



Colorado Measures of Academic Success



Technical Report

Math and English Language Arts (ELA)

2018

Colorado Measures of Academic Success (CMAS)
Mathematics & ELA (including CSLA)
Technical Report
2018

Table of Contents

Purpose of the Document.....	1
Overview of the 2018 Test Abbreviation.....	1
PART I: HISTORICAL OVERVIEW AND SUMMARY OF PROCESSES	3
CHAPTER 1: INTRODUCTION	3
Requirements	3
Background.....	4
Assessment Development Partners	5
Composition of the Assessments and Overview of the 2018 Abbreviation	7
Composition of the Assessments	7
Claim Structure for Mathematics.....	7
Claim Structure for English language arts.....	8
Score Structure.....	8
Test Structure	10
Test Structure for Mathematics.....	10
Test Structure for English Language Arts	10
Test Blueprints	11
Timing of Tests.....	11
CHAPTER 2: ITEM DEVELOPMENT AND ITEM BANKING	13
Item Development.....	13
Operational Items.....	13
Paper-based TEIs	13
Field Test Items.....	14
Item Banking System.....	19
CHAPTER 3: TEST CONSTRUCTION.....	21
Online Forms	22
Accommodated Test Forms	22
Paper	23
Braille.....	23
Large Print	24
CHAPTER 4: TEST ADMINISTRATION PROCEDURES	25
Manuals.....	25
CMAS Test Administrator Manual for Computer-Based Testing and the CMAS Test Administrator Manual for Paper-Based Testing	25

PearsonAccess ^{next} Online User Guide.....	25
CMAS and CoAlt Procedures Manual.....	25
Administration Training.....	26
Accessibility and Accommodations.....	26
Test Security	27
CHAPTER 5: SCORING THE ASSESSMENT	29
Machine Scoring.....	29
Human Scoring	30
Operational Scoring	30
Rangefinding.....	32
Backreading	33
Calibration.....	33
Automated Scoring	34
CHAPTER 6: STANDARD SETTING.....	39
Performance Standards vs. Cut Scores	39
CMAS and CSLA Standard Setting.....	40
CHAPTER 7: REPORTING.....	41
Description of Scores.....	41
Scale Scores	41
Performance Levels	41
Percentile Ranking.....	42
Percent Earned	42
Score Reports	42
Student Performance Reports	43
Aggregate Reports	43
CHAPTER 8: CALIBRATION, EQUATING, AND SCALING	44
IRT Models	44
CMAS Mathematics and ELA	44
CSLA	46
Equating and Scaling	46
Operational Equating and Scaling	47
Comparability of Online and Paper Forms for CMAS Mathematics and ELA.....	50
Field Test Equating.....	51
Ability Estimates.....	51

Overall and Subscale Scale Scores	52
Steps in the Calibration, Equating, and Scaling Process	52
Data Preparation.....	52
CHAPTER 9: RELIABILITY	55
Coefficient Alpha.....	55
Standard Error of Measurement.....	56
Conditional Standard Error of Measurement.....	56
Decision Consistency and Accuracy.....	57
Inter-Rater Agreement	57
CHAPTER 10: VALIDITY	59
Sources of Validity Evidence.....	59
Evidence Based on Test Content	59
Evidence Based on Internal Structure.....	61
Evidence Based on Relationships to Other Variables.....	63
Evidence Based on Response Processes.....	63
Evidence Based on the Consequences of Testing.....	64
Fairness	64
Sample Items.....	64
Universal Design.....	65
DIF	65
Accessibility Tools and Accommodations.....	65
Special Studies to Support the Abbreviated Test Design	66
Blueprint Adjustment Comparison	66
Sample Invariance Study	67
Pattern Scoring Analysis.....	68
PART II: STATISTICAL SUMMARIES FOR 2018.....	70
CHAPTER 1: SPRING 2018 OPERATIONAL ASSESSMENT	71
Administration Summary.....	71
Equating Results	71
Calibration and Anchor Set Evaluation	71
Item Statistics.....	72
IRT Curves.....	72
Performance Results	72
Reliability Statistics	73

Coefficient Alpha.....	73
SEM	74
CSEM.....	74
Decision Consistency and Accuracy.....	74
Inter-Rater Agreement	74
Validity Statistics	75
Intercorrelations	75
Reliability.....	75
Factor Analysis	75
Special Studies to Support the Transition.....	75
CHAPTER 2: SPRING 2018 EMBEDDED FIELD TEST	78
Field Test Forms	78
Inter-Rater Agreement	78
Data Review.....	78
REFERENCES	80
APPENDIX A. ELA AND CSLA SCORING RUBRICS.....	A-2
APPENDIX B. SAMPLE SCORE REPORTS.....	A-9
APPENDIX C. TEST CHARACTERISTIC CURVES, TEST INFORMATION CURVES, & CONDITIONAL STANDARD ERROR OF MEASUREMENT (CSEM) CURVES.....	A-14
APPENDIX D. SPECIAL STUDIES TO SUPPORT THE TRANSITION.....	A-37
Appendix D1. PARCC Mathematics Adjustment for Colorado	A-38
Appendix D2. PARCC ELA/L Adjustment for Colorado.....	A-111
Appendix D3. CSLA Blueprint Adjustment Study	A-221
Appendix D4. Colorado Equating Sample Invariance Study.....	A-235

PURPOSE OF THE DOCUMENT

The purpose of the *CMAS Mathematics & ELA (including CSLA) Technical Report 2018* is to describe the fourth administration of the state assessment, with spring 2018 being the first administration of comparable abbreviated assessments. This manual addresses the approach to the abbreviation as well as the development and construction, administration, scoring, reporting, classical item analysis results, reliability results, evidence of validity, item response theory (IRT) calibrations and scaling, and other technical characteristics of the 2018 assessments.

This report is intended for individuals with a background in the technical issues and requirements associated with the development, scoring and reporting of high stakes, state-wide assessments. Familiarity with psychometric and measurement standards, processes and procedures as found in the essential text, *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014) is assumed.

The *CMAS Mathematics & ELA (including CSLA) Technical Report 2018* 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 2018 administration. Results are provided for both the operational items and the embedded field test items.

This 2018 manual does not repeat all of the information provided in prior years' technical manuals. Technical information regarding 2015-2017 administrations, as well as the development and execution of foundational documents, activities, procedures and supporting research that served as the basis of the 2018 CMAS assessments may be found in the prior technical manuals, including the *PARCC Final Technical Report for 2017 Administration* (Pearson, 2018).

Overview of the 2018 Test Abbreviation

This technical report describes the strategies employed to establish comparability between the full versions of the assessments administered prior to 2018 and their abbreviated counterparts.

In consultation with the Colorado Department of Education's Technical Advisory Committee, deliberate approaches were taken to ensure the proportionately reduced assessments directly reflected their 2017 equivalents. Abbreviated test blueprints, including a comparison to 2017 test blueprints, are in Figures 1-4.

When compared to previous versions of the ELA and math assessments, the 2018 abbreviated assessments:

- Included the same content areas, item types, item development specifications, and overall unit structure (see Chapter 2 – Item Development)
- Maintained proportional coverage in test construction (i.e., including, not limited to, blueprint, construct, claim level total points, subclaim level, content, evidence statements, task and item types, and cognitive complexity) (see Chapter 2 – Item Development and Item Banking and Chapter 3 – Test Construction)
 - Included comparable items and anchor sets selected from the consortium item bank used as the source of items in prior administrations
- Preserved scoring rules from prior years for machine scored items (e.g., multiple choice, drag-and-drop, etc.) and constructed response items (i.e., used prior years’ rubrics, anchor papers, rules and scoring methods with the exception of paper-converted technology-enhanced items). A validity check was performed on the anchor items that were human scored. (see Chapter 5 – Scoring the Assessment)
- Retained overall scale scores and performance levels from prior years and continued to report at the claim and subclaim levels (see Chapter 7 – Reporting and Chapter 8 – Calibration, Equating, and Scaling). Writing claim and subclaim reporting was changed to percent correct due to the reduction in points available at those levels (see Part II, Chapter 1 – Special Studies to Support the Reduced Test Design).
- Continued to apply technically sound psychometric procedures related to item parameters, scaling and equating with input from external technical experts. Equated to CMAS 2017 scale to maintain score comparability to previous years (see Chapter 8 – Calibration, Equating, and Scaling). To increase differentiation, student scores were estimated using pattern scoring rather than number correct scoring (see Part II, Chapter 1 - Special Studies to Support the Reduced Test Design).
- Exhibited test characteristic curves for both the anchor sets and full forms that matched 2017 in shape and position (distribution of items by difficulty and discrimination) and maintained a target difference maximum of 0.05 (see Part II, Chapter 1 – Equating Results).

Grounded in the methodical assessment development, test construction, scoring, and psychometric strategies and techniques outlined above, analyses comparing full-length and reduced versions of the 2017 assessments as well as the observed 2018 results provide further empirical evidence of the comparability of the 2017 and 2018 CMAS assessments (see Appendix D). CMAS ELA and Math mean scale scores, standard deviations, and performance level distributions are comparable to prior versions of the assessments (see Appendix D and Tables 95 and 96).

PART I: HISTORICAL OVERVIEW AND SUMMARY OF PROCESSES

CHAPTER 1: INTRODUCTION

Requirements

All public-schools in Colorado are required by state law to administer a standards-based summative assessment each year in specified content areas and grade levels. Every student, regardless of ability or language background, must be provided with the opportunity to demonstrate their content knowledge through the state assessments. The Colorado Measures of Academic Success (CMAS) assessments in mathematics and, English language arts (ELA) are Colorado's end-of-year standards-based assessments designed to measure students' achievement of the grade-level Colorado Academic Standards in those content areas at grades 3-8.

As a requirement of Colorado School Law C.R.S. §22-7-1006.3 (4) (a) and (b), English learners with Spanish as their home language in grades 3 and 4 who meet established eligibility criteria may take the Colorado Spanish language arts (CSLA) forms of the CMAS ELA assessment. The Spanish forms are considered accommodated versions of CMAS. As a result, CSLA forms are developed to be parallel and comparable to the other CMAS forms in test design, item type, scoring, and reporting. To maintain this comparability, the revised CMAS ELA blueprints were used to develop the CSLA forms administered during the spring 2018 administration.

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

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

Intended Population

The CMAS assessments are intended to be taken by all students enrolled in public schools in grades 3-8 with the exception of some students with the most significant disabilities who may take the Colorado Alternate Assessment as determined by the student's IEP or other educational team. In addition English learners in their first year in the United States are exempt from the ELA assessment. Students with disabilities and English learners may take the CMAS assessments with or without accommodations that do not change the construct of the assessment. Accommodations are determined based on classroom experience and educational team decisions.

Purpose of CMAS

CMAS assessments were designed from the start to be used for a variety of purposes, including serving as one uniform indicator informing parents and educators about individual student achievement of the grade-level CAS and allowing comparisons to other students across the state. Results are also used as a piece of information in the evaluation of educator, school and district performance.

CMAS is a source of data that:

- may be used as a prompt for further investigation at the student, classroom, school, and district levels
- supports districts/schools in reviewing and developing goals for the performance of their students, including subgroups
- may indicate that a review of programs, curricula, materials and/or scope and sequence may be appropriate
- may inform the evaluation of district/school approaches

Assessment results also support a range of data-driven stakeholder conversations, activities, and decisions, including school selection, program evaluation, investigative research, and policy/legislation formation and review.

Background

Prior to spring 2018, Colorado developed CMAS mathematics and English language arts (ELA) assessments in collaboration with the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium. For information on the background of the consortium and the development and administration of the 2015-2017 assessments, see prior years' *PARCC Final Technical Reports*.

In 2017, the state board of education provided direction to the department to decrease testing time. CDE began exploring the use of abbreviated versions of the prior years' test blueprints with the goal of decreasing testing time while retaining comparability to the CMAS mathematics and ELA assessments previously administered in Colorado in order to maintain longitudinal trend data. In response, assessment forms based on abbreviated blueprints were developed during the fall of 2017 and administered beginning with the spring 2018 administration (See *Overview of the 2018 Test Reduction* in this document for additional information).

Assessment Development Partners

2018 operational assessment items were licensed from New Meridian and selected from the bank of items developed in collaboration with the PARCC consortium. Consortium test and item development activities were conducted by Pearson and WestEd under the guidance and oversight of consortium leadership, including the Governing Board, the K-12 state leads, the Higher Education Leadership Team, the Technical Advisory Committee, the Operational Working Group members from each of the member states, the State Text and State Content Item Review Committees, and staff members from Parcc, Inc., the consortium's project management partner.

Activities specific to the CMAS assessments, including abbreviation of the full-length test blueprints and development of new items to be field tested in spring 2018, were conducted collaboratively by the Colorado Department of Education (CDE), the Colorado educator community, and the assessment contractor, Pearson. In addition, input and advice were provided by the Colorado Technical Advisory Committee (TAC).

Colorado Department of Education

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

Colorado Educator Community

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

- Item Writing: After receiving item writing assignments based on the academic standards, educators create assessment items. Items that successfully move through the entire item development process will eventually appear on the operational assessments. (Field test items that were written by Colorado educators will be included on the CMAS English language arts and mathematics assessments for the first time in Spring 2019.)
- Content and Bias Review: Educators review items to ensure content alignment and identify potential bias and sensitivity concerns before items are included on the embedded field test.

- Range-finding: Educators review student responses to field tested items and define the score point ranges for the scoring rubrics that are used to score student responses.
- Data Review: Before field tested items are included on operational assessments, educators review data to identify potential construct-irrelevant explanations for statistical flags.

Pearson

As the primary contractor responsible for end-to-end of the 2018 assessment cycle services and products, Pearson worked closely with CDE throughout the CMAS (all content areas) and CoAlt (science and social studies) assessment development and administration processes. This included item and test development, online and paper forms creation, enrollment, packaging and distribution, online test delivery, processing, scoring, customer service, standard setting, scoring, reporting, and psychometric services.

Tri-Lin Integrated Services, Inc.

As a subcontractor to Pearson, Tri Lin was responsible for CSLA content and test development. This included passage development, item development, and test form construction.

Technical Advisory Committee (TAC)

The CMAS TAC was comprised of psychometric, assessment, and special populations experts tasked with providing high-level consulting and expert advice regarding validity and reliability issues, including psychometric topics of the CMAS assessments. Topics for which the TAC provided input included blueprint design, scaling and equating, scoring, reporting, and comparability. The TAC included the following members:

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

Composition of the Assessments and Overview of the 2018 Abbreviation

Composition of the Assessments

CMAS assessments are standards-based tests designed to measure what students should know and be able to demonstrate at the end of each grade at the elementary and middle school levels (grades 3-8). Evidence statements reflecting college and career ready standards were developed to guide the development of the assessments. The spring 2018 CMAS mathematics and ELA (including CSLA) assessments were aligned to these evidence statements and subsequently mapped to the Colorado Academic Standards (CAS) in mathematics and reading, writing, and communicating:

- Math CAS: <http://www.cde.state.co.us/comath/statestandards>
- Reading, Writing, and Communicating CAS: <http://www.cde.state.co.us/coreadingwriting/statestandards>
- Evidence Statements: https://www.cde.state.co.us/assessment/cmas_testdesign

Claim Structures

The claim structures (master, major and subclaims) for the mathematics and ELA assessments, including Spanish forms, were grounded in the academic standards and informed the design and development of the assessments.

