Colorado Spanish Language Arts

Technical Report

2016

Table of Contents

Table of Contents ..... **i**

PART I: HISTORICAL OVERVIEW AND SUMMARY OF PROCESSES **3**

CHAPTER 1: INTRODUCTION AND BACKGROUND 3
Purpose of the Document *3*
Overview of CSLA *4*
Assessment Development Partners *5*
CHAPTER 2: ITEM DEVELOPMENT AND ITEM BANKING 7
Item-Writing Process *7*
Item Reviews *9*
Item Banking *11*
CHAPTER 3: TEST CONSTRUCTION 12
CHAPTER 4: TEST ADMINISTRATION PROCEDURES 14
Manuals *14*
Training *14*
Accessibility and Accommodations *14*
Test Security *15*
CHAPTER 5: SCORING THE ASSESSMENTS 16
CHAPTER 6: STANDARD SETTING 19
CHAPTER 7: REPORTING 21
Description of Scores *21*
Score Reports *22*
CHAPTER 8: CALIBRATION, EQUATING, AND SCALING 23
Calibration *23*
Equating and Scaling *23*
Steps in the Calibration and Scaling Process *25*
CHAPTER 9: RELIABILITY 27
Cronbach's Alpha *27*
Standard Error of Measurement *28*
Conditional Standard Error of Measurement *28*
Decision Consistency and Accuracy *28*
Inter-Rater Agreement *29*
CHAPTER 10: VALIDITY 30
Sources of Validity Evidence *30*
Fairness *31*

PART II: STATISTICAL SUMMARIES **33**

CHAPTER 1: OPERATIONAL ITEMS 34
Administration Summary *34*
Calibration Results *34*
Performance Results *34*
CHAPTER 2: EMBEDDED FIELD TEST ITEMS 35
Field Test Items *35*
Data Review *35*

REFERENCES **36**

CSLA Tables 1–28 ..... **38**

Appendices **71**

APPENDIX A: CSLA ELIGIBILITY FLOWCHART 72
APPENDIX B: CSLA TEST BLUEPRINTS 74
APPENDIX C: CSLA CONSTRUCTED RESPONSE RUBRICS 77

APPENDIX D: CSLA STANDARD SETTING REPORT 82
APPENDIX E: CSLA SAMPLE SCORE REPORTS 123

PART I: HISTORICAL OVERVIEW AND SUMMARY OF PROCESSES

CHAPTER 1: INTRODUCTION AND BACKGROUND

All public school students enrolled in Colorado are required by state law to take a standards-based assessment each year in specified content areas and grade levels. Every student, regardless of language background or academic ability, must be provided with the opportunity to demonstrate their content knowledge of the Colorado Academic Standards (CAS). The CAS were fully implemented in the 2013–2014 school year and outline the concepts and skills that students need in order to be successful in the current grade as well as to make academic progress from year to year. To measure students' mastery of more rigorous standards, Colorado has implemented a set of common assessments known as the Colorado Measures of Academic Success, or CMAS.

CMAS are the state's common measurement of students' progress at the end of the school year in English language arts, mathematics, science, and social studies. CMAS encompasses the Colorado-developed science and social studies assessments as well as the Partnership for Assessment of Readiness for College and Careers (PARCC)-developed English language arts (ELA) and mathematics assessments. The CMAS: Science and Social Studies assessments were first administered in 2014 in grades 4, 5, 7, 8, and high school. The CMAS: PARCC ELA and mathematics assessments were administered for the first time in spring 2015 in grades 3 through 10.

Beginning with the Spring 2016 administration, eligible Spanish-speaking students in grades 3 and 4 participated in the new Colorado Spanish Language Arts (CSLA) assessment in place of the CMAS: PARCC ELA assessment. CSLA is considered an accommodated form of the CMAS: PARCC ELA assessment. The CSLA assessments are aligned to the skills and concepts in the CAS and mirror the CMAS: PARCC ELA assessments. CSLA was developed because Colorado School Law §22-7-409 (3.5) (a) and (b) require a Spanish language arts assessment in grades 3 and 4.

Purpose of the Document

The purpose of the CSLA Technical Report is to inform users and other interested parties about the technical characteristics of this assessment. This technical report provides information about the Spring 2016 CSLA assessments, including content, assessment development, administration, scoring, and technical attributes.

The Spring 2016 CSLA Technical Report is divided into two parts. Part I presents an overview and summary of the components of the assessment. Information regarding the planning and administration of the assessment as well as details regarding item development, test construction, administration procedures, scoring, reporting, reliability, and validity are included in Part I of the document. Part II provides a statistical summary of the Spring 2016 administration, including results for both the operational items and the embedded field test items.

Overview of CSLA

Purposes of the CSLA Assessment

The primary purpose of CSLA is to provide high-quality linguistically accommodated Spanish assessments that align to the CMAS: PARCC ELA assessments. As part of the CMAS program, CSLA also seeks to achieve the goals of the Colorado Assessment System, which are to measure and support student progress toward the content standards; to provide students, parents, and other stakeholders with information regarding student achievement that can be used to help improve instruction and inform professional development; and to gauge the quality and efficiency of educational programs in public schools.

The Student Population

English language learners (ELLs) are a diverse group of students. These students come from a variety of cultural and educational backgrounds. Factors such as the number of years in school, amount of literacy and academic skills in their native language, access to language instruction, practice using academic English, mobility, and degree of family support can affect students' success in learning the English language (Breiseth, 2015).

Students who are eligible for CSLA are ELLs who have participated in an English language development program for five years or less and received academic instruction in Spanish within the past nine months. Districts must determine if the CSLA assessment is the best choice for the student. District assessment leadership should collaborate with ELL staff at schools to evaluate appropriateness and eligibility of a student to take CSLA. The CSLA eligibility flowchart can be found in Appendix A and is also available online at the following location: http://www.cde.state.co.us/assessment/csladecision

Description of CSLA

The 2015–2016 school year was the first administration of the CSLA assessment for eligible Spanish-speaking students in grades 3 and 4. CSLA was field tested in Colorado schools in the Spring and Fall of 2015. CSLA was administered in a paper-based format. The tests were created using blueprints that mirror the CMAS: PARCC ELA assessments and are intended to be an accommodated version of CMAS: PARCC ELA.

The CSLA assessments consist of several tasks and passage sets. There are three task types: Literary Analysis Task (LAT), Research Simulation Task (RST), and Narrative Writing Task (NWT). For these tasks, students are asked to read one or more texts, answer comprehension and vocabulary questions, and write an essay response based on the text(s) they read. There are also literary and informational reading passages on the tests with comprehension and vocabulary questions students must answer.

A specific claim structure is used in the design and development of the CSLA assessment. The test is designed to obtain evidence from students that support the claims about the degree to which students have mastered the content standards. To support such claims, CSLA is designed

to measure and report student performance for multiple claims and subclaims. Student performance is provided for Reading and Writing claims and five subclaims: 1) Reading Literary Text, 2) Reading Informational Text, 3) Reading Vocabulary, 4) Written Expression, and 5) Writing Knowledge and Use of Language Conventions.

The items administered on the assessment are developed to gather specific evidence to support the inferences, or claims, about what students know and can do in relation to the content standards. The CSLA assessment contains two item types: Evidence-Based Selected Response (EBSR) items and Constructed Response (CR) items. The EBSR items are machine-scored items and ask students to provide evidence from the text that led them to a previous answer. The CR items are human-scored items and ask students to provide an extended written response to an essay prompt.

The CR items can be categorized as Prose Constructed Response (PCR) items or Narrative Prose Constructed Response (NPCR) items. PCR items are administered as part of the LAT and RST tasks, and NPCR items are administered as part of the NWT task. The various tasks and passage sets and their associated items are combined into three units, which compose the operational items on the assessment. In addition to the operational units, an embedded field test unit is also included on the assessment. Including field test items on the operational test reduces the need for future stand-alone field tests and allows newly developed test items to be field tested with a relatively large participation count.

