Colorado Measures of Academic Success

Technical Report

Science and Social Studies

2019

Table of Contents

Table of Contents i

Part I: Historical Overview and Summary of Processes 1

CHAPTER 1: INTRODUCTION AND BACKGROUND 1
Purpose of the Document 2
Assessment Development Partners 2
COMPOSITION OF THE ASSESSMENTS 5
SCORE STRUCTURE 7
TEST STRUCTURE 9
TIMING OF TESTS 9
CHAPTER 2: ITEM DEVELOPMENT AND ITEM BANKING 10
Item Development 10
Item Banking Systems 19
CHAPTER 3: TEST CONSTRUCTION 20
Online Forms 21
Accommodated Test Forms 21
CHAPTER 4: TEST ADMINISTRATION PROCEDURES 23
Manuals 23
Training 23
On-site Preparation 24
Accessibility and Accommodations 24
Test Security 25
CHAPTER 5: CONSTRUCTED-RESPONSE SCORING 27
Backreading 28
Calibration 28
CHAPTER 6: STANDARD SETTING 29
CHAPTER 7: REPORTING 30
Description of Scores 30
Score Reports 31
CHAPTER 8: CALIBRATION, EQUATING, AND SCALING 32
Calibration 32
Equating and Scaling 34
Steps in the Calibration, Equating, and Scaling Process 38
CHAPTER 9: RELIABILITY 40
Cronbach's Alpha 40
Standard Error of Measurement 41
Conditional Standard Error of Measurement 41
Decision Consistency and Accuracy 41
Inter-Rater Agreement 42
Sources of Validity Evidence 44
Fairness 47

Part II: Statistical Summaries for 2018–2019 49

CHAPTER 1: SPRING 2019 OPERATIONAL EXAM 50
Administration Summary 50
Equating Results 51
Performance Results 52
Reliability Statistics 52
Validity Statistics 53
CHAPTER 2: SPRING 2019 EMBEDDED FIELD TEST 54
Field Test Forms 54
Inter-Rater Agreement 54

Data Review *54*

References **55**

PART I: HISTORICAL OVERVIEW AND SUMMARY OF PROCESSES

CHAPTER 1: INTRODUCTION AND BACKGROUND

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

CMAS Science and Social Studies assessments were originally developed as online assessments. Colorado legislation (C.R.S. §22-7-1006.3 (1) (d)) requires that a paper-based version be available for all online assessments which may be selected by local educational providers to be administered to their students. These decisions may be made at the local level by grade and content area. The comparable paper-based test forms may also be administered to students with disabilities and English learners (ELs) as appropriate in schools that otherwise are administering the online test forms of the assessments.

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

Intended Population

The CMAS Science and Social Studies assessments are intended to be taken by all students enrolled in public schools, with the exception of some students with the most significant cognitive disabilities who may take the Colorado Alternate Assessment (CoAlt): Science and Social Studies assessments. Eligibility for CoAlt: Science and Social Studies is determined by the student's Individualized Education Program (IEP) or other educational team. Students with disabilities and ELs may take the CMAS Science and Social Studies assessments with or without accommodations that do not change the construct of the assessment. Accommodations are determined based on classroom experience and educational team decisions.

Purpose of CMAS Science and Social Studies

The CMAS Science and Social Studies assessments are designed to be used for a variety of purposes, including to serve as one uniform indicator that will inform parents and educators about individual student achievement of the grade-level CAS in science and social studies, and allow for comparisons to be made across the state. Results are also used as a piece of evidence in the evaluation of educator, school and district performance relative to the CAS.

CMAS Science and Social Studies is a source of data that:

- may be used as a prompt for further investigation at the student, classroom, school, and district levels;

- may support local education agencies in reviewing and developing goals for the performance of their students, including subgroups;
- may indicate that a review of programs, curricula, materials and/or scope and sequence may be appropriate; and that
- may inform the evaluation of district/school approaches.

CMAS Science and Social Studies results also support a range of data-driven stakeholder conversations, activities, and decisions. These may include, but are not limited to, school selection, program evaluation, investigative research, and policy/legislation formation and review.

Purpose of the Document

The purpose of the *CMAS Science and Social Studies Technical Report* is to inform users and other interested parties about the development, content, and technical characteristics of the CMAS Science and Social Studies assessments. This technical report provides information about the planning and administration of the spring 2019 exams.

The *CMAS Science and Social Studies Technical Report* is divided into two parts. Part I presents an overview and summary of the components of the program. Information regarding the planning and administration of the assessments as well as details regarding item development, item banking, test construction, administration procedures, scoring, reporting, reliability, and validity are included in Part I of the document. Part II provides a statistical summary of the spring 2019 administration. Results are provided for both the operational items and the embedded field test items.

Assessment Development Partners

The CMAS Science and Social Studies assessments are collaboratively developed by the Colorado Department of Education (CDE), the Colorado educator community, and the assessment contractor, Pearson. In addition, input and advice is provided by the Colorado Technical Advisory Committee (TAC).

Colorado Department of Education

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

Colorado Educator Community

Educator participation in the CMAS Science and Social Studies development process is critical to ensuring that the assessments are aligned to the CAS, appropriate for Colorado students at the assessed grade level, and free from potential bias and sensitivity issues. Throughout item and assessment development, educators participate in the following assessment development activities:

- Item Writing: After receiving item writing assignments based on the academic standards, educators create assessment items. Items that successfully move through the entire item development process will eventually appear on the operational assessments.

- Content and Bias Review: Educators review items to ensure content alignment and identify potential bias and sensitivity concerns before items are included on the embedded field test.

- Rangefinding: Educators review student responses to field tested items and define the score point ranges for the scoring rubrics that are used to score student responses.

- Data Review: Educators review student performance data associated with field tested items to identify potential construct-irrelevant explanations for statistical flags.

More detailed descriptions of these activities may be found in Chapter 2.

The majority of the field test items that appeared on the 2019 CMAS Science and Social Studies assessments were created and reviewed by committees comprised exclusively of Colorado educators. For each of these meetings, an effort was made to involve educators who were representative of the entire state of Colorado (geographic location, gender, and race) and familiar with the CAS and related instruction. The educators were also familiar with the assessment interaction and demonstration of achievement of different groups of students taking the CMAS Science and Social Studies assessments, including students with disabilities and ELs.

Pearson

As the primary contractor responsible for end-to-end of the 2019 assessment cycle services and products, Pearson worked closely with CDE throughout the CMAS Science and Social Studies assessment development and administration processes. This included item and test development, online and paper test forms creation, enrollment, packaging and distribution, online test delivery, processing, scoring, customer service, standard setting, score reporting, and psychometric services.

Colorado Technical Advisory Committee (TAC)

The Colorado TAC was comprised of psychometric, assessment, and special populations experts tasked with providing high-level consulting and expert advice regarding the validity and reliability of the CMAS Science and Social Studies assessments. The TAC also provided advice

on psychometric topics such as test blueprint design, scaling and equating, scoring, score reporting, and comparability. The TAC included the following members:

- Dr. Jamal Abedi, Professor, University of California, Davis
- Dr. Elliot Asp, Senior Partner, The Colorado Education Initiative
- Dr. Jonathan Dings, Executive Director of Student Assessment and Program Evaluation, Boulder Valley School District
- Dr. Lisa Escarcega, Executive Director, Colorado Association of School Executives
- Dr. Michael Kolen, Consultant
- Dr. Martha Thurlow, Director, National Center on Educational Outcomes

Composition of the Assessments

CMAS Science and Social Studies assessments are standards-based and designed to measure what students should know and be able to demonstrate at the end of each assessed grade.

- Science CAS: https://www.cde.state.co.us/coscience/statestandards
- Social Studies CAS: https://www.cde.state.co.us/cosocialstudies/statestandards

The subject and grade combinations for CMAS Science and Social Studies are shown in Table 1. This report pertains to the fourth operational administration for the High School Science assessment and the fifth operational administration for the Elementary/Middle School Science and Social Studies assessments in spring 2019.

CMAS Science and Social Studies is designed to be administered online via Pearson's TestNav platform. Each assessment contains selected-response (SR) items, technology-enhanced items (TEIs), and constructed-response (CR) items. Each assessment is comprised of three sections and all sections contain a combination of SR items, TEIs, and CR items.

The CMAS Science and Social Studies assessments cover the content standards, which are outlined below. Scientific Investigations and the Nature of Science is also included as a Science reporting category. Items in this reporting category are also aligned to one of the three content standards (Physical Science, Life Science, Earth Systems Science).

- Science
    - Physical Science: Students know and understand common properties, forms, and changes in matter and energy.
        - Observe, explain, and predict natural phenomena governed by Newton's laws of motion, acknowledging the limitations of their application to very small or very fast objects.
        - Apply an understanding of atomic and molecular structure to explain the properties of matter, and predict outcomes of chemical and nuclear reactions.
        - Apply an understanding that energy exists in various forms, and its transformation and conservation occur in processes that are predictable and measurable.
    - Life Science: Students know and understand the characteristics and structure of living things, the processes of life, and how living things interact with each other and their environment.
        - Analyze the relationship between structure and function in living systems at a variety of organizational levels, and recognize living systems' dependence on natural selection.
        - Explain and illustrate with examples how living systems interact with the biotic and abiotic environment.

        - Analyze how various organisms grow, develop, and differentiate during their lifetimes based on an interplay between genetics and their environment.
        - Explain how biological evolution accounts for the unity and diversity of living organisms.

    - Earth Systems Science: Students know and understand the processes and interactions of Earth's systems and the structure and dynamics of Earth and other objects in space.
        - Describe and interpret how Earth's geologic history and place in space are relevant to our understanding of the processes that have shaped our planet.
        - Evaluate evidence that Earth's geosphere, atmosphere, hydrosphere, and biosphere interact as a complex system.
        - Describe how humans are dependent on the diversity of resources provided by Earth and Sun.

    - Scientific Investigations and the Nature of Science: Students understand the processes of scientific investigation and design, conducting and evaluating, as well as communicating about, such investigations. Students understand that the nature of science involves making meaning of the natural world.

- Social Studies

    - History: History develops moral understanding, defines identity and creates an appreciation of how things change while building skills in judgment and decision-making. History enhances the ability to read varied sources and develop the skills to analyze, interpret, and communicate.

    - Geography: Geography provides students with an understanding of spatial perspectives and technologies for spatial analysis, and an awareness of interdependence of world regions and resources and how places are connected at local, national, and global scales.

    - Economics: Economics teaches how society manages its scarce resources, how people make decisions, how people interact in the domestic and international markets, and how forces and trends affect the economy as a whole. Personal financial literacy applies the economic way of thinking to help individuals understand how to manage their own scarce resources.

    - Civics: Civics teaches the complexity of the origins, structure, and functions of governments; the rights, roles, and responsibilities of ethical citizenship; the importance of law; and the skills necessary to participate in all levels of government.

CMAS Science and Social Studies item development began in 2012. Items were field-tested in 2013 in order to collect student performance data on all newly-developed items. The goal of the stand-alone field tests were twofold: (1) to allow for the evaluation of item quality through the review of traditional item performance data to support test construction: item difficulty, item-total correlations, item fit statistics, etc., and (2) to explore the use of Knowledge Technologies' (KT) automated-scoring engine with newly-developed CR items.

