.99
24	SR	Rasch	0.221	0	0	–	–	1.24	1.27
25	SR	Rasch	-0.207	0	0	–	–	0.97	1.05
26	SR	Rasch	-0.245	0	0	–	–	1.09	1.09
27	SR	Rasch	0.511	0	0	–	–	1.07	1.06
28	SR	Rasch	0.128	0	0	–	–	0.93	0.93
29	SR	Rasch	0.050	0	0	–	–	0.95	0.91
30	SR	Rasch	0.468	0	0	–	–	0.95	0.92
31	SR	Rasch	-0.219	0	0	–	–	0.89	0.86
32	SR	Rasch	0.248	0	0	–	–	0.91	0.90
33	SR	Rasch	-0.118	0	0	–	–	1.11	1.11
34	SR	Rasch	-0.028	0	0	–	–	1.16	1.15
35	SR	Rasch	-0.865	0	0	–	–	1.11	1.11
36	SR	Rasch	0.384	0	0	–	–	1.02	1.03

Table G.2. Operational Item Parameter Estimates—Science Grade 8

Item	Item Type	Model	B	D1	D2	D3	D4	Infit	Outfit
1	SPT	Rasch	0.188	0	-0.920	0.158	0.762	1.09	1.11
2	SPT	Rasch	-1.119	0	-0.115	0.819	-0.704	0.73	0.57
3	SPT	Rasch	-0.122	0	-0.725	-0.017	0.743	1.34	1.36
4	SR	Rasch	0.057	0	0	–	–	0.94	0.93
5	SR	Rasch	0.111	0	0	–	–	1.04	1.05
6	SR	Rasch	0.564	0	0	–	–	0.93	1.05
7	SR	Rasch	-0.007	0	0	–	–	1.06	1.09
8	SR	Rasch	-0.526	0	0	–	–	0.89	0.87
9	SR	Rasch	0.485	0	0	–	–	1.16	1.28
10	SR	Rasch	0.004	0	0	–	–	1.05	1.05
11	SR	Rasch	0.318	0	0	–	–	1.17	1.19
12	SR	Rasch	0.899	0	0	–	–	1.07	1.11
13	SR	Rasch	-0.799	0	0	–	–	0.81	0.73
14	SR	Rasch	0.187	0	0	–	–	0.97	0.97
15	SR	Rasch	0.252	0	0	–	–	1.02	1.03
16	SR	Rasch	-0.776	0	0	–	–	0.88	0.89
17	SR	Rasch	-0.362	0	0	–	–	0.85	0.80
18	SR	Rasch	-0.039	0	0	–	–	0.86	0.83
19	SR	Rasch	-0.384	0	0	–	–	1.09	1.08
20	SR	Rasch	0.428	0	0	–	–	0.97	0.95
21	SR	Rasch	0.813	0	0	–	–	0.99	0.98
22	SR	Rasch	-0.847	0	0	–	–	1.01	0.99
23	SR	Rasch	-1.183	0	0	–	–	1.00	0.94
24	SR	Rasch	1.189	0	0	–	–	1.07	1.12
25	SR	Rasch	-0.449	0	0	–	–	1.00	1.08
26	SR	Rasch	-0.125	0	0	–	–	0.95	0.91
27	SR	Rasch	-0.661	0	0	–	–	1.03	1.01
28	SR	Rasch	-0.515	0	0	–	–	0.96	0.93
29	SR	Rasch	-0.918	0	0	–	–	0.96	0.94
30	SR	Rasch	1.347	0	0	–	–	1.04	1.01
31	SR	Rasch	-0.847	0	0	–	–	0.96	0.89
32	SR	Rasch	-0.071	0	0	–	–	0.86	0.83
33	SR	Rasch	1.533	0	0	–	–	1.06	1.11
34	SR	Rasch	-0.254	0	0	–	–	0.83	0.79
35	SR	Rasch	-1.237	0	0	–	–	0.92	0.81
36	SR	Rasch	0.462	0	0	–	–	1.22	1.29
37	SR	Rasch	0.133	0	0	–	–	0.97	0.95
38	SR	Rasch	0.307	0	0	–	–	1.01	0.99
39	SR	Rasch	1.259	0	0	–	–	1.03	1.05
40	SR	Rasch	0.230	0	0	–	–	1.30	1.40
41	SR	Rasch	-0.753	0	0	–	–	0.86	0.78
42	SR	Rasch	1.231	0	0	–	–	1.07	1.10