Assessment Development Partners

The CSLA assessments are collaboratively developed by the Colorado Department of Education (CDE), the Colorado educator community, and the assessment contractors, Pearson and Tri-Lin Integrated Services, Inc. Additional input and advice are provided by a Technical Advisory Committee (TAC).

Colorado Department of Education

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

Colorado Educator Community

Throughout the assessment development process, educators provide input through participation in content and bias review, data review, and standard setting meetings. For each meeting, an effort is made to involve educators who teach ELLs and educators who are familiar with the instruction and needs of the students in an English language development program. In addition to classroom teachers, school administrators, program directors, and post-secondary educators are also recruited to participate in the assessment development process.

Pearson

Pearson is the primary contractor, holding the responsibility for the administration and psychometric analysis of the CSLA assessments. This includes enrollment, packaging and

distribution, scoring, customer service, standard setting, score reporting, and psychometric services.

Tri-Lin Integrated Services, Inc.

Tri-Lin is a subcontractor and is responsible for content and test development. This includes passage development, item development, and test form construction.

Technical Advisory Committee

The TAC is composed of psychometric and assessment experts tasked with providing high-level consulting and expert advice regarding the creation of a reliable and valid assessment. Input is received on topics such as blueprint design, score reports, scaling and equating, and standard setting. The TAC members are as follows:

- Dr. Jamal Abedi, Professor, University of California, Davis
- Dr. Elliot Asp, Senior Fellow, State Policy and Implementation Support, Achieve
- Dr. Jonathan Dings, Executive Director of Student Assessment and Program Evaluation, Boulder Valley School District
- Dr. Lisa Escarcega, Colorado Association of School Executives
- Dr. Michael Kolen, Professor, University of Iowa
- Dr. Martha Thurlow, Director, National Center on Educational Outcomes

CHAPTER 2: ITEM DEVELOPMENT AND ITEM BANKING

The CSLA item development process involves various steps. It is structured in a manner to develop a variety of item types that align directly to the CAS. To the extent possible, CSLA follows a similar item development process as the CMAS: PARCC ELA assessment. When developing the passages and items, the CSLA item development process considers the purpose of the assessment, specifically, that the test is intended to be a linguistically accommodated version of the CMAS: PARCC ELA. Throughout the assessment development process, CDE relies greatly on input from Colorado educators who teach Spanish language arts and who are language development experts to ensure that the CSLA assessments are equitable for the intended population of students and that the assessments accurately measure the content.

The validity of a state assessment relies on the methodology that frames the development and design of the assessment. In support of that claim, Tri-Lin and Pearson have upheld these considerations as the cornerstones of CSLA item and test development:

- The test specifications ensure that the CSLA items align to the evidence statements they are intended to measure.
- The CSLA item development plan is designed to produce and maintain a robust item bank; items were written to address the scope of essential measured standards, grade-level difficulties, and cognitive complexity.
- The CSLA item and test development processes are compliant with industry standards.

Item-Writing Process

Developing high quality Spanish language arts assessment content with authentic stimuli that measures rigorous standards is a complex process that starts with item writing. Item writing is a tiered, inter-related process that begins with the development of the item development plan (IDP), based on the test blueprints for each grade level.

Test Blueprint

The CSLA test blueprints mirror the CMAS: PARCC ELA blueprints. Therefore, CSLA mirrors CMAS: PARCC ELA in terms of content, standards measured, item types, and score points. The CSLA test blueprints can be found in Appendix B.

Item Development Plan

The IDP is used to forecast the targeted number of items and associated passages needed to create a robust item bank that would be refreshed over time. The CSLA item bank supports the administration of the assessments along with practice tests.

Item development for CSLA began in the fall of 2014. CSLA passage and item development was conducted by Tri-Lin under the guidance and oversight of CDE and Pearson. The CSLA items are written to measure concepts and skills found in the CAS and go through multiple rounds of review, including content and bias review and data review.

The item-writing process included the following steps:

Passage Development

Using the CMAS: PARCC ELA Passage Selection Guidelines, Tri-Lin Spanish language arts content specialists and assessment developers were trained to develop appropriate passages that met the requirements of the text complexity framework and a variety of text types that allow for a range of standards/evidences to be demonstrated to meet the CMAS: PARCC ELA assessment claims. Tri-Lin applied the CMAS: PARCC ELA Task Generation Models and Cognitive Complexity framework to select passage tasks that most accurately assessed the content and cognitive and linguistic demands required at each grade level.

Tri-Lin assessment specialists conducted fact checking and reviewed the passages to ensure 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 prior to item writing as readily accessible, moderately complex, or very complex.

Tri-Lin Spanish editors checked passages for clarity, correctness of language, appropriateness of language for the grade level, and adherence to style guidelines.

After the CSLA passages were approved by CDE, Tri-Lin began the item development process.

Item Development

After the passages were approved by CDE, teams of Spanish item writers were trained and began developing items that mirrored the released CMAS: PARCC ELA items in the areas of command of textual evidence, response mode, processing demand, and text complexity. The CMAS: PARCC Item Guidelines for ELA/Literacy Summative Assessment and the Cognitive Complexity Framework guided item development to ensure that text complexity and item/task complexity interacted to determine the overall complexity of a task.

Three main sources of item complexity were identified:

1. Command of textual evidence – amount of text students must process in order to respond correctly to an item (low complexity was associated with items targeting a single piece of information; moderate to high complexity was associated with items requiring synthesis of ideas and details either from a single text or across texts).
2. Response mode – how students are required to respond to an item (low complexity was associated with selecting a correct answer from a series or list of options; moderate to high complexity was associated with selecting multiple correct answers, citing text evidence to support a response, and writing an extended constructed response).

3. Processing demand – linguistic demands and reading load in item stems, item directions, and response options. Three contributing features were identified with values ranging from low to moderate complexity.

Item Reviews

Item Reviews for Quality Assurance

After items were written, Tri-Lin's team of content specialists, assessment developers, and editors conducted rigorous reviews of items for content accuracy, alignment to the standards, range of difficulty, equitability for all student populations based on the principles of universal design, bias and sensitivity, and alignment with CMAS: PARCC format, style, and complexity. Reviewers also ensured that the items required students to find text-based evidence for generalizations, conclusions, or inferences drawn consistent with CMAS: PARCC's Cognitive Complexity Framework.

Tri-Lin conducted a universal design review to assess item accessibility irrespective of diversity of background, cultural tradition, and viewpoints; to appraise the role of language in setting; to appraise contributions of diverse groups to the history and culture of the United States, and to edit for inappropriate language usage or stereotyping with regard to sex, race, culture, ethnicity, class, or geographic region.

After the Tri-Lin internal reviews were completed, the items were reviewed and approved for presentation to the CDE by the lead assessment specialist. Prior to the educator committee reviews, CDE reviewed and approved the CSLA items.

Educator Content and Bias Review Meetings

CDE experts, Colorado educators, and postsecondary faculty with diverse backgrounds from across the state conducted rigorous reviews of every passage and item developed for the CSLA system to ensure all test items are of the highest quality, aligned to the standards, and fair for all student populations. The purposes of an educator review are to identify any potential bias or stereotype in test items and to ensure that the items are properly aligned to the content standards, accurately measure the intended content, and grade-appropriate. The educator reviews also provide feedback to Tri-Lin, Pearson, and CDE on the quality, accuracy, alignment, and appropriateness of the test passages and items developed. The meetings were conducted either in person or virtually and 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 passage and item:

- Used clear, unambiguous, and grade-level appropriate language
- Avoided complex sentence structure
- Had one correct answer

- Contained plausible distractors
- Represented the range of cognitive complexities and included challenging items for students performing at all levels
- Was appropriate for students in the assigned grade in terms of reading level, vocabulary, interest, and experience
- Had scoring guidelines that capture exemplar responses at each score point
- Included appropriate and clear graphics that are relevant
- Was free of ethnic, gender, political, and religious bias

In addition to reviewing all passages and items, committee members were given the opportunity to recommend edits and accept or reject items based on grade-level appropriateness, content, and potential bias concerns. The committee made one of three recommendations on every item: "accept," "accept with modifications," or "reject."

