Colorado Measures of Academic Success

# Science and Social Studies


# Technical Report

# 2016

# Table of Contents

# 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 language background or ability, must be provided with the opportunity to demonstrate their content knowledge. The Colorado Measures of Academic Success (CMAS): Science and Social Studies is Colorado's standards-based assessment designed to measure the Colorado Academic Standards (CAS) in the content areas of Science and Social Studies.

## Purpose of the Document

The purpose of the *CMAS Technical Report* is to inform users and other interested parties about the development, content, and technical characteristics of the CMAS assessments. The technical report provides information about the planning and administration of the exams during spring 2016.

The *CMAS 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 2016 administration. Results are provided for both the operational items and the embedded field test items.

## Overview of the Exams

The CMAS is a standards-based assessment designed to measure what students should know and be able to demonstrate at each grade level. The CMAS is aligned with the CAS for Science and Social Studies:

http://www.cde.state.co.us/coscience/statestandards

http://www.cde.state.co.us/cosocialstudies/statestandards

The subject and grade combinations for CMAS are shown in Table 1. This report pertains to the second operational administration for High School (HS) and the third operational administration for Elementary School and Middle School (ES/MS), in April 2016.

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

simulations and in the items to answer the questions or respond to the prompts. The simulation-based items may be SR, TEI, or CR.

Likewise, a subset of the Social Studies assessment includes Performance Events. The items within each Performance Event all relate to a collection of sources about a Social Studies topic. Performance Events 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 Performance Event may be SR, TEI, or CR.

The Science and Social Studies CMAS assessments cover the content standards outlined below. Additionally, Scientific Investigations and the Nature of Science is included as a Science reporting category. This reporting category is composed of items that 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.
    - 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.
    - 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.
    - 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 a particular way of building knowledge and 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 item development began in 2012 and items were field-tested in 2013 in order to collect response data on all newly developed CMAS 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, 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 that survived comprised the item pool that supported test construction. Following the first operational administration for ES/MS in spring 2014, performance standards were set and final cut scores were used for reporting purposes. Following the first operational administration for HS in fall 2014, performance standards were recommended but no social studies cut scores were approved and science cut scores were approved for one year only and used for student-level reporting purposes.

## Assessment Development Partners

The CMAS 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 a Technical Advisory Committee (TAC).

### Colorado Department of Education

CDE staff work closely with Pearson on each facet of the assessment with CDE serving as the ultimate approver.

### Colorado Educator Community

Throughout assessment development, educators contribute to item and assessment development through participation in item writing, content and bias review, data review, rangefinding, and standard setting meetings. For each meeting, an effort is made to involve educators who are representative of the entire state of Colorado.

### Pearson

Pearson is responsible for the content development, administration, and psychometrics of the CMAS assessments. This includes item and test development, online and paper forms creation, enrollment, packaging and distribution, online test delivery, processing, scoring, customer service, standard setting, score reporting, and psychometric services.

### Technical Advisory Committee

The TAC is comprised of psychometric and assessment experts tasked with providing high-level consulting and expert advice regarding the creation of the CMAS assessments. Input is received on topics such as blueprint design, score reports, scaling and equating, automated scoring, and standard setting. The TAC includes the following members:

- Dr. Jamal Abedi, Professor, University of California, Davis
- Dr. Elliot Asp, Special Assistant to the Commissioner, Colorado Department of Education
- 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, Professor, University of Iowa
- Dr. Martha Thurlow, Director, National Center on Educational Outcomes

# CHAPTER 2: ITEM DEVELOPMENT AND ITEM BANKING

The item development process for the CMAS involves following a spectrum of prescribed steps in order to develop a broad diversity of items that align directly to the Colorado Academic Standards (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, Pearson upheld these considerations as the cornerstones of the CMAS item and test development:

- The test specifications ensure the CMAS items align to the evidence outcomes (EOs) and grade-level expectations they are intended to measure.
- The CMAS 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).
- The item and test development processes promote the equivalency of the online and the paper-and-pencil assessments.
- The CMAS item and test development processes are compliant with industry standards.

Pearson's proprietary software, Item Tracker Test Builder (ITTB), is used to support the item and test development process. As described in the following section, items can be moved into various "buckets," each representing a step in the item development process.

## Item Development

The item-writing process is a tiered, inter-related process that begins with the development of the test blueprints for each grade level within each subject, continues with designing the item development plan (IDP), and uses the IDP to forecast the targeted number of item and associated stimulus across EOs needed to create a robust item bank that would be refreshed over time. Once written, an item goes through multiple rounds of review, including content and bias review and data review.