Claim Structure for Mathematics

- **Master Claim** – The master claim makes a statement about the degree to which a student is on track to being ready for the next grade in mathematics based on their achievement of the grade-level CAS.
- **Subclaims** – The subclaims are intended to provide more granular information about student demonstration of the knowledge and skills within the math content area as reflected in the CAS. The content reflected in each of the subclaims is provided below:
 - **Subclaim A:** Major Content with Connections to Practices.
 - **Subclaim B:** Additional and Supporting Content with Connections to Practices.
 - **Subclaim C:** Highlighted Practices with Connections to Content – Mathematical Reasoning
 - **Subclaim D:** Highlighted Practice with Connections to Content – Modeling and Application

Claim Structure for English language arts

- **Master Claim** – The master claim makes a statement about the degree to which a student is on track for being ready for the next grade in English language arts based on their achievement of the grade-level CAS.
- **Major Claim** – The major claim provides information on a student’s achievement of reading and comprehending a range of sufficiently complex texts independently.
- **Subclaims** – The subclaims are intended to provide more granular information about student demonstration of the knowledge and skills within the ELA content area as reflected in the CAS. The content reflected in each of the subclaims is provided below:
 - **Vocabulary, interpretation, and Use**
 - **Reading Literature**
 - **Reading Informational Text**
 - **Written Expression**
 - **Knowledge of Language and Conventions**

Score Structure

Master claim: The degree to which a student demonstrated the concepts and skills represented in the master claim is reported through both a performance level and a scale score. There are five performance levels based on a scale score range of 650-850. Both policy and specific grade-level performance level descriptors are reported. The policy level performance levels and associated scale score ranges are provided below.

CMAS Policy Level Performance Level Descriptors and Associated Overall Scale Scores					
	Did Not Yet Meet Expectations	Partially Met Expectations	Approached Expectations	Met Expectations	Exceeded Expectations
Performance Level Descriptor	Students who do not yet meet academic expectations for the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They will <i>need extensive academic support</i> to engage successfully in further studies in this content area.	Students who demonstrate a limited command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They will <i>need additional academic support</i> to engage successfully in further studies in this content area.	Students who demonstrate a moderate command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They will <i>likely need additional academic support</i> to engage successfully in further studies in this content area.	Students who demonstrate a strong command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They are <i>academically prepared</i> to engage successfully in further studies in this content area.	Students who demonstrate a distinguished command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They are <i>academically well prepared</i> to engage successfully in further studies in this content area.
Scale Score	650-699	700-724	725-749	750-varies*	varies*-850

*Varies by grade and content area.

Major claim: (ELA only) - The degree to which a student demonstrated the concepts and skills represented in the reading major claim is reported through a scale score with a range of 110-190.

Subclaims: The degree to which a student demonstrated the concepts and skills represented in the subclaims is reported as percent earned. These percentages are not comparable across years.

Test Structure

Test Structure for Mathematics

The mathematics assessments contain selected-response (SR) items, technology-enhanced (TE) items, and constructed-response (CR) items and include three types of items:

Type I:

- Tasks assessing concepts, skills and procedures
- Subclaims A and B
- One point and two point items (grades 3-8) and four points (grade 6-8)
- SR and TE items
- Calculator (grade 6-8) and noncalculator (grade 3-8)

Type II:

- Tasks assessing expressing mathematical reasoning
- Subclaim C
- 3 or 4 point items
- SR, TE, and CR parts, all items will have at least one CR part
- Calculator (grades 6-8), noncalculator (3-5)

Type III:

- Tasks assessing modeling/application
- Subclaim D
- 3 or 6 point items
- SR, TE, and CR parts, all items will have at least one CR part
- Calculator (6-8), noncalculator (3-5)

Test Structure for English Language Arts

The ELA assessments are passage based with a combination of literary and informational passages. Multiple passages may be used to respond to some questions. The ELA assessments contain selected-response (SR) items, technology-enhanced (TE) items, and Prose Constructed Response (PCR) tasks. For PCRs, students receive a prompt, respond to reading questions and write an extended response, which is scored on a multi-trait rubric (knowledge of language and conventions, and written expression) (see Appendix A). ELA PCRs include three types of tasks: literary analysis, research simulation, and narrative writing.

CSLA forms are developed to be parallel and comparable to the other CMAS paper forms in test design, item type (SR and PCR), scoring, and reporting. To maintain this comparability, the abbreviated CMAS ELA blueprints were used to develop the CSLA forms administered during the spring 2018 administration.

Test Blueprints

CDE and Pearson collaborated in designing the 2018 CMAS mathematics and ELA grade-specific abbreviated blueprints. The blueprints were designed to measure the same constructs as, and provide content comparability to, the previous year's assessments. Eligible content continued to reflect the Colorado Academic Standards and Evidence Statements used in prior years. With the intent to reduce testing time, the 2018 blueprints are a proportionate abbreviation of the 2017 forms. 2018 CSLA forms reflected the 2018 third and fourth grade ELA blueprints respectively.

The 2018 blueprints achieved comparability through establishing proportionality in the following areas:

- **Master Claim level (total points)**
- **Reading (Major Claim) and writing**
- **Subclaim level**
 - ELA
 - Reading:
 - Informational text versus literary texts and vocabulary
 - Text complexity/word count
 - Writing:
 - Written expression and knowledge of language and conventions
 - Types of writing within grade spans
 - Math
 - Major content; additional and supporting content; expressing mathematical reasoning; and modeling and application
 - Item types (Type 1, Type 2, Type 3)
 - Calculator use

The 2018 high level test blueprints found in Figures 1 through 4 provide the following information for both 2017 and 2018 to facilitate comparisons:

- The percentage targets for the number of points per Major Claim and Subclaims
- The total number of weighted (ELA and CSLA only) and unweighted points overall
- The percentage targets for the number of points per task type (mathematics only)

Timing of Tests

Each 2018 assessment was composed of three sections with field test items embedded. The timing of the sections varied by grade and content area as indicated below:

2018 CMAS Testing Times		
	ELA	Math
Grades 3-5	Sections 1-3: 90 minutes Total time: 270 minutes	Sections 1-3: 65 minutes Total time: 195 minutes
Grades 6-8	Sections 1-3: 120 minutes Total time: 360 minutes	Sections 1-3: 65 minutes Total time: 195 minutes

As evidenced by the following table, the 2018 abbreviated test design reduced the combined ELA and math testing times when compared with 2017. For the first time since the beginning of the assessment program, the abbreviated test design for each ELA assessment included three units of equal duration, featuring embedded field testing rather than separate field test units. Prior to 2018, separate ELA field test units were administered to sampled schools. The 2018 testing times reflected all testing including new embedded field test items for future use. In addition to a reduction in timing, the equal session lengths within grades and across grade spans provided schools with more flexibility in combining content area sessions during administration.

CMAS ELA and Math Testing Times: 2018 compared to 2017			
	ELA	Math	Total (ELA and Math)
CMAS: Grade 3 Testing Times			
2017: Student without ELA stand-alone field test	255 minutes	240 minutes	495 minutes
2017: Student with ELA stand-alone field test	345 minutes	240 minutes	585 minutes
2018: Student with embedded ELA field test	270 minutes	195 minutes	465 minutes
CMAS: Grades 4 and 5 Testing Times			
2017: Student without ELA stand-alone field test	270 minutes	240 minutes	510 minutes
2017: Student with ELA stand-alone field test	360 minutes	240 minutes	600 minutes
2018: Student with embedded ELA field test	270 minutes	195 minutes	465 minutes

	CMAS: Grades 6 – 8 Testing Times		
2017: Student without ELA stand-alone field test	310 minutes	240 minutes	550 minutes
2017: Student with ELA stand-alone field test	420 minutes	240 minutes	660 minutes
2018: Student with embedded ELA field test	330 minutes	195 minutes	525 minutes

CHAPTER 2: ITEM DEVELOPMENT AND ITEM BANKING

For details about item and task design and development that occurred prior to the development of the Spring 2018 CMAS field test items, refer to the *PARCC Final Technical Report for 2017 Administration* (Pearson, 2018).

Item Development

Operational Items

Computer-based Items

As previously discussed, 2018 CMAS forms were built to the abbreviated blueprints. Blueprints were modified and assessments were built to measure the same construct as the 2017 CMAS assessments but in a shorter period of time. (See Appendices D1-D3 for additional details.) The operational items used to build spring 2018 forms were items that were previously used.

Colorado educators joined educators from other consortium states for each of the item reviews held for operational items that were included on the 2018 CMAS assessments.

Paper-based TEIs

Beginning in spring 2018, the CMAS paper form was developed to be parallel to the online form, meaning the same passages and items appear on both the paper and computer-based forms. To support that, parallel paper-based items were developed for TEIs in a way that was comparable in terms of student interaction. 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.

Field Test Items

While the spring 2018 operational items were previously-used items, spring 2018 field test items were developed solely by Colorado. Field test items that appeared on the 2018 assessment were reviewed by committees of Colorado educators exclusively. For each of these meetings, an effort was made to involve educators who were representative of the entire state of Colorado (geographic, gender, and race) and familiar with the Colorado Academic Standards, related instruction, and the assessment interaction and demonstration of achievement of the CAS of different groups of students, including students with disabilities and English learners, taking the CMAS assessments. The following section describes the item development process for the spring 2018 CMAS, including CSLA, field test items.

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 (including CSLA) item and test development:

- The item development process ensures the CMAS (including CSLA) items align to the ESs.
- The CMAS (including CSLA) item development plans were designed to produce and maintain a robust item bank; items are written to address the scope of measured standards, grade-level difficulties, and cognitive complexity.
- For CMAS (excluding CSLA), the item and test development processes promote the equivalency of the online and the paper-and-pencil assessments.
- The CMAS items were developed with the intention of being administered on multiple testing platforms.
- The CMAS (including CSLA) item and test development processes are compliant with industry standards.

The item-writing process used for developing the 2018 field test items was a tiered, inter-related process that began with the development of the test blueprints for each grade level within each subject, continued with designing the item development plan (IDP), and used the IDP to forecast the targeted number of items and associated stimuli across ESs needed to create a robust item bank that would be refreshed over time. Once written, the items went through multiple rounds of review, including contractor and department reviews, content and bias review and data review.

Item Development Plan (IDP)

The IDP was designed to determine the number of passages (ELA, including CSLA, only) and items for each Evidence Statement (ES) needed to construct the assessment based on the blueprint requirements. The item bank was analyzed, and ES, task type, and cognitive complexity gaps were identified. The process is structured in a manner to develop a variety of item types that align directly to the ESs that are aligned to the Colorado Academic Standards.

Each IDP will be updated at the beginning of each item development cycle with development targets that address any task model, passage type, ES, item/task type, and cognitive complexity shortages.

Passage Selection/Development

The initial step of development for CMAS ELA began with the research and selection of high-quality literary and informational texts. Due to availability of appropriate passages and challenges with acquiring permissions, passages to be used on the CSLA forms were commissioned. The number and types of needed passages were determined by the CMAS test construction specifications, a gap analysis of the pool of available passages, and the item development plan. The text/passage selection (and writing) guidelines, task model descriptions and cognitive complexity framework defined the number of texts/passages by text type, genre, length, and complexity. Contractor assessment specialists trained passage searchers to find (or write for CSLA items) relevant and rich texts that permitted a range of content to be developed. The guidelines and descriptions were consistent with what had been used in prior years' development.

Passage searchers and writers submitted the passages for contractor assessment specialists to review and evaluate using approved criteria, including adherence to the cognitive demand, relevance, and purpose of the test and the appropriate use of graphics as needed to improve text comprehension. Test passages were analyzed and rated for text complexity. Contractors checked passages for clarity, correctness of language, appropriateness of language for the grade level, and adherence to style guidelines.

Accepted passages were then presented to CDE for review. Once the passages were accepted by CDE, committees of educators reviewed them for content and bias. The committees were comprised of educators from throughout the state representing a variety of student populations including students with disabilities and students with limited English proficiency. Passages accepted by both CDE and recommended by the educator committees were then prepared for item writing.

Item Writing

Upon approval of the passages for ELA, including CSLA passages, and the IDP for ELA, including CSLA and mathematics, Item Writer Workshops were conducted and facilitated by contractor assessment specialists. Item writing assignments were given to the professional item writers. For CSLA items, item writers proficient in written academic Spanish were trained and developed the CSLA items.

These item writers for CMAS (including CSLA) developed a variety of items, across task types and across ESs. The item writers worked with Pearson and/or Tri-Lin assessment specialists when clarification was needed. The item writers used the ESs; the Colorado Academic Standards (CAS); and secure item specification documents, including item-writing guidelines (universal

design guidelines, bias and sensitivity guidelines, and editorial guidelines), and the item writing checklist to guide them in completing their assignments. These resources were consistent with what had been used in prior years.

Item writers authored the items in ABBI, where Pearson or Tri-Lin assessment specialists completed their initial review. The assessment specialists reviewed and suggested revisions to the items and metadata for the item authors. The item writers made the revisions and resubmitted the items within ABBI.

Contractor Item Review for Quality Assurance

After items are written, Pearson and Tri-Lin assessment specialists evaluated each item specifically for content correctness; grade appropriateness; and ES, CAS, and cognitive complexity alignment. The assessment specialists focused 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 performed additional fact checking to ensure accuracy.

Pearson and Tri-Lin copy editors checked items for clarity, correctness of language, appropriateness of language for the grade level, adherence to style guidelines, and conformity with acceptable item-writing practices.

When appropriate, CR items also were 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. Equation editor items are reviewed for their scorability by a digital content development specialist.

Pearson and Tri-Lin assessment specialists also performed 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, disability, or geographic region. The universal design review also included 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, disability or geographic region. In addition, items were reviewed for visual bias, accessibility for students with disabilities, and convertibility to braille and text-to-speech.

Once the internal reviews were completed, each item’s status was updated in ABBI, and the lead assessment specialist conducted a final content review. Approved items had their status updated again in ABBI and were presented to CDE for review.

Adhering to these resources ensured that each Colorado item measured the ES and standard, was content- and grade-appropriate, was factually accurate, had appropriate answers and distractors, was accessible to all populations required to take the assessments, was free from any bias, and followed the Colorado style.

CDE Pre-Review

CDE reviewed items in ABBI to ensure that the content was correct, the alignment was sound, the cognitive complexity was appropriate, the language and content were grade-appropriate, the graphics were clear and relevant to the item, and free of bias/sensitivity issues.

