

Colorado Alternate (CoAlt) Science and Social Studies Technical Report 2024–2025

Prepared by Pearson for the Colorado Department of Education (CDE)

Foreword

This technical report documents the evidence of reliability and validity to support test users in evaluating the intended purposes, uses, and interpretations of the test scores for the administration of the Colorado Alternate (CoAlt) Science and Social Studies assessments. The evidence includes descriptions of the test design, development, and administration procedures; the student test results; and psychometric analyses including calibration, equating, and scaling to ensure that the test results can be compared across different test forms and administrations. The report adheres to industry best practices and follows the guidelines of the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

The Colorado Department of Education's vision is to create an equitable educational environment where all students and staff in Colorado thrive. Their role is to improve student outcomes and ensure that students and families across Colorado have access to high-quality schools by providing actionable support to local educational agencies, implementing policy and legislation in an effective way, and sharing the experiences of local educational agencies and students.

Colorado Department of Education
201 East Colfax Ave., Denver, CO 80203
303-866-6600
https://www.cde.state.co.us/



Copyright © 2025 by the Colorado Department of Education (CDE). All rights reserved. This report was created by Pearson under contract with CDE. The content and format of this report is determined by CDE.

Table of Contents

Chapter 1: Introduction 7
1.1. Assessment Overview 7
1.2. Background 8
1.3. Purpose of CoAlt 9
1.4. Testing Requirements 9
1.5. Assessment Development Partners 10
Chapter 2: Test Design 12
2.1. Alternate Academic Achievement Standards 12
2.2. Item Types 13
2.3. Test Frameworks and Blueprints 14
2.3.1. Science 14
2.3.2. Social Studies 15
2.4. Cognitive Complexity 16
2.5. Test Composition 16
Chapter 3: Test Development 17
3.1. Content Management Tool 18
3.2. Item Writing 18
3.3. Item Review 18
3.3.1. Internal Review 18
3.3.2. External Content and Bias Review 19
3.4. Form Construction 19
3.4.1. Operational Forms 21
3.4.2. Field Test Forms 21
3.5. Data Review 21
Chapter 4: Test Administration 23
4.1. Manuals 23
4.2. Test Materials 24
4.3. Administration Training 24
4.4. Practice Resources 25
4.5. Accessibility Features and Accommodations 25
4.6. Test Security 26
4.7. Test Monitoring 27
4.7.1. Training 27
4.7.2. Process 27
4.7.3. Participation 27
4.7.4. Results 28
Chapter 5: Scoring 30
5.1. SR Scoring 30
5.2. SPT Scoring 30
Chapter 6: Reporting 31
6.1. Description of Scores 31
6.1.1. Scale Scores 31
6.1.2. Performance Levels 31
6.1.3. Percent Earned 32

6.2. Score Reports 32
Chapter 7: Standard Setting 33
Chapter 8: Test Results and Analysis 34
8.1. Student Participation 34
8.2. Performance Results 34
8.3. Classical Item Analysis 35
8.4. Subclaim Correlations 36
Chapter 9: Calibration, Equating, and Scaling 38
9.1. IRT Model 38
9.2. Data Preparation 39
9.3. Calibration 39
9.4. Equating 39
9.4.1. Operational Equating 39
9.4.2. Field Test Equating 40
9.5. Item-Level IRT Statistics 40
9.6. Scaling 40
Chapter 10: Reliability 42
10.1. Internal Consistency (Coefficient Alpha) 42
10.2. Standard Error of Measurement (SEM) 43
10.3. Conditional Standard Error of Measurement (CSEM) 44
10.4. Decision Consistency and Accuracy 44
Chapter 11: Validity 46
11.1. Evidence Based on Test Content 46
11.2. Evidence Based on Response Processes 47
11.3. Evidence Based on Internal Structure 47
11.4. Evidence Based on Relations to Other Variables 47
11.5. Evidence for Validity and Consequences of Testing 48
11.6. Fairness 49
References 50
Appendix A: CoAlt Eligibility Guidelines 52
Appendix B: Sample Student Performance Report 54
Appendix C: Scale Score Distributions 56
Appendix D: Scale Score Distribution Histograms 59
Appendix E: Performance Results by Demographic Subgroup 61
Appendix F: Classical Item-Level Statistics 64
Appendix G: IRT Item-Level Statistics 67
Appendix H: Test Characteristic Curves (TCCs) 70
Appendix I: Test Information Curves (TICs) and CSEM Curves 73
Appendix J: Test Administrator Survey Responses 79

List of Tables

Table 1.1. Student Participation Requirements 9
Table 1.2. Assessment Development Partners 10
Table 3.1. Item Development Activities for the 2024–2025 Cycle 18
Table 3.2. Number of Test Forms Constructed in Spring 2025 20
Table 3.3. Item Statistical Flagging Criteria 21
Table 3.4. Data Review Results 22
Table 4.1. Test Administration Activities 23
Table 4.2. Test Materials 24
Table 4.3. Number of Participating Schools in Test Monitoring by Geographic Region 28
Table 4.4. Number of Participating Students (Observations) in Test Monitoring 28
Table 4.5. Test Monitoring Percent Agreement Rates between Transcribers 28
Table 5.1. SPT Scoring Rubric 30
Table 7.1. Performance Levels and Policy Claims 31
Table 6.1. Performance Level Cut Scores 33
Table 8.1. Student Participation N-Count Demographic Distribution 34
Table 8.2. Scale Score Performance Summary and Performance Level Distributions 35
Table 8.4. Summary Statistics for Points Earned by Subclaim 35
Table 8.5. Summary of *P*-Values and Item-Total Correlations 36
Table 8.6. Correlations Between Subclaims—Science 36
Table 10.1. Coefficient Alpha—Science 43
Table 10.2. Coefficient Alpha—Social Studies 43
Table 10.3. SEM—Science 43
Table 10.4. SEM—Social Studies 44
Table 10.3. Kappa Values 45
Table 10.4. Decision Accuracy and Consistency Estimates 45
Table 10.5. Decision Accuracy and Consistency of Cut Scores 45
Table 11.1. Correlation Between Test Validity Questions and Student Scores 48
Table 11.2. Correlation Between CMAS Science and DLM ELA and Mathematics 48
Table 11.3. Student Performance Over Time 49
Table C.1. Scale Score Distribution—Science Grade 5 56
Table C.2. Scale Score Distribution—Science Grade 8 57
Table C.3. Scale Score Distribution—Science Grade 11 57
Table E.1. Scale Score Summary Statistics by Demographic Subgroup—Science Grade 5 61
Table E.2. Scale Score Summary Statistics by Demographic Subgroup—Science Grade 8 61
Table E.3. Scale Score Summary Statistics by Demographic Subgroup—Science Grade 11 62
Table F.1. SR Item Classical Statistics—Science Grade 5 64
Table F.2. SPT Item Classical Statistics—Science Grade 5 64
Table F.3. SR Item Classical Statistics—Science Grade 8 65
Table F.4. SPT Item Classical Statistics—Science Grade 8 65
Table F.5. SR Item Classical Statistics—Science Grade 11 66
Table F.6. SPT Item Classical Statistics—Science Grade 11 66
Table G.1. Operational Item Parameter Estimates—Science Grade 5 67
Table G.2. Operational Item Parameter Estimates—Science Grade 8 68
Table G.3. Operational Item Parameter Estimates—Science Grade 11 69

List of Figures

Figure D.1. Scale Score Distribution Histogram—Science Grade 5 59
Figure D.2. Scale Score Distribution Histogram—Science Grade 8 59
Figure D.3. Scale Score Distribution Histogram—Science Grade 11 60
Figure H.1. TCC—Science Grade 5 70
Figure H.2. TCC—Science Grade 8 71
Figure H.3. TCC—Science Grade 11 72
Figure I.1. TIC—Science Grade 5 73
Figure I.2. TIC—Science Grade 8 74
Figure I.3. TIC—Science Grade 11 75

Chapter 1: Introduction

This technical report documents the evidence of reliability and validity for the spring 2025 administration of the Colorado Alternate (CoAlt) assessments in science and social studies to support test users in evaluating the intended purposes, uses, and interpretations of the test scores. The evidence includes descriptions of the test design, development, and administration procedures; the student test results; and psychometric analyses including calibration, equating, and scaling to ensure that the results are comparable across different test forms and administrations. The report ends with a synthesis of the evidence from throughout the document to construct the validity argument for the assessments.

1.1. Assessment Overview

The CoAlt assessments are Colorado's end-of-year alternate summative assessments administered annually each spring in grades 5, 8, and 11 in science and grades 4 and 7 in social studies (on a sampling basis) 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, response methods, attention span, and short-term memory. The alternate assessments are administered as an alternative to the Colorado Measures of Academic Success (CMAS) general assessments to ensure participation of all students in Colorado's state assessment program, designed to measure student achievement of Colorado's alternate academic achievement standards known as the Extended Evidence Outcomes (EEOs) of the Colorado Academic Standards (CAS).[1]

The CoAlt assessments are administered by the test administrator using paper-based test books in a one-on-one setting 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.

The CoAlt Science assessments include 1-point selected response (SR) and 3-point supported performance task (SPT) item types, while the CoAlt Social Studies assessments include 1-point SR items only. The test administrator marks a student's responses to the 1-point SR items and indicates their assigned scores for the 3-point SPT items in a scannable answer document that is then returned to Pearson. Student results are reported as an overall scale score and accompanying performance level, with the percentage of points earned provided for each reporting category in science; subscores are not reported for social studies. The assessments have four performance levels: *Emerging, Approaching Target, At Target,* and *Advanced*. Students in the top two performance levels are considered ready for continuing study in the content area. (Results for social studies in 2025 were reported as percentile ranks only, with overall scale scores and performance levels reported beginning in 2026.)

[1] The CoAlt English Language Arts (ELA) and Mathematics assessments are administered by the Dynamic Learning Maps (DLM) consortium and documented in a separate technical report located online at https://dynamiclearningmaps.org/publications.

1.2. 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 the Every Student Succeeds Act of 2015 (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 and social studies were originally adopted in December 2009. On August 3, 2011, the State Board of Education adopted the EEOs of the CAS for students who qualify for an alternate assessment. In partnership with Colorado educators and Pearson, CDE developed the CoAlt Science and Social Studies assessments to evaluate student mastery of the EEOs for students with the most significant cognitive disabilities.

The first operational administration of the CoAlt assessments was in spring 2014 for grades 5 and 8 in science and grades 4 and 7 in social studies and fall 2014 for high school based on the 2009 CAS. The CoAlt Social Studies high school assessment was deferred in 2015 and eliminated in 2020. The grades 4 and 7 assessments were also eliminated in 2020 per legislation. Thus, the social studies assessments have not been administered since 2014 for high school and 2019 for grades 4 and 7. The spring 2020 CoAlt Science administration was cancelled due to the COVID-19 pandemic. In 2021, Colorado received a partial waiver of the federal assessment requirements due to COVID-19 conditions in Colorado. Except for students with a parent/guardian excusal, only students in grades 8 and 11 took the CoAlt Science assessment.

In 2022, newly revised CAS were implemented for science. Senate Bill 212 (also known as CAP4K), implemented by Colorado in 2008, 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. 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 Science assessment is aligned.

Full implementation of the new three-dimensional science standards took place in the 2021–2022 school year. The new CoAlt Science assessment was administered to all tested students for the first time in 2022, 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. The 2022 science assessment reported percentile ranks only. Standard setting was completed in October 2022, and overall scale scores and performance levels were reported beginning in 2023.

New social studies standards were adopted in 2022, with full implementation in the classroom expected by fall 2024. The first administration of the new CoAlt Social Studies assessments in grades 4 and 7 occurred in spring 2025. High school was not included in this return. The new EEOs for CoAlt were significantly different from the previous versions and required all new item development. As a result, all the CoAlt items in the initial year of test construction were new, rendering spring 2025 an operational field test administration. Standard setting was conducted in October 2025, and the new cut scores will be applied starting in spring 2026. The 2025 results were reported as percentile ranks only.

1.3. Purpose of CoAlt

The goals of the Colorado Assessment System, including CoAlt, 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. The primary purpose of the CoAlt assessment 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 classroom instruction;
- inform district and school leaders about potential programming and instruction priorities;
- provide 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 academic achievement standards;
- provide aggregated data for the state's accountability system; and
- allow students to demonstrate their mastery of skills and concepts in the EEOs.

1.4. Testing Requirements

Table 1.1 summarizes the student participation requirements for the CoAlt Science and Social Studies assessments. Eligible students take CoAlt in place of CMAS Science and Social Studies assessments. All public schools in Colorado are required by state law to administer the standards-based summative assessment each year in the specified grade levels in science to comply with the federal accountability requirements as stated in ESSA. ESSA also specifies 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, and 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 social studies assessments are not included in the federal accountability system but are administered to measure students' knowledge and skills of the CAS for Social Studies and help support social studies instruction in Colorado. Beginning in 2016 based on Colorado Revised Statute 22-7-1006.3 (1)(a)(III), the social studies assessments are administered on a sampling basis, with schools participating once every three years. Social studies assessments are not planned to be included in the 2025 district/school performance frameworks or beyond.

Table 1.1. Student Participation Requirements

Assessment	Student Participation
CoAlt Science	Students in grades 5, 8, and 11 take the grade-level science assessment for their enrolled grade.
CoAlt Social Studies	Students in grades 4 and 7 take the grade-level social studies assessment for their enrolled grade at selected schools only.

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 most significant cognitive disability, although Intellectual Disability does not have to be the student's primary disability label for IDEA eligibility. 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.

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. However, every student, regardless of ability or language background, must be provided with the opportunity to demonstrate their content knowledge through the state assessments.

1.5. Assessment Development Partners

CoAlt assessment activities were conducted collaboratively by CDE, the Colorado educator community, and Pearson, with input and advice from the Colorado Technical Advisory Committee (TAC), as shown in Table 1.2. Each contributor plays a vital role in ensuring that the assessments yield valid and reliable test results. Educator participation in the test development process is critical to ensuring that the assessments are aligned to the EEOs, are appropriate for Colorado students with the most significant cognitive disabilities at the assessed grade level, and are free from bias and sensitivity issues. Recommendations from the TAC have been reviewed, addressed, and incorporated into the assessments.

Table 1.2. Assessment Development Partners

Organization/Group	Roles and Responsibilities
Colorado Department of Education (CDE)	• Implements state and federal education laws as the administrative arm of the State Board of Education • Works closely with Colorado school districts, educators, community stakeholders, and other assessment development partners to develop and administer the state assessments, focusing on creating assessments that serve students, schools, districts, and the community while complying with state and federal legal requirements • Works with Pearson on each facet of the assessment, with CDE serving as the ultimate approver of the services and products provided
Colorado Educator Community	• Review items to ensure content alignment and identify potential bias and sensitivity concerns before items are field tested

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

Organization/Group	Roles and Responsibilities
	• Participate in data review to review field tested items with statistical parameters outside of normal ranges to determine if the items are acceptable for inclusion in the operational item bank
Pearson	• Primary contractor responsible for the end-to-end assessment cycle services and products • Works closely with CDE throughout the CMAS and CoAlt Science assessment development and administration processes, including item and test development, forms creation, enrollment, packaging and distribution, test delivery, scoring, customer service, standard setting, scoring, score reporting, and psychometric services
Technical Advisory Committee (TAC)	• A group of psychometric, assessment design, and special populations experts who provide consulting and expert advice for the assessment • Provides input on the blueprint design, scaling and equating, scoring, reporting, alignment studies, peer review, standard setting and validity and reliability issues

Chapter 2: Test Design

This section describes the assessment framework from which the CoAlt Science and Social Studies assessments are developed, including the content standards that provide the basis for test development, the item types, the test blueprints that guide the construction of the test forms, and the overall structure of each assessment. This information is important for providing transparency into the intended construct and rationale behind the assessments to support the valid interpretations of the test scores. More details about the CoAlt assessment test design are available online at https://www.cde.state.co.us/assessment/coaltsss.