A Content and Bias Reconciliation Meeting was conducted following the educator meeting. The reconciliation meeting included CDE, Pearson, and Tri-Lin staff. At this meeting, committee comments were reviewed, proposed edits were reconciled, and item outcomes were finalized. The approved passages and items were then placed in the CSLA item bank, thereby becoming eligible for future field testing.

Data Review

After development of the CSLA items, selected items were administered in a stand-alone field test in the Spring and Fall of 2015. The goal of a field test is to allow for the evaluation of the quality of the items through a review of traditional item performance data to support test construction. A committee of educators who teach ELLs at grades 3 and 4 were convened to review the newly developed items along with the student performance data. The data review committee members were provided passages, item images, and content metadata along with classical statistics and Differential Item Functioning (DIF) statistics to review.

The classical statistics included item sample size, p-value, point biserial, item mean score, item-total correlation, and response distribution. DIF analyses were conducted by gender using the Mantel & Haenszel and the Mantel method, which is a polytomous extension of the Mantel-Haenszel statistic (Mantel, 1963; Mantel & Haenszel, 1959). Classification rules were used to classify items as having either negligible, moderate, or large DIF. Items that were classified as having moderate or large DIF were reviewed by the data review committee.

During the data review meetings, educators were trained to interpret the statistical information, and while the committee used the data as a tool to inform their judgments, the committee was instructed not to base their final assessment of the appropriateness or fairness of items solely on these data. Committee members reviewed each item and made a recommendation as to whether to "accept" or "reject" the item.

Following the data review meetings in the Spring and Fall of 2015, Data Review Reconciliation meetings were held which included CDE, Tri-Lin, and Pearson staff. At the reconciliation meeting, the assessment specialists and psychometricians discussed the committee comments

from the data review meeting as well as any concerns they have about the items. After the item outcomes were finalized during reconciliation, accepted field test items were included in the item pool used to build the Spring 2016 operational test. Items that were not used on the operational test were classified as available for use on future operational assessments. Items that were rejected were re-classified to eliminate them from use on an operational test. These items may be modified and field tested again on future test forms.

Following the first operational administration of the CSLA assessments in Spring 2016, a subsequent data review meeting was convened to review the item performance of the embedded field test items included on the operational assessments. The committee evaluated the items and the student performance data and provided recommendations as to whether to "accept" or "reject" the items. After the data review meeting, item outcomes were finalized during a reconciliation meeting and the field test items were classified accordingly in the item bank.

Item Banking

The CSLA item bank houses passages and items at each grade level. The item bank supports the administration of the assessments. Items that passed all stages of the development process (e.g., item review, content and bias review, and data review) were placed in the operational item bank to become eligible for use in future assessments. Prior to each operational administration, the item bank is evaluated to determine the item development needs for future operational administrations.

Item Bank Statistics

The metadata for each item are included in the item bank, which includes the item image, test date, the assessed content standards, the form on which the item appeared, the item position on the form, the item type, the correct key, and the maximum number of points possible for a correct answer.

The item summary statistics include the item sample size, p-value, point biserial, item mean score, item-total correlation, the response distribution that presents the percentage of students achieving each score point both overall and by ability level, and DIF classification by gender.

CHAPTER 3: TEST CONSTRUCTION

Test forms are constructed through an iterative process between Tri-Lin and Pearson staff. CDE then reviews the forms, provides feedback, and gives final approval as described below.

When building operational test forms, the Tri-Lin assessment specialists select a set of operational items in accordance with the test blueprint and test construction specifications. 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 are used during form construction:

- adherence to the test blueprint
- review of the item statistics and adherence to the statistical criteria found in the test construction specifications
- balance of gender, ethnicity, geographic regions, and relevant demographic factors
- selection of items with various stimuli types throughout the test form to enhance the test-taker experience by providing variation in the items presented
- efficient and deliberate use of varied content representative of the knowledge and skills in the content standards
- review of the full test form, including field test items, for instances of clueing and/or content overlap

After the initial operational items are selected, the test form is reviewed by Tri-Lin assessment specialists. The assessment specialists verify that the form meets the test blueprint (i.e., the required number of passages, items, and item types). The form is then presented to Pearson psychometrics for analysis; the psychometrician verifies that the form falls within the established psychometric and blueprint parameters.

After the form is reviewed by Tri-Lin and Pearson, the form is presented to CDE for review. If needed, CDE, Tri-Lin, and Pearson collaborate to finalize the form. This can be an iterative process with the end result being CDE's form approval.

After the operational form is approved, field test items are selected from the item bank. The assessment specialists assemble field test item sets so that they comprise the appropriate distribution of the required number of passages, items, and item types. They also review item replacement for future years to ensure appropriate item rotation. Field test items chosen are embedded on the operational form in a designated location.

The specific responsibilities for Tri-Lin, Pearson, and CDE during test construction are outlined below:

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

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

CHAPTER 4: TEST ADMINISTRATION PROCEDURES

This chapter provides information related to the CSLA test administration procedures. Prior to the test administration, CDE provided training for Colorado districts, schools, and teachers to ensure that schools and students were prepared for the assessments and that test administration procedures were standardized. Test administration procedures were communicated as described below.

Manuals

Several manuals were created to support the CSLA administration. These manuals include the following:

- *CSLA Test Administrator Manual*
- *CSLA Data Supplement*
- *PearsonAccess^{next} User Guide*

Training

CDE conducted in-person administration trainings for District Assessment Coordinators in Colorado. In addition, Pearson customer service center staff were trained to answer questions thoroughly and knowledgably about the CSLA administration and to escalate inquiries as necessary. CDE also hosted WebEx training sessions covering topics such as CSLA eligibility requirements, test design, accommodations, distribution of materials, and test security.

Accessibility and Accommodations

The CSLA assessments were developed to be accessible for eligible Spanish-speaking students. Linguistic accessibility was considered from the beginning of the test development process and is inherent within the CSLA assessment and administration. Even though the assessments are designed to be linguistically accessible, students taking the assessments may require changes to the assessment procedures, or accommodations, in order to accurately demonstrate their knowledge and skills of the content.

Accommodations provide a student with an opportunity to engage with the assessment while not affecting the reliability or validity of the assessment. Accommodations can be adjustments to the test presentation, materials, environment, or response mode of the student and are based on student need. Accommodations should not provide an unfair advantage to any student. Providing an accommodation for the sole purpose of increasing test scores is not ethical. Accommodations must be documented in the student's Individualized Education Plan (IEP) or 504 Plan 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. Available accommodations for the CSLA assessment include a large print version and an oral script version. Other allowable accommodations align with CMAS: PARCC's allowable accommodations for students with an IEP or 504 Plan participating in the ELA/literacy paper-based assessment. CMAS: PARCC's linguistic accommodations do not apply because the CSLA form is the linguistic accommodation.

Test Security

Districts were trained on assessment security to ensure that security procedures were maintained during the test administration. Materials used during the administration of the assessment were to be kept in locked storage locations when not under the direct supervision of approved assessment coordinators or test administrators. All state, district, and/or school personnel were required to sign a security agreement prior to handling test materials. By signing the security agreement, personnel agreed to a set of security guidelines that required them to follow all procedures set forth in the manuals. Personnel could not divulge the contents of the assessment, copy any part of the assessment, except for students with allowable accommodations, or review test questions with students. They also could not allow students to remove test materials from the room where testing takes place or interfere with the independent work of any student taking the assessment.