After field testing, items went through an educator data review and those items that survived comprised the item pool that supported test construction. Following the first operational administration of the Elementary/Middle School assessments in spring 2014, performance standards (i.e., cut scores) were set and final cut scores were approved and used for reporting purposes. The same process was undertaken for the High School Science assessment following the first operational administration in fall 2014. The Colorado State Board of Education formally adopted CMAS Science and Social Studies performance standards between 2014 and 2016.

Score Structure

Master Claim: The degree to which a student demonstrated the concepts and skills represented in the grade-level CAS in science and social studies is reported through both a performance level and a scale score. There are four performance levels based on a scale score range of 300-900. The policy level performance level descriptors and associated scale score ranges are provided below.

CMAS Policy Level Performance Level Descriptors and Associated Overall Scale Scores				
	Partially Met Expectations	**Approached Expectations**	**Met Expectations**	**Exceeded Expectations**
Performance Level Descriptor	Students who demonstrate a limited command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They will *need additional academic support* to engage successfully in	Students who demonstrate a moderate command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They will *likely need additional academic support* to engage successfully in further studies in this content area.	Students who demonstrate a strong command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They are *academically prepared* to engage successfully in further studies	Students who demonstrate a distinguished command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They are *academically well prepared* to engage successfully in further studies in this content area.

	further studies in this content area.		in this content area.	
Scale Score	Grade 4: 300-556 Grade 5: 300-545 Grade 7: 300-591 Grade 8: 300-555 HS SC: 300-542	Grade 4: 557-698 Grade 5: 546-649 Grade 7: 592-700 Grade 8: 556-651 HS SC: 543-672	Grade 4: 699-792 Grade 5: 650-770 Grade 7: 701-769 Grade 8: 652-784 HS SC: 673-773	Grade 4: 793-900 Grade 5: 771-900 Grade 7: 770-900 Grade 8: 785-900 HS SC: 774-900

Test Structure

Test Structure for CMAS Science

The CMAS Science assessment contains SR items, TEIs, and CR items. A subset of the science assessment includes simulation-based item sets, which are groups of items that all relate to a scientific investigation or experiment. Students use the information in the Science Simulations (SIMs) and in the items to answer the questions or respond to the prompts. The simulation-based items may be SR items, TEIs, or CR items.

Test Structure for CMAS Social Studies

The CMAS Social Studies assessment contains SR items, TEIs, and CR items. A subset of the social studies assessment includes Performance Events (PEs). The items within each PE all relate to a collection of sources about a social studies topic. PEs will have either a history or geography theme but may also incorporate economics and civics. Students use the information in the sources to answer the questions or respond to the prompts. The items associated with the PE may be SR items, TEIs, or CR items.

Timing of Tests

Each 2019 assessment was composed of three sections with operational items and field test items embedded. The timing of the sections varied by grade and content area as indicated below:

2019 CMAS Science and Social Studies Testing Times		
	Science	**Social Studies**
Grades 3-5	Sections 1-3: 80 minutes Total time: 240 minutes	Sections 1-3: 80 minutes Total time: 240 minutes
Grades 6-8	Sections 1-3: 80 minutes Total time: 240 minutes	Sections 1-3: 80 minutes Total time: 240 minutes
High School	Sections 1-3: 50 minutes Total time: 150 minutes	Not administered in 2019

CHAPTER 2: ITEM DEVELOPMENT AND ITEM BANKING

The item development process for the CMAS Science and Social Studies assessments involves following prescribed steps in order to develop a diverse bank of items that align directly to the CAS. All items are developed with the intention of being administered on multiple testing platforms: online, online-accommodated, and paper-and-pencil assessments.

The validity of a state assessment relies on the methodology that frames the development and design of the assessment. In support of that claim, CDE and Pearson upheld these considerations as the cornerstones of the CMAS Science and Social Studies item and test development:

- The item development process ensures the CMAS Science and Social Studies items align to the Evidence Outcomes (EOs) and Grade Level Expectations (GLEs) they are intended to measure.
- The CMAS Science and Social Studies item development plan was designed to produce and maintain a robust item bank; items were written to address the scope of essential measured standards, grade-level difficulties, and cognitive complexity (i.e., Depth of Knowledge [DOK]).
- The item and test development processes promote the comparability of the online and the paper-and-pencil assessments with regard to item interaction and formatting.
- The CMAS Science and Social Studies item and test development processes are compliant with industry standards, intended to ensure interoperability between different test platforms.

Pearson's proprietary software, ABBI (Assessment Banking and Building solutions for Interoperable assessments), is used to facilitate the item and test development processes. As described in the following section, items can be moved into various statuses in ABBI. Each status represents a step in the item development process.

Item Development

The item development process is a tiered, inter-related process that begins with the development of the test blueprints for each grade level within each subject. The item development process continues with designing an item development plan (IDP), developed intentionally after bank analysis to determine where the item bank needs to be more robust. The IDP forecasts the number of item types and associated stimulus across GLE, EO, and DOK, which are needed to create a robust item bank that is refreshed over time.

Test Blueprint

CDE and Pearson collaboratively developed test blueprints for each assessed grade or grade band for the CMAS Science and Social Studies assessments. Each test blueprint specifies the number of test items by content standard/reporting category, Prepared Graduate Competency (PGC), and items associated with a SIM or PE. Ranges of points or items are set by PGC, GLE, EO, item type, and DOK. The specificity of the test blueprints ensures the CMAS Science and Social Studies assessments cover the breadth of the content indicated by the CAS within the assessed grade or grade band. CMAS Science and Social Studies test blueprints can be found in Figures 1–5.

Item Development Plan

The IDP is designed to determine the number of items needed to construct the CMAS Science and Social Studies assessments based on the test blueprint requirements. To construct the IDP, the item bank is analyzed and gaps are identified. Each subject and grade-level IDP is updated at the beginning of each item development cycle to inform development targets that address any stimulus, item type, GLE, EO, and DOK shortages.

Item-Writing Process

After a thorough analysis of the item bank and development of the IDPs, Pearson uses the following resources to plan for item development:

- content specifications
- language accessibility/bias and sensitivity guidelines
- editorial style guidelines
- universal design guidelines

The initial step of development is the preliminary conception and composition of the Science SIMs and Social Studies PEs. These ideas are presented in the form of storyboards to CDE for review and feedback, along with suggested EOs that the SIMs and PEs address. CDE provides feedback on how to move forward with the development of the SIMs and PEs. The SIMs and PEs are then fully developed and presented to educators for review. After the SIMs and PEs are approved, items are written to a variety of EOs, either internally or by educators.

Item Writer Workshop

The Item Writer Workshop is conducted after the initial design of the SIMs and PEs. This is a two-part meeting facilitated by Pearson content assessment specialists in conjunction with CDE. CDE invites educators from across the state of Colorado, and every effort is made to ensure that the educators who participate are representative of the population of the state in terms of geographic location, gender, and race/ethnicity. Item writers are placed into committees based upon their background and educational expertise. Item writing assignments are informed by the

IDP, then the assignments are split up and given to the Colorado educator item writers. The item writers use the CAS, Assessment Frameworks, item specifications documents, item writing guidelines, and the item writing checklists to guide them in completing their assignments (examples of these documents can be found in Appendix A). The item writers also work with the Pearson content assessment specialist if any clarification is needed. Content specialists from CDE are present to provide assistance.

During the first part of the Item Writer Workshop, (IWW1), educators concurrently consult the CAS and the item type required by the IDP to begin the process of writing an item. Using their background and educational expertise, they determine how best to measure the construct and work backwards to write an item that measures the intended construct and is free of construct-irrelevant distractions, such as extraneous language and unnecessary background information. The requirements for a high-quality item are communicated through the educator checklists, which are intended for educators to consult at each stage of the item development process.

For SR items and TEIs, educators write the stem of the item which asks the question measured by the construct. They also write the key and associated distractors. A quality distractor contains a plausible error that would indicate a common misconception in student understanding. Distractors are accompanied by a rationale which indicates the common misunderstanding.

For CR items, educators write the stem of the item which asks the question measured by the construct. They also write the rubric and a sample student response for each score point. The rubrics for each item are intended to be comprehensive of all possible student responses that would demonstrate knowledge of the standard, however, the sample student response for each score point may not be comprehensive of all of the ways a student could respond and demonstrate knowledge of the standard. The intention of the score points indicated by the rubric are to indicate the range of all correct answers to the question asked in the item. The intention of the sample student response is to guide scorers as to what a response would look like in student language. The range of all responses that demonstrate knowledge of the standard is not finalized until after the item is field tested so that actual student responses can be taken into consideration (for more detailed information regarding this process see “Rangefinding” later in this chapter).

Item writers author the items, enter the items in ABBI and submit the items to Pearson content assessment specialists for review during the first part of the Item Writer Workshop. The meeting then adjourns for a period of time, typically about two weeks. During this time, the Pearson content assessment specialists review and suggest revisions to the items and metadata for the Colorado educator item writers. This feedback is recorded in ABBI for Colorado educator item writers to review during the second part of the workshop and is maintained for the life of the item. The item writers make the revisions suggested by the Pearson content assessment specialist and meet again in person for a feedback session surrounding the items (Item Writer Workshop 2, IWW2). During this feedback session, educators who participated on the same committee are provided the opportunity to review other educator’s items and provide feedback with regard to the Item Writer educator checklist. Feedback is recorded in ABBI and maintained for the life of the item. Pearson content assessment specialists facilitate this feedback session and make additional suggestions for refinement during group discussion. Upon the completion of the IWW2, item writers resubmit the items in ABBI to the Pearson content assessment specialists.

Pearson Review Process

The content assessment specialists first evaluate each item after the item writer workshop concludes for alignment to the CAS, grade appropriateness, and DOK alignment. They then apply the CMAS Science and Social Studies style guide guidelines.

CMAS Science and Social Studies style guidelines are a universal set of rules which are applied to all CMAS Science and Social Studies items. These rules are set around art specifications, language specifications, layout, and text-to-speech. The intent of the CMAS Science and Social Studies style guidelines is to ensure that each student taking a science or social studies assessment has the same experience.

During the second part of this review, the content assessment specialists focus on the overall quality of the items, relevance to the purpose of the test, and appropriateness of graphics. Research librarians perform additional fact-checking to ensure accuracy.

The Editorial Department checks items for clarity, correctness of language, appropriateness of language for the grade level, adherence to style guidelines, and conformity with acceptable item-writing practices. Rubrics for CR items are reviewed for their ability to be scored by a performance scoring director, and items and/or scoring guidelines (rubrics) with score points deemed "difficult to score" are revised in collaboration with the assessment specialist(s) at this point in the process.

Pearson performs a universal design review (UDR) to assess item accessibility irrespective of diversity of background, cultural tradition, and viewpoints; to evaluate changing roles and attitudes toward various groups; to review the role of language in setting and changing attitudes toward various groups; to 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 to edit for inappropriate language usage or stereotyping with regard to sex, race, culture, ethnicity, class, or geographic region.

During the UDR, Pearson's Assessment Development Services also focuses on reviewing items for potential bias to ensure that all test items are fair, and that all students have an equal opportunity to demonstrate achievement regardless of their gender, ethnic background, religion, socio-economic status, or geographic region. In addition, items are reviewed for visual bias, accessibility for students with disabilities, and convertibility to Braille and text-to-speech.