Table G.3. Operational Item Parameter Estimates—Science Grade 11

Item	Item Type	Model	B	D1	D2	D3	D4	Infit	Outfit
1	SPT	Rasch	0.818	0	-1.279	0.311	0.968	1.34	1.32
2	SPT	Rasch	-0.402	0	-0.799	0.006	0.793	0.95	1.00
3	SPT	Rasch	0.468	0	-0.607	0.188	0.419	1.24	1.25
4	SR	Rasch	0.231	0	0	–	–	1.00	0.97
5	SR	Rasch	0.192	0	0	–	–	1.07	1.05
6	SR	Rasch	0.139	0	0	–	–	1.05	1.09
7	SR	Rasch	-1.508	0	0	–	–	0.86	0.79
8	SR	Rasch	1.012	0	0	–	–	1.19	1.49
9	SR	Rasch	0.899	0	0	–	–	1.10	1.12
10	SR	Rasch	-0.890	0	0	–	–	0.95	0.90
11	SR	Rasch	0.192	0	0	–	–	1.12	1.16
12	SR	Rasch	1.826	0	0	–	–	1.07	1.13
13	SR	Rasch	-0.581	0	0	–	–	0.90	0.84
14	SR	Rasch	0.702	0	0	–	–	1.16	1.24
15	SR	Rasch	0.297	0	0	–	–	0.96	0.92
16	SR	Rasch	0.061	0	0	–	–	0.93	0.94
17	SR	Rasch	-0.362	0	0	–	–	0.96	0.92
18	SR	Rasch	-0.660	0	0	–	–	0.98	0.97
19	SR	Rasch	-0.890	0	0	–	–	0.86	0.80
20	SR	Rasch	-0.016	0	0	–	–	0.97	0.98
21	SR	Rasch	0.791	0	0	–	–	1.15	1.49
22	SR	Rasch	-0.595	0	0	–	–	1.03	1.05
23	SR	Rasch	0.572	0	0	–	–	1.04	1.44
24	SR	Rasch	-0.490	0	0	–	–	0.84	0.81
25	SR	Rasch	-1.794	0	0	–	–	0.89	0.75
26	SR	Rasch	1.637	0	0	–	–	1.08	1.52
27	SR	Rasch	0.297	0	0	–	–	0.98	1.01
28	SR	Rasch	-0.727	0	0	–	–	0.81	0.76
29	SR	Rasch	-0.131	0	0	–	–	1.02	0.99
30	SR	Rasch	0.218	0	0	–	–	1.01	0.97
31	SR	Rasch	0.761	0	0	–	–	1.06	1.36
32	SR	Rasch	-0.055	0	0	–	–	0.94	0.92
33	SR	Rasch	-0.157	0	0	–	–	0.91	0.88
34	SR	Rasch	-0.336	0	0	–	–	0.92	0.88
35	SR	Rasch	-0.003	0	0	–	–	0.91	0.88
36	SR	Rasch	0.048	0	0	–	–	0.95	0.92
37	SR	Rasch	-0.687	0	0	–	–	1.02	0.99
38	SR	Rasch	0.600	0	0	–	–	1.03	1.02
39	SR	Rasch	0.572	0	0	–	–	0.95	0.90
40	SR	Rasch	-0.555	0	0	–	–	0.94	0.92
41	SR	Rasch	-0.119	0	0	–	–	0.89	0.85
42	SR	Rasch	-1.377	0	0	–	–	0.94	0.86

Appendix H: Test Characteristic Curves (TCCs)