CHAPTER 5: SCORING THE ASSESSMENTS

The CSLA assessment contains two item types: EBSR items and CR items. The EBSR items are machine-scored items and ask students to provide evidence from the text that led them to a previous answer. The selected response items are scored on a 0−2 point scale.

The CR items are human-scored items and ask students to provide an extended written response to an essay prompt. The CR items can be categorized as PCR items or NPCR items. PCR items are administered as part of the LAT and RST tasks, and NPCR items are administered as part of the NWT task.

Both the PCR and NPCR items have two trait dimensions. The PCR traits are 1) Reading Comprehension and Written Expression (RCWE) and 2) Writing Knowledge of Language and Conventions (WKLC). The NPCR traits are 1) Written Expression (WE) and 2) Writing Knowledge of Language and Conventions (WKLC). For the PCR LAT and RST tasks, the RCWE trait is worth 0−3 points for grade 3 and 0−4 points for grade 4. The PCR trait of WKLC is worth 0−3 points for both grades 3 and 4. For the NPCR items, all traits are worth 0−3 points. Weighting is also applied to the RCWE and WE traits as part of the test design. Written expression is weighted by 3 to give it more emphasis in the total score. The holistic rubrics used to score the PCRs and NPCRs mirror the rubrics developed for the CMAS: PARCC ELA assessment and can be found in Appendix C.

Pearson's Performance Scoring team implemented the CR scoring process. The CR scoring process is described below.

Scoring Model

Each operational test is scored using a Regional Scoring model. Regional Scoring includes several components that together provide a comprehensive performance scoring model.

- Scorers are trained using comprehensive training materials developed by scoring experts. These materials include student responses scored by participants at the rangefinding meetings.
- Scorers must pass a qualifying test for the item types that they will score.
- Student responses are converted to electronic images at Pearson facilities. They are then transmitted for computer-based scoring.
- Scorers work from the San Antonio, TX, Pearson Scoring Services facility. Their computers are set up for image-based scoring. A comprehensive set of scoring and monitoring tools is integrated into the scoring system.

Pearson's processes and tools provide a replicable quality system that strengthens consistency across projects and locations within Pearson's Scoring Services operations. Pearson's Scoring Services team uses a comprehensive system for continually monitoring and maintaining the accuracy of scoring on both group and individual levels. This system includes daily analysis of a comprehensive set of statistical monitoring reports, as well as regular "backreading" of scorers.

Embedded field test scoring was completed using regional scorers. Regional scoring took place in San Antonio, TX. All scorers were required to have a four-year college degree. The following sections describe the rangefinding process and the major components of the quality assurance system, including backreading and calibration.

Rangefinding

Rangefinding meetings are held following the administration in which an item is field tested. The purpose of rangefinding is to define the range of performance levels within the score points of the rubrics using student responses. Each rangefinding committee includes Pearson's Scoring Services staff, CDE content representatives, and grade level teachers with relevant content expertise and experience with special populations. Participants create consensus scores for student responses that are subsequently used to develop effective training materials for scoring of CR items.

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

Following an introductory session presented by a Pearson Assessment Creation Services content specialist, the rangefinding committee is divided into several break-out groups. Each group is assigned a range of field test items to be reviewed, following the process outlined below:

1. The scoring director introduces each item. The committee reviews the item and corresponding rubric.

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

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

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

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

Backreading

Backreading is the method of immediately monitoring a scorer's performance, and, therefore, an important tool for Pearson's scoring supervisors. Backreading is performed in conjunction with the statistics provided by reader performance reports and as indicated by scoring directors, allowing scoring supervisors to target particular readers and areas of concern. Scorers showing low inter-rater agreement or those showing anomalous frequency distributions are given immediate, constructive feedback and monitored closely until sufficient improvement is demonstrated. Scorers who demonstrate through their agreement rates and frequency distributions that they are scoring accurately will continue to be spot-checked as an added confirmation of their accuracy. Rater agreement information for the Spring 2016 administration can be found in Part II of this report.

Calibration

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

CHAPTER 6: STANDARD SETTING

To support the interpretation of student results, student performance on the CSLA assessments is described in terms of five performance levels: Exceeded Expectations, Met Expectations, Approached Expectations, Partially Met Expectations, and Did Not Yet Meet Expectations. After the first operational administration of the CSLA assessments in Spring 2016, a standard setting meeting was held to determine the performance standards. Performance standards specify what level of performance on a test is required for a test taker to be classified in a given performance level.

The Modified Extended Angoff approach (Cizek, 2012; Cizek, Bunch, & Koons, 2004; Hambleton & Plake, 1995) was used to set performance standards on the CSLA assessments. With this methodology, standard setting panelists review the content of each test item, and considering the content the item is measuring and the content knowledge of the students at the cut scores (i.e., borderline students), the panelists make a judgment about what score a borderline student would receive on the item to be considered "just barely" in a performance level. Panelists use the PLDs to conceptualize "borderline" students (those students just barely in a particular performance level) in order to determine the score the borderline student would obtain on each item. The individual item-level cut scores for each particular performance level are then summed for each panelist to obtain the recommended test-level cut scores that are used to define the performance levels.

One committee was convened to recommend performance standards for both grades 3 and 4. The CSLA standard setting committee consisted of ten panelists. Panelists were grouped into tables of three with three to four panelists per table. The CSLA panelists included educators who teach ELL students at grades 3 and 4, are content experts with knowledge of the subject-area curriculum, and are familiar with the instruction and specific needs of the students in an English language proficiency program. In addition to classroom teachers, educators in higher education and school administrators and/or directors who are familiar with instruction in classrooms where the Spanish language is used also participated in the meeting.

The CSLA standard setting was held on June 27–29, 2016. During the three-day meeting, panelists received training on the assessment and the standard setting process, reviewed the grade-level PLDs, reviewed the Spring 2016 operational items, reviewed the borderline student descriptors, and applied the Modified Extended Angoff method to establish cut score recommendations across three rounds of rating. During the process of establishing cut score recommendations, panelists also reviewed the content standards assessed by the CSLA items, reviewed CMAS: PARCC ELA external data, engaged in table level and whole group discussions, and considered the impact of their cut scores on student performance when making their CSLA cut score recommendations.

Once the performance standards were recommended for the grade 3 and the grade 4 assessments, the standard setting panelists made cross-grade comparisons during vertical articulation. The purpose of vertical articulation was to review the impact data associated with the recommended cut scores across both grades to determine if the trend of the impact data is reasonable given the PLDs, the test-taking population, and the concepts and skills presented on the assessments. At

the completion of vertical articulation, the cut score recommendations were then reviewed by CDE to ensure that the performance standards contributed to a well-articulated and coherent assessment program.

For grade 3, an estimated 22% of students were in the top two performance levels, 20% for Met Expectations, and 2% for Exceeded Expectations. For grades 4, an estimated 14% of students were in the top two performance levels, 13% for Met Expectations, and 1% for Exceeded Expectations. More details about the CSLA standard-setting meeting and the final cut scores can be found in the full standard setting report in Appendix D.

CHAPTER 7: REPORTING

Several score reports are generated to communicate student performance on the CSLA assessment. The information below describes the types of scores given on reports and the types of reports available. For additional details on score reports, see the Spring 2016 Score Interpretive Guide at http://www.cde.state.co.us/assessment/2016cslainterpguid.

Description of Scores

CSLA reports provide information about student performance in terms of scale scores, performance levels, and subclaim performance indicators.

Scale Scores

A scale score is a conversion of a student's total test score (i.e., the total number of points earned on a test) onto a scale that is common to all test forms for that assessment. Scale scores are particularly useful for comparing assessment scores across years from different test administrations. For CSLA, students receive an overall test scale score that determines a student's performance level. CSLA scale scores ranges from 650 to 850. Conditional standard error of measurement (CSEM) is provided as an indicator of the range of scale scores a student would likely receive if the assessment was taken multiple times. Additionally, CSLA reports separate scale scores for the Reading and Writing claims, also called reporting categories. CSLA Reading scale scores range from 10 to 90 and CSLA Writing scale scores range from 10 to 60. Chapter 8 provides technical details related to scale development.