Once the internal reviews within each department are completed, the items are moved into ABBI to the final content review status for the lead content assessment specialist to review and approve the item to present to CDE.

Adhering to these processes ensures that each item created for the CMAS Science and Social Studies assessments measures the intended EO/standard, is content and grade appropriate, is accessible to all populations required to take the assessments, is free from any bias, and follows the Colorado style.

CDE Review Process

CDE's Review is completed by multiple different members of the department. Their titles and what their review includes are below:

- Assessment Unit Content Development Specialist
    - Ensures that the items align to the intended EO, DOK level, and Nature of Science (if applicable). Ensures that the text of the item is appropriate for the item type and is not extraneous to the construct being measured. Checks for cuing and clanging throughout the item to ensure that the key is not obvious or cued. Reviews scoring information and verifies that computer-scored items are scoring correctly. Verifies that all metadata is complete and correct. Finally, ensures that the layout of the item is appropriate and complies with CMAS Science and Social Studies style guidelines.
- Standards and Instructional Support Content Specialist
    - Conducts a second check to verify EO alignment, DOK level, and Nature of Science (if applicable).
- Assessment Unit Special Education Specialist
    - Reviews items, including all graphics, to ensure that the items are accessible for all populations, including those students with visual impairments. Ensures that the item does not contain any extraneous text which would cause an undue burden on a student accessing the item via the text-to-speech platform. Conducts a UDR to ensure that items comply with the principles of universal design.
- Assessment Unit English Learner Specialist
    - Considers each item from the standpoint of an EL to ensure that the items do not contain any construct-irrelevant vocabulary words which would impede an EL's ability to access the item. Verifies that the item can be translated into Spanish for accommodated versions of the assessment.

These stakeholders at CDE record comments in ABBI with feedback which pertains to their review. This feedback is accompanied with a vote of, "Accept," "Accept with Edits," or "Reject."

- Items marked "Accept" require no further revisions. Those items are moved to "Ready for Content and Bias" status, indicating they are ready to be taken to the Content and Bias meeting.
- Items marked "Accept with Edits" are revised per CDE's feedback and, if necessary, reviewed again by the content editors/research librarians/UDR/Art at Pearson. These items are then reviewed again by CDE and reconciled with Pearson's content assessment specialist and deemed either "Accept" or "Reject."
- Items marked "Reject" are indicated "Do Not Use" in ABBI. The Pearson content assessment specialists write replacement items, which go through the same rigorous review process as other new items.

When CDE stakeholders have completed their review of the item, the CDE Assessment Unit Content Development Specialist reconciles all comments to ensure that all feedback is taken into consideration by Pearson for any required edits to an item. CDE alerts the Pearson content assessment specialists when this review is complete so that they may begin working on any required edits to the item.

Content and Bias Review

A Content and Bias committee, which includes Colorado educators with content expertise and special population expertise, review newly-developed items. The overarching purpose of the educator review is to (1) identify any potential bias or stereotype in test items, and (2) ensure the items are properly aligned to the content standards and GLEs, accurately measure intended content, and are grade appropriate. The Content and Bias Review checklist can be found in Appendix B.

The committee members are trained and instructed to verify that each stimulus and item (list is non-exhaustive):

- uses clear, unambiguous, and grade-level appropriate language;
- avoids complex sentence structure;
- uses everyday words to convey meaning when vocabulary is not part of the tested construct;
- has one correct answer (depending on the item type);
- contains plausible distractors that represent feasible misunderstandings of the content (depending on the item type);
- contains a rubric which is comprehensive of what students need to do to demonstrate understanding of the content (depending on the item type)
- represents the range of cognitive complexities and includes challenging items for students performing at all levels;
- is appropriate for students in the assigned grade in terms of reading level, vocabulary, interest, and experience;
- has scoring guidelines that capture exemplar responses at each score point (for CR items);
- includes appropriate and clear graphics/art/photos that are relevant to the item and are accessible to all testing populations;
- is free of ethnic, gender, political, religious and any other bias;
- avoids construct-irrelevant content that may unfairly advantage or disadvantage any student subgroup; and

- considers access issues at the time of item writing (example: determine how students with visual disabilities will access items with needed visuals/graphics/animation).

The committee makes one of three recommendations on every item based on the content and bias review: "Accept," "Accept with Edits," and "Reject."

- Items marked "Accept" require no further revisions. Those items are moved to "Ready for Field Test" status, indicating they are ready to be on the embedded field test.
- Items marked "Accept with Edits," are discussed as a group to determine what content improvements can be made to the item. Those content improvements are then noted by the Pearson content assessment specialist facilitating the meeting to be discussed with CDE during the reconciliation period.
- Items marked "Reject" are indicated "Do Not Use" in ABBI.

Following the educator meeting, a Content and Bias reconciliation period begins. The reconciliation period includes meetings with CDE staff and Pearson's content assessment specialists. At this meeting, committee comments are reviewed those committee comments are reconciled with regard to the intended outcome of the item, and educator feedback is taken into consideration for editing purposes. At this time, the items are edited with respect to the committee feedback by Pearson content assessment specialists. The items again move through the Pearson editing process and the CDE review process for approval to be placed on the embedded field test.

Embedded Field Test

Newly-developed field test items are embedded within the operational assessment forms that are administered to students. These items do not contribute to student scores. Embedded field testing allows Pearson and CDE to gather student response data about the functioning of all newly-developed items before determining whether they should be included as part of the operational assessment. Student responses to field test items inform the rangefinding process and provide item statistics which are used in data review.

Depending on the grade, between 8 and 11 field test forms were administered in 2019. Within a grade, each field test form was parallel; that is, each student received the same number of each item type and in the same location on the form. Table 55 shows the number of field test forms and field test items per grade.

Rangefinding

Rangefinding meetings are held following the administration in which an item is field tested. The purpose of rangefinding is to define the range of performance levels within the rubrics' score points using student responses.

Rangefinding committees include Colorado educators who are grade-level teachers with relevant content expertise, an educator with special education expertise, and an EL educator. Participants create consensus scores for student responses that are subsequently used to develop effective training materials for scoring of CR items.

Rangefinding is facilitated by Pearson's Scoring Services staff. CDE content assessment specialists, Pearson's item development staff, and Pearson Sales and Contract Management representatives are present for support.

During item development, rubrics are created which demonstrate, to the item writer's ability, the full range of student responses that would indicate that the student knows and understands the standard(s) in the CAS. During the field test, students may use additional or alternate methods to demonstrate knowledge of the standard. The rangefinding meeting aims to identify all possible student responses which would receive credit at each score point.

A rangefinding set is a collection of field test student responses from the Pearson Scoring Services team that are organized by presumed score point, based on the rubric, prior to the committee assigning a score. This helps to provide consistency for the committee, and possibly identify subtle differences in responses that may show the distinction between score points on the item.

For social studies items, Pearson's Scoring Directors construct one rangefinding set per item, which includes 30 responses. For science items, pre-constructed sets with additional responses are brought to the meeting. Responses included in these sets represent the full spectrum of scores to the greatest extent possible. For each item, the responses are ordered based on estimated score from high-scoring to low-scoring; however, actual scores are not revealed to committee members. Each set includes responses clearly earning each available score point for each type of question. The set also includes samples of responses that may have been challenging to score (i.e., the score points earned were not necessarily clear).

Following an introductory session presented by a member of Scoring Services group, the rangefinding committee is divided into grade and subject area breakout groups. Each group is assigned a range of field test items to be reviewed, following the process outlined below:

1. The Scoring Director introduces each item. The committee reviews the item and corresponding rubric.

2. The committee reads student responses—individually or as a group—and then discusses and decides the most appropriate score point for each response.

3. The Scoring Director records committee members' comments as well as the final consensus score for each student response. Consensus is reached when a majority of committee members agree upon a particular score point for a response and all members agree to accept the score of the majority.

4. A designated committee member records consensus scores. After reviewing responses for each item, the committee member compares his or her notes with those kept by the Scoring Director and provides sign-off to indicate agreement with the recorded scores.

Should additional information need to be included in the rubric at the conclusion of this process, the scoring director presents the proposed change to CDE for final approval. Field test scoring of human-scored items occurs at the conclusion of the Rangefinding meeting.

Data Review

Following the conclusion of scoring field tested items, a committee of educators convenes to review items, computer and human scored, along with student performance data for items that do not meet certain statistical criteria. Separate Data Review committees convene for CMAS Science and Social Studies. Data Review committee members are provided item images and metadata, along with classical statistics and differential item functioning (DIF) statistics.

Classical statistics include item means, item–total correlations, and distribution of responses across answer options or score points, depending on item type. Items are flagged based on several statistical criteria (e.g., very low or very high item mean, low item–total correlation, few or no students achieving a certain score point, etc.), and flagged items are taken to data review.

DIF analyses for CMAS Science and Social Studies items are conducted on various subgroups (gender, ethnicity, free and reduced lunch, IEP, and ELs) using Mantel–Haenszel Delta DIF statistics (Dorans & Holland, 1992).

Classification rules derived from National Assessment of Educational Progress (NAEP) guidelines (Allen, Carlson, & Zelenak, 1999) are used to classify items as having either negligible, moderate, or significant DIF. Items that are classified as moderate or significant DIF are taken to data review.

During the Data Review meeting, educators are trained to interpret the statistical information and judge the appropriateness of the items presented for data review. The committee members use the data as a tool to direct them toward potential content flaws in an item and discuss whether there were construct-irrelevant reasons for a data flag. A data flag, by itself, is not the sole reason an item is rejected. Committee members are instructed that their final judgments about the appropriateness or fairness of an item for any individuals and subgroups 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 made a recommendation as to whether to “accept” or “reject” it. An accepted item meant that the educators, through their varying expertise, determined that there was not a construct-irrelevant reason for the data flag within the item. 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 be issues such as language that is above grade-level or content that is biased against a particular group. Construct-relevant reasons for data flags could be simply difficult content that is part of the standards or distractors that

reflect a very common misunderstanding of the concept covered by the item. Following the meeting and CDE determinations, ABBI is updated by moving accepted items into "Ready for Operational" or "Do Not Use" status.

Item Banking Systems

ABBI is Pearson's proprietary software that supports the item and test development processes, from initial content authoring through the content review cycles. ABBI is the authoritative source for all content, data, and functionality for all CMAS system components.

ABBI serves as the repository where the item bank is housed, item revisions are catalogued, and assessment specialists upload and revise items and item metadata. Here, the items and associate stimuli (SIM and PE storyboards) are tracked and revisions are recorded from creation through retirement in a secure environment.

Custom reports can be generated out of ABBI. This feature allows content assessment specialists (and clients who have access to ABBI) to generate Excel reports that capture metadata (unique item number, EO, item type, DOK, associated stimulus, item status, item statistics and comments) useful for analyzing the item bank. ABBI is the source of reference for how and when changes to the item and the metadata have been implemented.

ABBI also supports the test construction process. "Forms" is the software component that allows the content specialists to build test forms in collaboration with the psychometricians. This is a cooperative and iterative process. The content assessment specialists and psychometricians work to construct test forms that meet blueprints, fall within the established statistical parameters, and adhere to the test design. The following information is available and visible in ABBI Forms:

- content information such as standard, EO, and DOK
- classical statistics such as item means and item-total correlations
- Item Response Theory (IRT) statistics such as difficulty, discrimination, guessing, and item-model fit

CHAPTER 3: TEST CONSTRUCTION

Pearson is responsible for the implementation and monitoring of all phases of the test construction process. Test forms are constructed through an iterative process between Pearson content and Pearson psychometric staff. CDE then reviews the test forms, provides feedback, and gives final approval as described below.