2.1. Alternate Academic Achievement Standards

Standards are broad goals articulating what students should know, understand, and be able to do over a given timeframe and are the foundation of teaching and learning. The EEOs are alternate academic achievement standards for students with the most significant cognitive disabilities, aligned to the grade-level 2020 CAS in science and 2022 CAS in social studies but reduced in depth, breadth, and complexity, available online at http://www.cde.state.co.us/CoExtendedEO/StateStandards. The EEOs were developed for a very small and specific group of students in all grades who are identified with a significant cognitive disability and meet specific eligibility criteria. They provide clear, rigorous learning outcomes to all students can learn the skills, knowledge, and confidence to succeed in postsecondary environments and the workforce, to be well-informed and responsible citizens, and to lead fulfilling personal lives.

The 2020 CAS for Science underwent significant changes based on the NGSS that were guided by *A Framework for K–12 Science Education* (National Research Council, 2012) and designed to reflect more recent research and thinking in science education based on the three dimensions of Disciplinary Core Ideas (DCIs), Science and Engineering Practices (SEPs), and Crosscutting Concepts (CCCs). The CAS for Social Studies were updated for 2018 although never fully implemented. The standards were further revised and adopted in 2022, with seven new evidence outcomes being added, one in grade 4 and six in grade 7. The changes to the general 2022 CAS for social studies were minimal enough that links to the previous standards and item bank could still be made.

Once the new standards were established, they were then modified to the EEOs for CoAlt by a revision team of general and special educators across grades levels, selected through an application process. Significant revisions to the EEOs occurred by pairing the first generation EEO with the revised Evidence Outcomes (EOs) in the CAS on a one-to-one basis, increasing the rigor for students with the most significant cognitive disabilities. Either the EEOs were retained and/or revised, or the team developed new ones for each Evidence Outcome in the CAS. The revised EEOs were then reviewed by content and special education experts before being approved by the State Board of Education. Learning progressions were also developed that break down the EEO into the critical component skills, providing a clearer picture of how students can access the content.

More guidance on the EEOs and learning progressions are provided in webinars by CDE, available at http://www.cde.state.co.us/CoExtendedEO/StateStandards under "Webinar Training Series For The 2020 Extended Evidence Outcomes."

2.2. Item Types

The CoAlt Science assessment includes 1-point selected response (SR) and 3-point supported performance task (SPT) item types, while the CoAlt Social Studies assessment includes 1-point SR items only.[3] The test administrator records student responses to the SR items (A, B, C, D, or NR when there is no response from the student) and their scores on the science SPT items (0, 1, or NR) on a scannable answer document included with the task manipulatives set provided for each test, which is then returned to Pearson for scoring.[4]

The social studies items are all standalone items with three answer options. The science 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 science SR 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.

The science 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 either "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. In these tasks, four answer cards are provided, but one is not used.

[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/.

2.3. Test Frameworks and Blueprints

The CoAlt frameworks were developed to better identify the content standards that may be assessed on the CoAlt assessments. The frameworks define the elements of the EEOs suitable for state testing and are available at https://www.cde.state.co.us/assessment/coalt_sss_pld_td. 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 and Social Studies test blueprints based on the CMAS blueprints. While the complete blueprints are used internally, this section presents the high-level CoAlt Science and Social Studies blueprints that summarize the number of items and percentage of score points on each test.

2.3.1. Science

Table 2.1 presents the high-level CoAlt Science test design, and Table 2.2 presents the blueprints for each grade. The 2022 CoAlt Science timing results showed that most students were completing the test in a reasonable amount of time. However, to address timing concerns from educators before the first operational administration, item reductions were made to the 2023 grade 5 blueprint. After the 2023 administration, CDE received more feedback from CoAlt educators expressing concern about the administrative load of the high school assessment. As a result, CDE implemented a reduced blueprint in grades 8 and 11 in 2024 to reduce administrative and testing burden for this population. The grade 8 blueprint was reduced given that it mirrors the high school blueprint.

This blueprint reduction is also in line with those made for CMAS Science in grade 11. Like CMAS Science, the 2024 CoAlt Science grades 8 and 11 blueprints were a proportional reduction of the 2023 blueprints. Teachers were still able to divide the test administration over as many sessions/days as appropriate for the student. In addition to the reduced blueprints for grades 8 and 11, all CoAlt Science assessments had fewer embedded field test items to allow teachers' concerns about administrative and testing burden to be addressed across all CoAlt Science assessments. The 2025 blueprint did not change from 2024.

Table 2.1. CoAlt Science Grades 5, 8 and 11 Test Design

	Test Design
Sampling	No sampling. All CoAlt Science students are eligible
Test Structure	1 section with embedded FT items
Mode	1:1 paper assessment
Item Types	1 pt. SR items and 3 pt. SPT items
Test Length	36 items, 42 pts. for grade 5 and 33 items, 39 pts. for grades 8 and 11
IRT Models	RPCM
Scoring	Raw-to-scale score (RSS) tables
Reporting Scale (test level)	Scale score range: 150–350; scale was set using two cut score points
Reporting Scale (subscores)	Subscores for Subclaims and SEP reported as percent of points earned
Performance Levels	Four levels: Emerging, Approaching Target, At Target, and Advanced

Table 2.2. 2025 CoAlt Science Test Blueprints

Grade	Subclaim	#Item Sets	Total #Items	#1-Point SR Items	#3-Point SPT Items	Total #Points	% of Total Points
5	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%
8	Physical Science	4	13	12	1	15	38%
	Life Science	3	10	9	1	12	31%
	Earth and Space Science	3	10	9	1	12	31%
	Total	10	33	30	3	39	100%
11	Physical Science	4	13 (14)	12 (13)	1	15–16	38–41%
	Life Science	3	10	9	1	12	31%
	Earth and Space Science	3	10 (9)	9 (8)	1	11–12	28–31%
	Total	10	33	30	3	39	100%

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

2.3.2. Social Studies

The revised 2022 standards prompted an update to the CoAlt Social Studies assessments. Table 2.3 presents the resulting test design, and Table 2.4 presents the high-level CoAlt Social Studies blueprints for each grade. The five standard areas of the social studies EEOs of the CAS are History, Geography, Economics, Civics, Personal Financial Literacy. The assessments are designed to allocate a certain number of points per subclaim (reporting category). In 2024-2025 these subscores were not provided, however they will be reported as percent of points earned starting in 2025-2026 when full score reports are provided. Economics and Personal Financial Literacy will be combined to meet the minimum necessary points for reporting.

Table 2.3. CoAlt Social Studies Grades 4 and 7 Test Design

	Test Design
Sampling	One-third of schools sampled
Test Structure	1 section with embedded FT items
Mode	1:1 paper assessment
Item Types	1 pt. SR items
Test Length	36 items, 36 pts.
IRT Models	Rasch
Scoring	Raw-to-scale score (RSS) tables
Reporting Scale (test level)	Scale score range: 150–350; scale was set using two cut score points (beginning in 2026)
Reporting Scale (subscores)	Subscores for subclaims are reported as percent of points earned (beginning in 2026)
Performance Levels	Three levels (beginning in 2026)

Table 2.4. 2025 CoAlt Social Studies Test Blueprints

Grade	Subclaim	Total #1-Point SR Items	Total #Points	% of Total Points
4	History	9	9	25%
	Geography	10	10	28%

Grade	Subclaim	Total #1-Point SR Items	Total #Points	% of Total Points
	Economics	4	4	11%
	Civics	9	9	25%
	Personal Finance Literacy	4	4	11%
	Total	36	36	100%
7	History	10	10	28%
	Geography	9	9	25%
	Economics	4	4	11%
	Civics	9	9	25%
	Personal Finance Literacy	4	4	11%
	Total	36	36	100%

Note. SR = selected-response,

2.4. Cognitive Complexity

All CoAlt 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. Level 4: Extended Thinking items are not included because the tests do not contain any extended-response items.

2.5. Test Composition

Table 25 presents the number of items on each test form, including the number of operational vs. embedded field test items and the total number of possible score points. The spring 2025 CoAlt Science 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. The spring 2025 CoAlt Social Studies assessment was considered an operational field test given that all items were newly developed and field tested in 2025.

Table 2.5. 2025 Test Composition

Assessment	#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
Science 5	2	50	33	3	12	2	42
Science 8	2	51	39	3	18	0	39
Science 11	2	51	39	3	18	0	39
Social Studies 4	1	42	36	N/A	6	N/A	42
Social Studies 7	1	42	36	N/A	6	N/A	42

Note. OP = operational, FT = field test, SR = selected-response, SPT = supported performance task, N/A = not applicable

Chapter 3: Test Development

This section describes the test development process for the spring 2025 CoAlt Science and Social Studies assessments. The operational test forms for both content areas in all grades were newly developed by Pearson. The new EEOs of the 2022 CAS for social studies were significantly different from the previous versions and required all new item development. As a result, 100% of the CoAlt items in the initial year of test construction for the spring 2025 operational field test were new. In future administrations, the same operational forms constructed for 2025 will be used across all three years for CoAlt Social Studies give the small population size, but with different embedded field test items each year.

Table 3.1 presents the item development activities during the 2023–2024 development cycle for items eligible for field testing beginning in spring 2025. The CoAlt Science and Social Studies item development follows the same process as the CMAS assessments to the extent possible, with modifications to reflect the unique characteristics of the alternate assessment program such as the item types and needs of the student population. CDE relies on input from both general and special education Colorado educators and alternate assessment specialists to ensure that the CoAlt Science assessment is equitable for students and accurately measures 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 plan used to forecast the targeted number of items needed to create a robust item bank that is refreshed over time. Items are then written by Colorado educators, followed by multiple rounds of review, including contractor, CDE, and Colorado educator content, bias, and data reviews. Each content area and grade-level assessment has its own item pool, or item bank, that houses all the stimuli and items and supports the development of the assessments.

Prior to the annual item development cycle, Pearson and CDE evaluate the strengths of the item pools and consider the needs for future tests to establish an item development plan (also referred to as an item bank gap analysis) 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 item development plans 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.

New items are developed based on this plan, followed by internal and external reviews. Accepted items are added to the item pool as field test items. Test forms are then constructed for both the operational and field test items. Following the test administration, the field tested items are reviewed during data review where they are either accepted into the operational item pool, accepted with edits, or rejected from any further use.

Table 3.1. Item Development Activities for the 2024–2025 Cycle

Event	Date(s)
Item Writer Workshop (Science)	Dec. 10–12, 2024
Item Writer Workshop (Social Studies)	Feb. 18–19, 2025
Content and Bias Review (Science)	June 17–18, 2025
Content and Bias Review (Social Studies)	June 25, 2025
Data Review (Science)	Oct. 7-8, 2025
Data Review (Social Studies)	July 22, 2025

3.1. Content Management Tool

The test development process is managed within Pearson's Assessment Banking and Building solutions for Interoperable assessments tool (ABBI) that acts as a content development and management tool, item bank, and publication system supporting both online and paper publication. The item development workflow moves items and assets from inception through a series of reviews to final publication. The system captures the outcomes at each review and maintains previous versions of each item. As items travel through the review process, every version of each asset is archived, along with each comment received in any review. Reviewers have immediate access to all older versions, providing version control throughout development. Forms are also built in ABBI. After items are used, ABBI stores the resulting statistics, including exposure statistics and classical and item response theory (IRT) statistics.

ABBI allows remote internet access by item writers and reviewers while ensuring security with individualized passwords, limited access for external users, and strong encryption of all information. ABBI manages item content throughout the entire lifecycle of an item, as well as item content beyond the operational life of the item, including items identified for use in sample tests or other training materials. ABBI provides on-demand reports of the content and item bank status. Each item is directed through a sequence of reviews and approvals by Pearson and CDE before it is identified for operational or field test use.

3.2. Item Writing

After the test blueprints and item development plans were developed, the teacher item writing process began. SR and SPT items 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.

3.3. Item Review

3.3.1. Internal Review

Once the 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.

3.3.2. 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.

3.4. Form Construction

Table 3.2 presents the number of test forms constructed for the spring 2025 administration. Each CoAlt test form is administered as a fixed-form assessment, meaning all students receive the same set of operational items in a predetermined order but different embedded field test items depending on which test form they receive. A core form refers to the set of operational items that each student receives, while the field test forms refer to the various versions of a core form with different sets of field test items. Individual test scores are based on the operational items (i.e., the core form) only. The purpose of field testing is to administer newly developed items to generate item statistics and assess their eligibility to become operational. The field test items do not count toward a student's score.

Table 3.2. Number of Test Forms Constructed in Spring 2025

Assessment	#Core Forms	#Field Test Forms
Science 5	1	2
Science 8	1	2
Science 11	1	2
Social Studies 4	1	1
Social Studies 7	1	1

Test form construction is the process of selecting and sequencing the operational and field test items for each test form, which is a complex, interactive task that requires both content and psychometric expertise. Once the test forms were constructed by Pearson, CDE reviewed the forms, provided feedback, and gave final approval. The following guidelines were used during the spring 2025 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

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

3.4.1. Operational Forms

After the initial operational items are selected, the assessment specialist verifies that the test form meets the blueprint and specifications (i.e., the required number of items, EEO coverage, and item types). The psychometrician then verifies that the form falls within the psychometric and blueprint parameters and identifies the anchor item set within each operational 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.

3.4.2. Field Test Forms

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.

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 3.3 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 3.3, 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 3.3. 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 alone was not the sole reason an item was 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. Accepted items were moved into "Ready for Operational" status, whereas 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 3.4 presents the results of the data review based on spring 2025 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 3.4. Data Review Results

Assessment	#Accepted	#Accepted for Revision and Re-field Test	#Rejected
Science 5	5	0	0
Science 8	3	0	0
Science 11	4	0	0
Social Studies 4	22	N/A	4
Social Studies 7	14	N/A	0

Note. N/A = not applicable

Chapter 4: Test Administration

The CoAlt Science and Social Studies assessments are paper-based assessments administered one-on-one by a test administrator who records student responses on a scannable answer document. The 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. Table 4.1 presents the test administration window, including the release of the CoAlt score reports. (See Chapter 7 for information on reporting.)

Table 4.1. Test Administration Activities

Event	Date(s)
DAC Administration Training	Beginning Nov. 11, 2024
Spring 2025 Administration Window	April 7–25, 2025
Reports Released	June 30, 2025

The District Assessment Coordinator (DAC) is responsible for establishing the administration schedule and ensuring that every student taking a CoAlt Science and Social Studies 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 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.

4.1. Manuals

The following manuals were sent to schools and DACs to support the CoAlt Science and Social Studies administrations:

- The *CoAlt Science and Social Studies Test Administrator Manual* provides instructions for administering and scoring the CoAlt Science and Social Studies assessments and 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.