Performance Levels

Performance levels are reported at the overall test level. Examinees are classified into performance levels based on their overall scale score as compared with the cut scores, which were obtained from standard setting. CSLA has five performance levels:

- Exceeded Expectations
- Met Expectations
- Approached Expectations
- Partially Met Expectations
- Did Not Yet Meet Expectations

Subclaim Performance Indicators

Within each Reading and Writing reporting category for CSLA are specific skill sets (subclaims) students demonstrate on the assessment. Five subclaims are reported: 1) Reading Literary Text, 2) Reading Informational Text, 3) Reading Vocabulary, 4) Written Expression, and 5) Writing Knowledge of Language and Conventions. Subclaim performance is reported indicating how the

student performed relative to the overall performance of students who met or nearly met expectations for the grade-level assessment. As with the overall test score and reporting category scores, a measure of student proficiency for each subclaim is estimated on a common, underlying measurement scale. Subclaim performance is reported using categories rather than scale scores. Performance in the Level 1–2 range of that scale is categorized as 'Did Not Yet Meet or Partially Met Expectations," performance in the Level 3 range is categorized as "Approached Expectations," and performance in the Level 4–5 range is categorized as "Met or Exceeded Expectations."

Score Reports

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

Student Performance Reports

The Student Performance Report provides information about the performance of a particular student on the CSLA assessment. The student's scale score, associated performance level, and subclaim performance indicators are displayed on a two-page report along with comparative information related to school, district, and state performance. Student Performance Reports are printed and shipped to districts for distribution to students and parents.

Aggregate Reports

Three types of aggregate reports are produced for CSLA:

- Content Standards Roster
- Performance Level Summary
- Evidence Statement Analysis

These reports are produced at the school, district, and state levels and provide summary information for a given school or district. State, district, and school reports are provided electronically through PearsonAccess^{next} Published Reports, and access to the reports is limited to authorized users.

CHAPTER 8: CALIBRATION, EQUATING, AND SCALING

Item Response Theory (IRT) was used to develop, calibrate, equate, and scale the CSLA assessments. The Rasch Partial Credit Model was the measurement model used for test construction, calibration, scaling, and equating and to maintain and build the item bank. All calibration, scaling, and item-model fit analyses were accomplished within the IRT framework. The initial administration of the CSLA assessments in Spring 2016 determined the base scale for the assessments.

Calibration

The Rasch Partial Credit Model

Calibration is the process used to obtain item parameter estimates and then place all items and students on a common scale. For each CSLA grade-level assessment, the Rasch Partial-Credit Model (RPCM) was used to place the items and student proficiency on the same Rasch scale. The model is an extension of the Rasch one-parameter IRT model attributed to Georg Rasch (1966), as extended by Wright and Stone (1979), Masters (1982), and Wright and Masters (1982). The RPCM was selected because of its flexibility in accommodating various item types (i.e., multiple-choice items and items with multiple response categories). The RPCM maintains a one-to-one relationship between scale scores and raw scores, meaning that each raw score is associated with a unique scale score. It is the underlying Rasch scale that allows for comparisons of student performance across years and facilitates the maintenance of equivalent performance standards across years.

The RPCM is defined by the following mathematical measurement model where, for a given item involving *m*+1 score categories, the probability of person *n* scoring *x* on question *i* is given by:

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

The RPCM provides the probability of a student scoring *x* on *m* steps of question *i* as a function of the student's proficiency level, θ_n (sometimes referred to as "ability"), and the step difficulties, δ_{ij}, of the *m* steps in question *i*.

Equating and Scaling

Equating involves adjusting for differences in the difficulty of test forms, both within and across assessment administrations. Equating makes certain that students taking one form of a test are neither advantaged nor disadvantaged when compared to students taking a different form. Each time a new test form is constructed, equating is used to allow scores on the new form to be comparable to scores on the previous form by placing the scores on both forms on the same scale. It is the underlying Rasch scale obtained from calibration that facilitates equating of test

forms. The Rasch scale can then be transformed to create scale scores to allow for the interpretation of test scores.

Equating and Scaling

The Spring 2016 administration of the CSLA assessments represents the first operational tests on the newly developed Rasch scale. In the following years, equating will be used to place the new operational test forms on this newly developed operational scale. To obtain Rasch item parameter estimates for the Spring 2016 CSLA assessments, the RPCM was applied to the operational and embedded field test items. Winsteps (Linacre, 2011) was used for all grade-level calibrations.

The calibration of the operational and embedded field test items for each grade-level assessment occurred in several steps. First, the operational items were calibrated. For the PCR and NPCR items, unweighted trait scores were used in the calibration. Next, the embedded field test items were calibrated with the operational items using fixed common item parameter calibration. With this calibration method, the embedded field test items are calibrated with the operational item parameters fixed at their previously estimated item parameter values in order to place the embedded field test items on the same scale as the operational items.

Ability Estimates

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

Scale Scores

Student proficiencies for each assessment were then transformed to scale scores. The CSLA scale scores represent linear transformations of the student proficiencies (θ). The transformation is made by first multiplying any given θ by a slope (a) and then adding an intercept (b). The following linear transformation was used to convert student proficiency estimates into scaled scores (SS):

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

The a and b values are referred to as scaling constants. These scaling constants are applied each year to the Rasch proficiency estimates for that year's set of operational items. In order to obtain the two scaling constants, two features of the desired CSLA scale score system were identified in advance. For CSLA, the proficiency estimate corresponding to the Level 2 cut score and the proficiency estimate corresponding to the Level 4 cut score were identified and used to obtain the a and b scaling constants. To generate the scale scores for the overall test and the Reading

and Writing claims, three sets of scaling constants were calculated for each grade-level assessment.

Once the scaling constants were obtained, student proficiencies for the overall test were then transformed to scale scores with a range from 650 to 850 where Level 2 is a scale score of 700 and Level 4 is a scale score of 750. Student proficiencies for Reading were transformed to scale scores with a range from 10 to 90 where Level 2 is a scale score of 30 and Level 4 is a scale score of 50. Student proficiencies for Writing were transformed to scale scores with a range from 10 to 60 where Level 2 is a scale score of 25 and Level 4 is a scale score of 35. After the scale scores were obtained, the lowest observable scale score (LOSS) and the highest observable scale score (HOSS) were applied. The 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 10 and 90; and for the Writing scale, LOSS and HOSS were set to 10 and 60.

Subclaim Performance Indicators

Subclaim performance is reported using categories rather than scale scores. The subclaim performance categories are 1) Met and Exceeded Expectations, 2) Approached Expectations, and 3) Did Not Yet Meet Expectations or Partially Met Expectations. In order to obtain the three categories, two raw score reference points were identified for each subclaim. To determine the reference points, student proficiency estimates were first generated for each subclaim. Only those operational items representing the specific subclaim were included in the anchored calibration to obtain the proficiency estimates. The calibration for the Written Expression subclaim included the weighting of the PCR and NPCR trait scores. The proficiency estimates corresponding to the Level 3 cut score and the Level 4 cut score on the overall scale were then located on each proficiency scale to determine the reference points. These score points were then used to determine the three performance indicators for each subclaim.

Steps in the Calibration and Scaling Process

The entire process previously described was conducted for each CSLA assessment. 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 in preparation.

- The data files containing student responses were verified and exclusion rules were applied.
- Traditional item analyses of all items were conducted prior to calibration.
- Incomplete data matrices (IDMs) were created.