The assessment specialists select a set of operational items in accordance with the test blueprints and test construction specifications (see Figures 1–5 for test blueprints). Items selected for operational use must meet the test blueprint with a variety of topics and contexts with specified psychometric targets.

The following guidelines for CMAS Science and Social Studies are used during test form construction:

- review of the constructs and content included within each content strand (or reporting category) to establish that items address the breadth of content within each strand
- balance of gender, ethnicity, geographic regions, and relevant demographic factors
- thorough review of individual items to establish the content within items are up to date and relevant
- adherence to established test construction specifications and blueprints
- selection of items with various stimuli type throughout the test form to enhance the test-taker experience by providing variation in the appearance of item types presented
- efficient and deliberate use of varied content representative of the knowledge and skills in the CAS
- review of full form, including field test items, for instances of clueing and/or content overlap

After the initial operational item pull is complete, the test form is reviewed by two content assessment specialists. Each content assessment specialist verifies that the test form meets the test blueprint and the test construction specifications (i.e., the required EO coverage, DOK allocation, item type). The test form is then presented to psychometrics for analysis and the psychometrician verifies that the test form falls within the established psychometric and test blueprint parameters. The psychometric lead also identifies the anchor item set within each operational test form.

Once the test form is vetted internally, the test form is presented to CDE for review. If needed, the content assessment specialists, psychometricians, and CDE collaborate to finalize the test form. This can be an iterative process with the end result being CDE's test form approval.

After the operational test form is approved, field test items are selected from the items that were developed, reviewed, and accepted by CDE, and reviewed and accepted by the educator committees. Items chosen for field testing are placed on a test form in a designated section and

sequence. The content assessment specialists assemble field test sets of items so that they comprise the appropriate distribution of standards, item types, topic coverage, expected item difficulty, cognitive levels, and key distributions.

Online Forms

The majority of students take the CMAS Science and Social Studies assessment online. The online format facilitates the inclusion of innovative item types and various integrated accessibility features and accommodations as described in Chapter 4.

Accommodated Test Forms

Accommodated test forms for the CMAS Science and Social Studies assessments are available for both the online and paper-based test forms. For online test forms, text-to-speech and color contrast are available in English and Spanish. The various options for paper test forms are described below. Additionally, oral scripts in both English and Spanish are available for online and paper test forms. English oral scripts are also available for local translation into languages other than Spanish.

Paper

Paper-based versions of the CMAS Science and Social Studies assessments are available as an accommodation or for schools that choose not to test online as allowed by state law. A Spanish transadaptation is also available on paper.

The paper test form is parallel to the online test form. Parallel paper-based items were developed for TEIs. In some cases, this was achieved with traditional SR items and in other cases, items were presented in a manner that required human-scoring. For example, a drag-and-drop item may have been converted to an item that asked the student to draw lines from the draggers to the drop bays.

For the spring 2019 administration, the operational items on the paper-based version and Braille test form of the CMAS Science and Social Studies assessments were the same as the operational items on the online test forms.

Braille

After approval of the paper test materials, a braille version of the assessment was created according to the process outlined below:

1. Pearson posts final test forms as PDFs for National Braille Press (NBP).

2. NBP reviews the items for brailleability. During this review, translation concerns for text and graphics are noted.

3. Brailleability review report is provided to Pearson.

4. Pearson and CDE review and provide solutions for brailleability concerns.

5. NPB translates the test form into braille.

6. The braille form is proofread twice by a braille proofreader who is National Library Service certified or a certified transcriber.

7. Edits are made based on the proofreader's feedback.

8. The braille form is sent to Pearson.

9. The braille form is reviewed by a committee of Pearson staff, CDE staff, NBP staff, and Colorado Teachers of the Visually Impaired (TVIs) who are certified in braille.

10. Notes from the committee review are verified by CDE staff and are sent to NBP for updates to the braille form.

11. The braille form is finalized and printed.

Large Print

Large print versions of the CMAS Science and Social Studies assessments are also created. The large print versions are a 50 percent enlargement of the regular paper form and are printed on 14" × 18" paper. The large print versions include a Visual Description booklet. The Visual Description booklet contains a description of artwork (maps, photographs) for which it may be difficult for a student with visual impairments to see the subtleties within the art. In collaboration with Pearson content and accessibility experts, CDE reviews the paper test forms and identifies which pieces of art need to be described in the Visual Description Test Booklet.

CHAPTER 4: TEST ADMINISTRATION PROCEDURES

This chapter of the report provides information related to the CMAS Science and Social Studies administration procedures. Prior to the administration of the assessments, districts, schools, and teachers (Test Administrators) were to ensure that their students and systems were prepared for the assessments. Such information was communicated to the appropriate individuals via manuals and in-person and recorded trainings as described below.

Manuals

Several manuals were created to aid with the 2019 CMAS Science and Social Studies administration, described in the following sections.

CMAS Test Administrator Manual for Computer-based Testing and CMAS Test Administrator Manual for Paper-based Testing

These manuals describe the procedures Test Administrators were to follow when administering the online and paper CMAS assessments. Prior to administering any CMAS assessment, Test Administrators were to carefully read these manuals. Test administration policies and procedures were to be followed as written so that all testing conditions were uniform statewide. The guidelines and test administration scripts in these manuals were provided to ensure every student in Colorado received the same standard directions during the administration of the CMAS Science and Social Studies assessments.

PearsonAccess^{next} Online User Guide

This guide provides guidance for District Assessment Coordinators (DACs), School Assessment Coordinators (SACs), District Technology Coordinators (DTCs), Test Administrators, and Student Enrollment personnel who utilize PearsonAccess^{next}.

CMAS and CoAlt Procedures Manual

This manual provides instructions for the coordination of the CMAS assessments. Instructions include the protocols that all school staff were to follow related to test security and test administration and providing accommodations and accessibility features to students with disabilities and ELs. The manual also includes the tasks that were to be completed by DACs, SACs, DTCs, and data specialists before, during, and after test administration.

Training

Administration training is intended to make sure all individuals involved in CMAS 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. Fidelity to standardized test administration processes and procedures helps to ensure the comparability of resulting scores and accurate interpretation of results. Thorough in-person regional trainings were conducted by CDE and Pearson personnel across the state. CDE and Pearson presented trainings to the DACs that contained information regarding proper procedures

for administration, security requirements, receiving and returning materials to Pearson, and the use of PearsonAccess^{next} with TestNav 8. Additionally, recorded versions of the live trainings were posted on the CDE Assessment Unit website. Administration training materials, including slide decks, manuals, and how-to guides were also available on the CDE Assessment Unit website for training SACs and Test Administrators. After CDE trained DACs, the DACs trained School Assessment Coordinators, Test Administrators, and any other individuals within the district who planned to participate in the 2019 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 Assessment Unit and Pearson Program Team 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 the Department.

On-site Preparation

Districts were instructed in site readiness preparations, TestNav, proctor caching, and use of the SystemCheck tool to configure their testing technology environment and evaluate their configuration for district readiness.

Districts were also provided with tools and resources to test their environment readiness status. Issues identified from site readiness evaluations were assessed by Pearson and CDE and appropriate corrective actions were developed and communicated to affected districts.

Accessibility and Accommodations

Accessibility features and accommodations provided in 2019 were consistent with those offered to students in 2018. Accessibility was considered from the beginning of the test development process and was inherent within the CMAS Science and Social Studies assessments and administration. For example, the CMAS Science and Social Studies online test engine, TestNav 8, includes tools and accessibility features, such as a text highlighter, that were made available to all students to increase the accessibility of the assessments. Also included was the text-to-speech accessibility feature, which allowed for text to be read to students by means of the embedded software audio feature. Although the accessibility features of text-to-speech and online color contrast were available to all students in the content areas of science and social studies, only those who needed text-to-speech or color contrast were assigned to these accessibility features in advance of testing.

Beyond the tools and accessibility features that were available for all students, assessment accommodations were available to the population of students who had IEP, 504, or EL plans. 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 individual 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.

Accommodations must be documented 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.

Some of the available accommodations included English oral scripts, Spanish oral scripts, oral scripts for signed presentation and local translation into languages other than English and Spanish, braille test forms, large print test forms, and Spanish test forms with and without text-to-speech.

Live webinar accommodations and accessibility features training was conducted by CDE for district level personnel. The intent of this training was to ensure all individuals providing these supports across the state follow the procedures associated with each accommodation and accessibility feature. Providing accessibility features and accommodations in a standardized manner helps to ensure the comparability of resulting scores and accurate interpretation of results. A recorded version of the live training, slide decks, and procedural information (*Section 6.0* of the *CMAS and CoAlt Procedures Manual*) were available on the CDE Assessment Unit website for training SACs and Test Administrators.

Test Security

Procedures described in this section were put in place to enhance the likelihood that security was maintained before, during, and after assessment administration. Materials used during the paper administration of the assessment were to be kept in locked storage locations when not under the direct supervision of Pearson or approved testing coordinators and administrators. All district and school personnel involved in the assessment administration were required to participate in annual local training on the CMAS Science and Social Studies assessments. DACs were responsible for overseeing training for the district, including verifying that the DTC and SACs were trained. SACs were responsible for ensuring that Test Administrators, Test Examiners, and all individuals involved in test administration at the school level were trained and subsequently acted in accordance with all security requirements. A chain of custody plan for materials was required to be written and implemented to ensure materials were securely distributed from DACs to SACs to Test Administrators/Test Examiners and securely returned from Test Administrators/Test Examiners to SACs and then to DACs. SACs were required to distribute materials to and collect materials from Test Administrators/Test Examiners each day of testing, and securely store and deliver materials to DACs after testing was completed in accordance with the instructions in the 2019 *CMAS and CoAlt Procedures Manual*.

All individuals involved in the administration of the assessments were required to sign a security agreement prior to handling test materials, which required them to follow all procedures set forth in the aforementioned manuals and prevented them from divulging the contents of the assessment, copying any part of the assessment, reviewing test questions with the students, allowing students to remove test materials from the room where testing was to take place, or interfering with the independent work of any student taking the assessment. During online testing, all computer functions not necessary to complete the test were disabled, and access was restricted to disallow activities in all applications outside the testing program.

The PearsonAccess^{next} online administration platform used during the administration included permissions-based user role access to all information within the system including accessing student information, setting up and delivering test sessions, administering tests, and accessing reports. Access to online assessments was tightly controlled before, during and after test administration, requiring a login ID and password to enter the system for each unit. Test content was locked and could not be accessed by students or district/school level user after the students submitted their answers. Each unit of the paper test required students to break the unit seal before accessing the test content. To enhance security during test administration, assessment forms were spiraled at the student level, decreasing the likelihood that a student would be working on the same items as their peers at the same time.

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

CHAPTER 5: CONSTRUCTED-RESPONSE SCORING

CR items are scored using holistic rubrics, which are generated for each unique item and finalized at rangefinding. The finalized rubrics, along with the training materials for each item, are maintained by Pearson's Scoring Services group.