4.2. Test Materials

Table 4.2 presents the paper-based test materials used by the test administrator during the administration of the CoAlt 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. The Science 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 4.2. Test Materials

Test Material	Description
CoAlt *Test Administrator Manual*	Provides information necessary for the administration and scoring of the CoAlt Science and Social Studies assessments for use by the test administrator. The manual contains the SPT Score Flow Chart for scoring the science 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.)

4.3. Administration Training

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. Administration training is intended to make sure all individuals involved in the CoAlt Science and Social Studies 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. CoAlt assessment administration and training procedures are standardized to ensure that students receive comparable test results while allowing flexibility to accommodate the unique needs of students in this population. Fidelity to standardized test administration processes and procedures helps to ensure the comparability of resulting scores and accurate interpretation of results.

Test administration procedures were communicated to the appropriate individuals via manuals and virtual and recorded trainings. 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 https://ed.cde.state.co.us/assessment/cmas/training-cmas. 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.

4.4. Practice Resources

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

4.5. Accessibility Features and Accommodations

The CoAlt Science and Social Studies 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 assessments and administration procedures. For example, the CoAlt assessments are read aloud to students, and all students 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 and Social Studies 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

4.6. 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 and 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.

4.7. Test Monitoring

During the spring 2025 administration, seven 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 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.

4.7.1. 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 *Procedures Manual*, the *Test Administrator Manual*, a test monitor checklist, and an answer document.

4.7.2. 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.

4.7.3. Participation

Pearson and CDE worked together to recruit schools to participate in test monitoring during the spring 2025 administration. 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 4.3, the participating school districts represented four of the eight geographic regions of the state. As shown in Table 4.4 that presents the number of observations conducted (counted once for each student) compared to the total student population, six students were observed for grade 5, one student was observed for grade 7, two students were observed for grade 8, and one student was observed for grade 11.

Table 4.3. Number of Participating Schools in Test Monitoring by Geographic Region

Assessment	Metro	North Central	Northeast	Northwest	Pikes Peak	Southeast	Southwest	West Central
Science 5	1	2	-	-	-	-	2	-
Science 8	2	-		1	-	-	-	-
Science 11	-	-	-	1	-	-	-	-
Social Studies 4	-	-	-	-	-	-	-	-
Social Studies 7	-	1	-	-	-	-	-	-

Table 4.4. Number of Participating Students (Observations) in Test Monitoring

Assessment	Population N	Male	Female	Sample N	Male	Female
Science 5	430	62%	38%	5	60%	40%
Science 8	396	59%	41%	3	33%	67%
Science 11	341	63%	37%	1	0%	100%
Social Studies 4	153	69%	31%	0	0%	0%
Social Studies 7	128	59%	41%	1	100%	0%

4.7.4. Results

In general, test monitors indicated that the testing environment was appropriate. Test administrators appeared comfortable with the students, were well prepared to administer the test, provided the necessary accommodations, and administered the assessment at an appropriate pace. The test monitors also noted that the testing rooms had adequate space, were free of visible materials that could aid in answering test items, and there were no significant interruptions or distractions in the testing rooms. There was one instance, however, in which CoAlt materials were not prepared ahead of testing in the North Central region. The test monitors also reported relatively limited opportunities to observe CoAlt sessions, with Cherry Creek noting an increase in student opt-outs from CoAlt testing this year. Overall, test security policies were followed consistently across the monitored districts, and secure test materials were not left unattended at any point.

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. As shown in Table 4.5 that presents the resulting agreement results for spring 2025, the perfect agreement rates indicate high levels of agreement between the sets of transcribed student responses.

Table 4.5. Test Monitoring Percent Agreement Rates between Transcribers

Assessment	Perfect Agreement	Non-Perfect Agreement
Science 5	93%	7%
Science 8*	90%	10%
Science 11	100%	0%
Social Studies 4	-	-
Social Studies 7	93%	7%

**Note: one student was omitted from this analysis due to nonresponse during the monitoring session.*

Chapter 5: 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.

5.1. 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.

5.2. 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 5.1. 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 5.1. 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 6: Reporting

6.1. Description of Scores

The CoAlt Science reports provide information on student performance in terms of overall scale scores, performance levels, and percent earned scores. Results for social studies in 2025 were reported as percentile ranks only, with overall scale scores and performance levels reported beginning in 2026.

6.1.1. Scale 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.

6.1.2. Performance Levels

Performance levels are reported at the overall assessment level. Students are classified into performance levels based on their overall scale score and the cut scores obtained from standard setting. CoAlt Science has four performance levels: *Emerging*, *Approaching Target*, *At Target*, and *Advanced*.

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: (a) 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 6.1, and (b) grade-level PLDs that provide detailed descriptions of performance levels by grade level, available online at https://www.cde.state.co.us/assessment/coalt_sss_pld_td and included on the Individual Student Performance Report and in the *CMAS and CoAlt Interpretive Guide to Assessment Reports*.

Table 6.1. 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

6.1.3. Percent Earned

To prevent incorrect interpretations and provide a more generally understood metric, student performance on the CoAlt Science assessment 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.

6.2. Score Reports

Two types of score reports are provided to communicate student performance on the CoAlt Science and Social Studies assessments: (a) the student-level Student Performance Report and (b) 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://coassessments.com/reporting/.

Chapter 7: Standard Setting

To support the interpretation of the overall student results, student performance on the CoAlt Science assessments is described in terms of performance levels. (Results for social studies in 2025 were reported as percentile ranks only, with overall scale scores and performance levels reported beginning in 2026.) 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 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 are used to report student results on the CoAlt Science assessments.

Standard setting for the CoAlt Social Studies assessment aligned to the EEOs of the 2022 CAS took place in October 2025 using the modified ID Matching method, with the new cuts to be applied starting in 2026.

Table 7.1 presents the resulting scale score cut scores for each grade that are used to report student results on the CoAlt Science assessments.

Table 7.1. Performance Level Cut Scores

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

Chapter 8: Test Results and Analysis

8.1. Student Participation

Table 8.1 presents a breakdown of the number of students who took the spring 2025 CoAlt Science and Social Studies assessments by various demographic characteristics. All forms were administered in paper format. A total of 1,448 students across grades took the assessment in spring 2025.

Table 8.1. Student Participation N-Count Demographic Distribution

Subgroup	Science 5	Science 8	Science 11	Social Studies 4	Social Studies 7
Total	430	396	341	153	128
No IEP	*	*	*	*	*
IEP	429	396	339	153	128
No Accommodation	*	*	*	*	*
Accommodation	430	396	341	153	128
Am. Indian/Alaska Native	*	*	*	*	*
Asian	21	*	19	*	*
Black	33	35	28	*	*
Hispanic	179	191	132	60	63
White	168	140	135	62	53
Hawaiian/Pacific Islander	*	*	*	*	*
Two or More Races	24	*	16	*	*
Missing	*	*	*	*	*
No Economic Disadvantage	176	177	176	73	60
Economic Disadvantage	254	219	165	80	68
Female	165	161	126	47	53
Male	265	235	215	106	75
Language Proficiency NA	343	302	284	127	96
Language Proficiency NEP	54	48	26	20	21
Language Proficiency LEP	18	*	*	*	*
Language Proficiency FEP	*	36	27	*	*
Not Migrant	430	396	341	153	128
Migrant	*	*	*	*	*

*n-count less than 16

8.2. Performance Results

Table 8.2 presents the scale score performance summary and performance level distributions (i.e., the percentage of students classified into each performance level), andTable 8.3 Table 8.4 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. Social studies is not included in the scale score and performance level

tables given that results for social studies in 2025 were reported as percentile ranks only, with overall scale scores and performance levels reported beginning in 2026.

Table 8.2. Scale Score Performance Summary and Performance Level Distributions

Assessment	N	Mean	SD	Median	%Emerging	% Approaching Target	% At Target	%Advanced
Science 5	430	243	34.3	243	27%	31%	26%	17%
Science 8	396	236	27.4	236	31%	35%	29%	5%
Science 11	341	239	38.4	243	38%	18%	26%	18%

Table 8.3. Percentile Rank Performance Summary

Assessment	N	Mean	SD	Median
Social Studies 4	153	50	28.8	53
Social Studies 7	128	50	29.0	50

Table 8.34. Summary Statistics for Points Earned by Subclaim

Subclaim	Assessment	Mean	SD	Min.	Max.	Average % Correct
Physical Science	Science 5	7.3	3.4	0	16	45.98%
	Science 8	7.5	3.2	0	14	50.00%
	Science 11	7.4	3.4	0	15	49.30%
Life Science	Science 5	4.5	2.0	0	9	49.60%
	Science 8	6.2	2.3	0	11	51.71%
	Science 11	5.8	2.7	0	12	48.30%
Earth and Space Science	Science 5	7.4	3.2	0	16	43.52%
	Science 8	5.3	2.3	0	11	44.39%
	Science 11	5.0	2.4	0	12	41.69%

8.3. Classical Item Analysis

Table 8.5Table 8.4 presents the overall item difficulty and item discrimination results across all items for each assessment, and Appendix F presents the item-level classical statistics for the spring 2025 CoAlt Science and Social Studies 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.

Table 8.45. Summary of *P*-Values and Item-Total Correlations

Statistic	Assessment	#OP Items	Mean	SD	Min.	Max.	Median
P-value	Science 5	36	0.44	0.12	0.22	0.68	0.44
	Science 8	33	0.49	0.16	0.19	0.81	0.51
	Science 11	33	0.46	0.15	0.20	0.80	0.47
Item-Total Correlation	Science 5	36	0.38	0.11	0.16	0.58	0.37
	Science 8	33	0.38	0.12	0.11	0.61	0.37
	Science 11	33	0.42	0.13	0.16	0.64	0.45

Note. SD = standard deviation, Min. = minimum, Max. = maximum

8.4. Subclaim Correlations

The CoAlt Science assessments have three subclaim scores (Physical Science, Life Science, and Earth and Space Science), while social studies has five (History, Geography, Economics, Civics, and Personal Financial Literacy). One way to assess the internal structure of a test is through the evaluation of correlations among subclaim subscores, as presented in Table 8.5 **Error! Reference source not found.**. There is evidence of unidimensionality if the components within a content area are strongly related to each other. For CoAlt science, the intercorrelations between the subclaims were between 0.649 and 0.760, 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 0.778 to 0.927. For CoAlt social studies, given that 2025 was the standard-setting year and student test scores will be reported for the first time in spring 2026, subclaim correlations were not included.

Table 8.56. Correlations Between Subclaims—Science

Assessment	Subclaim	Life Science	Earth and Space Science	Total Test
Science 5	Physical Science	0.621	0.655	0.900
	Life Science	–	0.586	0.802
	Earth and Space Science	–	–	0.881
Science 8	Physical Science	0.592	0.615	0.892
	Life Science	–	0.560	0.824

Assessment	Subclaim	Life Science	Earth and Space Science	Total Test
	Earth and Space Science	–	–	0.835
Science 11	Physical Science	0.703	0.622	0.911
	Life Science	–	0.607	0.876
	Earth and Space Science	–	–	0.828

Chapter 9: Calibration, Equating, and Scaling

The item response theory (IRT) Rasch one-parameter (1PL) and partial credit model (RPCM) were used to develop, calibrate, equate, and scale the CoAlt Science and Social Studies 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. Equating was conducted to place the spring 2025 test forms on the spring 2023 base scale for the CoAlt Science assessments. For the CoAlt Social Studies assessments, since 2025 was the standard-setting year, scaling was conducted to establish the baseline scale. Equating was not required for 2025 but will be implemented in subsequent years. All steps in the calibration, equating, and scaling processes were repeated for each 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.

9.1. IRT Model

For each grade-level assessment, the RPCM was used to place the CoAlt Science items and student proficiency on the same scale. For CoAlt Social Studies, the Rasch 1PL IRT model was used because the assessment includes 1-point SR items only.

The RPCM 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 performance level, θ_n (also referred to as "ability"), and the step difficulties, δ_{ij} of the m steps in question i as follows:

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

The formula for the Rasch 1PL IRT model is as follows, where the probability of student *n* scoring correctly on the item *i* is a function of the student's ability level, θ_n and the item difficulty level, δ_i.

$$P_{in}(\theta) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}$$

9.2. Data Preparation

Prior to any analyses, several steps were completed in preparation: (a) the data file containing student responses was verified and exclusion rules were applied, (b) traditional item analyses of all items were conducted prior to calibration, and (c) 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, and item-total correlation.

9.3. Calibration

Calibration refers to the estimation of item parameters that places items and students on a common scale. To obtain the Rasch item parameter estimates for the spring 2025 assessments, the RPCM and Rasch 1PL models were applied to the operational and embedded field test items for science and social studies, respectively. Winsteps (Linacre, 2021) was used for all grade-level calibrations. For CoAlt Science assessments, all operational items within a grade are calibrated concurrently. For CoAlt Social Studies assessments, since 2025 was the standard-setting year, all operational and field-test items were calibrated concurrently to establish the baseline scale.

9.4. Equating

9.4.1. Operational 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 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, item calibration 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 test administrations.

A fixed common items approach was used to equate the spring 2025 CoAlt Science assessments. The operational items used to post-equate the assessments to the base scales are called anchor items. The anchor items are a set of common items that are already equated to the base scale and are placed on forms from adjacent administrations. This set of items represents the test blueprint in terms of content and item types and represents approximately 30% of a full test form. To obtain

equated Rasch parameter estimates, anchor item parameter estimates were fixed to their previously equated item parameter estimates before calibrating the remaining non-anchor operational items. This method placed the non-anchor operational items on the same scale as the anchor items.

The stability check for the anchor items was conducted using classical item analysis, scatter plots of item difficulties, and displacement estimates from Winsteps to compare the anchor item performance across years. Displacement estimates greater than or equal to ±0.30 were used as the flagging criteria. Items flagged from the stability check are examined and consideration is given to the impact of flagged item(s) on the content representativeness of the resulting anchor set. A flag alone is not the sole criteria for removing an item from the anchor item set. It is important to also make sure that the remaining anchor set continues to be representative of the overall content and structure of the test. The final anchor sets for grades 5, 8, and 11 represented 33% of the total test points.

For CoAlt Social Studies, no equating was conducted in 2025 because it was the standard-setting year. Equating will be performed starting from spring 2026.

9.4.2. Field Test Equating

For CoAlt Science, the field test equating process is similar to operational equating, except the anchor items are the operational items. With this calibration method, the embedded field test items are calibrated with the operational item parameters fixed at their previously estimated values. This process places the field test item parameter estimates onto the operational base scale. All field test items are calibrated concurrently. For CoAlt Social Studies, since 2025 was the standard-setting year, all operational and field-test items were calibrated concurrently to establish the baseline scale.

9.5. Item-Level IRT Statistics

Appendix G presents the item parameter estimates for each grade. The item numbers are merely identifiers and do not reflect the sequence of items as they were presented to students. The "Model" refers to the IRT model under which the item was estimated (Rasch for all CoAlt assessments). The "B" column shows the item parameter estimate for difficulty, and "D1" through "D4" for the RPCM category threshold estimates for CoAlt Science. The last two columns reflect the infit and outfit statistics generated from Winsteps. Fit values were reviewed, and no items were removed due to misfit. For CoAlt Social Studies, item-level IRT statistics were not included for 2025 because it was the standard-setting year, and 2026 will be the first year that scores are reported.

9.6. Scaling

After the item parameter estimates were obtained for the 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.

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. After equating and scaling, a test characteristic curve (TCC) for each grade-level test was created 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.