### Test Blueprint

Pearson designed the Science and Social Studies grade-specific blueprints with input and approval from CDE. Each blueprint contains the number of test items by content standard, item type, and cognitive demand. During this phase, Pearson created an IDP delineating the targeted number of items per EO, grade-level expectation (GLE), depth of knowledge (DOK) level, and item type for development.

Test blueprints provide the following information:

- the number of operational items
- the total number of score points a student can earn
- the total number of score points within each reporting category
- the number of points associated with a performance event or simulation
- the appropriate distribution of items across EOs and GLEs
- the DOK distribution for the specified grade level

Blueprints can be found in Figures 1–5.

**Item Development Plan**

The IDP is designed to determine the number of items at each EO/standard needed to construct the assessment based on the blueprint requirements. The item bank is analyzed and EO, item type, and DOK gaps are identified. Each IDP is updated at the beginning of each item development cycle with development targets that would address any stimulus, EO, item type, and DOK shortages.

**Item-Writing Process**

Pearson uses the following resources during item development:

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

The initial step of development is the Social Studies Performance Events (PEs) and Science Simulation (SIMs) preliminary conception and composition. These ideas and storyboards are presented to CDE for review and feedback, along with suggested EOs that the PEs and SIMs address. CDE provides input on how to move forward with the development of the PEs and SIMs. The PEs and SIMs are then developed and items are written to a variety of EOs, either internally or by educators during the item writing workshop (IWW).

After the IDP is developed, the IWW is conducted, facilitated by Pearson assessment specialists. Item writing assignments are given to the Colorado educator item writers. These educators write a variety of items, across item types and across EOs. The item writers use the standards, frameworks, item specifications document, item writing guidelines, and the item writing checklist to guide them in completing their assignments. The item writers also work with the

Pearson Assessment Specialist if any clarification is needed.

When the item writers have completed and submitted their items to Pearson, Pearson assessment specialists upload the items into the item banking system, Tracker. Upon upload, the items are assigned a unique identification number (UIN). In Tracker, the assessment specialists review and, if needed, make revisions to the items and metadata. This is an iterative process.

The assessment specialists evaluate each item specifically for content correctness, grade appropriateness, EO and DOK alignments, and also apply the CMAS style guide guidelines. The assessment specialists focus on the quality of the items, adherence to the principles of universal design, cognitive demand, 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. CR items are reviewed for their scorability 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 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 universal design review, 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 would 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 have been completed, the items are moved into Tracker to the final content review "bucket" for the lead assessment specialist to review and approve the item to present to the client.

Adhering to these resources ensures that each Colorado item measures the intended EO/standard, is content and grade appropriate, is accessible to all populations required to take the assessments, and is free from any bias, and, as importantly, follows the Colorado style.

**CDE Pre-Review**

CDE has access to Tracker and can use Tracker to review items in the item banking system.

Once items have been revised and deemed acceptable to present to CDE, items are moved in

Tracker into the CDE Pre-Review "bucket." CDE reviews items in the item banking system to ensure that the content is correct, the EO alignment is sound, the DOK is appropriate, the language and content are grade-appropriate, and the graphics and art are clear and relevant to the item.

When CDE has completed its review of the items, CDE moves the items to the "From CDE Pre-Review" bucket, which allows the assessment specialists to move the items to the appropriate level in Tracker. CDE's comments are recorded in Tracker and CDE outcome options are "Accept," "Accept with Modifications," and "Reject."

- For items marked "Approved," no more revisions are needed and those items are moved into the Content and Bias bucket.
- For items marked "Accept with Modifications," items are revised per CDE's feedback and, if necessary, re-reviewed by the content editors/research librarians/UDR/Art. These items are re-reviewed by CDE and reconciled with Pearson's assessment content specialist and either deemed "Approved" or "Reject."
- Items marked "Reject" are rejected/DNUed in Tracker. Replacement items are written by an assessment content specialist and uploaded into Tracker. New UINs are assigned and the item goes through the same rigorous review process as a new item.

**Content and Bias**

Following the completion of the internal Pearson and CDE reviews, the items are reviewed by a Content and Bias Committee, which is comprised of Colorado educators from across the state with diverse backgrounds, including content expertise and special population expertise. The 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 committee members are trained and instructed to verify that each stimulus and item (list 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);
- 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, and religious 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 Modifications," and "Reject."

A Content and Bias Reconciliation Meeting is conducted within the week following the educator meeting. The reconciliation meeting includes CDE staff and Pearson's assessment specialists. At this meeting, committee comments are reviewed, proposed edits are reconciled, and item outcomes are finalized. Tracker is updated to reflect the edits and outcomes. The approved items are then moved to the Ready for Field-Testing bucket in Tracker.