Each operational exam is scored using either a Distributed or Regional Scoring model depending upon content area. Scoring includes several components that together provide a comprehensive performance scoring model.

Scorers are trained using comprehensive training materials developed by scoring experts. These materials include student responses scored by Colorado educators at the rangefinding meetings.

- Following the rangefinding meetings, Scoring Services' personnel create training material with an anchor set (up to 12 responses) and a full practice set (up to 10 responses). Each CR item is then scored with the associated training material.

- Scorers must pass a qualifying test for the item types that they will score.

- Student responses are converted to electronic images at Pearson facilities. They are then transmitted for computer-based scoring.

- Distributed scorers are located across the United States and work from their homes. Their computers are set up for image-based scoring. A comprehensive set of scoring and monitoring tools is integrated into the scoring system. In addition to the systematic tools, content supervisory staff is available by phone to help answer any training or scoring related questions that may arise. With distributed scoring, scorers may score 7 days per week with extended evening hours.

- Regional scorers are located within a physical scoring location site. As with distributed scoring, regional scoring also utilizes a comprehensive set of scoring and monitoring tools integrated into the scoring system. In addition to the systematic tools, content supervisory staff is physically on site to help answer any training or scoring related questions that may arise. Unlike distributed scoring, regional scoring is traditionally only offered Monday through Friday during normal business hours.

Pearson's processes and tools provide a replicable quality system that strengthens consistency across projects and locations within Pearson's Scoring Services operations. Pearson's Scoring Services team uses a comprehensive system for continually monitoring and maintaining the accuracy of scoring on both group and individual levels. This system includes daily analysis of a comprehensive set of statistical monitoring reports as well as regular "backreading" of scorers. Reliability statistics are monitored during scoring and interventions are applied if a scorer or item is not meeting minimum requirements. A detailed description of these measures is included in Chapter 9.

Embedded field test scoring was completed using regional scorers. Regional field test scoring took place in San Antonio, TX and Mesa, AZ for science and Austin, TX for social studies. All scorers were required to have a four-year college degree.

Backreading

Backreading is the method of immediately monitoring a scorer's performance, and is, therefore, an important tool for Pearson's scoring supervisors. Backreading is performed in conjunction with the statistics provided by reader performance reports and as indicated by scoring directors, allowing scoring supervisors to target particular readers and areas of concern. Scorers showing low inter-rater agreement or those showing anomalous frequency distributions are given immediate, constructive feedback and monitored closely until sufficient improvement is demonstrated. Scorers who demonstrate through their agreement rates and frequency distributions that they are scoring accurately will continue to be spot-checked as an added confirmation of their accuracy. An explanation of rater agreement statistics can be found in Chapter 9 and rater agreement statistics for the spring 2019 administration can be found in Part II of this report. The CMAS Science and Social Studies agreement rate requirements are as follows:

- 2-point item: 90% perfect and 95% perfect plus adjacent agreement
- 3-point item: 80% perfect and 95% perfect plus adjacent agreement

Calibration

Calibration sets are responses selected as examples that help clarify particular scoring issues, define more clearly the lines between certain score points, and reinforce the scoring guidelines as presented in the original training sets. They can be applied to groups, a subset of groups, or individual scorers, as needed. These sets are used to proactively promote accuracy by exploring project-specific issues, score boundaries, or types of responses that are particularly challenging to score consistently. Scoring directors administer calibration sets as needed, particularly for more difficult items.

CHAPTER 6: STANDARD SETTING

To support the interpretation of student results, student performance on the CMAS Science and Social Studies assessments is described in terms of four performance levels: Partially Met Expectations, Approached Expectations, Met Expectations, and Exceeded Expectations. The performance level labels were updated in 2015-2016 to match the labels used for the CMAS Math and English Language Arts (ELA) assessments. Only the performance level labels were updated, the Performance Level Descriptors and cut scores set during standard setting were not changed. Performance standards were set for grades 4 and 7 Social Studies and grades 5 and 8 Science after the first operational administration in spring 2014. Details of the Elementary/Middle School standard setting can be found in the *2013–2014 CMAS Technical Report*. Performance standards were set for High School Science after the first operational administration in fall 2014. Details of the High School Science standard setting can be found in the *2014–2015 CMAS Technical Report*.

CHAPTER 7: REPORTING

Several score reports are generated to communicate student performance on the CMAS Science and Social Studies assessments. The reports contain a variety of score types at different levels of the blueprint, as described in this section. For additional details on score reports, see the *CMAS and CoAlt Interpretive Guide 2019* at https://www.cde.state.co.us/assessment/cmas_coalt_interpretiveguide_2019.

Description of Scores

CMAS Science and Social Studies reports provide information on student performance in terms of scale scores, performance levels, and percent correct scores.

Scale Scores

A scale score is a conversion of a student's response pattern to a common scale that allows for a numerical comparison between students. Scale scores are particularly useful for comparing test scores over time and creating comparable scores when a test has multiple forms. For CMAS Science and Social Studies, students receive scale scores in each of the following areas.

- Overall test
- Content Standards
    - Science: Physical Science, Life Science, and Earth Systems Science
    - Social Studies: History, Geography, Economics, and Civics
- Scientific Investigation and the Nature of Science (for science assessments only)
- Selected-Response and Technology-Enhanced items
- Constructed-Response items

Each of these scales range from 300 to 900. Chapter 8 provides technical details related to scale development for CMAS Science and Social Studies.

Performance Levels

Performance levels are reported at the overall assessment level. Examinees are classified into performance levels based on their scale score as compared with the cut scores, which were obtained from standard settings. There are four performance levels: Partially Met Expectations, Approached Expectations, Met Expectations, and Exceeded Expectations.

Percent Earned

Percent earned scores are provided at the PGC and GLE levels. The percent of points earned refers to the number of points a student earned out of the total number of points possible within a reporting category. Unlike scale scores, percent earned scores cannot be compared across years because individual items change from year to year. In addition, they cannot be compared across PGCs or GLEs because the number of items and the difficulty of the items may not be the same.

Score Reports

Sample score reports can be found in Appendix C. Two types of score reports are provided: student level and aggregate.

Student Performance Reports

Student Performance Reports provide information about the performance of a particular student. The student's various scale scores, associated performance level, and percent earned scores are displayed on a four-page report along with comparative information related to the student's school, district, and state performance. In addition, performance level descriptions are provided.

Two copies of Student Performance Reports are printed and shipped to districts.

Aggregate Reports

Three types of aggregate reports are produced:

- Content Standards Report
- School Performance Level Summary
- Item Analysis Report

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} Published Reports and access to the reports is limited to authorized users.

CHAPTER 8: CALIBRATION, EQUATING, AND SCALING

IRT was used to develop, calibrate, equate, and scale the CMAS Science and Social Studies assessments. The two-parameter logistic (2PL) (Birmbaum, 1968), three-parameter logistic (3PL) (Birmbaum, 1968), and generalized partial credit (GPC) (Muraki, 1992) were applied to CMAS Science and Social Studies. These measurement models are routinely used for test forms construction, calibration, scaling and equating, and maintaining and building item banks. All test analyses, including calibration, scaling, and item-model fit, were accomplished within the IRT framework. SR items were fit to the 3PL model, TEIs were fit to either the 2PL or 3PL model depending on the guessing factor of the item, and CR items were fit to the GPC model. IRTPRO (SSI, Inc., 2011) was used for calibration and the calibration of the first operational administration determined the base scale. The program STUIRT (Kim & Kolen, 2004) was used to obtain the Stocking and Lord transformation constants for equating purposes.

Calibration

The 2PL, 3PL, and GPC IRT models

The item response function (IRF) of the 2PL, 3PL, and GPC IRT models relates examinee ability to the probability of observing a particular item response given the item's characteristics. The item characteristic function (ICF) relates examinee ability to the expected examinee score. The 2PL model (Birmbaum, 1968), uses two item parameters to relate the probability of person *i* correctly answering a dichotomously scored item *j*:

$$P_{ij}(\theta) = \frac{1}{1 + \exp\left[-Da_j\left(\theta_i - b_j\right)\right]},$$

where D is set equal to 1 when defined on the logistic scale, as IRTPRO parameterizes all models. The item discrimination parameter is a_j; and the item difficulty parameter is b_j. The 3PL model (Birmbaum, 1968) adds an item parameter to the model:

$$P_{ij}(\theta) = c_j + \frac{1 - c_j}{1 + \exp\left[-Da_j\left(\theta_i - b_j\right)\right]},$$

where c_j is the item pseudo-guessing parameter.

The GPC model (Muraki, 1992) has three item parameters to relate the probability of person *i* responding in the *x*-th category, to a polytomous scored item *j*:

$$P_{ij}(\theta) = \frac{\exp\left[\sum_{v=0}^{x} Da_j\left(\theta - b_j + d_{jv}\right)\right]}{\sum_{k=0}^{M_i} \exp\left[\sum_{v=0}^{k} Da_j\left(\theta - b_j + d_{jv}\right)\right]}, x = 0, 1, \ldots, M_i,$$

where all parameters are as they were before and d_{jv} is the category parameter for category *v* of item *j* and M_i is the maximum score on item *j*.

The graphical representation of the IRF and ICF are the item response curves (IRC) and item characteristic curves (ICC), respectively. For dichotomous items the IRF and ICF are equal, but for polytomous items the IRC and ICF are different.

As an example, consider Figure 6, which depicts a 2PL item that falls at approximately 0.85 on the ability (horizontal) scale. When a person answers an item at the same level as their ability, then that person has a roughly 50% probability of answering the item correctly. Another way of expressing this is that in a group of 100 people, all of whom have an ability of 0.85, about 50% of the people would be expected to answer the item correctly. A person whose ability was above 0.85 would have a higher probability of getting the item right, while a person whose ability is below 0.85 would have a lower probability of getting the item right.

Figure 7 shows IRCs of obtaining a wrong answer or a right answer. The dotted-line curve (j=0) shows the probability of getting a score of "0" while the solid-line curve (j=1) shows the probability of getting a score of "1." The point at which the two curves cross indicates the transition point on the ability scale where the most likely response changes from a "0" to a "1." At this intersection, the probability of answering the item correctly is 50 percent.

Figure 8 shows IRCs of obtaining each score category for a polytomously scored item. The dotted-line curve (j=0) shows the probability of getting a score of "0." Those of very low ability (e.g., below -2) are very likely to be in this category and, in fact, are more likely to be in this category than the other two. Those receiving a "1" (partial credit) tend to fall in the middle range of abilities (the thick, solid-line curve, j=1). The final, thin, solid-line curve (j=2) represents the probability for those receiving scores of "2" (completely correct). Very high-ability people are more likely to be in this category than in any other, but there are still some of average and low ability who can get full credit for the item.

The points at which lines cross have a similar interpretation as that for dichotomous items. For abilities to the left of (or less than) the point at which the j=0 line crosses the j=1 line, indicated by the left arrow, the probability is greatest for a "0" response. To the right of (or above) this point, and up to the point at which the j=1 and j=2 lines cross (marked by the right arrow), the most likely response is a "1". For abilities to the right of this point, the most likely response is a "2." Note that the probability of scoring a "1" response (j=1) declines in both directions as ability decreases to the low extreme and increases to the high extreme. These points then may be thought of as the difficulties of crossing the *thresholds* between categories.