Figure H.1. TCC—Grade 5

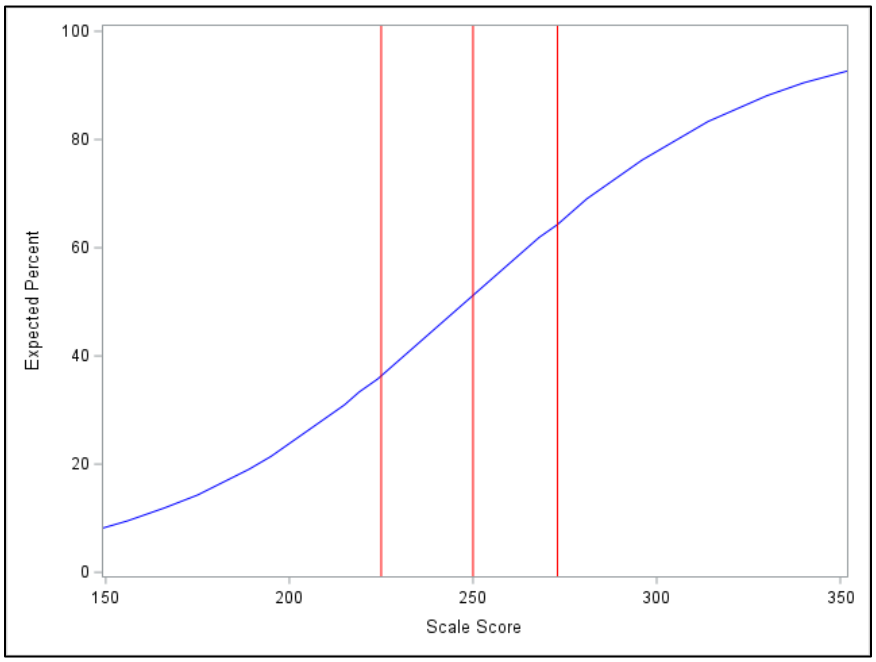


Figure H.2. TCC—Grade 8

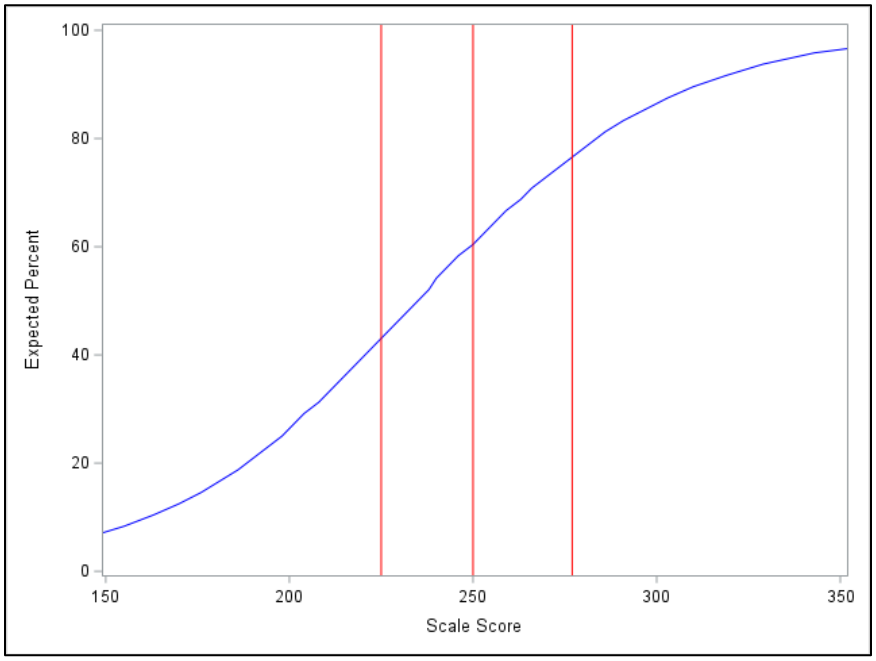
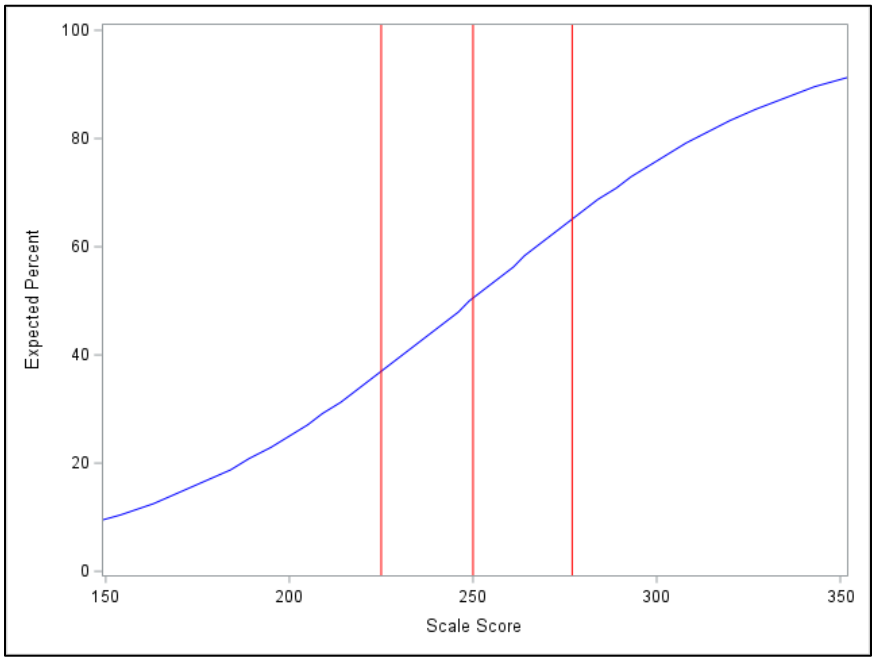