Chapter 10: 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 and Social Studies assessments was evaluated with the following analyses:

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

10.1. 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 are provided for the overall test and by subclaim for CoAlt Science and for the overall test for CoAlt Social Studies, as shown in Table 10.1 and Table 10.2. The coefficient alpha for the total group ranged from 0.81 for grade 8 to 0.85 for grade 11 (across the science and social studies assessments. Given the differences in length, it is expected that the coefficient alpha for the overall test will be higher than that of the subscales. For CoAlt Social Studies, a combined subclaim for Economics and Personal Financial Literacy was created, as each individually accounted for only a small portion of the test. Appendix E presents the coefficient alphas by demographic subgroup.

Table 10.1. Coefficient Alpha—Science

Assessment	Total Test	Physical Science	Life Science	Earth and Space Science
Science 5	0.83	0.7	0.55	0.62
Science 8	0.81	0.66	0.55	0.47
Science 11	0.85	0.73	0.6	0.59

Note. For grade 5, the subclaim is Physical Science/Life Science.

Table 10.2. Coefficient Alpha—Social Studies

Assessment	Total Test	History	Geography	Civics	Economics & Personal Financial Literacy
Social Studies 4	0.83	0.45	0.69	0.57	0.46
Social Studies 7	0.82	0.53	0.54	0.42	0.51

10.2. 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:

$$SEM = s_x\sqrt{1 - p_{XX}}$$

where S_x is the standard deviation of test scores, and p_{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 10.3 and Table 10.2 present the SEM results by subclaim for the CoAlt Science and Social Studies assessments. The SEM values for the total group ranged from 2.57 to 3.11.

Table 10.33. SEM—Science

Assessment	Total Test	Physical Science	Life Science	Earth and Space Science
Science 5	3.11	1.87	1.37	1.98

Assessment	Total Test	Physical Science	Life Science	Earth and Space Science
Science 8	2.92	1.84	1.57	1.71
Science 11	2.88	1.76	1.68	1.56

Note. For grade 5, the subclaim is Physical Science/Life Science.

Table 10.4. SEM—Social Studies

Assessment	Total Test	History	Geography	Civics	Economics & Personal Financial Literacy
Social Studies 4	2.57	1.21	1.38	1.33	1.2
Social Studies 7	2.77	1.48	1.37	1.41	1.3

10.3. 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.

Social studies is not included given that results for social studies in 2025 were reported as percentile ranks only, with overall scale scores and performance levels reported beginning in 2026.

10.4. Decision Consistency and Accuracy

The CoAlt Science scale is 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 10.5 presents the kappa interpretations. Table 10.6 presents the decision accuracy and consistency results, and Table 10.7 presents the accuracy and consistency estimates at each cut score. Social studies is not included given that results for social studies in 2025 were reported as percentile ranks only, with overall scale scores and performance levels reported beginning in 2026.

Table 10.5. 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 10.6. Decision Accuracy and Consistency Estimates

Assessment	Accuracy	Consistency	PChance	Kappa
Science 5	0.64	0.54	0.26	0.39
Science 8	0.69	0.59	0.30	0.41
Science 11	0.70	0.61	0.27	0.46

Table 10.7. Decision Accuracy and Consistency of Cut Scores

Statistic	Assessment	*Approaching Target* Cut	*At Target* Cut	*Advanced* Cut
Accuracy	Science 5	0.90	0.85	0.88
	Science 8	0.89	0.86	0.95
	Science 11	0.90	0.88	0.91
Consistency	Science 5	0.86	0.80	0.83
	Science 8	0.84	0.80	0.93
	Science 11	0.85	0.84	0.87

Chapter 11: 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 assessments that are validated but rather the interpretations of the scores. The purpose of the CoAlt Science and Social Studies 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 and Social Studies assessments mature, validity evidence supporting the assessments' interpretations will continue to be collected and documented.

11.1. 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 and Social Studies 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 Chapter 2 and Chapter 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.

11.2. 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.

11.3. 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 and Social Studies assessments is evaluated using reliability coefficients as provided in Section 10.1.

11.4. 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 11.1 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 11.1, 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 opportunities to engage with the CAS in the classroom setting.

Social studies is not included given that results for social studies in 2025 were reported as percentile ranks only, with overall scale scores and performance levels reported beginning in 2026.

Table 11.1. Correlation Between Test Validity Questions and Student Scores

Question	Assessment	N	Correlation
Familiarity with the Student	Science 5	422	0.05
	Science 8	384	0.03
	Science 11	321	-0.04
Hours Per Week in Instruction on the Content Area	Science 5	421	0.18
	Science 8	386	0.25
	Science 11	320	0.17
How Much Instructional Time in the Content Area Is in the General Education Classroom	Science 5	422	0.32
	Science 8	384	0.27
	Science 11	323	0.19

Students who are eligible to take the alternate assessments take the Dynamic Learning Maps (DLM) consortium assessments for ELA and mathematics. Table 11.2 presents the correlations between the CoAlt science scale scores and DLM performance.

Table 11.2. Correlation Between CMAS Science and DLM ELA and Mathematics

Grade	Correlation with DLM ELA	Correlation with DLM Mathematics
5	0.74	0.68
8	0.72	0.66
11	0.63	0.59

11.5. Evidence for Validity and Consequences of Testing

Because state tests are administered "in the expectation that some benefit will be realized from the intended use of the scores" (AERA et al., 2014, p. 19), validity evidence supporting the use and interpretation of CoAlt results may be investigated as a consequence of testing. One intended consequence of testing is that more students will demonstrate mastery over the CAS over time, as evidenced by more students achieving in the top performance levels, if the data are used appropriately to make improvements in programming at the school and district levels.

Table 11.3 presents the percentage of students who have reached proficiency on the CoAlt Science assessments over the years. As shown in the table, student performance has improved since the first administration where students' scale scores and performance levels were first reported.

Table 11.3. Student Performance Over Time

Assessment	First Administration % *Approaching Target or At Target*	2025 % *Approaching Target or At Target*	% Change, First Admin. to 2025
Science 5	35.2	42.7	7.5
Science 8	32.5	34.1	1.6
Science 11	33.0	43.7	10.7

Note. The first administration for science was spring 2022. The first administration for which scale scores and performance levels were generated for science was spring 2023.

11.6. 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 4.4)
- 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 (Chapter 2, Chapter 3, Section 4.5)

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)? ☐ Is a current Individualized Education Program (IEP) in place or being developed for the student?	☐ **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 ☐ **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***. ***Empirical evidence includes, but is not limited to, formal testing results, multi-disciplinary team evaluations, and other evaluative data.***	☐ **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.** ☐ **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
Spring 20XX

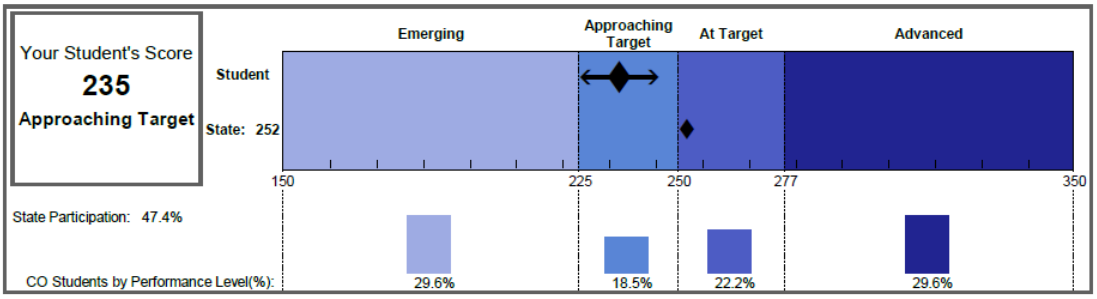
Colorado Alternate Assessment
Student: **FIRSTNAME LASTNAME**
SASID: 9999999999 Birthdate: **MM/DD/YYYY**
School: **SAMPLE SCHOOL NAME (9999)**
District: **SAMPLE DISTRICT NAME (9999)**
Watch a video about this report!

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 to see a video that will talk you through your student's report.

Your student's performance is shown as:
- A number on a scale between 150 and 350
- A performance level that is described below

As you review this report:
- Review arrows around the large diamond to see 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

FIRSTNAME008 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

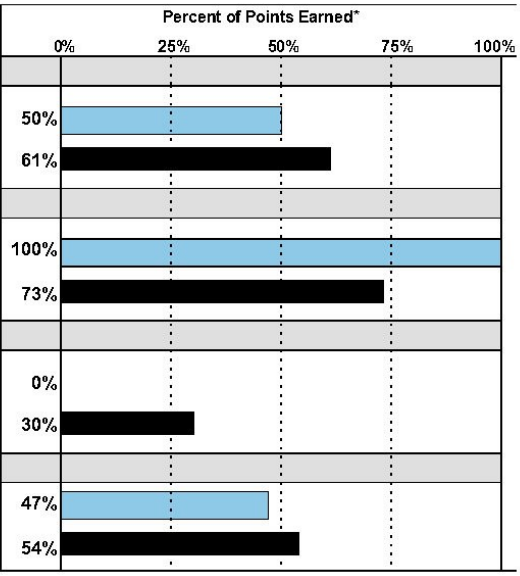
03192024-Z9999999-9999-9999 - 0000000

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	0	15	0% 30%
Science and Engineering Practices			
Making sense of the natural world through investigation and problem solving	14	30	47% 54%



*Percent of points earned cannot be compared across years because individual test questions change from year to year. They also cannot be compared across specific areas of science because the number and difficulty of questions 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 .1. Scale Score Distribution—Science Grade 5

Scale Score	Freq.	%	Cum. Freq.	Cum. %
150	13	3.02	13	3.02
160	6	1.40	19	4.42
170	3	0.70	22	5.12
178	1	0.23	23	5.35
186	3	0.70	26	6.05
192	6	1.40	32	7.44
198	5	1.16	37	8.60
204	7	1.63	44	10.23
209	12	2.79	56	13.02
214	18	4.19	74	17.21
218	16	3.72	90	20.93
223	24	5.58	114	26.51
225	11	2.56	125	29.07
231	24	5.58	149	34.65
235	29	6.74	178	41.40
239	28	6.51	206	47.91
243	19	4.42	225	52.33
247	21	4.88	246	57.21
250	21	4.88	267	62.09
255	18	4.19	285	66.28
259	22	5.12	307	71.40
263	19	4.42	326	75.81
267	17	3.95	343	79.77
271	15	3.49	358	83.26
273	11	2.56	369	85.81
280	16	3.72	385	89.53
284	7	1.63	392	91.16
289	8	1.86	400	93.02
294	7	1.63	407	94.65
299	8	1.86	415	96.51
304	6	1.40	421	97.91
310	1	0.23	422	98.14
317	4	0.93	426	99.07
324	2	0.47	428	99.53
343	1	0.23	429	99.77
350	1	0.23	430	100.00

Note: Freq. = frequency, Cum. Freq. = cumulative frequency, Cum. % = cumulative percentage

Table .2. Scale Score Distribution—Science Grade 8

Scale Score	Freq.	%	Cum. Freq.	Cum. %
150	5	1.26	5	1.26
152	1	0.25	6	1.52
163	2	0.51	8	2.02
171	1	0.25	9	2.27
179	3	0.76	12	3.03
185	3	0.76	15	3.79
191	6	1.52	21	5.30
196	5	1.26	26	6.57
201	11	2.78	37	9.34
205	12	3.03	49	12.37
210	18	4.55	67	16.92
214	27	6.82	94	23.74
218	16	4.04	110	27.78
222	12	3.03	122	30.81
225	25	6.31	147	37.12
229	21	5.30	168	42.42
233	22	5.56	190	47.98
236	18	4.55	208	52.53
240	23	5.81	231	58.33
244	12	3.03	243	61.36
247	18	4.55	261	65.91
250	20	5.05	281	70.96
255	20	5.05	301	76.01
259	21	5.30	322	81.31
263	16	4.04	338	85.35
267	15	3.79	353	89.14
271	13	3.28	366	92.42
276	10	2.53	376	94.95
277	7	1.77	383	96.72
286	8	2.02	391	98.74
292	1	0.25	392	98.99
299	2	0.51	394	99.49
315	2	0.51	396	100.00

Note: Freq. = frequency, Cum. Freq. = cumulative frequency, Cum. % = cumulative percentage

Table .3. Scale Score Distribution—Science Grade 11

Scale Score	Freq.	%	Cum. Freq.	Cum. %
150	10	2.93	10	2.93
161	3	0.88	13	3.81
170	1	0.29	14	4.11
178	9	2.64	23	6.74
185	10	2.93	33	9.68
192	8	2.35	41	12.02
198	10	2.93	51	14.96

Scale Score	Freq.	%	Cum. Freq.	Cum. %
204	14	4.11	65	19.06
209	19	5.57	84	24.63
214	22	6.45	106	31.09
219	13	3.81	119	34.90
224	11	3.23	130	38.12
225	22	6.45	152	44.57
234	8	2.35	160	46.92
239	10	2.93	170	49.85
243	14	4.11	184	53.96
248	8	2.35	192	56.30
250	16	4.69	208	61.00
257	28	8.21	236	69.21
262	20	5.87	256	75.07
267	13	3.81	269	78.89
272	11	3.23	280	82.11
277	14	4.11	294	86.22
283	8	2.35	302	88.56
288	17	4.99	319	93.55
294	6	1.76	325	95.31
300	4	1.17	329	96.48
307	5	1.47	334	97.95
314	1	0.29	335	98.24
323	2	0.59	337	98.83
332	1	0.29	338	99.12
350	3	0.88	341	100.00

Note: Freq. = frequency, Cum. Freq. = cumulative frequency, Cum. % = cumulative percentage