Item Fit

Item fit is evaluated using Yen's (1981) Q_1 statistic. The Q_1 statistic allows for the evaluation of an item's IRT model fit to observed student performance. In the calculations of Q_1, the observed and expected (based on the model) frequencies were compared at 10 intervals, deciles, along the scale. Yen's Q_1 fit statistic was computed for each item using the following formula:

$$Q_{1_i} = \sum_{j=1}^{10} \frac{N_{ij}(O_{ij}-E_{ij})^2}{E_{ij}(1-E_{ij})},$$

where N_{ji} is the number of students in interval j for item i, and O_{ij} and E_{ij} are the observed and expected proportions of students in interval j for item i.

The Q_1 then was transformed so that the value could be evaluated using the chi-square distribution:

$$Z_{Q_{1_i}} = \frac{Q_{1_i} - df}{\sqrt{2df}},$$

where *df* is the degree of freedom for the statistic (*df* = 10–the number of parameters estimated; *df* = 7 for SR items in a 3PL model). If $Z_{Q_{1_i}}$ is greater than Z_{crit} then the item is flagged for "poor" model fit:

$$Z_{crit} = \frac{N_i * 4}{1500},$$

where N_i is the sample size.

Equating and Scaling

Equating of operational test forms involves adjusting for differences in the difficulty of test forms, both within and across assessment administrations. Equating makes certain that students taking one form of a test were neither advantaged nor disadvantaged when compared to students taking a different test form. Each time a new test form is constructed, equating is used to allow scores on the new test form to be comparable to scores on the previous test form.

Calibration is used to obtain item parameter estimates and, in the process, puts all items and examinees on a common scale. A scale transformation can then be applied to create meaningful scale scores.

Operational Equating and Scaling

Equating is used to place new test forms on the operational scale. Spring 2014 and fall 2014 were the first operational administrations for the Elementary/Middle School and High School Science assessments, respectively, so those administrations were used to establish the base scale for the assessments. For spring 2019, equating was used to place the spring 2019 test forms on the 2014 base scale.

Calibrations

In order to obtain item parameter estimates, the 2PL, 3PL, or GPC model was applied to the items. SR items were fit to the 3PL model; TEIs were fit to either the 2PL or 3PL model, depending on whether the guessing factor was higher or lower than .05, respectively; and CR items were fit to the GPC model. IRTPRO (SSI, Inc., 2011) was used for all calibrations.

Anchor Items

A common items approach is used for equating operational test forms for the CMAS Science and Social Studies assessments. Test forms from adjacent administrations contain a set of items that are the same across the two administrations. This set of items represents the test blueprint in terms of content and represents roughly 30% of a full test form. Due to the relatively high percentage of points coming from CR items, both SR and CR items are included in the anchor set.

Consistency of Constructed-Response Scoring Check

The CMAS Science and Social Studies assessments include a high percentage of CR items and therefore, to be more reflective of the construct being measured, the anchor sets include CR items. For accurate equating, it is important that the items in the anchor sets be consistently scored across administrations. With SR items, scoring is exactly the same each time the item is administered (e.g., answer option 'A' is always scored as the correct answer) such that changes in item performance across administrations can be solely attributed to changes in student performance. With CR items, scoring is done by human raters so it is important that scoring be monitored both within an administration and across administrations to maintain consistent scoring throughout. Such procedures are in place including consistency in training and the use of validity papers throughout scoring. As an additional check, prior to equating, the consistency of the CR scoring was examined via the rescoring of a subset of the previous year's papers to remove any items that exhibit statistics drift in scoring characteristics so that the accuracy of the equating is not jeopardized. If a CR item appeared to lack consistency across the administrations, considerations were given to removing the item from the anchor set.

Stability Check

The item parameter stability check for the anchor items was conducted using classical item analyses, scatter plots of item parameter estimates, and ICC comparison. For the ICC comparison, old and new ICCs were compared using the *z*-score approach based on D^2 (Wells, Hambleton, Kirkpatrick, & Meng, 2011) as outlined below:

1. Obtain the theoretically weighted estimated posterior theta distribution using 31 quadrature points (-3 to 3).
2. Compute the slope and intercept constants using Stocking & Lord in STUIRT with all anchor items in the linking set.
3. Place the original anchor item parameter estimates onto the baseline scale by applying the constants obtained in Step 2.
4. For each anchor item, calculate D^2 between the ICCs based on old (x) and new (y) parameters at each point in this theta distribution:

$$D_i^2 = \sum^{k} \left[P_{ix}(\theta_k) - P_{iy}(\theta_k)\right]^2 \bullet g(\theta_k)$$

   where i = item, x = old form, y = new form, k = theta quadrature point, and g = theoretically weighted posterior theta distribution.
5. Compute the mean and standard deviation of the D^2 values.

6. Flag the items with a D^2 more than 2 standard deviations above the mean.

Final Anchor Sets

Items flagged from the consistency of constructed-response scoring check and 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 linking 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.

STUIRT

Using the item parameter estimates for the anchor set from the previous year and the current year, the program STUIRT (Kim & Kolen, 2004) was used to obtain the Stocking and Lord transformation constants to place the current administration's items on the operational scale. The scale transformation constants, slope A and intercept B, were applied to the item parameter estimates to place the new test items (new, N) on the operational scale (old, O) (Kolen & Brennan, 2004).

$$\alpha_{jO} = \alpha_{jN}/A$$

$$b_{jO} = A * b_{jN} + B$$

$$c_{jN} = c_{jN}$$

$$d_{jvO} = A * d_{jvN}$$

Paper Forms

Online and paper items were developed to be nearly identical except for a very small number of items. Operational paper items deemed identical to operational online items were assumed to have the same item parameter estimates. IRTPRO was used to estimate paper items with no online counterparts, while paper items with an online counterpart were fixed to their online counterparts' item parameter estimates. This process produced item parameter estimates for all paper items.

Comparability of Paper and Online Test Forms

The scale score distributions were examined using a matched samples approach to investigate the extent to which the online and paper test forms produce comparable scores. Multiple variables were used for determining the matched groups to result in "equal" groups of paper and online examinees. The variables included sex, race/ethnicity, free/reduced lunch eligibility, language proficiency, IEP, district setting, and past test scores (CMAS ELA for Elementary/Middle School Social Studies, CMAS Math for Elementary/Middle School Science, and PSAT for High School Science). Scale score distributions of CMAS scores between the matched samples were

compared to quantify the mode effect. To quantify the differences between the two distributions, the effect size of the differences between the two distributions was calculated:

$$d = \frac{M_{group1} - M_{group2}}{SD_{pooled}}$$

Suggested interpretations of Cohen's d (Cohen, 1977) are as follows:

.2 = a 'small' effect size
.5 = a 'medium' effect size
.8 = a 'large' effect size

A threshold for a possible mode effect was set of an effect size of .1 or greater and a matched sample size of at least 1,000 students. The number of students taking the paper test form can be found in Table 23. Based on the paper sample sizes, the effect size was calculated for all three science grades and grade 4 social studies. CDE made the final decision on whether to make an adjustment for mode differences for each assessment.

For assessments where an adjustment was deemed necessary, scores from the paper test form were adjusted using a linear transformation to match the mean and standard deviation of the online form. The conversion was applied to the overall, reporting category, and item type scores.

Field Test Equating

The process for field test equating is similar to that of operational equating although instead of an anchor set being used to equate, the operational items are used as anchors to place the field test items on the operational base scale.

Ability Estimates

Examinee ability was estimated using IRT pattern scoring based on examinee responses and the operational item parameter estimates. Examinee ability was estimated at the overall test level and at each subscale. Estimates were obtained via the maximum likelihood method (MLE) applied within the ISE software program (Chien & Shin, 2012). Pattern scores use the examinee's response (overall or subscale) to determine his or her ability estimate, which may lead to different theta estimates for the same raw score. One item on the grade 5 science assessment was removed and suppressed from scoring before determining examinee ability estimates.

Overall and Subscale Scale Scores

Examinee ability estimates were then transformed to scale scores ranging from 300 to 900 with a mean of 600 and standard deviation of 100. This was done not only at the overall test level but also for the following subscales:

- Content Standards
    - Science: Physical Science, Life Science, and Earth Systems Science
    - Social Studies: History, Geography, Economics, and Civics

- Scientific Investigation and the Nature of Science (for science only)
- Selected-Response and Technology Enhanced items
- Constructed-Response items

The following linear transformation was used to convert examinee theta estimates into scaled scores:

$$SS = 100 * \theta + 600$$

LOSS and HOSS were set to 300 and 900, respectively, for each scale.

Steps in the Calibration, Equating, and Scaling Process

The calibration, equating, and scaling process was repeated for each subject/grade. All steps were independently replicated by at least two members of the Pearson psychometric team to ensure the accuracy of the processes.

Data Preparation

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

- A traditional item analysis (TRIAN) and adjudication were completed on all items.
- The data file containing student responses was verified and exclusion rules were applied.
- Incomplete data matrices (IDMs) were created.

A TRIAN of all SR items was conducted prior to calibration. The purpose of this review is to use classical statistics to identify potential test administration and score issues. Specifically, SR items having one or more of the following characteristics are flagged:

- P-value <= 0.15
- Item-total score correlation < 0.20
- Incorrect option selected by 40 percent or more examinees

A list of flagged items is communicated to the content specialists for review and confirmation that the correct key has been applied. A sample TRIAN report is provided in Figure 9.

All TEIs are put through an adjudication process. For each item, the frequency distribution of responses that are scored correctly is created along with the frequency distribution of responses that are scored as incorrect. Content specialists review each response in the frequency reports

and indicate whether the response should be scored as correct. The content specialists' indications are then cross-referenced with how the responses are scored to confirm that scoring is accurate. A sample adjudication spreadsheet is provided in Figure 10.

Calibration, Equating, and Scaling

For the spring 2019 administration, several different analyses were done to obtain item parameter estimates for online operational and field test items and ability estimates for examinees.

- Online operational items
    - Used IRTPRO control files and IDM to obtain online operational item parameter estimates
    - Used ISE to estimate student abilities
    - Calculated item fit statistics and plotted expected vs. observed IRFs for each operational item
    - Evaluated consistency of scoring and stability of anchor items
    - Used STUIRT to scale 2019 operational items to operational base scale
- Online field test items
    - Used IRTPRO control files and IDM to obtain item parameter estimates of operational and field test items
    - Used STUIRT to scale field test items to operational base scale using the online operational items as the anchor set
    - Calculated item fit statistics and plotted expected vs. observed IRFs for field test item

CHAPTER 9: RELIABILITY

A variety of statistics can be calculated that pertain to the reliability of the CMAS Science and Social Studies assessments. In this report, Cronbach's alpha, standard error of measurement (SEM), conditional standard error of measurement (CSEM), decision consistency and accuracy, and inter-rater agreement are provided as described below. For these statistical estimates for the spring 2019 administration, see Part II of this document.

Cronbach's 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.

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). There are several methods for estimating reliability; however, in this report, an internal consistency method is used. In this method, a single test form is administered to the same group of subjects to determine whether examinees 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 (two score values) and polytomous (two or more score values) test items and is computed using the following formula:

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

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.

Cronbach's alpha ranges in value from 0.0 to 1.0, where higher values indicate a greater proportion of observed score variance is true 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 examinees will respond consistently across items within the test.