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

- Item sample size
- P-value
- Point biserial
- Item mean score
- Item-total correlation
- Response distribution

Prior to calibration, the classical statistics for the parts of the EBSR items that are key-based were also evaluated to identify potential test administration or scoring issues. A list of flagged items identified using flagging criteria was communicated to the assessment specialists for review and confirmation that the correct key had been applied.

Calibration

Several different calibrations were conducted to obtain item parameter estimates for the operational and embedded field test items.

- Operational Items
    - Used Winsteps control files and IDM to obtain operational item parameter estimates
        - Obtained operational Rasch item difficulty values, step deviation values, and item fit values
- Embedded Field Test Items
    - Used Winsteps control files and IDM to scale the embedded field test item parameter estimates to the operational scale by fixing the item parameter estimates of the operational items
        - Obtained embedded field test Rasch item difficulty values, step deviation values, and item fit values

CHAPTER 9: RELIABILITY

A variety of statistics can be calculated that pertain to the reliability of the CSLA assessments. In this report, Cronbach's alpha, standard error of measurement (SEM), conditional standard error of measurement (CSEM), decision consistency and accuracy, and inter-rater agreement will be described. For these statistical estimates, see Part II of this document.

Cronbach's Alpha

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

The reliability coefficient of a measure is the proportion of variance in observed scores accounted for by the variance in true scores. The coefficient can be interpreted as the degree to which scores remain consistent over parallel forms of an assessment (Ferguson & Takane, 1989; Crocker & Algina, 1986). There are several methods for estimating reliability; however, in this report, an internal consistency method is used. In this method, a single form is administered to the same group of subjects to determine whether examinees respond consistently across the items within a test. A basic estimate of internal consistency reliability is *Cronbach's Coefficient Alpha* statistic (Cronbach, 1951). Coefficient alpha is equivalent to the average split-half correlation based on all possible divisions of a test into two halves. Coefficient alpha can be used on any combination of dichotomous (two score values) and polytomous (two or more score values) test items and is computed using the following formula:

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

where n is the number of items,

S_j^2 is the variance of students' scores on item j, and

S_X^2 is the variance of the total-test scores.

Cronbach's alpha ranges in value from 0.0 to 1.0, where higher values indicate a greater proportion of observed score variance is true score variance. Two factors affect estimates of internal consistency: test length and homogeneity of items. The longer the test, the more observed score variance is likely to be true score variance. The more similar the items, the more likely examinees will respond consistently across items within the test. The coefficient alpha estimates can be found in Tables 1–2 and Table 4.

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 (such as coefficient alpha) to estimate the number of score points that a student's test score would be expected to vary if the student were 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. SEM values can be found in Table 3.

Conditional Standard Error of Measurement

While the SEM provides an estimate of precision for an assessment, the CSEM considers how measurement error likely varies across the scale score. In other words, the CSEM provides a measurement error estimate at each score point on an assessment. Because there is typically more information about students with scores in the middle of the score distribution where scores are most frequent, the CSEM is usually smallest, and thus the scores are most reliable, in the middle of the score distribution.

An IRT method for estimating score-level CSEM is used because test- and item-level difficulties for the CSLA assessments were calibrated using the Rasch measurement model. By using CSEMs that are specific to each scale score, a more precise error band can be placed around each student's observed score. CSEM values are provided in Tables 19–24.

Decision Consistency and Accuracy

The overall test-level scales for CSLA are divided into five performance levels: Exceeded Expectations, Met Expectations, Approached Expectations, Partially Met Expectations, and Did Not Yet Meet 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 the CSLA assessments. The probability of an accurate classification (PA) is the probability that the performance level classification a student received is correct and is based on the agreement between the observed classification on the actual test form and true classification. The probability of a consistent classification (PC) is the probability that the performance level classification the student received is consistent with the classification that the student would have received on a parallel form. The probability of consistent classification by chance is the probability that the performance level the student received is accurate and occurred by chance. Kappa describes the agreement between classifications on two parallel forms. Consistency and accuracy estimates are provided in Table 26.

Inter-Rater Agreement

For the 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. For this method, two raters score the CR item using the appropriate rubric. The two independent ratings are then compared to determine the consistency of the ratings. Perfect, adjacent, and non-adjacent agreement rates were calculated. Rater agreement statistics are provided in Table 27.

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 CSLA assessments that are validated but rather the interpretations of the CSLA scores. The purpose of the CSLA assessments is to provide information about a student's level of mastery of the CAS. In support of that purpose, the previous chapters of this report describe processes that were implemented throughout the CSLA assessment cycle with validity and fairness considerations in mind; this chapter provides information regarding specific sources of validity evidence as well as fairness. Furthermore, validation is a process. As the 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

It is important to examine the extent to which the items on an assessment measure the intended construct. The CSLA assessments intend to measure the content standards of the CAS and steps are put in place throughout the development process with focus on this goal, as outlined in Chapter 2 of this report. For example, there are numerous reviews that an item goes through to confirm that it adequately aligns to the evidence statement that it is intended to measure. In addition, with the field testing of items, statistical bias analyses (i.e., DIF analyses) are conducted to identify any items that may be measuring a dimension unrelated to the intended construct. The test blueprints were carefully developed with specificity at multiple levels in an attempt to most optimally measure the content standards.

Evidence Based on Response Processes

Evidence based on response processes pertains to the cognitive aspect behind how students respond to items and the processes by which judges or observers evaluate student performance. On CSLA, responses from selected response items and written response items are obtained. Both item types were developed to more effectively measure the rigorous content standards. With the selected response items, a student must provide an answer to one part of the item and then provide evidence from the text that led them to the previous answer. The written response items require students to write a response to an essay prompt, which provides an authentic means for evaluating how well students can compose a written response across different types of genres. Evidence about how students are interacting with and responding to the EBSR and CR items was gathered from student field test responses (i.e., statistics such as item difficulty, response distribution, correlations, and DIF) and from feedback from educators who reviewed the statistics during data review.

Evidence Based on Internal Structure

The internal structure of an assessment pertains to the degree to which the items on an assessment measure one underlying construct. When assessments are designed to measure one underlying construct, the internal components of the assessments should exhibit a high degree of homogeneity that can be measured in terms of the internal consistency estimates of reliability. As a result, the internal consistency for the CSLA assessments is evaluated using reliability coefficients. In addition, the correlations between the claims and subclaims are provided. The internal consistency estimates are described in Chapter 9 and provided for the overall test, claims, and subclaims, as well as various subgroups in Part II of this report.

Evidence Based on Relations to Other Variables

Another measure of validity evidence is the relationship between test performance and performance on another measure, called criterion-related validity. This can be the relationship between two assessments taken at the same time (i.e., concurrent validity) or the relationship between assessments that measure the same or similar construct (i.e. convergent validity) or unrelated constructs (i.e., discriminant validity). Other available assessment scores that can be used for criterion-related validity evidence are being evaluated for CSLA.

Evidence for Validity and Consequences of Testing

As the CAS become more fully integrated into the classroom, and with additional administrations of the CSLA assessments, it is intended that information around the consequences of the assessment will be collected. Data regarding the intended and unintended consequences of the CSLA assessments will be collected and provided when data become available.

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 have been woven into the development and administration of the CSLA assessments.

Universal Design

The CSLA development process adheres to the principles of universal design, as described in Chapter 2, with the goal of avoiding construct-irrelevant aspects of the assessment.

Differential Item Functioning

Items are analyzed for DIF in order to identify any items that appear to be unfairly favoring one subgroup over another. All DIF-flagged items are then reviewed by assessment specialists to investigate whether there may be a flaw with the item.

Accessibility and Accommodations

As described in Chapters 3 and 4, the CSLA assessments were developed to be linguistically accommodated Spanish tests. In addition to incorporating accessibility into the assessment, accommodations are also available to those students who need additional changes to the test administration in order to access the assessment.

Practice Tests

Practice tests 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. Teachers and students were provided the opportunity to experience a sample test prior to the first operational administration of CSLA.