Appendix D: Scale Score Distribution Histograms

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

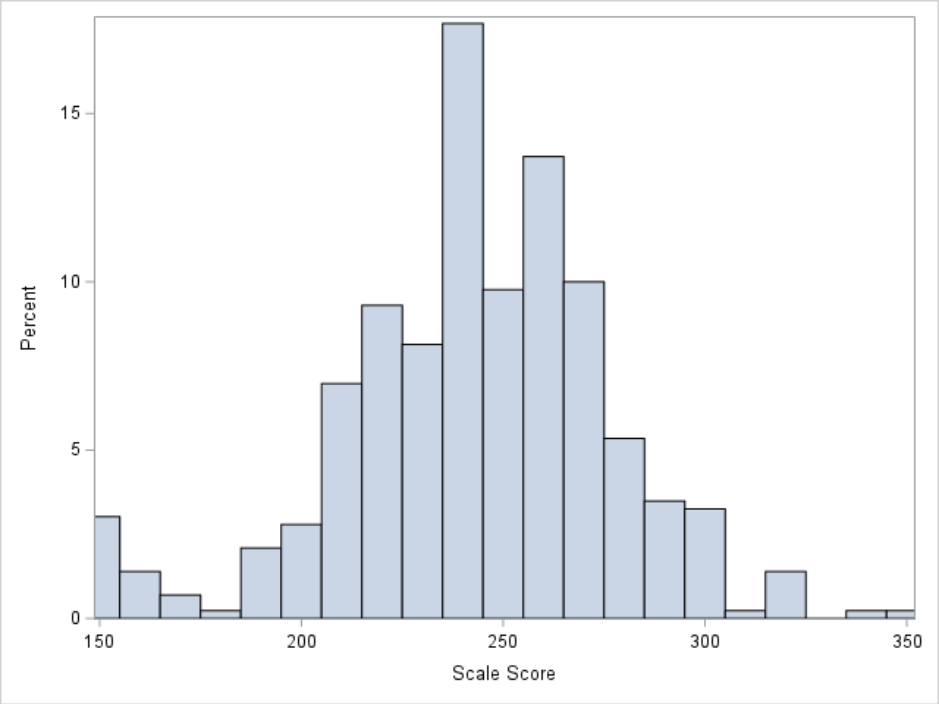


Figure .2. Scale Score Distribution Histogram—Science Grade 8

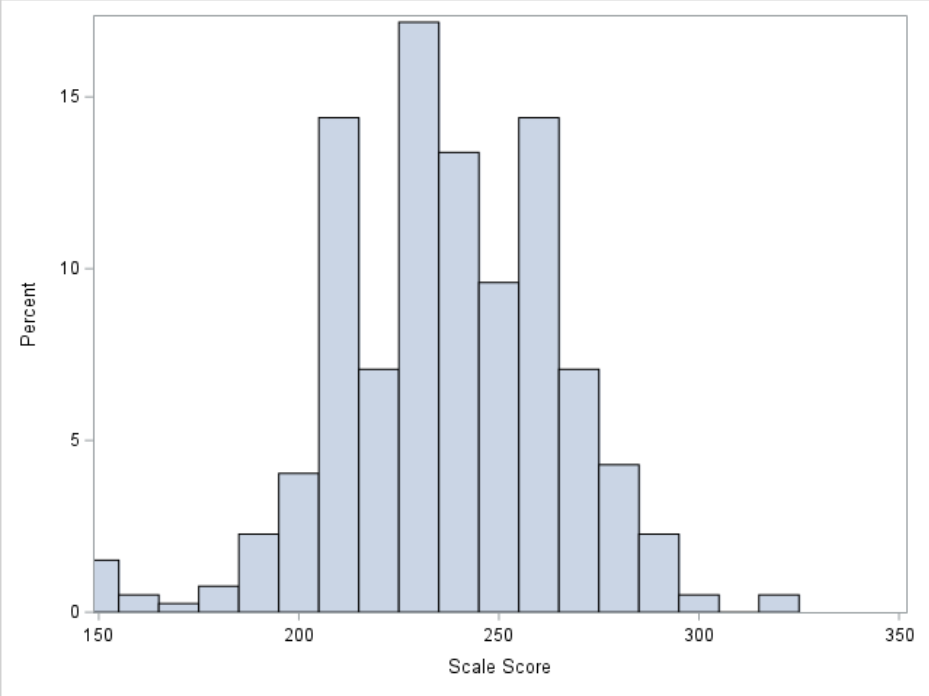
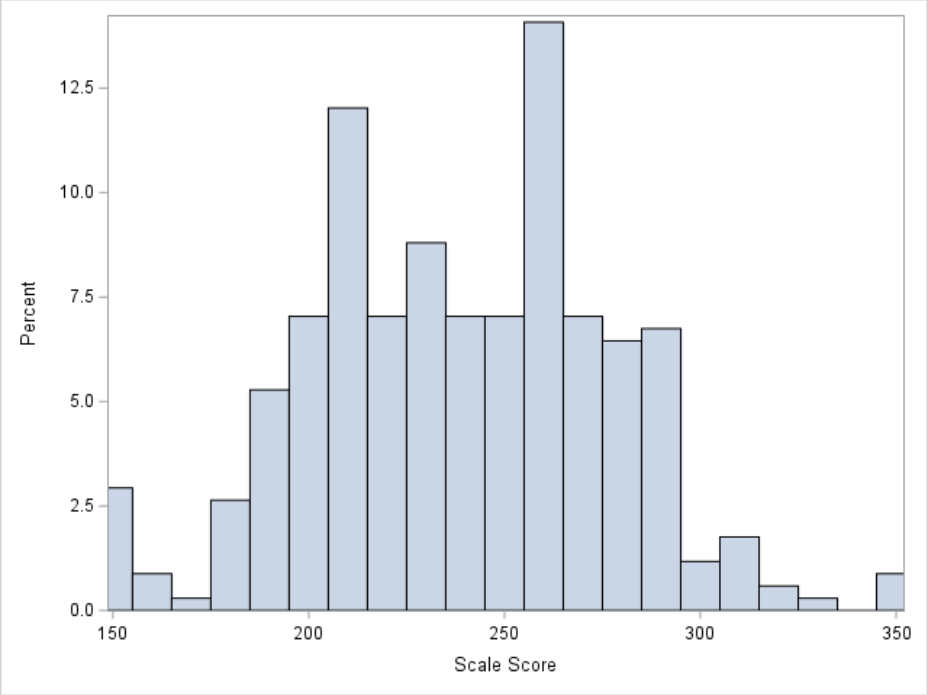


Figure .3. Scale Score Distribution Histogram—Science Grade 11



Appendix E: Performance Results by Demographic Subgroup

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

Subgroup	N	Mean	SD	Min.	Max.	Alpha
No IEP	*	*	*	*	*	*
IEP	429	242.86	34.20	150	350	0.83
No Accommodation	*	*	*	*	*	*
Accommodation	430	242.72	34.27	150	350	0.83
Am. Indian/Alaska Native	*	*	*	*	*	*
Asian	21	224.95	42.35	150	304	0.88
Black	33	241.27	31.20	150	324	0.82
Hispanic	179	241.73	34.58	150	350	0.84
White	168	245.54	33.49	150	324	0.83
Hawaiian/Pacific Islander	*	*	*	*	*	*
Two or More Races	24	247.33	30.66	150	294	0.77
Missing	*	*	*	*	*	*
No Economic Disadvantage	176	236.47	31.10	150	304	0.80
Economic Disadvantage	254	247.06	35.73	150	350	0.85
Female	165	244.84	30.38	150	343	0.79
Male	265	241.41	36.47	150	350	0.85
Language Proficiency NA	343	244.48	33.85	150	350	0.83
Language Proficiency NEP	54	226.59	33.37	150	299	0.83
Language Proficiency LEP	18	242.78	32.43	160	294	0.82
Language Proficiency FEP	*	*	*	*	*	*
Not Migrant	430	242.72	34.27	150	350	0.83
Migrant	*	*	*	*	*	*

*n-count less than 16

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

Subgroup	N	Mean	SD	Min.	Max.	Alpha
No IEP	*	*	*	*	*	*
IEP	396	236.07	27.45	150	315	0.81
No Accommodation	*	*	*	*	*	*
Accommodation	396	236.07	27.45	150	315	0.81
Am. Indian/Alaska Native	*	*	*	*	*	*
Asian	*	*	*	*	*	*
Black	35	243.63	26.57	191	299	0.82
Hispanic	191	235.64	26.61	150	315	0.79
White	140	236.46	28.17	150	315	0.82
Hawaiian/Pacific Islander	*	*	*	*	*	*
Two or More Races	*	*	*	*	*	*
Missing	*	*	*	*	*	*
No Economic Disadvantage	177	231.16	28.15	150	315	0.81
Economic Disadvantage	219	240.03	26.26	150	315	0.79
Female	161	234.45	25.63	152	315	0.78
Male	235	237.17	28.63	150	315	0.82

Subgroup	N	Mean	SD	Min.	Max.	Alpha
Language Proficiency NA	302	238.74	27.14	150	315	0.81
Language Proficiency NEP	48	220.98	27.07	150	267	0.79
Language Proficiency LEP	*	*	*	*	*	*
Language Proficiency FEP	36	235.33	26.95	150	286	0.78
Not Migrant	396	236.07	27.45	150	315	0.81
Migrant	*	*	*	*	*	*

*n-count less than 16

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

Subgroup	N	Mean	SD	Min.	Max.	Alpha
No IEP	*	*	*	*	*	*
IEP	339	238.46	38.53	150	350	0.85
No Accommodation	*	*	*	*	*	*
Accommodation	341	238.54	38.43	150	350	0.85
Am. Indian/Alaska Native	*	*	*	*	*	*
Asian	19	230.74	44.11	178	350	0.88
Black	28	247.86	31.94	185	294	0.79
Hispanic	132	236.11	36.49	150	350	0.83
White	135	238.41	39.51	150	332	0.86
Hawaiian/Pacific Islander	*	*	*	*	*	*
Two or More Races	16	240.50	49.25	150	350	0.91
Missing	*	*	*	*	*	*
No Economic Disadvantage	176	239.20	37.32	150	350	0.84
Economic Disadvantage	165	237.84	39.68	150	350	0.86
Female	126	237.78	37.07	150	350	0.83
Male	215	238.99	39.29	150	350	0.86
Language Proficiency NA	284	239.31	39.15	150	350	0.86
Language Proficiency NEP	26	226.92	39.47	150	323	0.85
Language Proficiency LEP	*	*	*	*	*	*
Language Proficiency FEP	27	243.37	29.55	185	314	0.74
Not Migrant	341	238.54	38.43	150	350	0.85
Migrant	*	*	*	*	*	*

*n-count less than 16

Table .4. Percentile Rank Summary Statistics by Demographic Subgroup—Social Studies Grade 4

Subgroup	N	Mean	SD	Min.	Max.	Alpha
No IEP	*	*	*	*	*	*
IEP	153	50.07	28.82	2	99	0.83
No Accommodation	*	*	*	*	*	*
Accommodation	153	50.07	28.82	2	99	0.83
Am. Indian/Alaska Native	*	*	*	*	*	*
Asian	*	*	*	*	*	*
Black	*	*	*	*	*	*
Hispanic	60	47.78	30.63	2	98	0.85
White	62	55.26	28.41	2	99	0.83

Hawaiian/Pacific Islander	*	*	*	*	*	*
Two or More Races	*	*	*	*	*	*
Missing	*	*	*	*	*	*
No Economic Disadvantage	73	50.78	30.32	2	98	0.86
Economic Disadvantage	80	49.41	27.55	2	99	0.79
Female	47	48.19	27.08	2	94	0.81
Male	106	50.90	29.64	2	99	0.84
Language Proficiency NA	127	53.06	28.24	2	99	0.81
Language Proficiency NEP	20	34.20	28.27	2	87	0.87
Language Proficiency LEP	*	*	*	*	*	*
Language Proficiency FEP	*	*	*	*	*	*
Not Migrant	153	50.07	28.82	2	99	0.83
Migrant	*	*	*	*	*	*

*n-count less than 16

Table .5. Percentile Rank Summary Statistics by Demographic Subgroup—Social Studies Grade 7

Subgroup	N	Mean	SD	Min.	Max.	Alpha
No IEP	*	*	*	*	*	*
IEP	128	50.13	29.00	1	99	0.82
No Accommodation	*	*	*	*	*	*
Accommodation	128	50.13	29.00	1	99	0.82
Am. Indian/Alaska Native	*	*	*	*	*	*
Asian	*	*	*	*	*	*
Black	*	*	*	*	*	*
Hispanic	63	50.37	28.96	2	99	0.80
White	53	50.06	30.03	1	99	0.85
Hawaiian/Pacific Islander	*	*	*	*	*	*
Two or More Races	*	*	*	*	*	*
Missing	*	*	*	*	*	*
No Economic Disadvantage	60	47.13	27.88	1	98	0.80
Economic Disadvantage	68	52.76	29.91	2	99	0.83
Female	53	49.83	30.50	2	99	0.84
Male	75	50.33	28.09	1	99	0.81
Language Proficiency NA	96	51.88	29.47	1	99	0.84
Language Proficiency NEP	21	46.52	29.07	7	92	0.77
Language Proficiency LEP	*	*	*	*	*	*
Language Proficiency FEP	*	*	*	*	*	*
Not Migrant	128	50.13	29.00	1	99	0.82
Migrant	*	*	*	*	*	*

*n-count less than 16

Appendix F: Classical Item-Level Statistics

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

Item	Omit %	*P*-value	Item–Total Correlation
1	0.95%	0.46	0.35
2	0.71%	0.35	0.28
3	0.71%	0.23	0.22
4	0.95%	0.52	0.51
5	0.95%	0.61	0.33
6	1.18%	0.22	0.23
7	0.47%	0.45	0.38
8	0.47%	0.46	0.53
9	0.71%	0.27	0.26
10	0.95%	0.61	0.41
11	0.95%	0.47	0.16
12	0.95%	0.41	0.22
13	0.95%	0.40	0.44
14	0.95%	0.59	0.34
15	0.95%	0.26	0.35
16	0.24%	0.30	0.34
17	0.00%	0.68	0.37
18	0.71%	0.34	0.46
19	0.00%	0.58	0.32
20	0.00%	0.38	0.43
21	0.00%	0.32	0.50
22	0.24%	0.50	0.37
23	0.24%	0.50	0.29
24	0.24%	0.34	0.29
25	0.24%	0.42	0.49
26	0.47%	0.50	0.47
27	0.95%	0.36	0.49
28	0.47%	0.53	0.52
29	0.00%	0.37	0.40
30	0.00%	0.41	0.23
31	0.00%	0.43	0.45
32	0.00%	0.40	0.42
33	0.00%	0.48	0.27

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

Item	Max. Points	Omit %	0%	1%	2%	3%	*P*-value	Item-Total Correlation
1	3	1.18	22.75	29.62	22.51	23.93	0.49	0.58
2	3	0.00	9.72	22.51	41.00	26.78	0.62	0.50
3	3	0.24	14.46	25.12	39.81	20.38	0.55	0.55

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

Item	Omit %	*P*-value	Item–Total Correlation
1	0.00%	0.57	0.26
2	0.26%	0.39	0.21
3	0.26%	0.28	0.31
4	0.00%	0.70	0.48
5	0.00%	0.44	0.35
6	0.00%	0.37	0.31
7	0.00%	0.51	0.48
8	0.00%	0.31	0.30
9	0.26%	0.41	0.11
10	0.00%	0.61	0.25
11	0.00%	0.35	0.38
12	0.00%	0.30	0.31
13	0.00%	0.65	0.37
14	0.00%	0.77	0.40
15	1.29%	0.29	0.28
16	0.00%	0.81	0.51
17	0.00%	0.62	0.45
18	0.00%	0.57	0.35
19	0.00%	0.57	0.34
20	0.00%	0.70	0.44
21	0.78%	0.22	0.27
22	0.00%	0.65	0.49
23	0.00%	0.51	0.42
24	0.00%	0.19	0.17
25	0.00%	0.53	0.53
26	0.00%	0.44	0.51
27	0.00%	0.27	0.30
28	0.00%	0.53	0.61
29	0.00%	0.73	0.46
30	0.26%	0.46	0.19

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

Item	Max. Points	Omit %	0%	1%	2%	3%	*P*-value	Item-Total Correlation
1	3	0.26	25.84	33.33	22.74	17.83	0.44	0.58
2	3	0.52	23.77	37.47	29.20	9.04	0.41	0.47
3	3	0.00	19.90	22.48	28.42	29.20	0.56	0.53