For CMAS Science and Social Studies, coefficient alpha estimates are provided for the overall test as well as each subscale (see Tables 3–8). Given the differences in length, it is expected that the coefficient alpha for the overall test will be higher than that of the subscales.

Standard Error of Measurement

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

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

where s_x is the standard deviation of test scores and

$\rho_{XX'}$ is the reliability coefficient.

There is an inverse relationship between the reliability coefficient (e.g., alpha) and SEM: the higher the reliability, the lower the SEM. SEMs can be found in Table 9.

Conditional Standard Error of Measurement

While the SEM provides an estimate of precision for an assessment, the CSEMs consider how measurement error likely varies across the scale score. For example, the CMAS Science and Social Studies assessments likely more accurately measure a student who scores a 600 (near the middle of the scale) than a student who scores either a 400 or an 800 (at the ends of the scale). During test construction, CSEMs are reviewed to ensure that they are minimized around the performance level cut scores.

The CSEM is defined as the standard deviation of observed scores given a particular true score and can be estimated using IRT. Plots of the CSEMs across the scale score are provided in Appendix D.

Decision Consistency and Accuracy

The CMAS Science and Social Studies scales are divided into four performance levels: Partially Met Expectations, Approached Expectations, Met Expectations, and Exceeded Expectations. 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 CMAS Science and Social Studies. For the overall test, consistency and accuracy estimates along with PChance and Cohen's Kappa (κ) coefficient (Cohen, 1960) are provided in Table 10 according to the following equation:

$$\kappa = \frac{P - P_c}{1 - P_c},$$

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

In addition, consistency and accuracy estimates at each cut score are provided in Tables 11 and 12.

Inter-Rater Agreement

For CR items, an additional form of reliability is assessed. Inter-rater agreement examines the extent to which examinees would obtain the same score if scored by different scorers. The following analyses will be conducted for each CR item, where R_1 is the first rater and R_2 is the second rater of the analyses.

1. Agreement rates
    a. Exact, which represents exact agreement between the two raters.
    b. Adjacent, which represents adjacent agreement between the two raters (i.e., a difference of 1 score point).
    c. Non-adjacent, which represents a difference of more than 1 score point between the two raters.

2. Quadratic kappa (Kappa)

    $KAPPA = \frac{E([X_1 - Y_1]^2)}{E([X_1 - Y_2]^2)}$, which is a comparison between the mean square error of rating pairs that are supposed to agree (X_1, Y_1) and those that are unrelated (X_1, Y_2).

3. Standardized mean differences (MD)

$$\bar{Z} = \frac{\left|\bar{X}_{R_1} - \bar{X}_{R_2}\right|}{\sqrt{\frac{sd_{R_1}^2 + sd_{R_2}^2}{2}}}$$

4. Correlations (CORR)

$$r_{R_1,R_2} = \frac{cov(R_1, R_2)}{sd_{R_1} * sd_{R_2}}$$

See Tables 13–22 for rater agreement statistics.

CHAPTER 10: VALIDITY

"Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, NCME, 2014). As such, it is not the CMAS Science and Social Studies assessments that are validated but rather the interpretations of the CMAS Science and Social Studies scores. The purpose of the CMAS Science and Social Studies assessments is to provide information about a student's level of mastery of the CAS. The CAS were designed such that mastery of the high school level standards should mean that a student is college and career ready. Mastery of the standards in the elementary and middle school grades indicates that a student is on track to being college and career ready at each grade level. In support of these ends, the previous chapters of this report described processes that were implemented throughout the CMAS Science and Social Studies assessment cycle with validity and fairness considerations in mind. This chapter provides information regarding specific sources of validity evidence as well as fairness.

Sources of Validity Evidence

The following sections describe various sources of validity evidence as outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).

Evidence Based on Test Content

It is important to examine the extent to which the items on an assessment measure the intended construct. The CMAS Science and Social Studies assessments intend to measure the CAS. The CAS are organized by standards (i.e., history, geography, civics, and economics). There are several levels of specificity below the standards. The first are the PGCs that represent the concepts and skills students need to master in order to be college and career ready by the time of graduation. Below PGCs are GLEs. GLEs are grade-specific expectations that indicate that students are making progress toward the PGCs. A number of EOs are included within each GLE to direct instruction toward particular topics and to provide more specific examples of topics for the GLE. As outlined in Chapter 2 of this report, targeted steps are included throughout the assessment development process to ensure that assessment items are appropriately measuring the CAS. For example, each item undergoes numerous reviews to confirm that it adequately aligns to the EO that it is intended to measure. In addition, with the field testing of items, DIF analyses are conducted 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 (e.g., spread by item type, SIM/PE associated items, DOK) in an attempt to most optimally measure the CAS.

In addition to these aforementioned internal processes, a formal alignment study was conducted by HumRRO in the fall of 2015. The overall results showed a strong alignment between the CAS and the content of the assessments but found some issues with alignment of DOK and the breadth of the High School Science assessment. It is important to note, however, that when the PGC was considered as the level of analysis rather than the GLE, the high school science issues discussed within the alignment study were resolved. The DOK issue was also resolved at the GLE level in later versions of the high school science assessment. In 2019, HumRRO calculated

the range criterion ratings based on the number of GLEs per PGC and found that all PGCs were adequately assessed.

Unlike the grade-specific Elementary/Middle School Colorado academic science standards, the High School standards identify the skills and concepts students are to have mastered by the end of high school. In order to avoid intruding on locally controlled scope and sequence determinations and to avoid signaling to the field that students should not receive instruction in one or more of the science domains, Colorado determined that the High School Science assessments would be comprehensive. The exams cover all three domains: Physical Science, Life Science and Earth Systems. In order to create assessments of reasonable length, in-spite-of the fact that the assessments essentially cover three years of content, the PGCs were intentionally made the focus of the assessments rather than the GLEs. The full alignment study report can be found in the *2015–2016 CMAS Technical Report*. HumRRO's revised CMAS High School Science calculations based on PGC as the level of analysis can be found in Appendix E.

Evidence Based on Response Processes

Evidence based on response processes pertains to the cognitive aspect behind how students respond to items. Since the CMAS Science and Social Studies assessments are Colorado's first online assessment, cognitive labs were held during the initial item development phase to evaluate whether students would find the nature of CMAS Science and Social Studies items (e.g., SIMs) or Pearson's TestNav browser-based testing platform challenging. In addition, the CMAS Science and Social Studies assessments include TEIs, a relatively new item type. To validate that students are responding as expected and that items are being scored as expected, an adjudication process is conducted for all TEIs once they have been administered.

Cognitive Labs

Cognitive labs were conducted with Colorado students in May 2013. Students were sampled from rural, urban, and suburban schools and asked to take between 7 and 16 items, depending on grade and subject. Students attempted a variety of item types on the TestNav platform and were asked to "think-aloud" as they worked through each item.

Students showed a high degree of facility in responding to the items, and only a small bit of supplemental training was speculated to be needed to acquaint them with the tools and navigation of the TestNav interface. Surveys were given to the students after completion of the assessment, which included a question that asked them to indicate whether they preferred paper or computer-based tests. The majority of students indicated that they preferred the computer-based version, and many commented that it had been an enjoyable experience. For a full report on the cognitive labs, see the *2013–2014 CMAS Technical Report*.

Adjudication

Since the CMAS Science and Social Studies assessments contain TEIs, it is important to validate that students are responding to the items as intended and that the scoring is accurate. As described in Chapter 8, every response for every TEI is reviewed by a content specialist to confirm that scoring is accurate. In addition, the adjudication indicates the frequency with which

each response was provided, which would likely identify any items where students were not interacting with the item as intended.

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. To analyze the internal structure of the CMAS Science and Social Studies assessments, a factor analysis is performed and scree plots are examined to investigate the number of dimensions that the CMAS Science and Social Studies assessments appear to be measuring. Given that a unidimensional IRT model is used for calibration and scaling, it is important that there be evidence to support its use. Scree plots for the spring 2019 administration can be found in Part II of this report. The scree plots for all assessments clearly support the use of a unidimensional IRT model.

Evidence Based on Relationships to Other Variables

It is important to explore the relationship between CMAS Science and Social Studies scores and other available assessment scores such as CMAS Math and ELA and ACT.

For grades 5 and 8, scores on the CMAS Science assessment were correlated with scores on the CMAS Math and ELA assessment scores for the same year.

Grade	Science with Math	Science with ELA	ELA with Math
5	0.78	0.82	0.74
8	0.84	0.82	0.77

The correlations between CMAS Science and CMAS Math and ELA are fairly high and are similar for Math and ELA. We would expect math to have a higher correlation with science; however, the correlation between math and ELA assessments are also quite high.

The correlation between CMAS Science and the ACT science component score for high school was included as impact data for the high school standard setting. Students in 12th grade did not test on CMAS ELA and Math in 2015. Their CMAS Science scores were correlated with science component scores on the ACT which was taken by the same students in spring of 2014. The correlation was .61 which is moderately high. The ACT assessment is not based on the CAS so we would expect only a moderate relationship between scores on these assessments.

Evidence for Validity and Consequences of Testing

Because state tests are administered "with the expectation that some benefit will be realized from the intended use of the scores" (AERA, APA, & NCME, 2014), validity evidence supporting the use and interpretation of CMAS Science and Social Studies assessment 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.

CMAS Science and Social Studies assessments have been administered to Colorado students since the spring of 2014. The percentage of students meeting or exceeding the standards' expectations has increased for most grade levels and content areas since the assessments were first administered. Over time, student performance has improved in grade 5 science (by 2.3%) and in grades 4 and 7 social studies (by 6.9% and 1.3%, respectively). The percentage of students meeting or exceeding expectations decreased slightly in grade 8 science (by 0.9%) and decreased by a greater percentage in High School science (by 3.9%) from 2014 to 2019. It is important to note, however, that direct comparisons between 2014 and 2019 should be made with caution for some grade levels in consideration of the potential impacts of participation rate fluctuation and social studies sampling.

Subject	Grade	2014 % Strong or Distinguished	2019 % Met or Exceeded	% Change 2014 to 2019
Science	5	33.6	35.9	2.3
	8	32.4	31.5	-0.9
	HS*	24.6	20.7	-3.9
Social Studies	4	17.0	23.9	6.9
	7	16.6	17.9	1.3

* Administered in fall 2014

Fairness

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

ePATs

Because the CMAS Science and Social Studies assessments are the first statewide assessments to be primarily online for Colorado students, it was important for students to have an opportunity to experience the online testing environment prior to the administration. ePATs are online practice tests that were developed to provide an opportunity for students to become familiar with the nature of the CMAS Science and Social Studies assessments. As the assessment system has progressed, ePATs are updated every year to reflect all current accessibility features and any updates to Pearson TestNav that will impact student interactions with the system. Versions of the ePATS with accessibility features that are not universal are also available so that students can practice using accommodations such as text-to-speech, color contrast, and Spanish text-to-speech.

Universal Design

The CMAS Science and Social Studies development process adheres to the principles of universal design with the goal of avoiding construct-irrelevant aspects of the assessment as described in Chapter 2 of this document.

DIF

As outlined in Chapter 2, all items were field tested and then analyzed for DIF in order to identify any items that appeared to be unfairly favoring one subgroup over another. All DIF-flagged items were then reviewed by educator committees to identify potential construct-irrelevant explanations for the flags.