When CDE completed its review of the CMAS (including CSLA) items, CDE alerted Pearson or Tri-Lin assessment specialists. CDE's comments and determinations regarding the status ("Accept," "Accept with Revisions," or "Reject") of the items were recorded in ABBI.

- Items marked "Accept" needed no more revisions and were ready for Content and Bias review.
- Items marked "Accept with Revisions" were revised per CDE's feedback and, if necessary, re-reviewed by content editors, research librarians, etc. These items were re-reviewed by CDE and reconciled with Pearson's assessment content specialist and either deemed "Accept" or "Reject."
- Items marked "Reject" were rejected and given a status of "Do Not Use" in ABBI. Replacement items were written by an assessment content specialist, and the items went through the same rigorous review process as new items.

Content and Bias Review

Following the completion of the internal Pearson, Tri-Lin, and CDE reviews, Content and Bias Review Committees comprising Colorado educators was convened. The purposes of this educator review were to (1) ensure the items are properly aligned to the content standards, accurately measure intended content, and are grade-appropriate, and (2) identify any potential bias or stereotype in test items. Separate committees were convened for CMAS mathematics, ELA, and accommodated CSLA items. Each committee was comprised of Colorado educators from across the state with diverse backgrounds and experience working with diverse (gender, race/ethnicity, income, geography, etc.) learners, standards and content expertise, and special population (students with disabilities and English learners) expertise. For accommodated CSLA items, an effort also was made to involve educators who teach ELs, were familiar with the instruction and needs of the students in an English language development program that utilizes native language instruction, and were proficient in written Spanish. The meetings were conducted either in person or virtually. They included group training on the expectations and processes of each meeting, followed by breakout groupings into grade/subject working committees where additional training was provided.

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

- uses clear, unambiguous, and grade-level appropriate language;
- avoids construct-irrelevant 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 (e.g., determine how students with visual disabilities will access items with needed visuals/graphics/animation).

The committee made one of three recommendations on every item based on the content and bias review: “Accept,” “Accept with Revisions,” or “Reject.”

Following the CMAS (including CSLA) educator meetings, CDE, Pearson, and Tri-Lin assessment specialists reviewed committee comments, reconciled proposed edits, and finalized item outcomes. ABBI was updated to reflect the edits and outcomes. The approved items and passages were then made ready for inclusion on 2018 forms as embedded field test items.

Data Review

After developments of the items, selected items were placed on the operational assessments in embedded field-test positions. The goal of the field test was to allow for the evaluation of the quality of the items through a review of item performance data to determine if the functioning of the items supported their inclusion in the item pool used for operational forms construction. Following the administration of items in a field-test environment, a committee of educators was convened to review them along with student performance data. Separate data review committees were convened for CMAS mathematics, ELA, and the accommodated CSLA items. Data review committee members were provided item images and metadata, along with classical statistics and differential item functioning (DIF) statistics.

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

DIF analyses for CMAS items were conducted on various subgroups (gender, ethnicity, free and reduced lunch, IEP, and ELs) using Mantel–Haenszel Delta DIF statistics (Dorans & Holland, 1992). The same analysis methods were used for CSLA items, but the DIF analyses are conducted by gender only, due to the population of students taking the form. Classification rules derived from National Assessment of Educational Progress (NAEP) guidelines (Allen, Carlson, & Zelenak, 1999) were used to classify items as having either negligible, moderate, or significant DIF. Items that are classified as moderate or significant DIF are taken to data review.

During the data review meeting, educators were trained to interpret the statistical information and judge the appropriateness of the items presented for data review. The committee members used the data as a tool to direct them toward potential flaws in an item and discussed whether there were construct-irrelevant reasons for a data flag. A data flag, by itself, was not the sole reason an item is rejected. Committee members were instructed that their final judgments about the appropriateness or fairness of an item for any individuals and subgroups encompassed by the data flag should be based on their expertise with their content area and experience as Colorado educators. Committee members reviewed each item and made a recommendation as to whether to “accept” or “reject” it. An accepted item meant that the educators, through their varying expertise, determined that there was not a construct-irrelevant reason for the data flag within the item. A rejected item indicated that the educators determined there was a construct-irrelevant reason for the data flag. Following the meeting and CDE determinations, ABBI is updated by moving accepted items into “Ready for Operational” status.

Item Banking System

Pearson’s proprietary software, ABBI (Assessment Banking and Building solutions for Interoperable assessments), is used to support the item and test development process from initial content authoring through the content review cycles. ABBI is the authoritative source for all content, data, and functionality for all CMAS system components.

ABBI serves as the repository where the item bank is housed, item revisions are catalogued, and items and item metadata are uploaded and revised by assessment specialists. Items can be moved into various statuses, each representing a step in the item development process. The items and associated stimuli are tracked and revisions are recorded from creation through retirement in a secure environment.

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

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 staff for CMAS mathematics and ELA, and Pearson and Tri-Lin staff for CSLA. Once test forms are constructed, CDE then reviews the forms, provides feedback, and gives final approval as described below.

When building test forms, Pearson or Tri-Lin assessment specialists select a set of operational items in accordance with the test blueprint and test construction specifications (see Figures 1 through 4 for test blueprints). Items selected for operational use must meet the blueprint and should include a variety of topics and contexts with specified psychometric targets.

The following guidelines for CMAS mathematics, ELA, and CSLA were used during 2018 form construction:

- adherence to established test blueprints and test construction specifications
 - exact match to blueprint for subclaims
 - same distribution of cognitive complexity as 2017 forms
 - same percentage of TEIs as 2017 forms
- review of the item statistics and adherence to the statistical criteria found in the test construction specifications
 - evaluation of item means, point biserial correlations, and score point distributions
 - evaluation of IRT item parameter estimates: a, b, d1, d2, d3, and d4
 - evaluation of fit statistics
 - mirroring of 2017 test characteristic curves and conditional standard errors of measurement curves
- balance in the representation of gender, ethnicity, geographic regions, and relevant demographic factors
- thorough review of individual items to establish the content within items is up to date and relevant
- selection of items with various stimulus types 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 ESs

- review of full form, including field-test items, for instances of clueing and/or content overlap

After the initial operational item pull is complete, the test form is reviewed by Pearson and Tri-Lin assessment specialists. The assessment specialists verify that the form meets the blueprint and test construction specifications (i.e., the required ES coverage, domain and subclaim coverage, cognitive complexity allocation, task type). The form is then presented to a Pearson psychometrician for analysis, and the psychometrician verifies that the form falls within the established psychometric and blueprint parameters. The psychometric lead also identifies the anchor item set within each operational form. Chapter 8 of this report provides more details about the anchor sets.

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

After the operational form is approved, field-test items are selected from the item bank. Items chosen for field-testing are placed on a form in a designated section and sequence. Pearson and Tri-Lin assessment specialists assemble field-test sets of items so that they comprise the appropriate distribution of standards, subclaims, task types, topic coverage, cognitive levels, and key distributions to meet the required item refresh rates in following years.

Online Forms

The majority of students take the CMAS mathematics and ELA assessments online. Using this format allows not only for the use of innovative item types but also for additional accessibility options and accommodations as described in Chapter 4 (e.g., text-to-speech).

Accommodated Test Forms

Accommodated test forms for CMAS mathematics are available for both the online and paper-based forms. For online forms, text-to-speech and color contrast are available in English and Spanish. 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. English oral scripts are also available for local translation into languages other than Spanish.

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

Paper

Paper-based versions of the CMAS mathematics and ELA assessments are published and are available if needed for an accommodation or for schools that choose not to test online, as allowed by state law. A Spanish transadaptation is also available on paper for CMAS mathematics.

As discussed in the previous chapter, the paper form is parallel to the online form, meaning paper and online forms of the 2018 CMAS mathematics and ELA assessments included the same operational items. Parallel paper-based items are developed for TEIs. In some cases, this is achieved with traditional SR items. In other cases, it requires an item that must be human-scored. For example, a drag-and-drop item may be converted to an item in which the student must draw a line between various draggers and drop bays. During equating, the statistics of the TEI are compared to the paper-based version to confirm equivalence.

CMAS CSLA is the accommodated version of CMAS ELA for eligible Spanish-speaking students in grades 3 and 4 and is administered on paper.

Braille

After approval of the CMAS mathematics and ELA paper test materials, a braille version of the assessments is created according to the process outlined below:

1. Pearson posts final test forms as PDFs to subcontractors (National Braille Press (NBP) for ELA and Region 4 for mathematics).
2. Subcontractor 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. Subcontractor 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, subcontractor staff, and 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 subcontractor for updates to the braille form.

11. The braille form is finalized and printed.

Large Print

Large print versions of the CMAS mathematics, ELA, and CSLA assessments are also created. The large print versions are a 50% enlargement of the regular paper form and are printed on 14" × 18" paper. When needed, the large print version includes a Visual Description booklet, which 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 CMAS (including CSLA) administration procedures. Prior to the administration of the assessments, districts, schools, and teachers (Test Administrators) were to ensure that their students and systems were prepared for the assessments. Such information was communicated to the appropriate individuals via manual and in-person and recorded trainings as described below.

Manuals

Several manuals were created to aid with the CMAS administration, described in the following sections.

CMAS Test Administrator Manual for Computer-Based Testing and the CMAS Test Administrator Manual for Paper-Based Testing

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

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/Sensitive Data personnel who utilize PearsonAccess^{next}.

CMAS and CoAlt Procedures Manual

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

Administration Training

Administration training is intended to make sure all individuals involved in CMAS assessment activities at the school and district levels are prepared to follow administration processes and procedures with fidelity, as well as support adherence to security procedures. Fidelity to standardized test administration processes and procedures helps to ensure the comparability of resulting scores and accurate interpretation of results. Thorough in-person regional trainings were conducted by CDE and Pearson personnel across the state. CDE and Pearson presented trainings to the DACs that contained information regarding proper procedures for administration, security requirements, receiving and returning materials to Pearson, and the use of PearsonAccess^{next} with TestNav 8. Additionally, recorded versions of the live trainings were posted on the CDE Assessment Unit website. Administration training materials, including slide decks, manuals, and how-to guides were also available through on the CDE Assessment Unit website for training SACs and Test Administrators. After CDE trained DACs, the DACs trained School Assessment Coordinators, Test Administrators, and any other individuals within the district who planned to participate in the 2018 administration.

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

Accessibility and Accommodations

Accessibility features and accommodations provided in 2018 were consistent with those offered to students in 2017. Accessibility was considered from the beginning of the test development process and is inherent within the CMAS assessment and administration. For example, the CMAS mathematics and ELA 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 for mathematics, 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. Similarly, CSLA was designed to be linguistically accessible for eligible Spanish-speaking students.

Beyond the tools and accessibility features, assessment accommodations were available to the population of students who had IEP, 504, or EL plans. Accommodations provide a student with an opportunity to engage with the assessment while not affecting the reliability or validity of the assessment. Accommodations can be adjustments to the test presentation, materials,

environment, or response mode of the student and are based on individual student need. Accommodations should not provide an unfair advantage to any student. Providing an accommodation for the sole purpose of increasing test scores is not ethical.

Accommodations must be documented and used regularly during classroom instruction and assessments prior to the assessment window to ensure the student can successfully use the accommodation. Although accommodations are used for classroom instruction and assessments, some may not be appropriate for use on statewide assessments. As a result, it is important that educators become familiar with the state assessment policies about the appropriate use of accommodations and that districts have a plan in place to ensure and monitor the appropriate use of accommodations. Certain accommodations are allowed only in special cases with CDE approval, due to being an inherent violation of the intended construct (e.g., auditory presentation for ELA and CLSA, which are intended to measure reading ability).

Some of the available accommodations for CMAS include CSLA in place of ELA (other linguistic accommodations do not apply as CSLA is the linguistic accommodation), English oral scripts (mathematics and with CDE approval for ELA), Spanish oral scripts (mathematics and with CDE approval for CSLA), oral scripts for signed presentation and local translation into languages other than English and Spanish, braille forms, large print forms, assistive technology forms for screen readers/non-screen readers, and Spanish text-to-speech forms for mathematics.

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

Test Security

Procedures described in this section were put in place to enhance the likelihood that security was maintained before, during, and after the 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. DACs were responsible for overseeing training for the district, including verifying that the DTC and SACs were trained. SACs were responsible for ensuring that Test Administrators, Test Examiners, and all individuals involved in test administration at the school level were trained and subsequently acted in accordance with all security requirements. A chain of custody plan for materials was required to be written and implemented to ensure materials were securely distributed from DACs to SACs to Test Administrators/Test Examiners and securely returned from Test Administrators/Test Examiners to SACs and then to DACs. SACs were required to distribute materials to and collect materials

from Test Administrators/Test Examiners each day of testing, and securely store and deliver materials to DACs after testing is completed in accordance with the instructions in the 2018 CMAS/CoAlt Procedures Manual.

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

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

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

CHAPTER 5: SCORING THE ASSESSMENT

The CMAS assessments contain various item types. CMAS ELA assessments contain selected response (SR); technology enhanced (TE), including parallel paper-based versions; and prose constructed-response (PCR) items. Since it is administered on paper, CSLA forms only contain SR and PCR items. CMAS mathematics assessments contain SR, TE and constructed response (CR) items. SR and TE items are machine-scored, with point values varying by item type and assessment. CMAS mathematics CR items are hand scored. The PCR items are scored on two trait dimensions using a combination of human and automated scoring. The holistic rubrics used to score the CMAS ELA and CSLA PCR items can be found in Appendix A. For CMAS ELA, a portion of PCR item responses are scored using an automated scoring engine that has been trained with human scored responses.

To maintain comparability with the scoring prior to 2018, scoring rules for SR and TE machine scored items (e.g., multiple choice, drag-and-drop, etc.) as well as CR items (i.e., used prior years' rubrics, anchor papers, rules and scoring methods with the exception of paper-converted technology-enhanced items) were preserved from previous years.

Pearson's Performance Scoring team implemented the CR and human PCR scoring process for CMAS mathematics and ELA, including CSLA. The CR scoring process is described below for operational scoring, including the Spanish language arts accommodated form and field test scoring, including the Spanish language arts accommodated form. In addition, there are sections addressing the rangefinding process and the major components of the quality assurance system including backreading, calibration, and validity papers.

Machine Scoring

To maintain comparability with the scoring prior to 2018, scoring rules and processes for operational SR and TE machine scored items (e.g., multiple choice, drag-and-drop, etc.) were preserved from previous years.

Machine scored items include key-based items and rule-based items. Key-based items tend to be a version of multiple choice and multiple select (students are selecting more than one correct answer). Rule-based items are machine scored technology-enhanced items.

Initial scoring expectations are developed during item development and are included as part of the item review process. The scoring rules and correct responses are included in the items' XML coding.