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

Item	Omit %	*P*-value	Item–Total Correlation
1	0.30%	0.70	0.55
2	0.30%	0.21	0.22
3	0.30%	0.28	0.27
4	0.30%	0.59	0.55
5	0.30%	0.30	0.16
6	0.30%	0.47	0.51
7	0.30%	0.72	0.61
8	0.30%	0.31	0.30
9	0.30%	0.39	0.46
10	0.00%	0.49	0.48
11	0.00%	0.55	0.42
12	0.00%	0.57	0.26
13	0.00%	0.80	0.46
14	0.00%	0.20	0.37
15	0.00%	0.39	0.37
16	0.00%	0.41	0.40
17	0.00%	0.35	0.25
18	0.59%	0.40	0.48
19	0.00%	0.29	0.16
20	0.00%	0.40	0.28
21	0.00%	0.48	0.58
22	0.00%	0.51	0.46
23	0.00%	0.51	0.51
24	0.30%	0.47	0.49
25	0.00%	0.51	0.62
26	0.00%	0.47	0.49
27	0.00%	0.73	0.45
28	0.00%	0.49	0.39
29	0.00%	0.34	0.30
30	0.30%	0.22	0.34

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

Item	Max. Points	Omit %	0%	1%	2%	3%	*P*-value	Item-Total Correlation
1	3	0.89	31.07	39.94	20.41	7.69	0.35	0.35
2	3	0.30	17.75	30.77	20.71	30.47	0.55	0.63
3	3	0.30	10.06	19.53	34.02	36.09	0.65	0.64

Appendix G: IRT Item-Level Statistics

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

Item	Item Type	Model	B	D1	D2	D3	D4	Infit	Outfit
1	SPT	GPC	-0.184	0	-0.694	0.309	0.385	1.03	1.07
2	SPT	GPC	-0.766	0	-1.162	-0.158	1.320	1.1	1.09
3	SPT	GPC	-0.515	0	-1.313	-0.284	1.597	1.19	1.2
4	SR	Rasch	-0.038					1.02	1.09
5	SR	Rasch	0.522					1.06	1.05
6	SR	Rasch	1.215					1.1	1.13
7	SR	Rasch	-0.232					0.88	0.86
8	SR	Rasch	-0.724					1.05	1.05
9	SR	Rasch	1.537					1.24	1.35
10	SR	Rasch	-0.226					1	0.98
11	SR	Rasch	0.130					0.88	0.84
12	SR	Rasch	0.840					1.04	1.05
13	SR	Rasch	-0.724					0.98	0.94
14	SR	Rasch	-0.060					1.19	1.23
15	SR	Rasch	0.220					1.13	1.15
16	SR	Rasch	0.254					0.93	0.94
17	SR	Rasch	-0.654					1.04	1.06
18	SR	Rasch	0.996					0.98	1.03
19	SR	Rasch	0.782					1.01	0.98
20	SR	Rasch	-1.118					1.02	1.01
21	SR	Rasch	0.570					0.91	0.85
22	SR	Rasch	-0.571					1.05	1.08
23	SR	Rasch	0.492					0.97	0.97
24	SR	Rasch	0.473					0.83	0.77
25	SR	Rasch	-0.289					1.01	1.01
26	SR	Rasch	-0.351					1.1	1.09
27	SR	Rasch	0.558					1.06	1.05
28	SR	Rasch	0.174					0.9	0.85
29	SR	Rasch	-0.226					0.92	0.9
30	SR	Rasch	0.439					0.89	0.85
31	SR	Rasch	-0.326					0.88	0.87
32	SR	Rasch	0.392					0.96	0.98
33	SR	Rasch	0.208					1.12	1.14
34	SR	Rasch	0.141					0.93	0.92
35	SR	Rasch	0.277					0.96	1
36	SR	Rasch	-0.093					1.09	1.11

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

Item	Item Type	Model	B	D1	D2	D3	D4	Infit	Outfit
1	SPT	GPC	0.109	0	-0.807	0.115	0.692	1	0.99
2	SPT	GPC	0.429	0	-1.262	-0.061	1.323	1.11	1.15
3	SPT	GPC	-0.235	0	-0.458	-0.094	0.552	1.11	1.19
4	SR	Rasch	-0.350					1.09	1.09
5	SR	Rasch	0.472					1.13	1.16
6	SR	Rasch	1.057					1.01	1.07
7	SR	Rasch	-1.026					0.91	0.8
8	SR	Rasch	0.187					1.01	1.02
9	SR	Rasch	0.560					1.04	1.04
10	SR	Rasch	-0.084					0.9	0.89
11	SR	Rasch	0.874					1.03	1.1
12	SR	Rasch	0.411					1.21	1.25
13	SR	Rasch	-0.547					1.09	1.14
14	SR	Rasch	0.701					0.97	0.97
15	SR	Rasch	0.749					0.97	0.96
16	SR	Rasch	-0.847					1.03	0.99
17	SR	Rasch	-1.183					0.87	0.82
18	SR	Rasch	1.189					1.12	1.19
19	SR	Rasch	-1.713					0.87	0.66
20	SR	Rasch	-0.585					0.93	0.93
21	SR	Rasch	-0.374					1.02	1.02
22	SR	Rasch	-0.146					1.02	1.05
23	SR	Rasch	-0.875					0.89	0.83
24	SR	Rasch	1.622					1.12	1.24
25	SR	Rasch	-0.847					0.92	0.88
26	SR	Rasch	-0.146					0.96	0.97
27	SR	Rasch	1.588					1.07	1.52
28	SR	Rasch	-0.192					0.86	0.84
29	SR	Rasch	0.228					0.88	0.85
30	SR	Rasch	1.086					1.02	1.03
31	SR	Rasch	-0.156					0.79	0.75
32	SR	Rasch	-1.203					0.9	0.85
33	SR	Rasch	0.155					1.15	1.14

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

Item	Item Type	Model	B	D1	D2	D3	D4	Infit	Outfit
1	SPT	GPC	0.544	0	-1.305	0.216	1.089	1.39	1.53
2	SPT	GPC	-0.407	0	-0.498	0.001	0.497	0.98	0.98
3	SPT	GPC	-0.921	0	-0.884	-0.087	0.971	0.98	1.06
4	SR	Rasch	-1.205					0.85	0.76
5	SR	Rasch	1.333					1.11	1.16
6	SR	Rasch	0.899					1.09	1.06
7	SR	Rasch	-0.708					0.88	0.82
8	SR	Rasch	0.764					1.19	1.27
9	SR	Rasch	0.127					0.92	0.9
10	SR	Rasch	-1.344					0.78	0.68
11	SR	Rasch	0.724					1.06	1.09
12	SR	Rasch	0.324					0.95	0.89
13	SR	Rasch	-0.152					0.93	0.97
14	SR	Rasch	-0.453					0.99	1.02
15	SR	Rasch	-0.554					1.16	1.2
16	SR	Rasch	-1.876					0.95	0.76
17	SR	Rasch	1.416					0.96	0.94
18	SR	Rasch	0.297					1.02	1.03
19	SR	Rasch	0.251					1.01	0.99
20	SR	Rasch	0.535					1.13	1.19
21	SR	Rasch	0.295					0.93	0.9
22	SR	Rasch	0.839					1.2	1.27
23	SR	Rasch	0.265					1.12	1.11
24	SR	Rasch	-0.095					0.83	0.79
25	SR	Rasch	-0.336					0.96	0.93
26	SR	Rasch	-0.289					0.91	0.88
27	SR	Rasch	0.048					0.93	0.89
28	SR	Rasch	-0.335					0.79	0.76
29	SR	Rasch	-0.119					0.91	0.89
30	SR	Rasch	-1.397					0.98	0.86
31	SR	Rasch	-0.152					1.02	1.03
32	SR	Rasch	0.566					1.08	1.16
33	SR	Rasch	1.273					0.99	1.06

Appendix H: Test Characteristic Curves (TCCs)

Figure .1. TCC—Science Grade 5

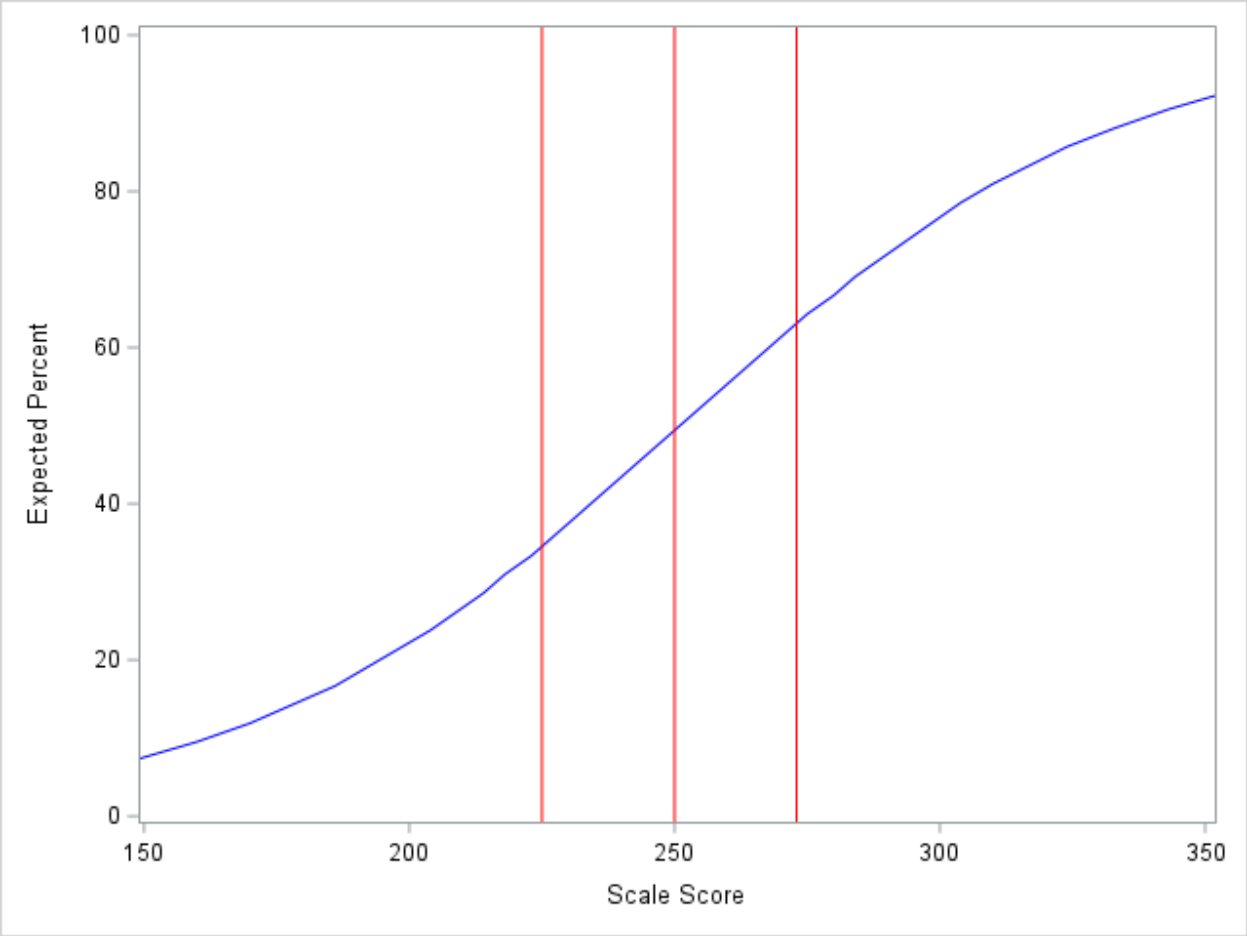


Figure A.2. TCC—Science Grade 8

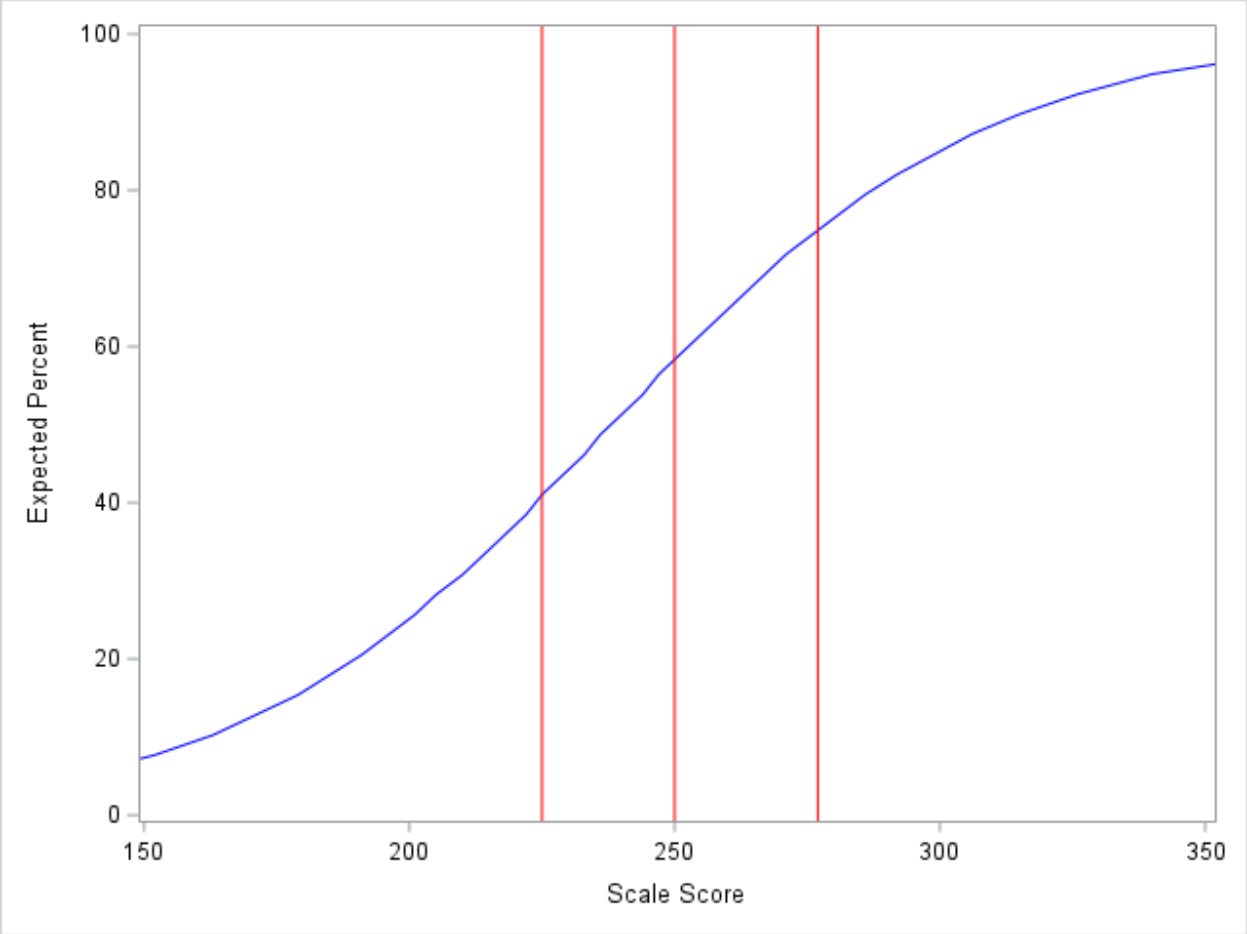
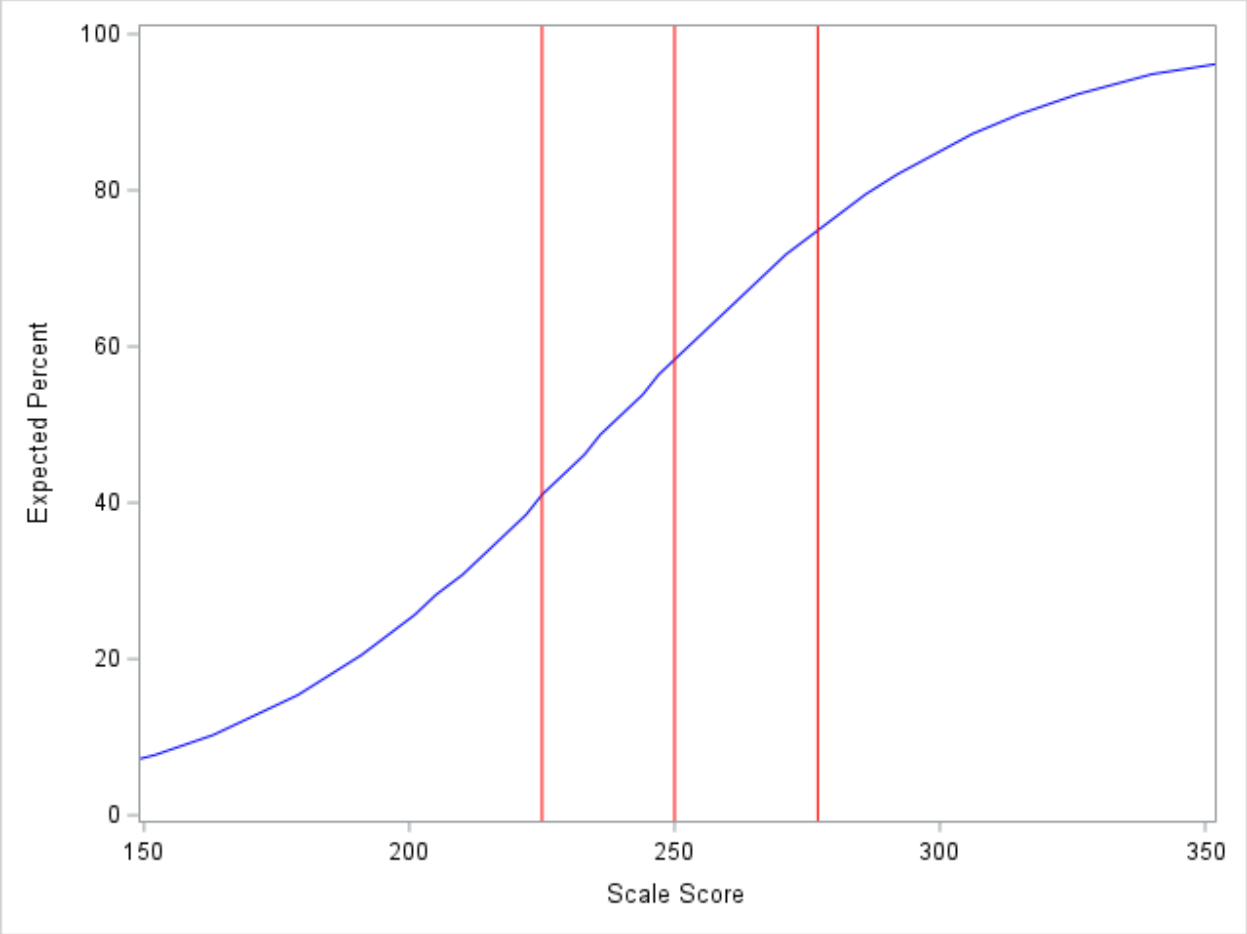