Figure H.3. TCC—Grade 11



Appendix I: Test Information Curves (TICs) and CSEM Curves

Figure I.1. TIC—Grade 5

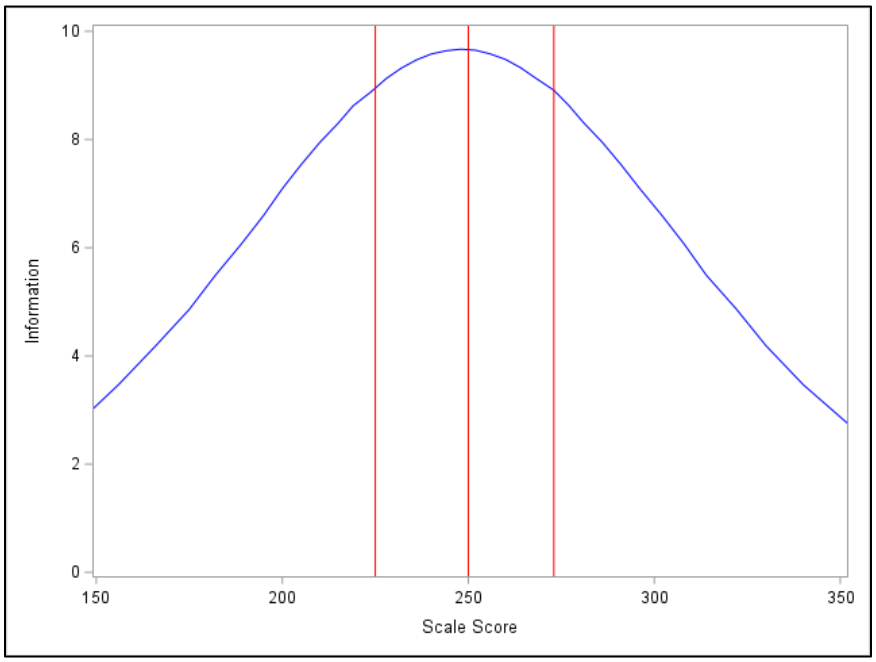