PART II: STATISTICAL SUMMARIES

This section contains an overview of the statistical summaries for the Spring 2016 administration. Administration summaries, calibration results, performance results, reliability evidence, and validity evidence are included for the operational items. Test form summaries and item performance review outcomes are provided for the embedded field test items.

CHAPTER 1: OPERATIONAL ITEMS

The following section provides high-level details about the CSLA assessments.

Administration Summary

Approximately 2,000 students took the CSLA assessments. Tables 1–4 show descriptive statistics for students and subgroups. The tables include descriptive statistics for the scale scores as well as reliability and SEM estimates. Descriptive statistics are also provided for the subclaims.

Calibration Results

Item Statistics

Tables 5–6 contain the classical item statistics. The "Type" column indicates the item type (i.e., Evidence-Based Selected item [EBSR] or Constructed Response item [CR]). Columns "% 0" through "% 4" contain the percentage of students at each score point for each operational item, and the "Mean Score" and "Item-Total Corr" columns contain the average score students earned on the item and the correlation between students' total test score and their item score.

Tables 7–8 contain the item parameter estimates for each grade-level assessment. The "Type" column indicates the item type. The "B" column contains the Rasch item difficulty estimates, columns "D1" through "D5" contain the category estimates, and the "Infit" and "Outfit" columns contain the item fit values.

See Chapter 8 for detailed information about the calibration process.

Performance Results

The cut scores, percent of students in each performance level, and the scale score ranges are provided in Tables 9–10. The percent of students in each subclaim performance category is provided in Table 11. The scale score distributions for each assessment are shown in Tables 12–17. Tables 19–24 are provided and include the raw score, scale score, and CSEM values. Correlations were calculated between the claims and subclaims for each assessment and are provided in Table 25.

Decision Consistency and Accuracy

Table 26 provides statistics related to decision consistency and accuracy. The table shows the consistency and accuracy estimates as well as the probabilities due to chance and kappa for both the assessments.

CHAPTER 2: EMBEDDED FIELD TEST ITEMS

The following section provides details around the field test items that were embedded within the CSLA assessments.

Field Test Items

Field test items were included on each operational test form. Forty-eight field test items were administered across the assessments. Each test form within a grade level was parallel; each student received the same number of each item type and in the same location on the form. Table 28 summarizes the number of field test forms and field test items per grade.

Data Review

Student performance data were obtained for all field test items and reviewed to determine if item performance was acceptable for the items to be used on future operational assessments. If any items were flagged for poor performance during the review process, the items would then go to data review to be reviewed by a committee of educators where they would decide whether to accept or reject the item. Table 28 summarizes the outcomes of the data review meeting where most items were accepted.

REFERENCES

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.

Breiseth, L. (2015). What You Need to Know About ELLs: Fast Facts. Retrieved from http://www.colorincolorado.org/

Cizek, G. J. (2012). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). New York, NY: Routledge.

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

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

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

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

Hambleton, R. K. & Plake, B.S. (1995). Using an Extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, *8*, 41–56.

Linacre, J.M. (2011). Winsteps® (Version 3.73.0) [computer program]. Beaverton, OR: 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.

Mantel, N. (1963). Chi-square tests with one degree of freedom; Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute, 22,* 719-748.

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

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

Research Associates.

Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

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

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

CSLA TABLES 1–28

Table 1. Grade 3 Performance by Subgroups

Group Type	Subgroup	N	Mean	SD	Min	Max	Alpha
Total Score		**1512**	**731**	**24.0**	**650**	**850**	**0.90**
Gender	Female	740	736	23.3	664	850	0.89
	Male	772	726	23.5	650	800	0.90
Ethnicity	American Indian	1	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	0	-	-	-	-	-
	Hispanic/Latino	1506	731	24.0	650	850	0.90
	White	5	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	0	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	86	739	27.2	664	850	0.90
	Economically Disadvantaged	1426	730	23.7	650	824	0.90
Students With Disabilities	504	11	-	-	-	-	-
	IEP	114	706	18.0	672	757	0.79
	No	152	731	24.0	650	792	0.88
	Missing	1235	733	23.1	653	850	0.89
Reading Score		**1512**	**42**	**8.8**	**10**	**80**	**0.90**
Gender	Female	740	44	8.7	21	80	0.89
	Male	772	41	8.7	10	67	0.89
Ethnicity	American Indian	1	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	0	-	-	-	-	-
	Hispanic/Latino	1506	42	8.8	10	80	0.90
	White	5	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	0	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	86	45	9.6	21	80	0.90
	Economically Disadvantaged	1426	42	8.8	10	75	0.89
Students with Disabilities	504	11	-	-	-	-	-
	IEP	114	34	5.4	17	49	0.66

Group Type	Subgroup	N	Mean	SD	Min	Max	Alpha
	No	152	42	8.6	10	61	0.88
	Missing	1235	43	8.7	17	80	0.89
Writing Score		**1512**	**30**	**8.9**	**10**	**60**	**0.72**
Gender	Female	740	32	8.6	10	60	0.71
	Male	772	28	8.7	10	49	0.70
Ethnicity	American Indian	1	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	0	-	-	-	-	-
	Hispanic/Latino	1506	30	8.9	10	60	0.72
	White	5	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	0	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	86	32	9.4	10	60	0.78
	Economically Disadvantaged	1426	29	8.8	10	56	0.72
Students with Disabilities	504	11	-	-	-	-	-
	IEP	114	20	9.2	10	44	0.75
	No	152	30	9.5	10	58	0.74
	Missing	1235	30	8.2	10	60	0.69

Table 2. Grade 4 Performance by Subgroups

Group Type	Subgroup	N	Mean	SD	Min	Max	Alpha
Total Score		**527**	**727**	**21.1**	**651**	**779**	**0.86**
Gender	Female	272	730	20.2	676	779	0.86
	Male	255	723	21.6	651	772	0.86
Ethnicity	American Indian	3	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	1	-	-	-	-	-
	Hispanic/Latino	518	727	21.3	651	779	0.86
	White	5	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-

Group Type	Subgroup	N	Mean	SD	Min	Max	Alpha
	Two or More Races	0	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	28	724	21.9	676	758	0.88
	Economically Disadvantaged	499	727	21.1	651	779	0.86
Students With Disabilities	504	0	-	-	-	-	-
	IEP	47	710	18.1	663	758	0.78
	No	114	725	20.9	651	768	0.85
	Missing	366	729	20.6	651	779	0.86
Reading Score		**527**	**41**	**7.3**	**17**	**59**	**0.83**
Gender	Female	272	42	7.3	17	59	0.84
	Male	255	40	7.2	17	59	0.82
Ethnicity	American Indian	3	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	1	-	-	-	-	-
	Hispanic/Latino	518	41	7.3	17	59	0.84
	White	5	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	0	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	28	41	7.1	26	53	0.84
	Economically Disadvantaged	499	41	7.3	17	59	0.83
Students with Disabilities	504	0	-	-	-	-	-
	IEP	47	35	4.9	21	52	0.56
	No	114	41	7.3	17	58	0.82
	Missing	366	42	7.2	17	59	0.83
Writing Score		**527**	**29**	**8.2**	**10**	**51**	**0.73**
Gender	Female	272	30	7.3	10	51	0.71
	Male	255	27	8.6	10	49	0.74
Ethnicity	American Indian	3	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	1	-	-	-	-	-
	Hispanic/Latino	518	29	8.2	10	51	0.73
	White	5	-	-	-	-	-

Group Type	Subgroup	N	Mean	SD	Min	Max	Alpha
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	0	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	28	26	9.9	10	40	0.80
	Economically Disadvantaged	499	29	8.1	10	51	0.73
Students with Disabilities	504	0	-	-	-	-	-
	IEP	47	23	8.7	10	38	0.74
	No	114	27	8.8	10	41	0.76
	Missing	366	30	7.6	10	51	0.71