For all items that are machine scored, prior to scoring, key checks are completed to verify that the machine is correctly identifying correct and incorrect responses. When there is a discrepancy in the scoring, content experts review the item and adjustments are made as needed. During testing, actual distribution of scores is compared to expected distribution. Further evaluation is completed when a discrepancy is identified.

Human Scoring

Operational Scoring

Human scored operational items utilized the same scorer credentials and qualifications; rubrics; training responses; qualification responses and processes; validity responses and reliability processes and statistics as in prior years.

Each operational assessment is scored using either a Distributed or Regional Scoring model depending upon content area. Items on the CSLA form and paper-based TEIs were regionally scored while all other human scored were completed relying on distributed scoring. Scoring includes several components that together provide a comprehensive performance scoring model.

- All scorers were required to pass a background check and sign a nondisclosure agreement agreeing to adhere to all security and confidentiality requirements.
- All scorers had at a minimum, a 4-year degree. Scorers were assigned to content areas based on their educational backgrounds, related fields of work and their demonstrated knowledge in the content area.
- Scorers of items appearing on the CSLA forms had to be proficient in written Spanish and English languages.
- Scorers were trained using comprehensive training materials developed by scoring experts. These materials rely on student responses scored at the rangefinding meetings by educators from the multi-state consortium, including from Colorado. Prior to qualifying for an item, scorers review an online training module that includes an overview of scoring; information specific to the item, like the prompt and rubric; and anchor sets. Scorers then scored multiple practice sets prior to attempting qualification. After successful qualification, scorers began scoring the item.
 - For CSLA items, stand-up training was led by a Pearson scoring director who presented item-specific materials, including the prompt and rubric. The scoring team then received training on anchor sets prior to moving into the online portion of training where scorers applied scores on multiple practice sets within the electronic scoring system. After each practice set, the scoring director reviewed the practice set results with the scorers prior to scorers taking qualification sets. After successful qualification, scorers began scoring the item.
- Scorers had to pass a qualifying test for the item types that they scored. Qualification sets were designed to test scorer accuracy across the range of score points for a given item.
- Student responses were converted to electronic images at Pearson facilities. They were then transmitted for computer-based scoring.
- Distributed scorers were located across the United States and worked from their homes. Their computers were set up for image-based scoring. A comprehensive set of scoring

and monitoring tools were integrated into the scoring system. In addition to the systemic tools, content supervisory staff were available by phone to help answer any training or scoring related questions that may arise. With distributed scoring, scorers were able to score 7 days per week with extended evening hours.

- Regional scorers were located within a physical scoring location site. As with distributed scoring, regional scoring also utilized a comprehensive set of scoring and monitoring tools integrated into the scoring system. In addition to the systemic tools, content supervisory staff were physically on site to help answer any training or scoring related questions that arose. Unlike distributed scoring, regional scoring was typically only completed Monday through Friday during normal business hours. Regional scorers were used for CSLA forms and paper-based TEIs.
- Additional security procedures were in place for distributed scoring. Data were securely transmitted through HTTPS and SSL technology using secure protocols for system authentication. Student responses were randomly routed through the scoring platform preventing scorer knowledge of student information, unless a student self-identified in the response. Scorers agreed not to use shared, institutional, or public computers to score and also not to save student responses or test materials. Scorer printing capabilities of materials, such as anchor papers, were only approved for printing after they had undergone and passed a personally identifiable information review by CDE. Scorers agreed to securely destroy or return to Pearson printed materials at the conclusion of scoring.

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

Field Test Scoring

Embedded field test scoring was completed using regional scorers. Regional field test scoring took place in San Antonio, TX, Austin, TX, Virginia Beach, VA, and Columbus, OH, for ELA; and San Antonio, TX, Austin, TX, Columbus, OH, Iowa City, IA, and Mesa, AZ, for mathematics. All scorers were required to have a four-year college degree.

Field test scorers received stand-up training led by a Pearson scoring director who presented item-specific materials, including the prompt and rubric. Scorers then reviewed the anchor sets in a group setting prior to scoring practice sets on paper.

Rangefinding

Constructed-response (CR) items were scored using rubrics. For mathematics, rubrics were generated for each unique item, while ELA used holistic rubrics for each item type. Rubrics were finalized at rangefinding and were maintained, along with the training materials for each item, by Pearson's Scoring Services group.

Rangefinding meetings were held following the administration in which an item was field tested. The purpose of rangefinding was to define the range of performance levels within the score points of the rubrics using student responses. Each rangefinding committee included Pearson's Scoring Services and content staff, state content representatives, and educators with relevant grade-level and content expertise and experience with special populations. Participants created consensus scores for a sample set of student responses that were subsequently used to develop effective training materials for scoring of CR items.

Pearson's scoring directors constructed one rangefinding set per item, which included approximately 30 responses. For multi-point items, pre-constructed sets with additional responses were brought to the meeting. Responses included in these sets represented the full spectrum of scores to the greatest extent possible. For each item, the responses were ordered based on estimated score from high-scoring to low-scoring; however, actual scores were not revealed to committee members. Each set included responses clearly earning each available score point for each type of question. The set also included 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 the Scoring Services group, the rangefinding committee was divided into several break-out groups based on educator expertise. Each group was assigned a range of field-test items to be reviewed, following the process outlined below:

1. The Scoring Director introduced each item. The committee reviews the item and corresponding rubric.
2. The committee read student responses—individually or as a group—and then discussed and decided the most appropriate score for each response.
3. The Scoring Director recorded committee members' comments as well as the final consensus score for each student response. Consensus was reached when a majority of committee members agreed upon a particular score point for a response and all members agreed to accept the score of the majority.
4. A designated committee member recorded consensus scores. After reviewing responses for each item, the committee member compared his or her notes with those kept by the Scoring Director and provided sign-off to indicate agreement with the recorded scores.

Following the rangefinding meetings, Scoring Services personnel created training material with an anchor set, which was used for initial training (up to 12 responses), and a full practice set (up to 10 responses). Each CR item was then scored with the associated training materials.

Backreading

Backreading is the method of immediately monitoring a scorer's performance and is, therefore, an important tool for Pearson's scoring supervisors. Backreading is performed in conjunction with the statistics provided by reader performance reports and as indicated by scoring directors, allowing scoring supervisors to target particular readers and areas of concern. Scorers showing low inter-rater agreement or those showing anomalous frequency distributions are given immediate, constructive feedback and monitored closely until sufficient improvement is demonstrated. Scorers who demonstrate through their agreement rates and frequency distributions that they are scoring accurately will continue to be spot-checked as an added confirmation of their accuracy. An explanation of rater agreement statistics can be found in Chapter 9, and rater agreement statistics for the spring 2018 administration can be found in Part II of this report. The targets agreement rates are as follows:

Math, Science, Social Studies:

1-point item: 90% perfect and 96% perfect plus adjacent agreement

2-point item: 80% perfect and 96% perfect plus adjacent agreement

3-point item: 70% perfect and 96% perfect plus adjacent agreement

4+-point item: 65% perfect and 96% perfect plus adjacent agreement

ELA and CSLA:

3-point and 4-point items: 65% perfect and 96% perfect plus adjacent agreement

Calibration

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

Validity Papers

Validity is a quality monitoring tool used during scoring. Validity papers are student responses chosen by Pearson scoring directors to measure accuracy of a scorer when applying the scoring rubric. Validity papers are blind to scorers, which means a scorer is not aware when they are scoring a validity paper. Scoring directors may choose to include an annotation with a validity

paper so that a scorer will receive immediate feedback if a validity paper is scored incorrectly. This is known as validity as review. Validity statistics are monitored by scoring directors throughout the life of a scoring project.

Automated Scoring

Consistent with prior years' scoring, automated scoring performed by Pearson's Intelligent Essay Assessor (IEA) was the default option for scoring the CMAS ELA assessments' online prose constructed-response (PCR) tasks. Under the default option, it was assumed that operational scores for approximately 67% of the online PCR responses would be assigned by IEA for the spring administration. The operational scores for the remaining online responses were assigned by human scorers. For the 2018 administration, all but one of the PCR items used scoring models developed and used in previous years, with the remaining item using 100% human scoring.

Items that used automated scoring were also given second scores through human scoring. Ten percent of responses were randomly selected and given a second reliability score to provide data for evaluating the consistency of scoring. Some responses received a second human score based on Smart Routing of particular score points. This procedure is described in more detail below.

The IEA scoring engine was trained on item data collected during the field testing phase and then further refined using operational test data where necessary. Human scorers provided scores, and the responses and scores were fed into IEA, which used them to develop an algorithm for how to score the item. For CMAS 2018, all items had been used operationally on previous assessments, so no training was done on the operational data.

Continuous flow

For IEA training for operational items completed in prior years, continuous flow scoring resulted in an integrated connection between human scoring and automated scoring. It refers to a system of scoring where either an automated score, a human score, or both could be assigned based on predetermined criteria.

For IEA training for operational items completed in prior years, continuous flow scoring facilitated the training of IEA using human scores assigned to operational online data collected early in the administration in cases where that was necessary. Once IEA obtained sufficient data to train and meet reliability criteria targets, it could be "turned on" and becomes the primary source of scoring (although human scoring continued for the 10 percent reliability sample and other responses that were routed accordingly).

Smart routing

The use of “smart routing” during operational scoring increased the quality of automated scoring by routing responses that based on IEA training were more likely to disagree with a human score to receive an additional human score.

When human scorers read a response, they typically apply integer scores based on a scoring rubric. For example, when there is strong agreement between two independent human scorers, they might both assign a score of 3, such that the average score over both raters is also a 3 (i.e., $(3+3)/2 = 3$). IEA simulates this behavior, but because its scores come from an artificial intelligence algorithm, it generates continuous (i.e., decimal-valued) scores. In the previous example, the IEA score might be a 2.9 or 3.1. Similarly, if the human scorers disagreed on a response and scored it as a 3 and a 4, for example, IEA would likely provide a score between 3 and 4 (e.g., 3.4 or 3.6). This continuous IEA score needs to be rounded to an integer score for reporting (i.e., a 3 or a 4, depending on rounding rules, in this example). Smart routing involves routing those responses with “in between” IEA scores to additional human scoring because the nature of the responses suggests there may be less confidence in the IEA score. Since these “in between” IEA scores are based on modeling human scores, it follows that human scores may be less certain as well. Therefore, responses are more likely to be double-scored and resolved if the IEA and human scores are non-adjacent.

Smart routing was utilized as needed to achieve targeted quality metrics (e.g., validity agreement or agreement with human scorers). Smart routing involved the application of the following steps:

1. The continuous IEA score for each of the two trait scores was rounded to the nearest score interval of 0.2, starting from zero. For example, IEA scores between 0 and 0.1 were rounded to an interval score of 0, scores between 0.1 and 0.3 were rounded to an interval score of 0.2, scores between 0.3 and 0.5 were rounded to an interval score of 0.4, etc.
2. Within each of these intervals, the percentage of exact agreement between IEA integer scores and the human scores was calculated for each trait.
3. For each prompt, agreement rates were evaluated for each interval and for each trait.
4. Responses within intervals for which IEA–human agreement on either trait was below a designated threshold were routed for additional human scoring.

For additional details about the development and training of IEA, see the *PARCC Final Technical Report for 2017 Administration* (Pearson, 2018).

Quality criteria for evaluating automated scoring

For operational scoring, consortium states had previously approved specific quality criteria for evaluating automated scoring at the time IEA was trained. The primary evaluation criteria for IEA was based on responses to validity papers with “known” scores assigned by experts. For each prompt scored, a set of validity papers was used to monitor the human-scoring process over

time. Validity papers were seeded into human scoring throughout the administration. The expectation is that IEA can score validity papers at least as accurately as humans can.

Additional measures of inter-rater agreement for evaluating automated scoring were proposed based on the research literature (Williamson, Xi, & Breyer, 2012). These measures were previously utilized in Pearson's automated scoring research and include Pearson correlation, kappa, quadratic-weighted kappa, exact agreement, and standardized mean difference. These measures are computed between pairs of human scores, as well as between IEA and humans, to evaluate how performance was the same or different. Criteria for evaluating the training of IEA given these measures include the following:

- Pearson correlation between IEA and human scores should be within 0.1 of human–human correlation.
- Kappa between IEA and human scores should be within 0.1 of human–human kappa.
- Quadratic-weighted kappa between IEA and human scores should be within 0.1 of human–human quadratic-weighted kappa.
- Exact agreement rate for IEA and human scores should be within 5.25% of the human–human exact agreement rate.
- Standardized mean difference between IEA and human scores should be less than 0.15.

The specific criteria for evaluating IEA included both primary and secondary criteria, described below.

Primary criterion. The performance of IEA was evaluated by comparing IEA scores with human scores for the set of validity papers. The primary criterion is stated as follows: *With smart routing applied as needed, IEA agreement is as good as or better than human agreement for each trait score.* For a given prompt, this criterion is operationalized as follows:

1. Determine agreement of the human scores with the validity papers for each trait.
2. Calculated agreement of the IEA scores with the validity papers for each trait.
3. Compare the IEA and human agreement on the validity papers.
4. If the IEA validity agreement is greater than or equal to the human agreement for each trait, IEA can be deployed operationally.

In addition to looking at overall validity agreement, conditional agreement was also examined. In general, it was desirable for IEA to exceed 65% agreement at every score point and to be close to or exceed the human validity agreement at each score point.

Contingent primary criterion. For many of the prompts trained in 2017, it was not possible to utilize human-scored validity responses in evaluating IEA performance. In these cases, IEA was evaluated based on IEA–human exact agreement for each trait score and compared to agreement based on responses that were double-scored by humans. IEA–human agreement was evaluated on a portion of the data, according to the following steps:

1. Determine exact agreement of the two human scores with each other for each trait.
2. Calculate agreement of the IEA scores with the human scores for each trait.
3. Compare the IEA–human agreement with the human–human agreement.
4. If the IEA–human agreement is within 5.25% of the human–human agreement, IEA can be deployed operationally.

In addition to the overall comparison, the following performance thresholds were targeted in the test data set: 1) at least 65% overall IEA–human agreement, and 2) 50% IEA–human agreement by score point (i.e., conditioned on the human score). These targets went beyond the contingent primary criteria approved by the state leads.

Secondary criteria. The secondary criteria involve comparing agreement indices for IEA–human scoring for various demographic subgroups, and can be stated as follows: *With smart routing applied as needed, IEA–human differences on statistical measures for each trait score are within the Williamson et al. (2012) tolerances for subgroups with at least 50 responses.* IEA–human agreement was evaluated according to the following steps:

1. Determine exact agreement of the two human scores with each other for each trait.
2. Calculate agreement of the IEA scores with the human scores for each trait.
3. Compare the IEA–human agreement with the human–human agreement.
4. For subgroups with at least 50 IEA–human scores and at least 50 human–human scores, compare agreement indices to the following criteria:
 - Pearson correlation between IEA–human should be within 0.1 of human–human.
 - Kappa between IEA–human should be within 0.1 of human–human.
 - Quadratic-weighted kappa between IEA–human should be within 0.1 of human–human.
 - Exact agreement between IEA–human should be within 5.25 percent of human–human.