Accessibility Tools and Accommodations

As described in Chapters 3 and 4, various accessibility tools and accommodations are available for students who take the CMAS Science and Social Studies assessments. The online testing format allows for accessibility features like text-to-speech and color contrast to be available to all students. In addition, accommodations are available for students who need them and include paper, large print, and braille forms as well as oral scripts. The test is also available with Spanish test-to-speech and paper transadaptions or oral scripts that can be translated into other languages. The purpose of these various options is to allow students to fully demonstrate their content knowledge without being hindered by non-construct related elements (e.g., vision challenges).

PART II: STATISTICAL SUMMARIES FOR 2018–2019

This section contains an overview of the statistical summaries for the following administrations:

- Spring 2019 Operational Exam
- Spring 2019 Embedded Field Test

For the operational administration, administration summaries, calibration results, performance results, reliability evidence, and validity evidence will be included. For the embedded field test, form summaries, rater agreement statistics, and data review outcomes will be provided.

CHAPTER 1: SPRING 2019 OPERATIONAL EXAM

The following section provides details on the spring 2019 administration of the CMAS Science and Social Studies assessments. For the social studies assessments, a sampled approach was implemented beginning in spring 2019. Approximately one-third of the students in each of grades 4 and 7 took the assessment. These groups were selected prior to testing and were representative of the state population with respect to various demographics.

Although there was a small percentage of Elementary/Middle School students whose parents excused them from taking the assessment in accordance with state law, the resulting group of students represented the state population in terms of demographics and achievement. For High School Science, there was a slightly higher percentage of students who were excused. Therefore, a sample was drawn for calibration purposes. Of the roughly 38,000 students who took the assessment, approximately 20,700 were selected to be included in the calibrations. This subset of students mirrored the state population in terms of demographics and achievement. It should be noted that tables and figures in this report related to calibrations and equating are based on the sample, while summary statistics from the administration are based on the entire group of students who took the assessment.

Administration Summary

Table 23 shows the breakdown by online test takers compared to those who took accommodated forms. Although a paper form was available to all students, the vast majority took it online.

For the social studies assessments (grades 4 and 7), roughly one-third of students (~20,000) students took the assessment as a result of the sampling approach implemented in 2019. Schools were divided into three groups that were each representative of the state and assigned a testing year intended to allow each school to test only once every three years. In addition, district impact was considered across grades. Specific sampling methodology steps were as follows:

1. Sample for grade 4.
    a. For each district configuration (1 school district, 2 school district, or 3+ school district), sort school list by school setting and then school average.
    b. Divide into three groups based on performance (1, 2, 3, 1, 2, 3, etc.).
2. Sample for grade 7.
    a. For districts with only one school for all grades, apply grade 4 group assignments to grade 7 school list.
    b. For districts with two or more schools, sort school list by school setting and then school average.
    c. Divide into three groups based on performance (1, 2, 3, 1, 2, 3, etc.).
3. For districts comprised of 2-10 schools, make sure that schools are included in two and only two groups (e.g., 1 and 3 but not 2). Modify as needed.
4. Evaluate demographic match to the overall state and revise as needed to match state student distribution on key demographics (gender, ethnicity, socioeconomic status,

students identified with a disability (on an IEP or 504 plan), students identified as ELs, charter school vs. non-charter, and district setting).

After administration, the final tested sample was reevaluated for demographic match to the state to account for parent excuses or potential changes in school populations. In 2019, the final sample matched our expectations so no further sample adjustment was necessary. Table 24 provides n-counts of various demographic characteristics for the students who took the CMAS Science and Social Studies assessments.

Equating Results

The initial calibrations revealed no items with problematic item parameter estimates. Fit plots were examined and no items were suppressed for statistical performance or item-model fit.

Review of anchor item stability analyses resulted in the dropping of zero to three items from the anchor set, depending on grade. For the grades where items were dropped, one to three additional items were designated as an anchor to counterbalance the items that were dropped. This resulted in anchor sets being between 28% and 33% of the points for all grades. One item was removed and suppressed from scoring in grade 5 science before determining examinee ability estimates.

As described in Chapter 8, the online and paper versions were virtually identical such that the item parameter estimates were assumed to be the same. The information provided for the curves and item statistics are based on the online estimates.

Item Statistics

Tables 25–29 provide the item parameter estimates for each grade. The "Item Type" uses the coding of SR for selected-response, XI for technology-enhanced, and CR for constructed-response. The "Model" refers to the IRT model under which the item was estimated (2PL, 3PL, and GPC). The "A" column shows the item parameter estimate for discrimination, "B" for difficulty, "C" for pseudo-guessing, and "D1" through "D4" for GPC category estimates. Not all item parameters apply to each item. For example, there is no "C" estimate for the GPC model.

The last column of the tables reflects whether an item was flagged for misfit based on Q1. There were no items flagged for grade 5 and two items were flagged for grades 4, 7, 8 and High School.

See Chapter 8 for detailed information about the calibration process.

Curves

The Test Characteristic Curves (TCC), Test Information Curves (TIC), and CSEM Curves are provided in Appendix D. It should be noted that the TCCs are provided in terms of percent correct rather than raw score. All three curves for grade 4 are presented first, followed by grades 5, 7, 8, and High School. Along with the curves, each of the three cut scores for a given grade is

indicated on the curves with a red vertical line. The middle vertical line corresponds to the cut for Met Expectations.

Performance Results

The cumulative scale score distributions for each grade are shown in Tables 30–34. Figures 11–15 display the same information in graphical form.

Table 35 provides summary statistics for overall scale scores. Means, standard deviations, and medians are provided. Each grade has a mean near 600 and standard deviation around 100, as expected based on the scaling methodology. See Chapter 8 for details.

The performance level distributions for each grade are shown in Table 36. It is noticeable that the distributions within each content area are comparable.

Summary statistics for content standard scale scores are shown in Table 37. Means, standard deviations, and medians are provided. The means across content standards are similar.

As described in Chapter 8, scales were developed for SR and CR items. Table 38 shows the means and standard deviations on students' scale scores based on those two scales. As expected, all means are near 600.

Tables 39–43 provide the means and standard deviations of raw scores for each GLE (for Elementary/Middle School) or PGC (for High School). In addition, the average percent correct is provided. These statistics should be interpreted cautiously because they may be based on a relatively low number of items, depending on GLE/PGC. In addition, items within one GLE/PGC may be more difficult than those in another GLE/PGC.

Tables 44–53 provide classical statistics at the item level. For SR items, the omit rate, p-value (the item mean and also the percentage of students correctly responding to an item), and item-total correlation is given. For CR items, the percentage of students earning each score point is provided in addition to the statistics included with the SR items.

Correlations were calculated between the various content standards of each assessment and are provided in Table 54.

Reliability Statistics

Coefficient Alpha

Coefficient alpha was calculated for both the content standards and the overall assessment, as shown in Table 3. As expected, the alphas for the content standards are lower than the overall assessment, likely due to differences in the number of items. The alphas for the full assessments ranged between .91 and .94.

Tables 4–8 display performance by various subgroups. The means, standard deviations, minimums, maximums, and alphas are provided.

SEM

Table 9 shows the SEMs that were calculated based on the alphas provided in Table 3.

CSEM

As previously noted, CSEM curves for each grade are included in Appendix D.

Decision Consistency and Accuracy

Tables 10–12 provide statistics related to decision consistency and accuracy. Table 10 shows accuracy and consistency estimates in addition to probabilities due to chance (PChance) and kappa for the entire assessment. Kappa describes the agreement between classifications on two parallel forms. The kappa value can be interpreted as follows (Altman, 1991):

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

Tables 11 and 12 provide the accuracy and consistency estimates at each of the cut scores.

Inter-Rater Agreement

For each operational item, approximately 10 percent of the responses were scored by a second reader, which allowed for rater agreement statistics to be calculated. Tables 13–22 provide the percentage of items with exact agreement, adjacent agreement, and non-adjacent agreement. In addition, the final columns show the kappa, mean difference, and correlation for each item. The target exact agreement rate is 90% for 2 point items and 80% for 3 point items. The target exact plus adjacent agreement rate is 95% for all items.

Validity Statistics

Factor analysis

A factor analysis was conducted for each grade and scree plots were constructed to display the relative size of each eigenvalue, as shown in Figures 16–20.

CHAPTER 2: SPRING 2019 EMBEDDED FIELD TEST

This section provides details on the field test items that were embedded within the spring 2019 administration of the CMAS Science and Social Studies assessments. Due to low n-counts, the items embedded in the paper and large print test forms were not scored; only items embedded in the online forms were analyzed.

Field Test Forms

Depending on the grade, between 8 and 11 field test forms were administered. Within a grade, each field test form was parallel; that is, each student received the same number of each item type and in the same location on the form. Table 55 shows the number of field test forms and field test items per grade.

Inter-Rater Agreement

For each CR item, approximately 1,500 responses were scored by highly qualified scorers, as described in Chapter 5. All of the responses were scored by two readers, which allowed for inter-rater reliability calculations. Rater agreement statistics can be found in Tables 18–22, where the percentage of items with exact agreement, adjacent agreement, and non-adjacent agreement are provided. In addition, the final columns show the kappa, mean difference, and correlation for each item. Items with poor rater agreement during field testing are reviewed by Pearson and CDE content experts for issues with the item or scoring rubric. These items may be field tested again after adjustments to the item or scoring rubric or removed from the bank.

Data Review

The Data Review meetings for the spring 2019 embedded field test items were held via WebEx in late July into early August. Field test data were analyzed and items were flagged based on classical statistics and DIF. Items that were flagged were taken through the data review process where committee members examined each item and recommended whether to accept or reject it. Table 56 summarizes the outcomes of the Data Review meetings where most items were accepted. It should be noted that although committee members were only given the choice of accepting or rejecting an item, there were a few cases where the committee recommended editing and re-field testing the item as reflected in Table 56.

REFERENCES

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

Allen, N. L., Carlson, J. E., & Zalanak, C. A. (1999). *The NAEP 1996 Technical Report.* Washington, DC: National Center for Education Statistics.

Altman, D. G. (1991). *Practical Statistics for Medical Research.* London, UK: Chapman and Hall/CRC Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M Lord & M.R. Novick, *Statistical theories of mental test scores* (pp. 392–479). Reading, MA: Addison-Wesley.

Chien, M. and Shin, D. (2012). *IRT Score Estimation Program, V1.3 [computer program]*. Iowa City, IA: Pearson.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–47.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: 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). Hillsdale, NJ: Erlbaum.

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

Kim, S. and Kolen, M. (2004). *STUIRT [computer program].* Iowa City, IA: The University of Iowa.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating: Methods and practices.* (2nd ed.). New York: Springer-Verlag.

Lee, W., Hanson, B. A., & Brennan, R. L. (2000, October). *Procedures for computing classification consistency and accuracy indices with multiple categories.* (ACT Research Report Series 2000–10). Iowa City, Iowa: ACT, Inc.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement,* 16, 159–176.

Scientific Software International, Inc. (2011). *IRTPRO [computer program].* Lincolnwood, IL.

Wells, C. S., Hambleton, R. K., Kirkpatrick, R., & Meng, Y. (2014). An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement in Education, 27,* 214–231.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245–262.