Figure I.2. TIC—Grade 8

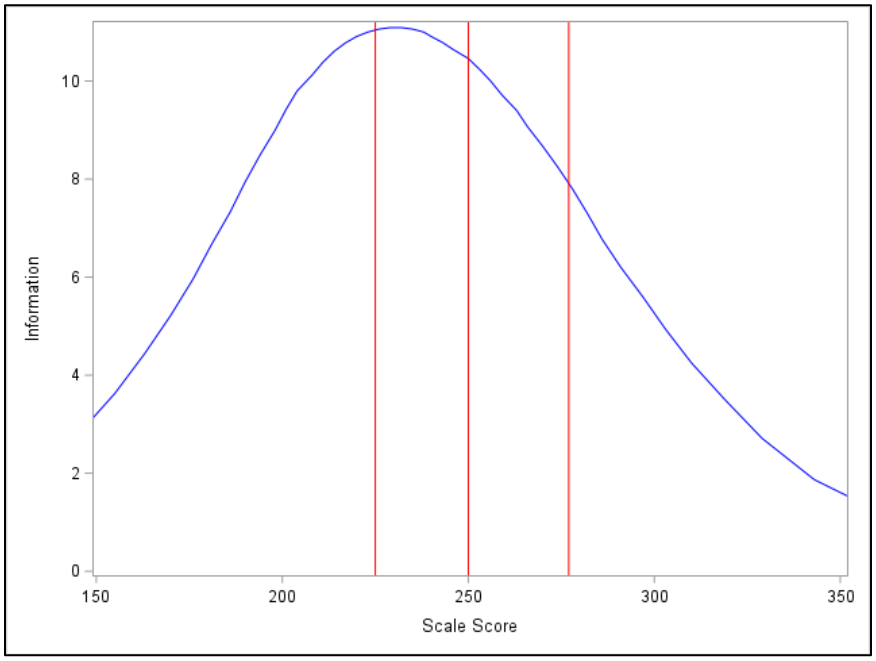


Figure I.3. TIC—Grade 11

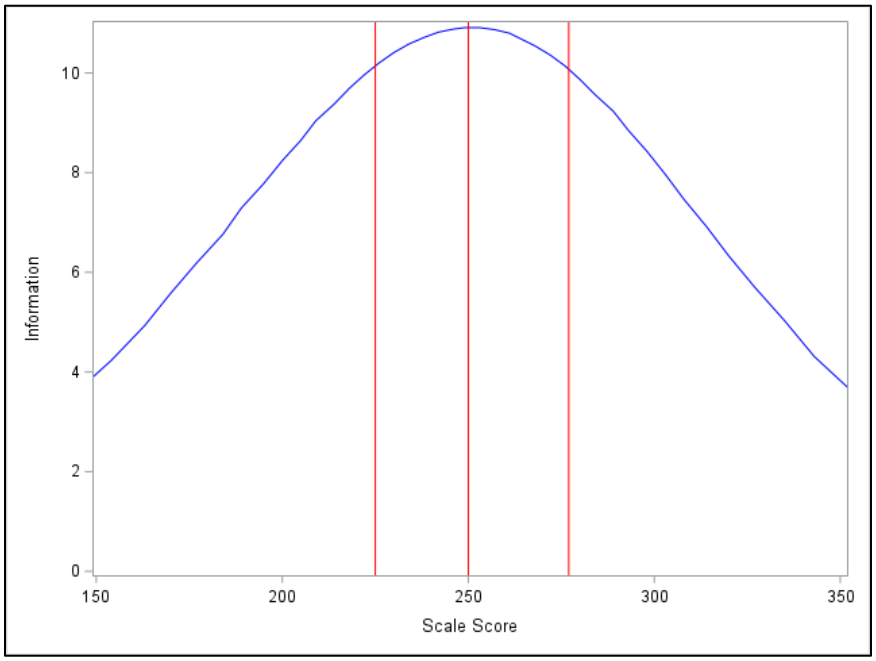


Figure I.4. CSEM Curve—Grade 5