- Standardized mean difference between IEA–human should be less than ± 0.15 (this criterion was applied to subgroups with at least 50 IEA–human scores).

Although it was not expected that these criteria would be met for all subgroups for all prompts, if results of the evaluation between IEA and human scoring for subgroups for any prompt indicated that IEA performance persistently failed on the criteria listed above, consideration would be given to resetting the responses scored by IEA and reverting to human scoring until such time that an alternate IEA model could be established with improved subgroup performance.

In addition to the secondary criteria above, the performance of IEA was also compared with the following targets on the various measures for subgroups with at least 50 responses:

- Pearson correlation between IEA–human should be 0.70 or above.
- Kappa between IEA–human should be 0.40 or above.
- Quadratic-weighted kappa between IEA–human should be 0.70 or above.
- Exact agreement between IEA–human should be 65 percent or above.

Hierarchy of assigned scores for reporting

When multiple scores are assigned for a given response, the following hierarchy determines which score was reported operationally:

- the IEA score is reported if it is the only score assigned
- if an IEA score and a human score are assigned, the human score is reported
- if two human scores are assigned, the first human score is reported
- if a backread score and human and/or IEA scores are assigned, the backread score is reported
- if a resolution score is assigned and an adjudicated score is not assigned, the resolution score is reported (note that if nonadjacent scores are encountered, responses are automatically routed to resolution)
- if an adjudicated score is assigned, it is reported (note that if a resolution score is nonadjacent to the other scores assigned, responses are automatically routed to adjudication)

CHAPTER 6: STANDARD SETTING

To support the interpretation of student results, student performance on the CMAS mathematics, ELA, and CSLA assessments is described in terms of five performance levels. The performance levels and their detailed descriptions are as follows:

Level 5: Exceeded expectations

Students who demonstrate a distinguished command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They are *academically well prepared* to engage successfully in further studies in this content area.

Level 4: Met Expectations

Students who demonstrate a strong command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They are *academically prepared* to engage successfully in further studies in this content area.

Level 3: Approached Expectations

Students who demonstrate a moderate command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They will *likely need additional academic support* to engage successfully in further studies in this content area.

Level 2: Partially Met Expectations

Students who demonstrate a limited command of the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They will *need additional academic support* to engage successfully in further studies in this content area.

Level 1: Did Not Yet Meet Expectations

Students who do not yet meet academic expectations for the concepts, skills, and practices embodied by the Colorado Academic Standards assessed at their grade level. They will need extensive academic support to engage successfully in further studies in this content area.

Performance Standards vs. Cut Scores

The performance levels and their descriptions provided above are policy definitions, developed to describe students' performance on the assessment directly in terms of the knowledge, skills, and practices the assessment is intended to measure. Standard setting is the process of translating those policy-driven performance standards into scores on the assessment. The purpose of a standard-setting study is to determine the boundaries—or cut scores—along the score scale that differentiate student performance among those levels (e.g., Cizek, Bunch, & Koons, 2004; Kane, 1994).

CMAS and CSLA Standard Setting

The standards and cut scores used for the CMAS 2018 mathematics and ELA assessments were set in 2015 in collaboration with the PARCC consortium. Details about the standard-setting process can be found in the PARCC *Performance Level Setting Technical Report* (Davis & Moyer, 2015).

CSLA standards were set in 2016 and details are available in the *CSLA Colorado Spanish Language Arts Technical Report* (Colorado Department of Education, 2016).

CHAPTER 7: REPORTING

Several score reports are generated to communicate student performance on the CMAS mathematics, ELA, and CSLA 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 *2018 CMAS and CoAlt Interpretive Guide* (Colorado Department of Education, 2018), available at http://www.cde.state.co.us/assessment/cmas_coalt_interpretive_guide_2018.

Description of Scores

CMAS mathematics, ELA, and CSLA reports provide information on student performance in terms of scale scores, performance levels, percentile ranks, and percent earned 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. Students receive scale scores in each of the following areas:

- Overall test
- Reading claim (ELA and CSLA only)

The overall scale for each test ranges from 650 to 850 which was retained from prior years of the CMAS math and ELA assessments. The Reading scale ranges from 110 to 190. The reading scale score is comparable to the scale scores reported in previous years, except that 100 was added to all scores to better differentiate it from a percent earned score. Chapter 8 provides technical details related to scale development.

Performance Levels

Performance levels and performance level descriptors (PLDs) 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-setting studies (see Chapter 6). Consistent with prior years of the CMAS Math and ELA assessments, there are five performance levels: Did not yet meet expectations, Partially met expectations, Approached expectations, Met expectations, and Exceeded expectations. Students in the top two categories (i.e., Met expectations and Exceeded expectations) are considered to be on track to being college- and career-ready in that content area. The 2018 performance levels, along with the performance level descriptors and the expectations within each performance level, were retained from previous

administrations of the assessments. Tables 1 and 2 list the scale score ranges associated with each performance level for each assessment.

Percentile Ranking

Percentile rankings based on the overall scale score are provided on student performance reports to indicate how the student performed compared with other students in the state. For example, a student with a percentile ranking of 70 performed better than 70 percent of students in Colorado.

Percent Earned

The blueprint reduction from CMAS 2017 lead to changes in how claims and subclaims were reported. Prior to 2018, the Writing Claim was reported as a scale score with a proficiency level cut. Special studies conducted to analyze the impact of reducing the blueprint on the consistency of scale scores and performance levels suggested that the writing claim scale score was not stable enough to continue with the scale score and maintain confidence that it was consistent with previous years. A similar finding was observed for the subclaims which previously were reported as performance levels based on the cut scores used on the overall scale. These were found to be inconsistent with previous years and had low reliability. The results of this study are discussed in Part II and the full studies are available in Appendix D?.

To prevent incorrect interpretations and give teachers and parents a metric that is more generally understood, students' performance on the Writing claim (ELA and CSLA) as well as the subclaims in ELA, CSLA, and mathematics are reported as the percentage of points earned within each reporting category. Unlike scale scores, percent of points possible scores cannot be compared across years, because individual items change from year to year and are not constructed to be comparable in difficulty at the claim or subclaim level. In addition, performance on different subclaims cannot be compared within an administration, because the number of items and the difficulty of the items within each subclaim may not be the same.

The percent of points possible score can be compared to aggregated state, district, and school performance on that reporting category. The student performance reports also include an indicator of how students who scored just above the Met Expectations cut score on the overall assessment performed on each category. This indicator gives similar information to the Met Expectations cuts that were used prior to 2018.

Score Reports

Sample CMAS mathematics and ELA student performance reports can be found in Appendix B. CSLA assessments are parallel and comparable to CMAS ELA assessments in scoring and reporting. Therefore, separate CSLA reports are not included (please refer to the CMAS ELA examples). Two types of score reports are provided: student level and aggregate. For a detailed

explanation of the information provided in the reports, refer to the *2018 CMAS and CoAlt Interpretive Guide* (Colorado Department of Education, 2018).

Student Performance Reports

Student Performance Reports provide information about the performance of a particular student. The student's scale score(s), associated performance level, percentile ranking, and percent of points possible scores are displayed on a two-page report, along with comparative information related to the student's school, district, and state performance. In addition, PLDs are provided.

In addition to the electronic versions made available to districts and schools, two copies of Student Performance Reports were printed and shipped to districts for distributing to parents/legal guardians and for maintaining locally.

Aggregate Reports

Several types of aggregate reports are produced for schools and districts.

- Performance Level Summaries
- Content Standards Rosters
- Evidence Statement Analysis Reports
- Student Rosters (school level only)
- District Summary of Schools (district level only)

These reports are produced at the school and/or district level and provide summary information for a given school or district. District and school reports are provided electronically through PearsonAccess Test Results, and access to the reports is limited to authorized users. Examples of each type of aggregate report and a detailed explanation are provided in the *2018 CMAS and CoAlt Interpretive Guide* (Colorado Department of Education, 2018).

CHAPTER 8: CALIBRATION, EQUATING, AND SCALING

Item Response Theory (IRT) was used to develop, calibrate, equate, and scale the CMAS mathematics, ELA, and CSLA assessments. The two-parameter logistic (2PL) (Birnbaum, 1968) and generalized partial credit (GPC; Muraki, 1992) models were applied to CMAS mathematics and ELA, and the Rasch partial credit model (RPCM) was applied to CSLA. 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. For CMAS mathematics and ELA the scales were equated to the previous CMAS (i.e., PARCC) base scale. The calibration of the first operational administration determined the base scale for CSLA.

IRT Models

CMAS Mathematics and ELA

The 2PL and GPC IRT models

The item response functions (IRFs) of the 2PL and GPC IRT models relate 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 (Birnbaum, 1968), uses two item parameters to relate the probability of person i correctly answering a dichotomously scored item j :

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

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 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 :

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

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 5, 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 6 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 7 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:

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

where N_{ij} 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 statistic was then transformed so that the value could be evaluated using the chi-square distribution:

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

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 $Z_{Q_{1i}}$ 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.

CSLA

Rasch Partial Credit Model

The RPCM is an extension of the Rasch one-parameter IRT model attributed to Georg Rasch (1966), as extended by Wright and Stone (1979), Masters (1982), and Wright and Masters (1982). The RPCM was selected because of its flexibility in accommodating various item types (i.e., multiple-choice items and items with multiple response categories).

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

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

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.

If the IRT models fit the data and the model assumptions are met, calibration of test items places both items and students on a scale that is independent of any particular sample of students up to a linear transformation. Equating is used to determine and apply a scale transformation that allows

for meaningful comparisons of student performance across different forms or administrations of the test.

Operational Equating and Scaling

Equating is used to place new forms onto the operational base scale. In order to maintain comparability with prior administrations, CMAS mathematics and ELA item parameter estimates for the spring 2018 administration were equated to the established base scale used in 2017. For CSLA, the spring 2018 item parameter estimates were equated to the spring 2016 CSLA base scale.

Calibrations

Calibration refers to the estimation of item parameters in the IRT framework, which places items and students on a common scale. In order to obtain item parameter estimates for CMAS mathematics and ELA, the 2PL or GPC model was applied to the items. Dichotomous items were fit to the 2PL model, and polytomous items were fit to the GPC model. IRTPRO (SSI, Inc., 2011) was used for all calibrations, and all operational item parameters were estimated in a single calibration (i.e., concurrent calibration) for each assessment. For CSLA, the RPCM was applied to all items in order to obtain item parameter estimates. All operational items within a grade were also calibrated concurrently. Winsteps (Linacre, 2011) was used for all CSLA calibrations.

For CMAS ELA and CSLA, PCR items were calibrated at the (unweighted) trait score level rather than as aggregated scores. To account for potential local dependence between the two trait scores, the item response matrix for CMAS ELA was modified before operational calibrations. For each PCR item, one of the two trait scores for each student was randomly selected, and the non-selected trait score was then removed from the dataset and treated as missing for calibration. The resulting item response dataset, known as a “Moulder” matrix, contained roughly half as many observations for each PCR trait score as for the non-PCR items. However, the datasets still contained an adequate number of student responses to conduct the calibrations. Due to the small population of students taking the CSLA assessment, trait scores were not removed from the data when conducting calibrations for CSLA.

Equating design. A common items approach is used for equating operational forms for the CMAS mathematics and ELA 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.

Consistency of constructed-response scoring check. The CMAS assessments include a high percentage of constructed-response (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 selected-response (SR) items, 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, the consistency of the CR scoring was examined prior to equating via the rescoring of a subset of the previous year’s papers to remove any items that exhibit statistical 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, 2014) as outlined below:

1. Obtain the theoretically weighted estimated posterior theta distribution using 31 quadrature points (-5 to 5).
2. Compute the slope and intercept constants using the Stocking & Lord method 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:

$$D_i^2 = \sum_k [P_{ix}(\theta_k) - P_{iy}(\theta_k)]^2 \bullet g(\theta_k)$$

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 stability check and consistency of constructed-response scoring check were examined, and consideration was given to the impact of flagged item(s) on the content representativeness of the resulting anchor set. A flag alone was not the sole criteria for removing an item from the linking set. It was important to also make sure that the remaining anchor set continued to be representative of the overall content and structure of the test.

Equating method. Using the item parameter estimates for the anchor set from the PARCC bank and the current administration, the computer program STUIRT (Kim & Kolen, 2004) was used to obtain the transformation constants to place the current administration's items on the operational scale using the Stocking & Lord (1983) method. 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), as follows:

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

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

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

Paper forms. Online and paper items were developed to be parallel to the online items. Operational paper items deemed identical to operational online items were assumed to have the same item parameter estimates. Paper items were fixed to their online counterparts' item parameter estimates. This process produced item parameter estimates for all paper items.

CSLA

Equating design. Two equating models were used for the spring 2018 assessments: a pre-equating model for grade 3 and a post-equating model for grade 4.

The pre-equating process involves linking the newly developed test form before it is administered to a set of items that appeared previously on one or more test forms. The pre-equated score tables that are created are then applied to the operational test administration results and no calibration of the current operational test is necessary. Using this process, the difficulty level of the newly developed test form is known prior to administration, and the anticipated raw scores that correspond to scale scores at performance standards can be identified.

For the post-equating process for grade 4, the spring 2018 student data for the operational items were calibrated using Winsteps v3.73. The Rasch item difficulties (RIDs) and the step deviation values for non-anchor operational items were placed on the 2016 base scale using anchor items. The anchors (i.e., common items) are operational items that are already equated to the base scale and placed on forms from adjacent administrations. The anchor items are placed in the same positions across all test forms within a grade, and anchor the scale between the new test form and

the base scale. The anchor items are selected to be representative of the blueprint in terms of content and item type. The anchors represent approximately 30% of the points on the test form.

Stability check. The stability check for the CSLA anchor items was conducted using classical item analysis, scatter plots of item difficulty, and displacement estimates from Winsteps. Items were flagged if the absolute value of the displacement estimate was greater than or equal to 0.30.

Final anchor sets. Items flagged from the stability check were examined, and consideration was given to the impact of flagged item(s) on the content representativeness of the resulting anchor set. A flag alone was not the sole criteria for removing an item from the linking set. It was important to also make sure that the remaining anchor set continues to be representative of the overall content and structure of the test.

Equating method. To obtain equated Rasch parameter estimates for the spring 2018 assessment, anchor item parameter estimates were fixed to their previously equated item parameter estimates before calibrating the remaining non-anchor operational items. This method placed the non-anchor operational items on the same scale as the anchor items.

Comparability of Online and Paper Forms for CMAS Mathematics and ELA

The scale score distributions for students taking online and paper assessments 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 matching variables included sex, race/ethnicity, free and reduced lunch status, language proficiency, IEP, and district setting, plus the prior year’s overall test score in the same content area. For grade 3, no prior test scores were available, so those samples were matched on the demographic variables only. Because the lack of prior test scores may have affected the quality of the match for grade 3, matched pairs in grade 3 for which either student scored the lowest obtainable scale score (LOSS) were removed prior to analysis.