**Data Review**

Following the administration of items in a field-test environment, a committee of educators is convened to review the newly developed items along with student performance data. 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.

DIF analyses are conducted on various subgroups (gender, ethnicity, free and reduced lunch, IEP, and ELL) using Mantel-Haenszel Delta DIF statistics (Dorans & Holland, 1992). Classification rules derived from National Assessment of Educational Progress (NAEP) guidelines (Allen, Carlson, & Zalanak, 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.

Educators are trained to interpret the statistical information and while the committee uses the data as a tool to inform their judgments, the committee is instructed not to base their final assessment of the appropriateness or fairness of items for all individuals and subgroups on solely the data. Committee members review each item and make a recommendation as to whether to "accept" or "reject" it. Following the meeting, Tracker is updated by moving accepted items into

"Ready for Operational" bucket.

## Item Banking System

ITTB is Pearson's proprietary software that supports the item and test development process.

The "Tracker" component 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 (PEs and simulation storyboards) are tracked and revisions are recorded from creation through retirement in a secure environment.

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

"Builder" 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 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 Builder:

- content information such as standard, EO, and DOK
- classical statistics such as item means and point biserial correlations
- IRT statistics such as difficulty, discrimination, guessing, and 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 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 will meet the blueprint with a variety of topics and contexts and meet specified psychometric targets.

The following guidelines are used during 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 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 form is reviewed by two assessment specialists. Each assessment specialist verifies that the form meets the blueprint (the required EO coverage, DOK allocation, item type). The form is then presented to psychometrics for analysis and the psychometrician verifies that the form falls within the established psychometric and blueprint parameters. The psychometric lead identifies the anchor item set within each operational form.

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

After the operational 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 form in a designated section and sequence. The 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 assessment online. Using this format allows not only for the use of innovative item types but also for various accessibility options and accommodations as described in Chapter 4.

## Accommodated Test Forms

Accommodated test forms for the CMAS assessments are available for both the online and paper-based forms. For online forms, text-to-speech, color contrast, and text-to-speech with color contrast are available. For paper forms, the various options are described below. In addition, oral scripts in both English and Spanish are available for online and paper forms.

### Paper

A paper-based version of the CMAS assessments is published and is available if needed for an accommodation or for schools who choose not to test online as allowed by state law.

The paper form is parallel to the online form. Parallel paper-based items were developed for TEIs. In some cases this was achieved with traditional selected-response items and in others it required an item that had to be human-scored. For example, a drag-and-drop item may have been converted to an item in which the student had to draw lines from the draggers to the drop bays.

For the spring 2016 administration, the operational items on the paper-based version of the CMAS assessments were the same as the operational items on the online forms. The Braille form was not the same as the 2016 online or paper forms.

### Braille

After approval of the paper test materials for the high school assessment, 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 (TVI) 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.

The elementary and middle school assessments used the same braille forms as the 2014 administration, which were developed following the same procedures.

### Large Print

Large print versions of the assessment 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 version includes 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. CDE reviews the paper form and identifies what 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 administration procedures. Prior to the administration of CMAS, 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 CMAS administration.

### CMAS Test Administrator Manual

This manual describes the procedures Test Administrators were to follow when administering the paper and online CMAS assessments. Prior to administering any CMAS assessment, Test Administrators were to carefully read this manual. 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 this manual were provided to ensure that every student in Colorado received the same standard directions during the administration of the test.

### CMAS and CoAlt: Science and Social Studies Data Supplement

The purpose of this document is to provide an overview of the data collection activities for the CMAS and CoAlt: Science and Social Studies assessments. The document provides a general overview of these processes along with accompanying procedures.

### CMAS: Science and Social Studies Accommodations Supplement

This document provided a supplement for the *2015-2016 PARCC Accessibility Features and Accommodations Manual* which included accommodation information for English language arts and mathematics assessments. Accommodations for CMAS science and social studies assessments were available to students identified with a disability (on an IEP or 504 plan) and/or identified as an English Learner (EL). Accommodations were available for these student populations for both the computer-based and paper-based forms of the 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 and CoAlt: Science and Social Studies assessments. Instructions include the protocols that all school staff were to follow related to test security and test administration. The manual also includes the tasks that were to be completed by DACs, SACs, and DTCs before, during, and after test administration.

## Training

Extensive onsite 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, the recorded versions of the live trainings were posted on Colorado's Avocet.pearson.com webpage (master index of program resources).