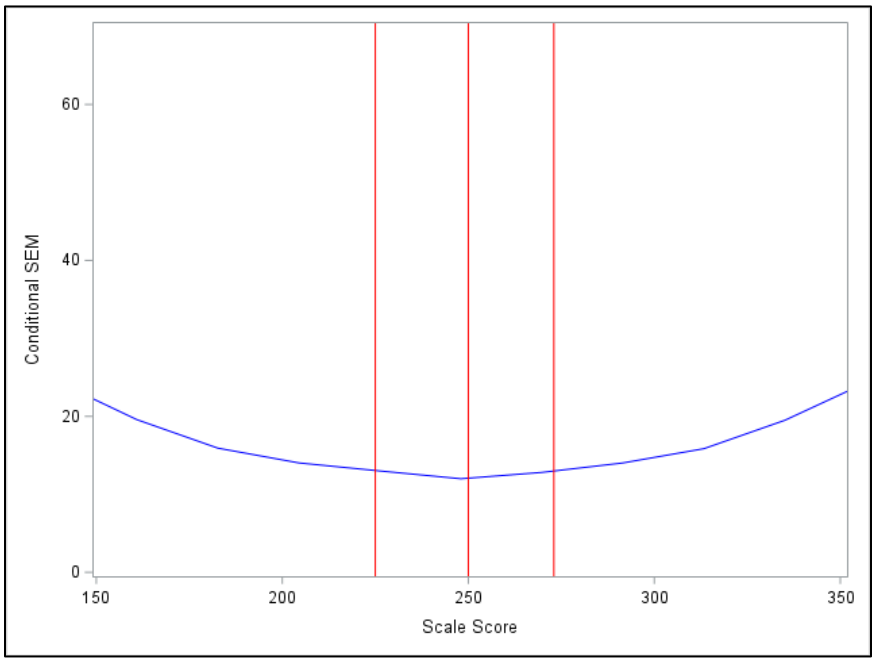


Figure I.5. CSEM Curve—Grade 8

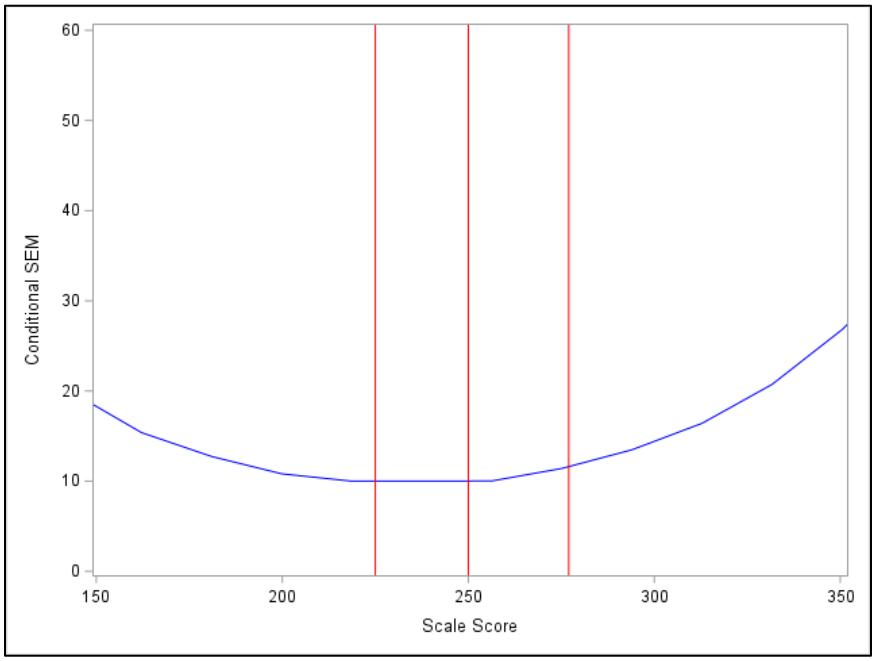
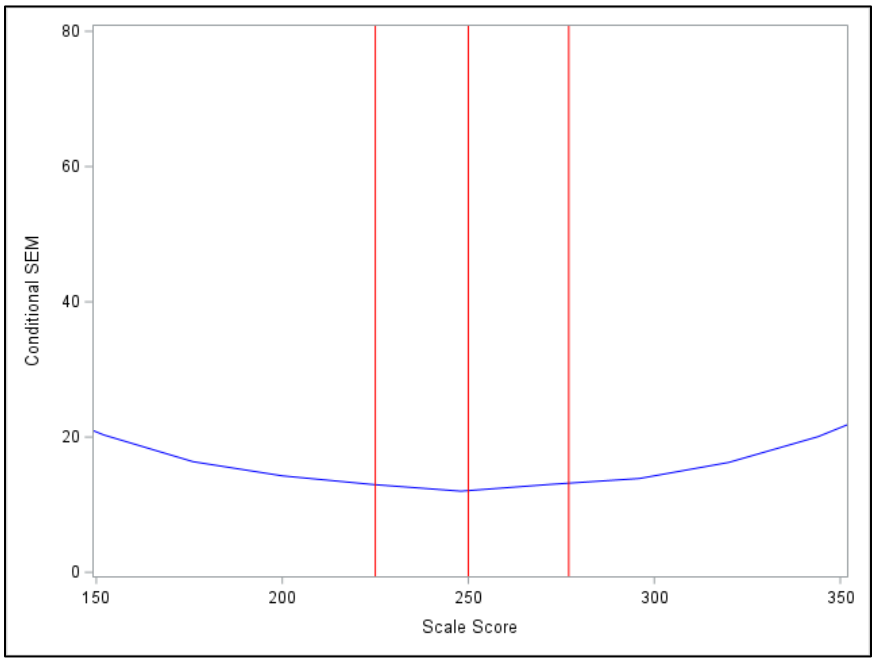