Scale score distributions of CMAS scores between the matched samples were compared to estimate the mode effect. To quantify the differences between the two distributions, the effect size of the differences between the two distributions was calculated as Cohen’s d (Cohen, 1977) using the mean scale score from each group and the pooled standard deviation:

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

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

- 0.2 = a ‘small’ effect size
- 0.5 = a ‘medium’ effect size
- 0.8 = a ‘large’ effect size

A threshold for a possible mode effect was set to an effect size of .1 or greater and a matched sample size of at least 1,000 students. The number of students taking the paper form can be found in Tables 50 and 51. The effect size was calculated for the mathematics and ELA in each grade, and the results were presented to CDE, which made the final decision on whether to make an adjustment for mode differences for each assessment.

For assessments where an adjustment was deemed necessary, scores from the paper form were adjusted using a linear transformation to match the mean and standard deviation of the online form. The conversion is applied to the overall scores. For ELA, the conversion is also applied to the Reading claim score.

Results of the spring 2018 mode adjustment analysis and the assessments for which an adjustment was applied are provided in Part II of this report.

Field Test Equating

The field test (FT) equating process for CMAS mathematics, ELA, and CSLA is similar to that of operational equating, except that the anchors are the operational items. This process placed the FT item parameter estimates onto the operational base scale. Items from all FT forms are calibrated concurrently.

For CMAS ELA, student responses to field-tested prose constructed-response (PCR) items are sampled for scoring and calibration. For each FT PCR item, a total of 6,000 responses are selected for scoring, with 50% scored on the WE trait, and 50% scored on WKL trait. The resulting item response matrix mirrors the Moulder matrix used for calibration of the operational PCR items.

Ability Estimates

CMAS Mathematics and ELA

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 for the Reading claim on the ELA assessment. Estimates were obtained via the maximum likelihood method (MLE) applied within the ISE software program (Chien & Shin, 2012). Pattern scores use the examinee's individual item response pattern (overall or Reading claim) to determine his or her ability estimate, which may lead to different ability estimates for the same raw score.

CSLA

After the item parameter estimates were obtained for the CSLA operational items, student abilities were estimated for each grade-level assessment by conducting an anchored calibration of the operational items' item parameter estimates. Student abilities were calculated for the overall test and the Reading claim. To obtain student ability estimates for the overall test, all the operational items were included in the anchored calibration. To obtain student ability estimates for the Reading claim, only those operational items representing the specific claim were included in the anchored calibration. The calibrations included the weighting of the PCR WE trait score. Student ability estimates were obtained via the joint maximum likelihood method (JMLE) applied within Winsteps.

Overall and Subscale Scale Scores

For CMAS mathematics, ELA, and CSLA, examinee ability estimates for the overall test were then transformed to scale scores ranging from 650 to 850 using the same scaling transformations as the prior year's administrations. For CMAS ELA and CSLA, the examinee ability estimates for the Reading claim were transformed to scaled scores ranging from 110 to 190.

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

$$SS = A * \theta + B$$

After the scale scores were calculated, the lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS) were applied. LOSS and HOSS were set to 650 and 850, respectively, for the overall test scale. For the Reading scale, LOSS and HOSS were set to 110 and 190.

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.

- The data file containing student responses was verified and exclusion rules were applied.

- A traditional item analysis (TRIAN) and adjudication, where applicable, were completed on all items.
- Incomplete data matrices (IDMs) were created.

Traditional Item Analysis (TRIAN) and Adjudication

A TRIAN of all 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, items having one or more of the following characteristics are flagged:

- P-value = 0
- Item-total score correlation < 0
- Incorrect option selected by more high-performing students (top 20%) than the keyed response
- Distractor-total score correlation > 0
- One or more score points earned by less than 5% of students

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 8.

All TE items 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 9.

Calibration, Equating, and Scaling

For the spring 2018 administration, several different analyses were performed to obtain item parameter estimates for online operational and field test items and ability estimates for examinees.

- CMAS mathematics and ELA
 - Online operational items
 - Used IRTPRO control files and IDM to obtain online operational item parameter estimates

- Evaluated consistency of scoring and stability of anchor items
- Used STUIRT to scale 2018 operational items to operational scale
- Calculated item fit statistics and plotted expected vs. observed IRFs for each operational item
- Used ISE to estimate student abilities
- 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
- CSLA
 - Operational items
 - Used Winsteps control files and IDM to obtain non-anchor operational item parameter estimates
 - Evaluated stability of anchor items to obtain the final anchor set
 - Used the final anchor set in Winsteps to scale the 2018 non-anchor items to the operational scale
 - Obtained item difficulty values, step deviation values, and item fit values for all items
 - Used Winsteps to estimate student abilities using the operational item parameter estimates
 - Embedded field test items
 - Used Winsteps control files and IDM to scale the embedded field test item parameter estimates to the operational scale by fixing the item parameter estimates of the operational items.
 - Obtained field test item difficulty values, step deviation values, and item fit values for the field test items

CHAPTER 9: RELIABILITY

A variety of statistics can be calculated that pertain to the reliability of the CMAS mathematics, ELA, and CSLA assessments. In this report, coefficient alpha (i.e., Cronbach's alpha), standard error of measurement (SEM), conditional standard error of measurement (CSEM), decision consistent and accuracy, and inter-rater agreement are provided, as described below. For these statistical estimates from the spring 2018 administration, see Part II of this document.

Coefficient Alpha

Within the framework of Classical Test Theory, an observed test score is defined as the sum of a student's true score and error ($X = T + E$, where X = the observed score, T = the true score, and E = error). A true score is considered the student's true standing on the measure, while the error score reflects a random error component. Thus, error is the discrepancy between a student's observed and true score.

In the CTT framework, 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.

Coefficient 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.

Coefficient alpha estimates are provided for the overall test, each subscale, and several demographic subgroups (see Tables 3 through 17). Given the differences in length, it is expected that the coefficient alpha for the overall test will be higher than that of the subscales.

Note that coefficient alpha is reported as a measure of internal consistency of the items that each scale comprises. However, because the reported scale scores for both the overall test and the Reading claim (ELA only) are determined using IRT pattern scoring, IRT-based conditional standard error of measurement (CSEM; see the “Conditional Standard Error of Measurement” section later in this chapter) is a more appropriate measure of the measurement error associated with these scale scores.

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 for the subclaims and the Writing claim are included in Table 18.

Conditional Standard Error of Measurement

While the SEM provides an estimate of precision for an assessment, CSEM gives an indication of how measurement error varies across the score scale. Each scale score has a CSEM estimate that indicates what the most likely range of scores would be for students receiving that score if they tested multiple times. The CMAS assessments measure more accurately at a scale score of 750 (near the middle of the scale) than at 650 or 850 (at the ends of the scale). During test construction, CSEMs are reviewed to ensure that they are minimized around the performance level cut scores.

The CSEM is defined as the standard deviation of observed scores given a particular true score and is estimated within the IRT framework as the inverse of the test information function. Plots of test information curves (TICs) and CSEM across the score scale range are provided in Appendix C for both the overall scale scores and Reading claim scale scores (ELA and CSLA only).

Decision Consistency and Accuracy

The CMAS mathematics, ELA, and CSLA scales are divided into five performance levels: (1) Did not yet meet expectations, (2) Partially met expectations, (3) Approached expectations, (4) Met expectations, and (5) Exceeded expectations. Based on a student's scale score, the student is classified into one of the five 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 mathematics, ELA, and CSLA. For the overall test, consistency and accuracy estimates along with PChance and Cohen's Kappa (κ) coefficient (Cohen, 1960) are provided in Table 19 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, Hanson, & Brennan, 2000).

In addition, consistency and accuracy estimates at each cut score are provided in Tables 20 and 21.

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.

- 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 points)
 - c. Non-adjacent, which represents a difference of more than 1 score point between the two raters

For the CMAS ELA and CSLA CR items (i.e., PCR task items), the following additional analyses are also conducted:

- 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)

- Standardized mean differences (MD)

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

- Correlations (CORR)

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

Rater agreement statistics for both operational and field test items are provided in Tables 22 through 49.

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 mathematics and ELA (including CSLA) assessments that are validated but rather the interpretations of the scores. The purpose of the CMAS mathematics and ELA (including CSLA) assessments is to provide information about a student’s level of mastery of the CCSS and the CAS. These standards 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 and CSLA assessment cycle with validity and fairness considerations in mind. This chapter provides information regarding specific sources of validity evidence as well as fairness. It also describes studies conducted to support the transition to using reduced versions of the test blueprints.

Furthermore, validation is a process. As the CMAS mathematics, ELA, and CSLA 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

Evidence based on the content of the assessment is supported by the degree of correspondence between test items and content standards. The degree to which the test measures what it claims to measure is known as construct validity. The CMAS mathematics, ELA, and CSLA assessments adhere to the principles of evidence-centered design, in which the standards to be measured (the CCSS and CAS) are identified, and the performance a student needs to achieve to meet those standards is delineated in the ESs. Test items are reviewed for adherence to universal design principles, which maximize the participation of the widest possible range of students.

The item development process for the assessments is driven by targets at the ES level. Before developing items, Pearson uses target spreadsheets to create an internal item development plan (IDP) aligned with the expectations of test design and with consideration of attrition rates at committee review and data review.

In addition to the ESs, content is aligned through the articulation of performance in the PLDs. At the policy level, the PLDs include policy claims about the educational achievement of students who attain a particular performance level, and a broad description of the grade-level knowledge,

skills, and practices that students performing at a particular achievement level are able to demonstrate. Those policy-level descriptors are the foundation for the subject- and grade-specific PLDs, which, along with the ES framework, guide the development of the items and tasks.

Gathering construct validity evidence for CMAS mathematics and ELA (including CSLA) is embedded in the process by which CMAS assessment content is developed and validated. At each step in the assessment development process, hundreds of educators, assessment experts, and bias and sensitivity experts were involved in review of text, items, and tasks for accuracy, appropriateness, and freedom from bias. See Chapter 2 for an overview of the content development process. In the early stages of development, Pearson conducted research studies to validate the item and task development approach. One such study was a student task interaction study designed to collect data on students' experience with the assessment tasks and technological functionalities, as well as the amount of time needed for answering each task. Pearson also conducted a rubric choice study that compared the functioning of two rubrics developed to score the prose constructed-response (PCR) tasks in ELA. Quantitative and qualitative evidence was collected to support the use of a condensed or expanded trait scoring rubric in scoring student responses.

CMAS items and tasks were field tested prior to their use on an assessment. During the initial field test administration in 2014, Colorado, along with other consortium states, collected feedback from students, test administrators, test coordinators, and classroom teachers on their experience with the assessments, including the quality of test items and student experience. A summary of the feedback can be found at <https://parcc-assessment.org/content/uploads/2017/11/PARCCCRDPolicyandPLDsFINAL.pdf>. The feedback from that survey was used to inform test directions, test timing, and the function of online task interactions. Performance data from the field test also informed the future development of additional items and tasks.

During item development, all item developers and item writers were provided an electronic version of the consortium's accessibility guidelines and the linguistic complexity rubric. Items and passages are reviewed internally by accessibility and fairness experts trained in the principles of universal design and who became well versed in the consortium's accessibility guidelines. Items received internal review for alignment to the evidence tables, task generation model, item selection guidelines, and accessibility and fairness reviews.

An important consideration when constructing test forms is recognition of items that may introduce construct-irrelevant variance. Such items should not be included on test forms to help ensure fairness to all subgroups of test takers. Committees reviewed all items for bias and sensitivity. Additionally, content experts facilitated reviews of all items. All reviewers were trained using bias and sensitivity guidelines, and the guidelines were used to review items and passages. Accommodations were made available based on individual need documented in the student's approved IEP or 504 Plan. Accessibility specialists worked to review forms and determine which forms should be used for students with accommodations.

The mathematics and ELA operational test forms were carefully constructed to align with the test blueprints and specifications that are based on the CCSS and CAS. During the fall of 2016, content experts representing various consortium states, along with other content experts, held a series of meetings to review the operational forms for ELA and mathematics. These meetings

provided opportunity to evaluate test forms in their entirety and recommend changes. Requested item replacements were accommodated to the extent possible while striving to maintain the integrity of the various linking designs required for the operational test analyses.

Psychometricians were available throughout this process to provide guidance with regard to implications of item replacements for the linking and statistical requirements.

Further information regarding the college- and career-ready content standards, performance level descriptors, and accessibility features and accommodations incorporated into the CMAS mathematics and ELA assessments is provided at <http://parcc-assessment.org/college-career-ready>.

Evidence Based on Internal Structure

Analyses of the internal structure of a test typically involve studies of the relationships among test items and/or test components (i.e., subclaims) in the interest of establishing the degree to which the items or components appear to reflect the construct on which a test score interpretation is based (AERA, APA, & NCME, 2014, p. 16). The term ‘construct’ is used here to refer to the characteristics that a test is intended to measure; in the case of the CMAS mathematics, ELA, and CSLA operational tests, the characteristics of interest are the knowledge and skills defined by the test blueprints.

The CMAS mathematics, ELA, and CSLA assessments provide a full summative test score and a Reading claim score, as well as percent-correct scores for the Writing claim and ELA and mathematics subclaims. The goal of reporting at this level is to provide criterion-referenced data to assess the strengths and weaknesses of a student’s achievement in specific components of each content area compared with other students taking the same assessment (for claim and subclaim scores) as well as students who took the assessment in prior years (for overall scores). This information can then be used by teachers to plan for further instruction, to plan for curriculum development, and to report progress to parents. The results can also be used as one factor in making administrative decisions about program effectiveness, teacher effectiveness, class grouping, and needs assessment.

Intercorrelations

The ELA and CSLA summative tests comprise two claim scores, Reading and Writing, and five subclaim scores—Reading Literature (RL), Reading Information (RI), Reading Vocabulary (RV), Writing Written Expression (WE), and Writing Knowledge Language and Conventions (WKL). The Reading claim score is a composite of RL, RI, and RV. The Writing claim score is reported only as a percentage of points earned. It is a composite of WE and WKL, and comprises only PCR items. The operational test analyses were performed by evaluating the separate trait scores of WE and WKL. Some PCR items also include RL or RI points; however, the reading points for those items were a duplicate of the WE score and were not included in calibrations.

The mathematics full summative tests have four subclaim scores—Major Content (MC), Mathematical Reasoning (MR), Modeling Practice (MP), and Additional and Supporting Content (ASC).

One way to assess the internal structure of a test is through the evaluation of correlations among subscores. For CMAS ELA and CSLA, these analyses were conducted between the Reading and Writing claim scores and the subclaims (RL, RI, RV, WE, and WKL). For CMAS mathematics, the analyses were conducted between the mathematics subclaim scores. If these components within a content area are strongly related to each other, this is evidence of unidimensionality.