## 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

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, 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. Some of the available accommodations included English oral scripts, Spanish oral scripts, oral scripts for local translation, paper forms, braille forms, large print forms, and Spanish text-to-speech forms.

## Test Security

Procedures described in this paragraph 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 state, district, and/or school personnel involved in the assessment administration were required to participate in annual local training on the CMAS assessment. In addition to the training, they were required to sign a security agreement prior to handling test materials. By signing the security agreement, personnel agreed to a set of security guidelines, 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 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.

# CHAPTER 5: CONSTRUCTED-RESPONSE SCORING

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 participants at the rangefinding meetings.

- 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. 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 is integrated into the scoring system. In addition to the systematic tools – staff is physically on site in 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 Performance Scoring operations. Pearson's Performance Scoring 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.

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

The following sections describe the rangefinding process and the major components of the quality assurance system including backreading and calibration.

## Rangefinding

Constructed-response 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 (SS).

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 score points of the rubrics using student responses. Each rangefinding committee includes Pearson's Scoring Service's staff, Pearson's scoring and item development staff, CDE content representatives, grade level teachers with relevant content expertise, an educator with special education expertise, an EL educator, and Pearson Client Services representatives. Participants create consensus scores for student responses that are subsequently used to develop effective training materials for scoring of CR items.

Pearson's Scoring Directors construct one rangefinding set per item, which includes 25 responses for items scored 0–2 and 30 responses for items scored 0–3. 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 were 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 break-out 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 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.

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

## 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. Rater agreement information for the spring 2016 administration can be found in Part II of this report.

## 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 is described in terms of four performance levels: Partially Met Expectations, Approached Expectations, Met Expectations, and Exceeded Expectations. The performance levels were updated in 2015-2016 to match the labels used for the ELA and Math assessments. Only the labels were updated, the Performance Level Descriptors and cut scores set during standard setting were not changed. Standards were set for grades 4 and 7 Social Studies and 5 and 8 Science after the first administration in spring 2014. Details of the elementary and middle school standard setting can be found in the 2013–2014 CMAS Technical Report. Standards were set for high school science and social studies after the first administration (fall 2014). Details of the high school 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 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 *Interpretive Guide* at http://www.cde.state.co.us/assessment/newassess-sum.

## Description of Scores

CMAS 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, 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.

### 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 Correct

Percent correct scores are provided at the Prepared Graduate Competency (PGC) and Grade Level Expectation (GLE) levels. Unlike scale scores, percent correct scores cannot be compared across years because individual items change from year to year. In addition, they cannot be compared across GLEs or PGCs 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 A. 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 correct scores are displayed on a four-page report along with comparative information related to the student's school, district, and state performance. In addition, PLDs 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 Test Results and access to the reports is limited to authorized users.

# CHAPTER 8: CALIBRATION, EQUATING, AND SCALING

Item Response Theory (IRT) was used to develop, calibrate, equate, and scale the CMAS assessments. The three parameter logistic (3PL) (Birmbaum, 1968), two-parameter logistic (2PL) (Birmbaum, 1968), and generalized partial credit (GPC) (Muraki, 1992) were applied. These measurement models are routinely used for 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 3PL or 2PL 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 that 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:

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

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 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 form. Each time a new form is constructed, equating is used to allow scores on the new form to be comparable to scores on the previous form.

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 forms onto the operational scale. Spring 2014 and fall 2014 were the first operational administrations for the elementary/middle school and high school assessments, respectively, so those administrations were used to establish the scale. For spring 2016, equating was used to place the spring 2016 forms on the 2014 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; TEI items were fit to either the 3PL or 2PL 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 forms for the CMAS assessments. Forms from adjacent administrations contain a set of items that are the same across the two administrations. This set of items represents the blueprint in terms of content and represents roughly 30% of a full 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 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, scoring is exactly the same each time the item is administered (e.g., 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, 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 were 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 Forms**

The scale score distributions were examined using a matched samples approach to investigate the extent to which the online and paper 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, language proficiency, IEP, district setting, and past test scores (CMAS ELA for social studies and CMAS Math for science). Scale score distributions of CMAS scores between the matched samples were compared to quantify the mode effect. As a way 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. For the spring 2016 administration, no mode effects were found for any of the grades by these criteria. Grades 4 and 7 had less than 1,000 paper testers. Grades 5 and HS had fewer than 1,000 students in the matched sample. Grade 8 had 1,016 students in the matched sample and an effect size of .09. The number of students taking the paper form can be found in Table 23.

**Field Test Equating**

The process for field test equating is similar to that of operational equating (page 25) 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 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.