Figure I.6. CSEM Curve—Grade 11



Appendix J: Test Administrator Survey Responses

How familiar are you with this student?

Grade	Very Familiar	Somewhat Familiar	Familiar	Somewhat Unfamiliar	Unfamiliar	Missing
5	86.01%	6.99%	3.63%	0.78%	0.52%	2.07%
8	87.10%	3.87%	4.09%	1.51%	0.22%	3.23%
11	83.50%	6.25%	3.25%	1.50%	0.25%	5.25%

How many hours per week does this student spend in instruction on this content area?

Grade	<1 Hour	1 to <2 Hours	2 to <3 Hours	3 to <4 Hours	4 to<5 Hours	≥5 Hours	Do Not Know	Missing
5	19.43%	33.42%	19.95%	14.51%	6.99%	3.63%	0.00%	2.07%
8	11.40%	12.90%	11.83%	19.78%	34.41%	6.24%	0.00%	3.44%
11	14.50%	15.75%	16.75%	23.50%	17.00%	6.25%	0.00%	6.25%

Approximately how much instructional time for this content area is in the general education classroom?

Grade	25%	50%	75%	100%	None	Missing
5	22.54%	6.48%	16.06%	25.65%	27.72%	1.55%
8	12.26%	8.39%	8.82%	36.13%	31.18%	3.23%
11	11.25%	4.00%	6.75%	17.75%	52.25%	8.00%

This student's primary receptive communication is:

Grade	Oral Language	Sign Language	Reading	Picture Communication	Tactile	Other	Do Not Know	Missing
5	89.90%	1.30%	0.00%	3.89%	0.00%	0.26%	0.00%	4.66%
8	87.96%	0.86%	1.51%	4.09%	0.22%	0.22%	0.00%	5.16%
11	87.25%	0.50%	1.00%	1.75%	0.00%	1.25%	0.00%	8.25%

This student's primary expressive communication is:

Grade	Oral Language	Sign Language	Writing	Picture Communication	Augmentative Communication Device	Tactile	Other	Do Not Know	Missing
5	75.13%	2.85%	0.26%	4.66%	10.10%	0.00%	2.85%	0.26%	3.89%
8	79.78%	1.94%	0.00%	4.09%	7.31%	0.22%	1.51%	0.00%	5.16%
11	76.50%	0.75%	0.75%	3.25%	8.50%	0.00%	2.25%	0.00%	8.00%

I feel that the student's responses accurately reflect their understanding of the material.

Grade	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Do Not Know	Missing
5	29.53%	38.86%	16.84%	6.74%	3.89%	1.30%	2.85%
8	34.19%	42.15%	11.18%	5.59%	1.29%	1.08%	4.52%
11	33.75%	37.75%	13.50%	4.25%	3.00%	0.50%	7.25%

How much time did this student take on the assessment?

Grade	0–30 Minutes	31–60 Minutes	61–90 Minutes	91–120 Minutes	121–150 Minutes	151–180 Minutes	≥181 Minutes	Missing
5	0.52%	30.57%	36.27%	16.84%	4.40%	3.11%	3.10%	5.18%
8	1.29%	28.82%	40.00%	13.55%	3.23%	2.58%	3.01%	7.53%
11	1.25%	28.25%	40.50%	13.50%	4.50%	3.00%	1.25%	7.75%