Reliability

The reliability analyses described in Chapter 8 of this report provide information about the internal consistency of the CMAS mathematics, ELA, and CSLA assessments. Internal consistency is typically measured via correlations amongst the items on an assessment and provides an indication of how much the items measure the same general construct. High reliability of test scores implies that the test items within a domain are measuring a single construct, which is a necessary condition for validity when the intention is to measure a single construct. Table 3 provides reliability estimates for the overall population for the full tests, the ELA and CSLA claims and subclaims, and the mathematics subclaims. Tables 4–17 also provide reliability estimates for subgroups of interest. The reliability estimates were computed using coefficient alpha, and were also used in the calculation of CTT-based estimates of SEM (described in Chapter 8).

Factor Analysis

A factor analysis is performed on the item response data for the CMAS assessments to analyze the number of dimensions the assessments appear to be measuring. Given that unidimensional IRT models are used for calibration and scaling, it is important that there be evidence to support their use. Scree plots for the spring 2018 administrations can be found in Figures 24 through 37. For most of the assessments, one factor explained the vast majority of the variance, which supports the use of a unidimensional IRT model. The scree plots indicate a strong underlying single factor for each assessment, although many of the tests suggest the presence of a less influential second factor. The second factor is more apparent for the ELA assessments, and the loadings for factor 2 for ELA were all much higher for the PCR trait items than any other items. This may indicate the influence of a writing construct that is separate from what is measured by the reading items.

Evidence Based on Relationships to Other Variables

To establish concurrent validity, the relationship between CMAS Mathematics and ELA scores and other available assessment scores were analyzed. One such score is from the CMAS Science assessment, which is administered to students in grades 5 and 8.

For grades 5 and 8, correlations were calculated between scores on the CMAS Science assessment and scores on the ELA and Math assessments. The results are below.

Grade	Science with Math	Science with ELA
5	.79	.82
8	.82	.83

The correlations between CMAS Science and the ELA and mathematics scores are fairly high and are similar for mathematics and ELA. We would expect mathematics to have a higher correlation with science. However, the correlations between ELA and mathematics scale scores are also fairly high for students who had valid scores on both assessments. This may be due to the number of constructed response items included on the CMAS Science assessment. Table 131 provides the correlations between ELA and mathematics.

Evidence Based on Response Processes

As noted in the AERA, APA, and NCME *Standards* (2014), additional support for a particular score interpretation or use can be provided by theoretical and empirical evidence indicating that test takers are using the intended response processes when responding to the items in a test. This type of evidence may be gathered from interacting with test takers in order to understand what processes underlie their item responses. Evidence may also be derived from feedback provided by test proctors/teachers involved in the administration of the test and raters involved in the scoring of constructed-response items. Evidence may also be gathered by evaluating the correct and incorrect responses to short constructed-response items (e.g., items requiring a few words to respond) or by evaluating the response patterns to multi-part items.

The PARCC consortium has undertaken research investigating the quality of the items, tasks, and stimuli, focusing on whether students interact with items/tasks as intended, whether they were given enough time to complete the assessments, and the degree to which scoring rubrics allow accurate and reliable scoring. In addition, PARCC has examined the accessibility of the test for students with disabilities and English learners. This research has included examining students' understanding of the format of the assessments and the use of technology. Although out of the purview of this technical report, several other PARCC research efforts have investigated questions relevant to response processes evidence. Various PARCC research studies are described at <https://parcc-assessment.org/research/>.

Evidence Based on the Consequences of Testing

Because state tests are administered “with the expectation that some benefit will be realized from the intended use of the scores” (AERA, APA, & NCME, 2014), validity evidence supporting the use and interpretation of CMAS mathematics and ELA (including CSLA) assessment results may be investigated as a consequence of testing.

One intended consequence of testing is that more students will demonstrate mastery over the Colorado Academic Standards over time, as evidenced by more students achieving in the top performance levels, if the data are used appropriately to make improvements in programming at the school and district levels. The CMAS mathematics and ELA assessments have been administered to Colorado students since the spring of 2015. The table below shows that with the exception of grade 6 math, student performance has improved since the first administration of the assessments. *Note: There have been changes in the available assessments by grade for 7th and 8th grade math across assessment administration so comparisons across years for those grades are not included.*

Subject	Grade	2015 % Met or Exceeded	2018 % Met or Exceeded	% Change 2015 to 2018
ELA	3	38.2	40.4	2.2
	4	41.7	46.1	4.4
	5	40.5	47.4	6.9
	6	39.1	42.8	3.7
	7	41.0	46.6	5.6
	8	40.9	43.8	2.9
Math	3	36.7	39.1	2.4
	4	30.2	33.9	3.7
	5	30.1	35.5	5.4
	6	31.7	30.4	-1.3

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.

Sample Items

Sample items provide the opportunity for teachers and students to become familiar with the test design and scoring of the assessments before experiencing the items on an operational test. Prior

to the operational administration, teachers and students are provided the opportunity to experience sample items for CMAS mathematics and ELA (including CSLA).

CMAS mathematics and ELA online sample items are included in practice environments called ePATs. As the assessment system progresses, ePATs are updated to reflect current accessibility features and any updates to Pearson TestNav that may impact student interactions with the system. Accommodated versions of the ePATs are also available so that students can practice using accommodations such as text-to-speech, color contrast, and Spanish text-to-speech. Paper sample items for students taking paper versions of the assessments (including CSLA) are available in PDF format for download and are accompanied by scoring and alignment documents.

Universal Design

The CMAS mathematics and ELA (including CSLA) 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 CMAS mathematics and ELA (including CSLA) items were field tested and then analyzed for DIF in order to identify any items that appeared to be unfairly favoring one subgroup over another. All DIF-flagged items were then reviewed by educator committees to identify potential construct-irrelevant explanations for the flags.

Accessibility Tools and Accommodations

As described in Chapters 3 and 4, various accessibility tools and accommodations are available for students who take CMAS mathematics, ELA, and CSLA. 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, braille forms, and oral scripts; as well as online forms designed to work with assistive technology such as screen readers. Students may also have extended time as required by their IEP or for students classified as English learners. The test is also available with Spanish test-to-speech and paper transadaptions or oral scripts that can be translated into other languages. Linguistic accommodations (i.e. transadaptation, oral presentation in any language, and scribe for constructed response items) require approval on the ELA assessment in order to preserve the intended construct of reading and writing in English according to the Colorado Academic Standards (see Chapter 4). The CSLA assessments were developed to be linguistically accommodated Spanish tests in accordance with Colorado state law. The purpose of these various options is to allow students to fully demonstrate their content knowledge without being hindered by non-construct related elements.

Special Studies to Support the Abbreviated Test Design

As described in Chapter 1, the spring 2018 CMAS mathematics, ELA, and CSLA assessments were the first CMAS assessments constructed using proportionally reduced versions of the full test blueprints used in previous assessments in order to decrease the total amount of testing time for Colorado students. The final reduced blueprints were approved in 2017, with the assessments based on the reduced blueprints developed during the fall of 2017.

The decision to implement assessments based on the reduced blueprints was informed by a number of research studies undertaken by CDE and Pearson, as well as expert input and advice from the CMAS Technical Advisory Committee. The goal of the studies was to examine whether the CMAS mathematics and ELA (including CSLA) assessments based on the reduced blueprints would continue to measure the same constructs, and whether the measurement properties of the assessments remained as close as possible to those of the full assessments. The overall conclusion of these studies was that the proposed changes had an inconsequential impact on overall student performance in terms of scale scores and performance level assignment. This section describes the design of some of these studies, with results summarized in Part II of this report. The full reports are included in Appendices D1 through D4.

Blueprint Adjustment Comparison

The purpose of this research was to evaluate the similarity between student scores from the 2017 assessments and those from an adjusted version of the assessment with a subset of items omitted and treated as missing.

The data for this research came from the grades 3 through 8 spring 2017 operational ELA and mathematics assessments and only included students from Colorado. For the purpose of these studies, the three CBT operational forms were used. The spring 2017 CSLA assessment data were used for the CSLA evaluation.

Assessment Adjustments

For each assessment, a subset of items were selected for exclusion based on timing targets and the number of resulting points across the reporting categories. The removal of items was intended to resemble the planned revision of the full-length test blueprint for the CMAS assessments, except that items were not able to be added for grade levels in which the adjustment specified additional items (i.e., because these data were from assessments that had already been administered). For CSLA, the subset of items selected was also based on the planned revision of the CMAS ELA blueprint.

In mathematics, the adjusted test forms were all reduced by 15 points (from 66 to 51); in ELA and CSLA, the adjusted test forms were reduced by between 29 and 43 points. A full description of the reduction in points by subclaim for each grade and subject is included in the reports

PARCC Mathematics Adjustment for Colorado, PARCC ELA/L Adjustment for Colorado, and CSLA Blueprint Adjustment, which are included in Appendices D1 through D3.

Calibration, Scaling, and Equating

To create the adjusted test forms, student responses on the items to be removed from each assessment were converted to missing or removed from the dataset entirely. The resulting dataset was then calibrated and item parameter estimates were equated following the operational calibration and equating procedures. Raw-score to scale-score conversion tables were also constructed and student scale scores and performance levels were assigned, again following the procedures used for operational scaling.

Results

The resulting datasets were compared with the operational 2017 data in order to determine how similar scores would be for Colorado students if the full-length 2017 Math, ELA and CSLA assessments had been reduced. The results were described in terms of scale score summary statistics, overall performance level agreement, subclaim performance level agreement, and overall test characteristic curves. A summary of the results is provided in Part II of this report. The full results are provided in the *PARCC Mathematics Adjustment for Colorado, PARCC ELA/L Adjustment for Colorado*, and *CSLA Blueprint Adjustment* research reports, included in Appendices D1 through D3.

Sample Invariance Study

In order to maintain the same scale for CMAS administrations that was used for reporting prior to 2018 administrations (including Colorado students), item parameter estimates should be stable between the full consortium sample and the Colorado-only subset. The purpose of this study was to examine the stability of item parameter estimates between Colorado and the full consortium by comparing the spring 2017 operational equating results with those using the subset of students from Colorado only.

The data for these studies came from the operational 2017 PARCC ELA/L assessment results for grades 3 through 8. Data from only the three CBT operational test forms were included.

Calibration, Scaling, and Equating

For each grade and subject, item responses from the Colorado-only subset were calibrated following operational consortium procedures. The parameter estimates were then equated to the established base scale following operational procedures. Because the different samples resulted

in different anchor stability results—which in turn could potentially cause some anchor items to differ between the full consortium and Colorado-only sample—the study was run using two different methods:

1. following consortium procedures as written, which allowed for a final Colorado-only anchor set that differed slightly from the final operational consortium anchor set, and
2. using the final operational 2017 anchor set, whether or not those items were stable between the base and target samples.

Using two methods for the ELA analysis allowed for the evaluation of the impact of anchor set choices. Because results between the two methods were very similar—and because using unstable anchor items for operational equating would be undesirable—the mathematics study used only Method 1.

Results

A summary of the results of this study is included in Part II of this report. The results were described through comparisons of equated item parameter estimates, comparisons between raw-score to scale-score conversion tables constructed using estimates from each sample, correlations between student scale scores from each table, and differences between student scale scores conditional on 2017 scale score.

A full description of the study methods and results is available in the *Colorado vs. PARCC ELA Sample Invariance Study* research reports, included in Appendix D4.

Pattern Scoring Analysis

Prior to 2018, mathematics and ELA assessments were scored based on number-correct scores (i.e., using IRT-based raw-score to scale-score conversion tables). An alternative scoring method is IRT pattern scoring (already used on the CMAS science and social studies assessments), wherein students' ability scores are estimated based on their entire pattern of item responses, not solely on the total number of points they earn. With pattern scoring, two students may both earn 56 points, for example, but receive different scale scores depending on the properties of the particular items they answered correctly.

After consultation with the CMAS Technical Advisory Committee, the decision was made to implement pattern scoring for the CMAS mathematics and ELA assessments.

One benefit of pattern scoring is an increase in precision. The reduction in blueprints for the spring 2018 CMAS mathematics and ELA assessments meant that fewer score points were used to estimate student abilities. A transition from number-correct scoring to pattern scoring compensated for the loss of precision that would be associated with reducing the number of

points. In addition, the transition to pattern scoring for mathematics and ELA created consistency across subject areas, since Colorado's science and social studies assessments already used pattern scoring.

A study was conducted to investigate the extent to which student ability estimates would change between number-correct and pattern scoring. The results indicated a high degree of consistency between the two methods.

One consideration with a transition to pattern scoring was how to address the weighting of the ELA assessment items. With pattern scoring, items are not assigned predetermined point values, and are instead weighted by how well they measure students' ability. Therefore, the weights assigned to certain items in the consortium blueprint (e.g., written expressions weighted by 3) would no longer apply when estimating pattern scores. With number-correct scoring, the written expression items intentionally accounted for a higher proportion of a student's total score than other items. With pattern scoring, however, items that carry the highest weight are those that measure students' ability most effectively. This can potentially increase measurement precision and the reliability of scores; however, each item's contribution to a student's overall score is different than it would be with number-correct scoring. For the purpose of estimating the ELA summative score, the Written Expressions trait is unweighted. However, the Writing claim and Written Expressions subclaim percent earned scores still reflect the weighted point values.

PART II: STATISTICAL SUMMARIES FOR 2018

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

- Spring 2018 Operational Exam
- Spring 2018 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 are provided.

CHAPTER 1: SPRING 2018 OPERATIONAL ASSESSMENT

The following section provides details on the spring 2018 administrations of the CMAS mathematics, ELA, and CSLA assessments.

Administration Summary

Tables 50 and 51 provide a breakdown of online test takers compared with those who took accommodated forms for mathematics and ELA. Although a paper form was available to all students, the vast majority took the assessments online.

Tables 52 through 54 provide n-counts for various demographic characteristics for the students who took the CMAS mathematics, ELA, and CSLA assessments.

Equating Results

Calibration and Anchor Set Evaluation

CMAS Mathematics and ELA

The initial calibration results were reviewed for problematic item parameter estimates, and fit plots were examined to detect items with poor model–data fit. Based on the initial calibration, one item on the grade 7 ELA assessment was removed from the final calibration and suppressed from scoring.

Review of anchor item stability analyses resulted in dropping zero to two items from the anchor set, depending on grade and subject. For certain assessments, additional items were designated as anchors to counterbalance the items that were dropped (i.e., to maintain the content balance of the anchor sets as compared with the full test form). The final anchor sets for mathematics represented between 24% and 33% of the total test points, and the final anchor sets for ELA represented between 36% and 47% of the unweighted total test points.

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

CSLA

The CSLA grade 3 assessment was pre-equated, so a post-administration calibration was not conducted after the operational administration. For the CSLA grade 4 post-equating, the

calibration results after the operational administration were reviewed and evaluated as part of the stability check. The final anchor set represented 47% of the unweighted total test points.