**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 2016 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 2016 operational items to operational 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 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 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 2016 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 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, 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 assessment likely more accurately measures 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).

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 B.

## Decision Consistency and Accuracy

The CMAS scale is 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. For the overall test, consistency and accuracy estimates along with PChance and Cohen's Kappa ($\kappa$) 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 assessment that is validated but rather the interpretations of the CMAS 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 assessment cycle with validity and fairness considerations in mind. This chapter provides information regarding specific sources of validity evidence as well as fairness.

Furthermore, validation is a process. As the CMAS assessments mature, validity evidence supporting the assessments' interpretations will continue to be collected and documented.

## 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 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 is 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 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 blueprint was carefully developed with specificity at multiple levels (e.g., spread by item type, simulation/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. The full report can be found in Appendix C.

## Evidence Based on Response Processes

Evidence based on response processes pertains to the cognitive aspect behind how students respond to items. Since the CMAS assessments are Colorado's first online assessment, cognitive labs were held during the item development phase to evaluate whether students would find the nature of CMAS items (e.g., simulations) or Pearson's TestNav browser-based testing platform challenging. In addition, the CMAS 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 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 assessments, a factor analysis is performed and scree plots are examined to investigate the number of dimensions that the CMAS 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 2016 administration can be found in Part II of this report.

### Evidence Based on Relationships to Other Variables

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

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

| Grade | Science with Math | Science with ELA | ELA with Math |
|---|---|---|---|
| 5 | .79 | .82 | .77 |
| 8 | .82 | .83 | .76 |

The correlations between CMAS Science and CMAS PARCC ELA and Math 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 PARCC 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 Colorado Academic Standards so we would expect only a moderate relationship between scores on these assessments.

### Evidence for Validity and Consequences of Testing

As the CAS become more fully integrated into the classroom and with additional administrations of the CMAS assessments, it is intended that information about the consequences of the assessment will be collected.

## 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 assessments.

### ePATs

Because the CMAS assessments are the first statewide assessment 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 assessments.

### Universal Design

The CMAS 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 investigate whether there was a flaw with the item.

### Accessibility Tools and Accommodations

As described in Chapters 3 and 4, various accessibility tools and accommodations are available for students who take CMAS. 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 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 2015–2016

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

- Spring 2016 Operational Exam
- Spring 2016 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 2016 OPERATIONAL EXAM

The following section provides details on the spring 2016 administration of the elementary, middle, and high school CMAS assessments. For the social studies assessments, a sampled approach was implemented beginning in spring 2016. 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 and middle students who opted out of taking the assessment, the resulting group of students represented the state population in terms of demographics and achievement. For high school, there was a slightly higher percentage of students who opted out. Therefore, a sample was drawn for calibration purposes. Of the roughly 34,000 students who took the assessment, approximately 25,000 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. For the social studies assessments (grades 4 and 7), roughly 20,000 students took the assessment. For science, the numbers were much higher. Although a paper form was available to all students, the vast majority took it online.

Table 24 provides n-counts of various demographic characteristics for the students who took the CMAS assessments.

## Equating Results

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

Review of anchor item stability analyses resulted in the dropping of one item from the anchor set for grades 4, 5, and 7. For grade 7, an additional item was designated as an anchor to counterbalance the item that was dropped. This resulted in anchor sets being 29% of the points for grades 4 and 5, and 26% of the points for grade 7. No items were dropped from grades 8 or HS resulting in anchors sets being 34% of the points for both grades.

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 was one item flagged for grades 4, three items flagged for grades 7 and 8, and no items flagged for grades 5 and HS.

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 B. 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 HS. Along with the curves, each of the three cut scores for a given grade is indicated on the curves with a red vertical line.

## 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. All means are near 600.

Tables 39–43 provide the means and standard deviations of raw scores for each GLE (for ES/MS) or PGC (for HS). In addition, the average percent correct is provided. These statistics should be interpreted cautiously since 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 .90 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 B.

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

### Inter-Rater Agreement

For each operational item, 5 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.

# 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 2016 EMBEDDED FIELD TEST

This section provides details on the field test items that were embedded within the spring 2016 administration of the CMAS assessments. Due to low n-counts, the items embedded in the paper and large print forms were not scored; only items embedded in the online forms were analyzed.

## Field Test Forms

Depending on the grade, between 6 and 12 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, 1,500 responses were scored by highly qualified scorers, as described in Chapter 5. A 5% subset of those 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.

## Data Review

The data review meeting for the spring 2016 embedded field test items was held via WebEx on August 10–12, 2016. 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 decided whether to accept or reject it. Table 56 summarizes the outcomes of the data review 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.

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.