Figure A.3. TCC—Science Grade 11



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

Figure A.1. TIC—Science Grade 5

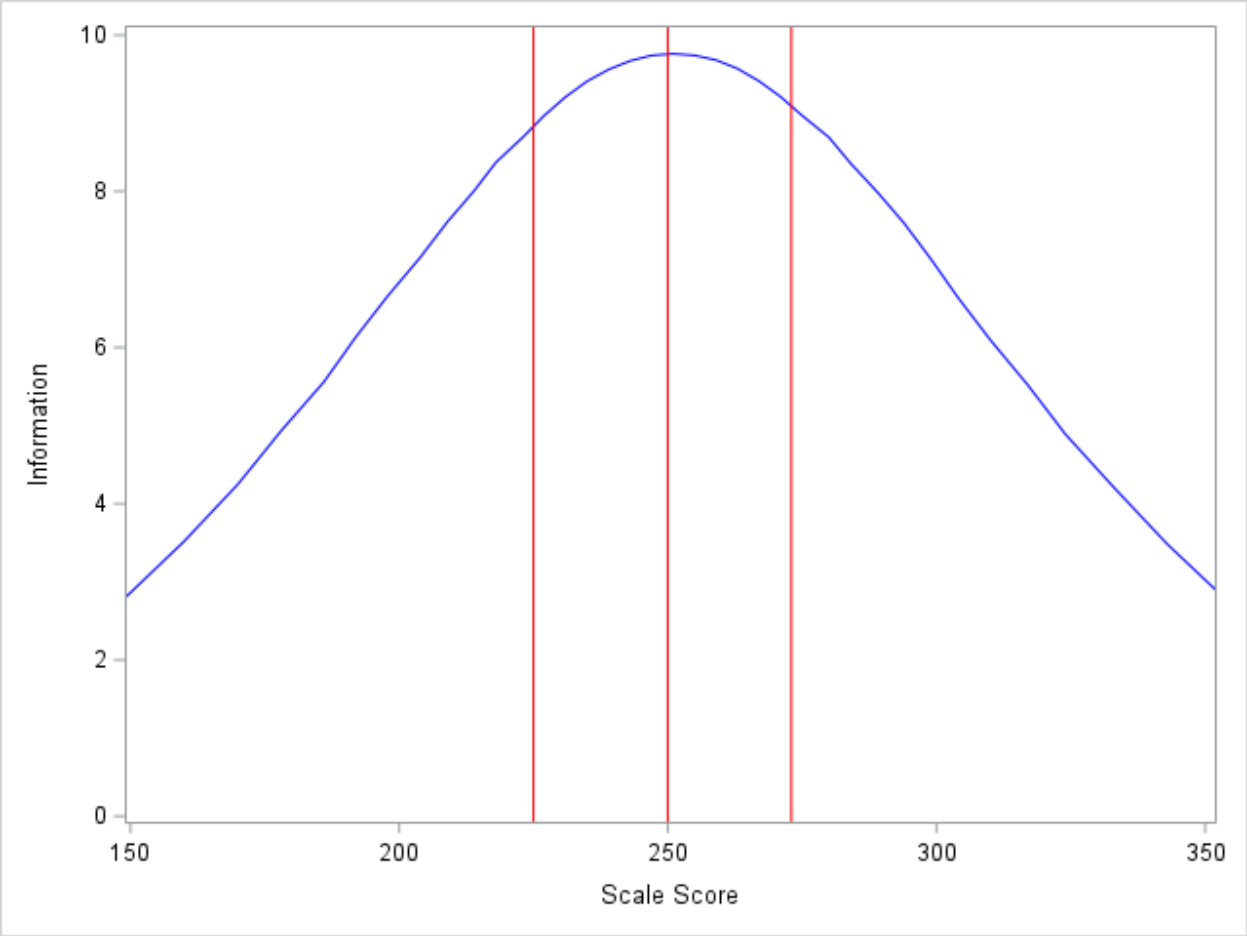


Figure A.2. TIC—Science Grade 8

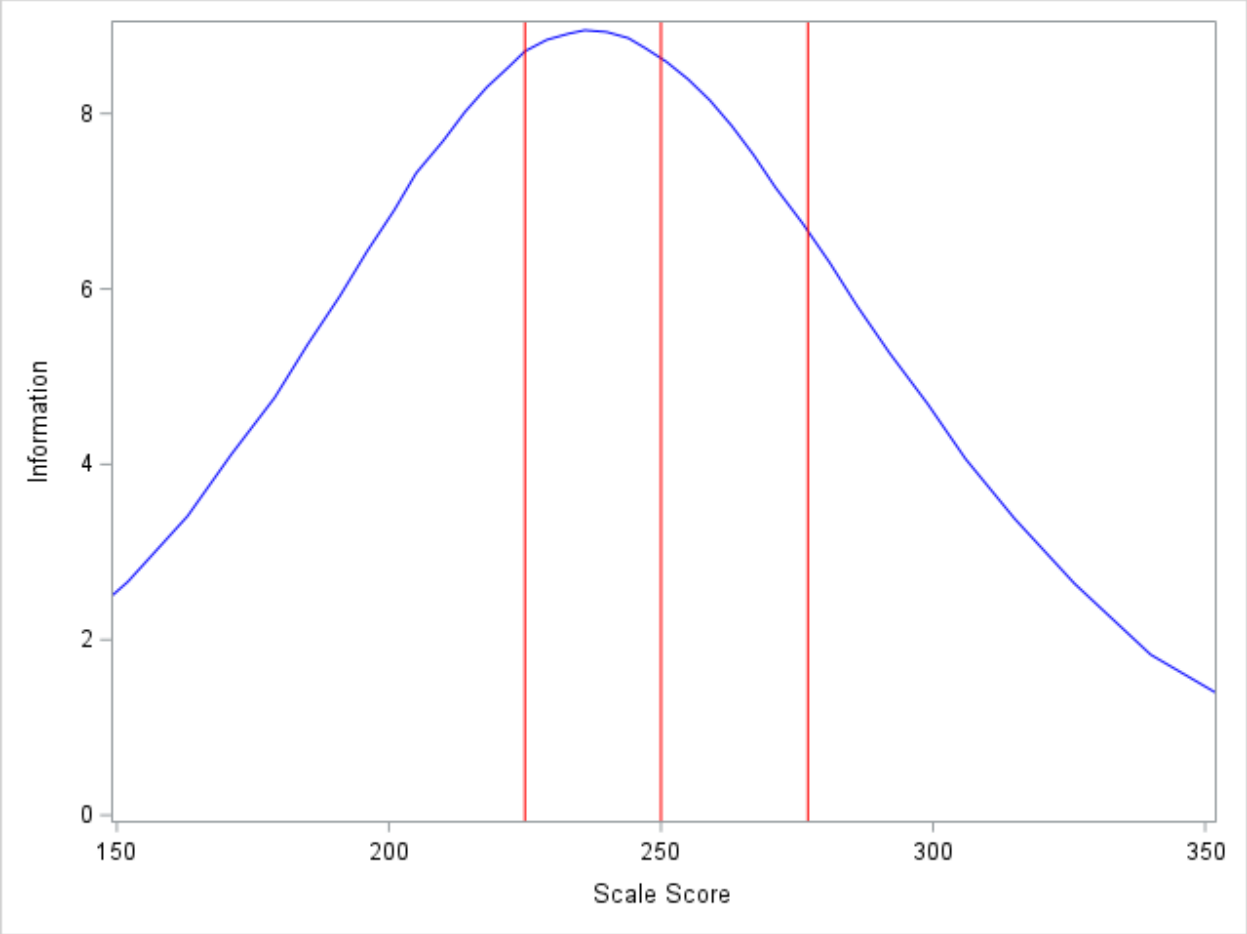


Figure A.3. TIC—Science Grade 11

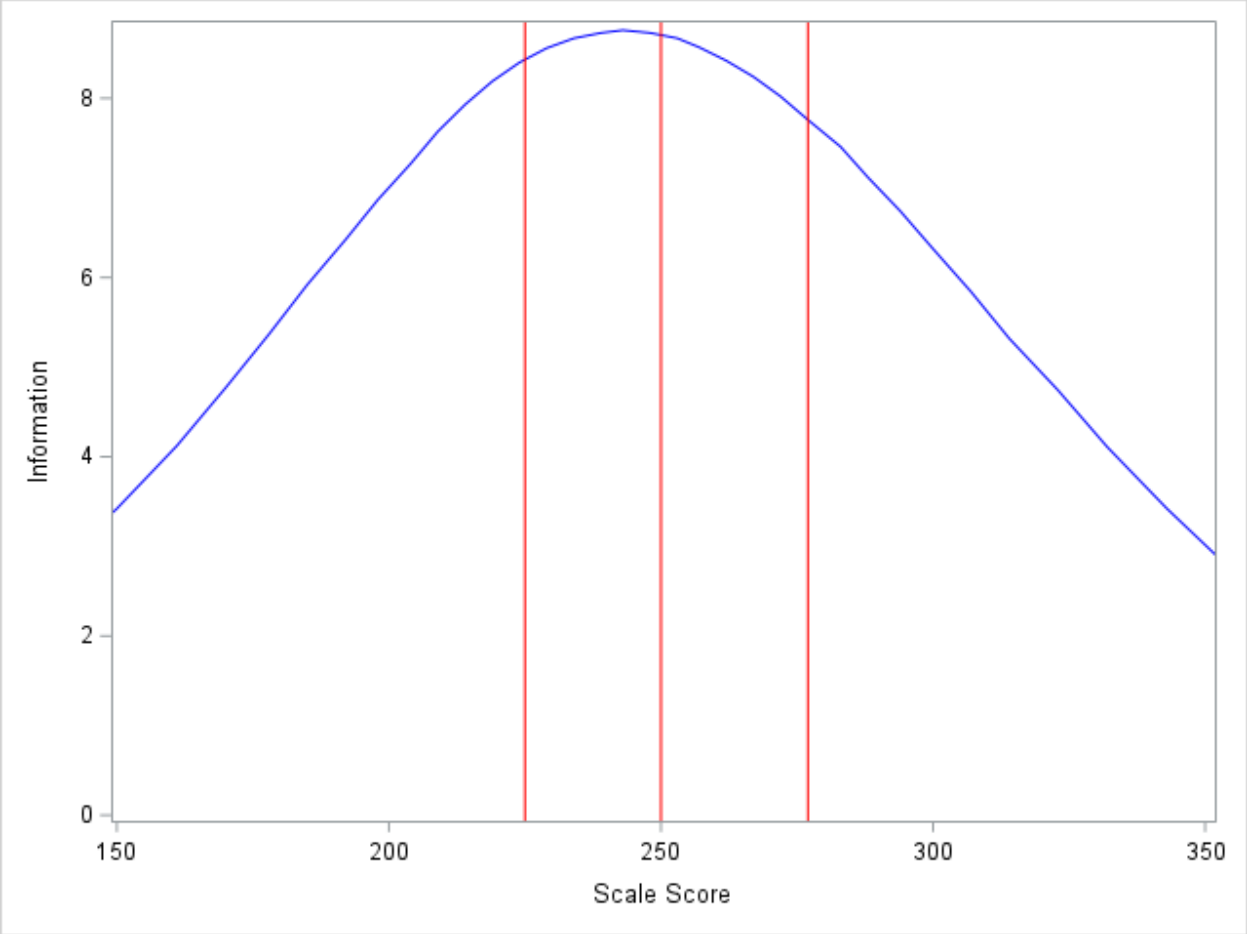


Figure A4. CSEM Curve—Science Grade 5

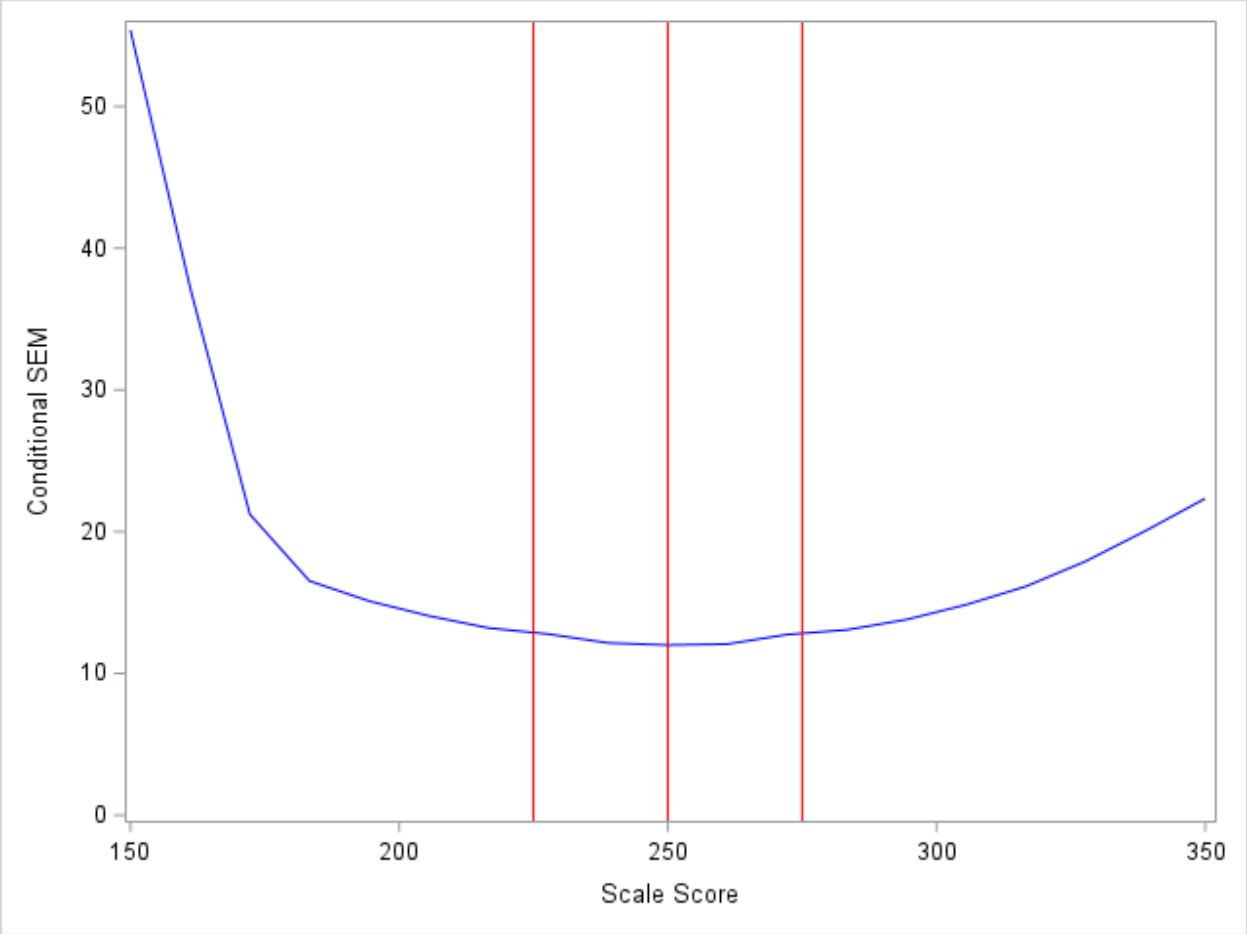


Figure A.5. CSEM Curve—Science Grade 8

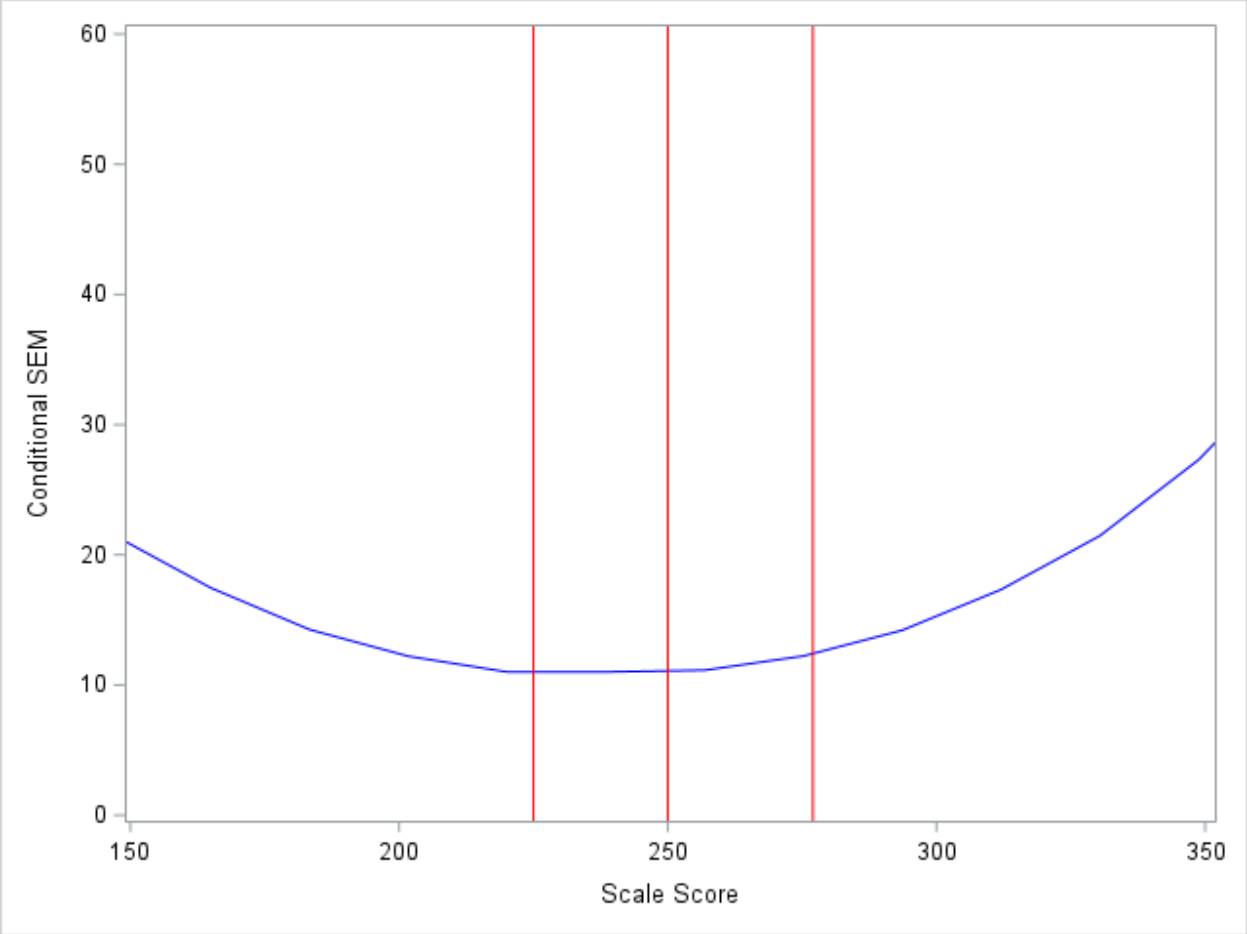
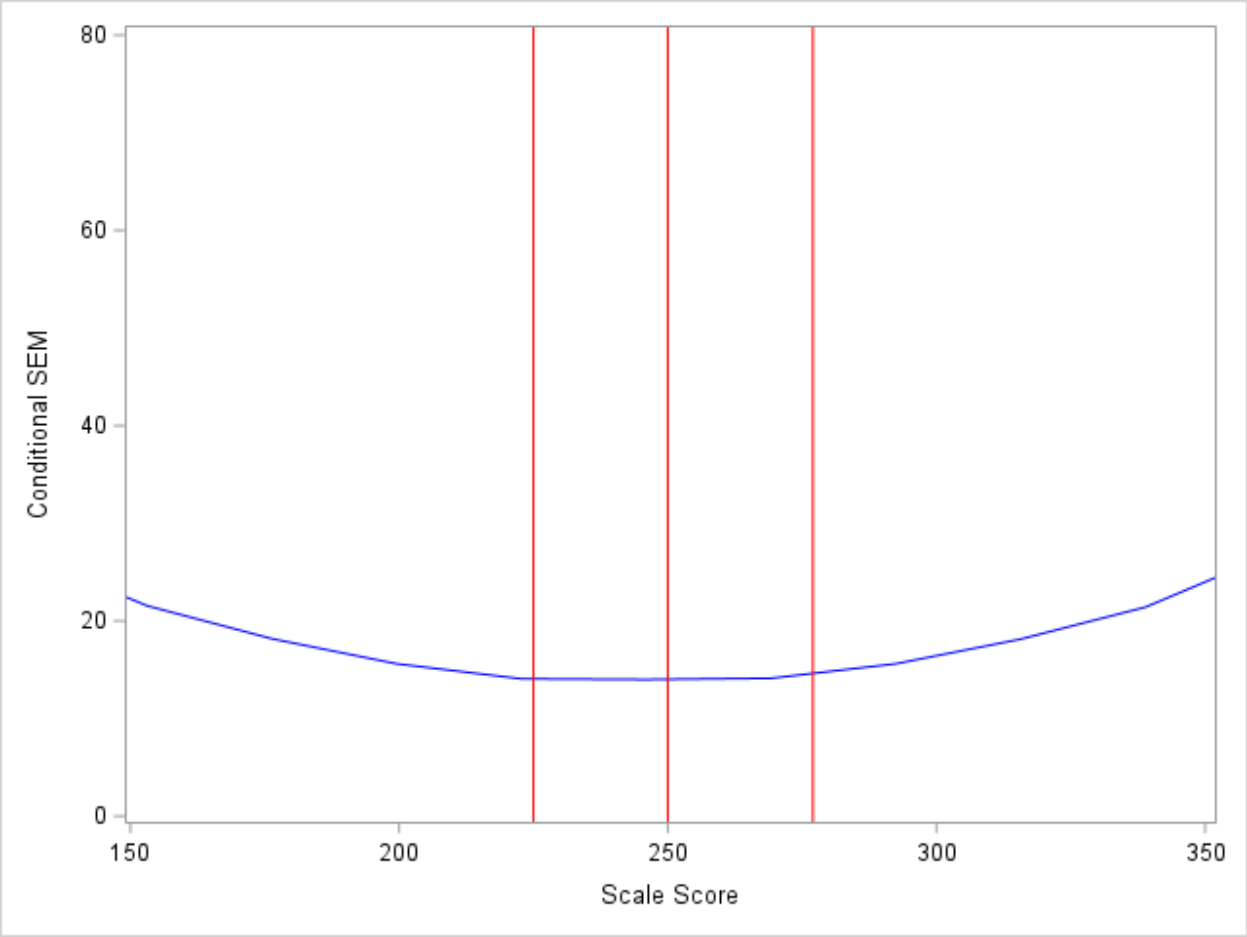


Figure A.6. CSEM Curve—Science Grade 11



Appendix J: Test Administrator Survey Responses

How familiar are you with this student?

Assessment	Very Familiar	Somewhat Familiar	Familiar	Somewhat Unfamiliar	Unfamiliar	Missing
Science 5	85.03%	5.44%	3.85%	0.91%	1.59%	3.17%
Science 8	85.61%	5.21%	1.99%	0.74%	1.99%	4.47%
Science 11	82.18%	6.03%	4.02%	0.29%	0.86%	6.61%
Social Studies 4	78.85%	11.54%	3.21%	0.64%	2.56%	3.21%
Social Studies 7	89.15%	4.65%	3.10%	1.55%	0.78%	0.78%

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

Assessment	<1 Hour	1 to <2 Hours	2 to <3 Hours	3 to <4 Hours	4 to<5 Hours	≥5 Hours	Do Not Know	Missing
Science 5	27.66%	27.89%	19.73%	12.02%	6.35%	2.95%	0.00%	3.40%
Science 8	12.66%	10.67%	11.66%	24.57%	30.27%	6.20%	0.00%	3.97%
Science 11	13.79%	14.66%	21.26%	21.84%	16.95%	4.60%	0.00%	6.90%
Social Studies 4	38.46%	26.92%	16.67%	9.62%	3.85%	1.28%	0.00%	3.21%
Social Studies 7	13.95%	18.60%	10.85%	27.13%	22.48%	5.43%	0.00%	1.55%

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

Assessment	25%	50%	75%	100%	None	Missing
Science 5	28.57%	21.09%	8.39%	8.84%	29.93%	3.17%
Science 8	32.75%	7.20%	6.45%	11.66%	37.47%	4.47%
Science 11	50.86%	8.05%	8.91%	4.60%	21.55%	6.03%
Social Studies 4	31.41%	19.87%	17.74%	9.62%	20.51%	3.85%
Social Studies 7	44.96%	12.40%	2.33%	6.20%	32.56%	1.55%

This student's primary receptive communication is:

Assessment	Oral Language	Sign Language	Reading	Picture Communication	Tactile	Other	Do Not Know	Missing
Science 5	89.53%	0.23%	0.00%	5.35%	0.00%	1.40%	0.47%	3.02%
Science 8	92.42%	0.25%	0.76%	2.27%	0.00%	0.25%	0.00%	4.04%
Science 11	87.39%	0.88%	2.05%	2.93%	0.00%	0.59%	0.00%	6.16%
Social Studies 4	81.05%	0.65%	0.00%	13.73%	0.65%	0.65%	0.00%	3.27%
Social Studies 7	88.28%	0.78%	0.78%	3.91%	0.78%	2.34%	0.78%	2.34%

This student's primary expressive communication is:

Assessment	Oral Language	Sign Language	Writing	Picture Communication	Augmentative Communication Device	Tactile	Other	Do Not Know	Missing
Science 5	79.77%	0.70%	0.00%	3.49%	9.30%	0.00%	1.86%	0.47%	4.42%
Science 8	80.30%	0.51%	0.76%	3.79%	6.31%	0.00%	2.78%	0.00%	5.56%
Science 11	80.06%	0.88%	1.47%	4.40%	5.87%	0.00%	0.88%	0.29%	6.16%
Social Studies 4	61.44%	0.00%	0.00%	12.42%	17.65%	0.65%	2.61%	0.00%	5.23%
Social Studies 7	70.31%	2.34%	0.00%	4.69%	13.28%	0.00%	4.69%	0.00%	4.69%

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

Assessment	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Do Not Know	Missing
Science 5	37.64%	31.75%	14.97%	5.90%	3.40%	2.72%	3.63%
Science 8	43.67%	33.50%	11.41%	4.22%	1.49%	0.25%	5.46%
Science 11	37.93%	37.36%	10.34%	5.17%	2.01%	0.29%	6.90%
Social Studies 4	28.21%	35.90%	17.95%	7.69%	2.56%	1.28%	6.41%
Social Studies 7	33.33%	31.01%	20.93%	7.75%	2.33%	1.55%	3.10%

How much time did this student take on the assessment?

Assessment	0–30 Minutes	31–60 Minutes	61–90 Minutes	91–120 Minutes	121–150 Minutes	151–180 Minutes	≥181 Minutes	Missing
Science 5	4.76%	48.98%	25.40%	9.52%	1.59%	1.36%	2.04%	6.35%
Science 8	8.19%	53.85%	22.58%	5.96%	0.99%	1.74%	0.99%	5.71%
Science 11	6.03%	55.17%	21.84%	6.61%	2.01%	0.86%	1.15%	6.32%
Social Studies 4	21.15%	58.97%	10.26%	1.92%	2.56%	0.64%	0.00%	4.49%
Social Studies 7	35.66%	45.74%	8.53%	3.10%	0.78%	0.00%	0.00%	6.20%