Item Statistics

Tables 55 through 68 provide the item parameter estimates for each grade. (Note that the Item numbers are merely identifiers, and do not reflect the sequence of items as they were presented to students.) The “Item Type” uses the coding of SR for selected-response, XI for technology-enhanced, and CR for constructed-response items. The “Model” refers to the IRT model under which the item was estimated (2PL, GPC, or RPCM). The “A” column shows the item parameter estimate for discrimination, “B” for difficulty, and “D1” through “D7” for GPC or RPCM category threshold estimates. Not all item parameters apply to each item. For example, there are no category threshold estimates for 2PL items.

The last column of the mathematics and ELA tables reflects whether an item was flagged for misfit based on Q1. Several items in each grade were flagged for misfit. Misfit plots for all items were reviewed, and misfit statistics were compared with data from the 2017 administration. Based on these reviews, no additional items were removed due to misfit flags. For CSLA, the “Infit” and “Outfit” columns contain the item fit values.

See Chapter 8 for detailed information about the calibration process.

IRT Curves

The test characteristic curves (TCCs), test information curves (TICs) and CSEM curves for both the overall scale scores and the Reading claim scale scores (ELA and CSLA only) are provided in Appendix C. The 2018 TCCs matched those from 2017 in terms of shape and position. The 2018 TCCs were reviewed across the distribution as well as at the cuts to ensure the match between years. Additionally, Colorado’s established maximum TCC difference of 0.05 was maintained between the 2017 full length and 2018 abbreviated forms. It should be noted that the TCCs are provided in terms of expected percent correct rather than expected raw score. Along with the curves, each of the four cut scores for a given grade is indicated with a red vertical line, as are the cut scores for the Reading claim. On the overall scale score TCCs, the vertical line at a scale score of 750 corresponds to the cut for Met Expectations for each assessment.

Performance Results

The cumulative scale score distributions for each grade are shown in Tables 69 through 94. Figures 10 through 23 display the same information in graphical form.

Table 95 provides summary statistics for overall scale scores. Means, standard deviations, and medians are provided. The 2017 scale score means and standard deviations are also included for comparison.

The performance level distributions for each grade are shown in Table 96. With the exception of grade 8 mathematics, the distributions within each performance level are similar across grades for both CMAS mathematics and ELA. Table 96 also lists the same distribution for the 2017 administration for comparison.

Summary statistics for points earned by subclaim are shown in Tables 97 through 99. Note that the assessments are constructed and scaled to produce overall scores and Reading claim scores that are comparable with overall scores and Reading claim scores from previous CMAS administrations (see Chapter 8 in Part 1 of this report). However, the assessments are not designed to permit meaningful comparisons across percent earned scores, such as subclaim scores, either within an assessment or across administration years, nor are they designed to permit comparisons of the Writing claim scores across administration years. The difficulty of the items that make up each subscore can vary across subscores and from year to year, making it inappropriate to make inferences based on percent-correct performance across subscores or based on subscore performance across years. The only percent earned subscore comparisons supported by the CMAS assessments are those comparing individual or group performance within one sub-category with the performance of other students or groups within the same sub-category and administration.

Tables 100 through 126 provide classical statistics at the item level for the CMAS mathematics, ELA, and CSLA assessments. For SR items, the omit rate, p-value (the item mean and, for 1-point items, the percentage of students correctly responding to an item), and item-total correlation are 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 subclaims for each assessment. For CMAS ELA and CSLA, correlations between the Reading and Writing claim were also calculated. These results are provided in Tables 127 through 130.

Reliability Statistics

Coefficient Alpha

Coefficient alpha was calculated for both the content subclaims and the overall assessment, as shown in Table 3. The internal consistency values for the full assessment ranged from 0.85 to 0.92.

Tables 4 through 17 display performance by various subgroups. The scale score means, standard deviations, minimums, and maximums, as well as overall coefficient alpha values, are provided.

SEM

Table 18 shows the SEMs that were calculated for each grade for each subclaim and for the Writing claim score (ELA and CSLA only), based on the reliabilities reported in Table 3. The classical SEM estimate is not reported for the overall test scale scores and the Reading claim scores, as those scores are based on IRT pattern scoring rather than the sum of item scores.

CSEM

IRT-based test information curves (TICs) and conditional standard error of measurement (CSEM) curves for the overall scale scores and the Reading claim scores are included in Appendix C.

Decision Consistency and Accuracy

Tables 19 through 21 provide statistics related to decision consistency and accuracy. Table 19 shows accuracy and consistency estimates in addition to probabilities due to chance (PChance) and kappa for the entire assessment. Kappa describes the agreement between classifications on two parallel forms. The kappa value can be interpreted as follows (Altman, 1991):

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

Tables 20 and 21 provide the accuracy and consistency estimates at each of the cut scores.

Inter-Rater Agreement

For each operational item, five percent of the responses were scored by a second reader, which allowed for rater agreement statistics to be calculated. Tables 22 through 35 provide the percentage of operational items with exact agreement, adjacent agreement, and non-adjacent agreement. Tables 36 through 49 provide rater agreement information for field test items. The target exact plus adjacent agreement rate is 95% for all items.

Validity Statistics

Intercorrelations

As described in Chapter 10, Tables 127 through 130 list correlations between the subclaims within each assessment, as well as between the Reading and Writing claims for the CMAS ELA and CSLA assessments. The intercorrelations for the mathematics subclaims were higher overall than the ELA and CSLA intercorrelations, although most values for all assessments were between 0.4 and 0.7. For CMAS ELA and CSLA, the two writing subclaims tended to have higher correlations with one another than they did with any of the reading subclaims.

Reliability

Reliability statistics for the overall test, the subclaim scores, and several demographic subgroups are presented in Tables 3–17. The full test reliabilities range from 0.85 to 0.92. The overall test reliability does not correspond directly with the overall student scale scores, as those are based on IRT pattern scoring. However, the overall estimates do provide evidence of unidimensionality of the assessments. Furthermore, the subgroup reliabilities were fairly consistent for the various demographic subgroups, with the exception of those based on language proficiency. The reliability of the test was lower for students identified as non-English proficient or limited English proficient, with the decrease in reliability being larger for the ELA and CSLA assessments than the mathematics assessments.

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 24 through 37. The results support the use of a unidimensional IRT model, although the ELA and CSLA scree plots do suggest that the Reading and Writing claims are distinct subscores.

Special Studies to Support the Transition

The decision to transition to reduced versions of the PARCC blueprint for the spring 2018 CMAS mathematics, ELA, and CSLA assessments was supported by the results of several studies conducted to examine potential differences based on different test designs and scoring methodologies. An executive summary of each report is included in the sections that follow. The full studies are described in the following reports, which are included in Appendices D1 through D4.

- *PARCC Mathematics Adjustment for Colorado*
- *PARCC ELA/L Adjustment for Colorado*

- *CSLA Blueprint Adjustment*
- *Colorado vs. PARCC ELA Sample Invariance Study*

Blueprint Adjustment Comparison

The purpose of this research was to evaluate the similarity between spring 2017 CMAS mathematics, ELA, and CSLA operational student scores and those calculated after adjusting the length of assessments.

A subset of operational items on the spring 2017 CMAS mathematics, ELA, and CSLA assessments were treated as omits or not administered items. Scoring tables or conversion tables were generated based on the adjusted number of operational items. Adjusted raw scores were computed based on student response strings and omitting the selected items. The adjusted scoring tables were applied to the student adjusted raw scores. Analyses compared the students' spring 2017 scale scores and performance levels based on the full assessments to the adjusted scale scores and performance levels based on the adjusted assessments.

The results included scale score summary statistics, overall performance level agreement, subclaim performance level agreement, correlations, and overall test characteristic curves.

The average scale scores were similar for the adjusted and full test forms. The percent of exact agreement in the overall performance level designation between the full assessment and the adjusted assessment ranged from 83.1% to 90.2% exact agreement across the grade levels for mathematics, from 80.9% to 85.1% for ELA, and 80.4% to 84.7% for CSLA. In addition, the correlations between full and adjusted scale scores were all greater than .98 in mathematics, .97 in ELA, and .95 in CSLA. The reading claim scale score correlations were all greater than .96 in ELA. The writing claim scale score correlation were lower, ranging from .91 to .97, depending on grade.

The subclaim performance levels were also compared across the overall and adjusted blueprints. The exact agreement varied quite widely by subclaim, with a range of 65.2% to 97.0% in mathematics and 60.3% to 95.9% in ELA. Due to the instability of the performance level categories, it was decided to report percent earned scores for the subclaims going forward.

Sample Invariance Study

The purpose of this research report was to examine the stability of parameter estimates between Colorado and the full PARCC consortium by comparing the spring 2017 operational equating results with those using only students from Colorado in 2017.

The consortium's operational scaling procedures from 2017 were replicated for the three online forms at each grade level. The base group in each case was the 2016 full consortium sample. The target group for this study was the subset of 2017 students from Colorado, while the target group in the operational equating was the full 2017 consortium sample (i.e., including Colorado students). New raw-to-scale score conversion tables were constructed based on the Colorado-

only equating, and the 2017 Colorado sample was rescored using these new conversion tables. The operational scoring results were compared with the Colorado-only results.

Changing the student samples from the full consortium to only Colorado students resulted in different anchor stability results. Because the PARCC procedures call for dropping unstable anchor items, this meant that the anchor set for the Colorado-only sample could differ from the operational PARCC equating. To determine the impact of such differences, the study was conducted using two different anchor set construction methods:

1. following the consortium procedures as written, which allowed for a final Colorado-only anchor set that differed slightly from the final operational consortium anchor set, and
2. using the final operational consortium anchor set, whether or not those items were stable between the base and target samples.

Using two methods for the ELA analysis allowed for the evaluation of the impact of anchor set choices. Because results between the two methods were very similar—and because using unstable anchor items for operational equating would be undesirable—this section focuses on the Method 1 results.

Most of the scores calculated using only Colorado students were within one or two scale score points of the operational scores. However, larger differences were observed for certain subjects and grades, particularly for very high-achieving students. The largest difference between scale scores across all grades and equating methods was 8 scale points, which was observed in grades 3, 6, and 8, at scale score values near 830 to 840.

Overall, the results of this analysis suggest that the Colorado sample is very similar to the full PARCC consortium sample in terms of item performance. The resulting scale score differences tended to be within two scale score points, although it is possible that any difference could be enough to change performance level classifications if the student's score is near the cut score. However, the total percentage of students whose performance level changed in this study was small. The largest difference for anchor set method 1 was 4.4%, which was observed in grade 5. All of the changes occurred for students whose operational score was within one point of the raw cut score, and the direction of changes was not consistent across grades.

CHAPTER 2: SPRING 2018 EMBEDDED FIELD TEST

This section provides details on the field test items that were embedded within the spring 2018 administration of the CMAS mathematics, ELA, and CSLA assessments. Due to low n-counts on other forms, only items embedded in the online English forms were analyzed for CMAS mathematics and ELA.

Field Test Forms

For CMAS mathematics, between 11 and 14 field test forms were administered, depending on grade level. For CMAS ELA, 8 online field test forms were administered for each grade. For CSLA, four field test forms were administered for grade 3 and two field test forms were administered for grade 4. For CMAS mathematics, ELA, and CSLA, each field test form within a grade was parallel; that is, each student received field test items worth the same number of points per item type, and the item locations were the same or matched as closely as possible across forms. Table 132 shows the number of field test forms and field test items per assessment and grade.

Inter-Rater Agreement

For each CR item, 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 36–49, where the percentage of items with exact agreement, adjacent agreement, and non-adjacent agreement are provided. In addition, the final columns for ELA and CSLA items show the kappa, mean difference, and correlation for each item. Items with poor rater agreement during field testing are reviewed by Pearson and CDE content experts for issues with the item or scoring rubric. These items may be field tested again after adjustments to the item or scoring rubric or removed from the bank.

Data Review

The data review meetings for the spring 2018 embedded field test items were conducted by web-based teleconference in late July and early August. For CMAS mathematics and ELA, the meetings were conducted by grade level and content area, with each committee reviewing items from one content area and two grade levels (e.g., ELA grades 3 and 4, ELA grades 5 and 6, and ELA grades 7 and 8). For CSLA, one committee was convened to review both grades 3 and 4. The committees for CMAS mathematics and ELA comprised educators with expertise in both the content area being assessed and Colorado students in the particular grade levels being reviewed. The CSLA committee comprised experts in bilingual instruction and English learners

at the elementary grades. Prior to the meetings, 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.

The data review meetings were facilitated by Pearson and Tri-Lin assessment specialists. Also attending each meeting were a Pearson psychometrician and representatives from CDE, including a CDE content specialist. Committee members received training on the purpose of the data review meeting and how to interpret the statistical information that would be provided. The committee members were notified that their decision on each item would serve as a recommendation to CDE, which was responsible for making the final accept/reject decision.

Table 133 summarizes the outcomes of the data review. It should be noted that although committee members were given only the choice between recommending to accept or reject an item, in several cases the committee suggested revisions to the item. These suggestions were recorded for potential re-field testing of the item.

REFERENCES

- Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report (NCES 1999–452)*. Washington, DC: National Center for Education Statistics, US Department of Education.
- Altman, D. G. (1991). *Practical Statistics for Medical Research*. London, UK: Chapman and Hall/CRC Press.
- 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.
- 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.
- Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, 23(4), 31-31.
- 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.
- Colorado Department of Education. (2016). *Colorado Spanish Language Arts technical report 2016*. Retrieved from <https://www.cde.state.co.us/assessment/2015-2016cslatechnicalreport>
- Colorado Department of Education. (2018). *Interpretive guide to assessment reports: A guide for parents and educators*. Retrieved from https://www.cde.state.co.us/assessment/cmas_coalt_interpretive_guide_2018
- 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.
- Davis, L. L. & Moyer, E. L. (2015). *Performance Level Setting Technical Report*. Retrieved from https://parcc-assessment.org/wp-content/uploads/2017/12/PARCC_PLS_TechReport_011316_toPARCC-final.pdf

- 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.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425–461.
- 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.
- Linacre, J. M. (2011). *Winsteps® Rasch measurement computer program [computer program]*. Beaverton, Oregon: Winsteps.com.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
- McClarty, K. L., Korbin, J. L., Moyer, E., Griffin, S., Huth, K., Carey, S., & Medberry, S. (2015). *PARCC benchmarking study*. Pearson Educational Measurement, Iowa City, IA: Pearson.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Pearson. (2018). *Final technical report for 2017 administration*. Retrieved from <https://parcc-assessment.org/wp-content/uploads/2018/03/PARCC-2017-Technical-Report-Final-03162018.pdf>
- Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 19, 49–57.
- Scientific Software International, Inc. (2011). *IRTPRO [computer program]*. Lincolnwood, IL.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201–210.

- 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.
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1), 2-13.
- Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.
- Wright, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5(2), 245–262.