Table 3. SEMs

Grade	Total Score SEM	Reading SEM	Writing SEM
3	7.6	2.9	4.7
4	7.9	3.0	4.2

Table 4. Subclaim Performance Summary

Grade	Subclaim	Max Possible Score	Mean	SD	Min	Max	Alpha
3	Reading Literary	25	9	5.7	0	25	0.80
	Reading Informational	21	7	4.7	0	21	0.74
	Reading Vocabulary	12	5	3.0	0	12	0.63
	Written Expression	27	7	5.3	0	27	0.60
	Writing Knowledge and Language Conventions	9	3	2.1	0	9	0.77
4	Reading Literary	26	9	4.9	0	23	0.74
	Reading Informational	26	6	3.9	0	21	0.59
	Reading Vocabulary	12	4	2.7	0	12	0.53
	Written Expression	33	9	6.8	0	27	0.65
	Writing Knowledge and Language Conventions	9	3	2.2	0	9	0.78

Table 5. Grade 3 Classical Statistics

ITEM	TYPE	% 0	% 1	% 2	% 3	% Omit	MEAN SCORE	ITEM-TOTAL CORR
1	EBSR	38.2	9.4	52.2		0.3	1.138	0.514
2	EBSR	41.7	42.1	15.7		0.5	0.734	0.427
3	EBSR	42.1	14.8	42.7		0.4	1.003	0.559
4	EBSR	37.3	22.0	39.9		0.8	1.018	0.506
5	EBSR	69.8	10.0	19.4		0.9	0.487	0.373
6	EBSR	44.0	20.9	34.2		0.9	0.893	0.479
7	CR	46.2	31.1	19.6	1.3	1.9	0.742	0.589
8	CR	28.7	39.9	27.3	2.2	1.9	1.013	0.615
9	EBSR	47.2	22.6	26.7		3.5	0.760	0.481
10	EBSR	47.2	13.6	35.3		4.0	0.841	0.614
11	EBSR	50.8	12.2	32.8		4.2	0.778	0.559
12	EBSR	50.3	16.3	29.2		4.2	0.747	0.543
13	EBSR	25.3	34.9	39.3		0.5	1.134	0.510
14	EBSR	39.9	16.5	42.9		0.7	1.023	0.621
15	EBSR	55.7	16.3	27.4		0.6	0.712	0.546
16	EBSR	60.9	21.9	16.6		0.6	0.551	0.279
17	EBSR	46.1	29.6	23.7		0.5	0.771	0.450
18	EBSR	54.0	14.4	31.0		0.6	0.765	0.544
19	CR	38.9	36.0	18.8	4.9	1.4	0.883	0.697
20	CR	27.7	33.7	34.1	3.2	1.4	1.113	0.591
21	EBSR	42.5	23.9	33.5		0.2	0.908	0.494
22	EBSR	37.1	37.6	25.0		0.3	0.876	0.536
23	EBSR	62.4	15.0	22.2		0.4	0.595	0.468
24	EBSR	48.0	28.1	23.5		0.4	0.751	0.427
25	CR	31.2	52.2	14.8	0.6	1.2	0.836	0.577
26	CR	27.4	39.1	30.0	2.2	1.2	1.059	0.573
27	EBSR	39.5	41.0	18.7		0.9	0.783	0.472
28	EBSR	51.3	25.7	21.7		1.3	0.691	0.294

ITEM	TYPE	% 0	% 1	% 2	% 3	% Omit	MEAN SCORE	ITEM-TOTAL CORR
29	EBSR	69.0	12.9	16.9		1.3	0.466	0.344
30	EBSR	46.0	15.3	37.6		1.0	0.906	0.521
31	EBSR	58.5	14.7	25.7		1.1	0.661	0.436
32	EBSR	64.5	17.5	16.9		1.2	0.512	0.316

Table 6. Grade 4 Classical Statistics

ITEM	TYPE	% 0	% 1	% 2	% 3	% 4	% Omit	MEAN SCORE	ITEM-TOTAL CORR
1	EBSR	44.8	22.6	31.7			0.9	0.860	0.445
2	EBSR	47.1	26.9	25.2			0.8	0.774	0.459
3	EBSR	36.8	30.9	31.5			0.8	0.939	0.490
4	EBSR	58.6	8.3	32.1			0.9	0.725	0.549
5	EBSR	72.1	15.9	11.0			0.9	0.380	0.289
6	EBSR	39.3	34.7	25.0			0.9	0.848	0.445
7	CR	28.1	34.5	29.4	5.3	0	2.7	1.093	0.737
8	CR	29.0	31.9	32.6	3.8	0	2.7	1.085	0.603
9	EBSR	79.7	6.3	10.6			3.4	0.275	0.072
10	EBSR	49.9	30.9	15.7			3.4	0.624	0.306
11	EBSR	43.5	32.3	20.5			3.8	0.732	0.400
12	EBSR	67.4	16.7	11.4			4.6	0.395	0.271
13	EBSR	64.9	18.8	15.9			0.4	0.507	0.394
14	EBSR	71.5	9.3	18.4			0.8	0.461	0.342
15	EBSR	51.2	26.2	21.8			0.8	0.698	0.296
16	EBSR	50.7	22.6	25.8			0.9	0.742	0.454
17	EBSR	60.3	20.7	18.6			0.4	0.579	0.349
18	EBSR	69.4	18.2	12.0			0.4	0.421	0.223
19	EBSR	50.5	22.4	26.8			0.4	0.759	0.472
20	EBSR	72.9	17.6	9.1			0.4	0.359	0.165

ITEM	TYPE	% 0	% 1	% 2	% 3	% 4	% Omit	MEAN SCORE	ITEM-TOTAL CORR
21	CR	47.2	29.2	18.0	4.7	0	0.8	0.795	0.636
22	CR	32.1	29.0	35.5	2.7	0	0.8	1.080	0.616
23	EBSR	35.7	17.8	46.1			0.4	1.101	0.536
24	EBSR	45.0	19.0	35.7			0.4	0.903	0.390
25	EBSR	49.9	29.6	20.1			0.4	0.698	0.179
26	EBSR	40.4	24.1	35.5			0.0	0.951	0.568
27	CR	37.6	18.6	26.9	15.9	0	0.9	1.203	0.704
28	CR	21.6	28.8	41.6	7.0	0	0.9	1.330	0.599
29	EBSR	60.7	22.2	16.1			0.9	0.545	0.413
30	EBSR	67.7	15.7	15.4			1.1	0.465	0.210
31	EBSR	57.1	17.5	23.7			1.7	0.649	0.284
32	EBSR	66.2	22.6	9.3			1.9	0.412	0.346
33	EBSR	69.4	15.0	14.0			1.5	0.431	0.252
34	EBSR	59.6	23.1	16.1			1.1	0.554	0.326

Table 7. Grade 3 Item Parameter Estimates

ITEM	TYPE	B	D1	D2	D3	D4	INFIT	OUTFIT
1	EBSR	-0.7705	0	1.3300	-1.3300		0.99	0.95
2	EBSR	0.0906	0	-0.7515	0.7515		1.02	1.02
3	EBSR	-0.5561	0	0.8237	-0.8237		0.93	0.89
4	EBSR	-0.5746	0	0.3322	-0.3322		1.00	0.99
5	EBSR	0.3508	0	1.0613	-1.0613		1.14	1.43
6	EBSR	-0.3640	0	0.3838	-0.3838		1.04	1.31
7	EBSR	-0.1168	0	0.2359	-0.2359		1.00	1.00
8	EBSR	-0.2850	0	0.9059	-0.9059		0.85	0.80
9	EBSR	-0.1933	0	1.0029	-1.0029		0.92	0.92
10	EBSR	-0.1151	0	0.6549	-0.6549		0.94	0.91
11	EBSR	-0.8168	0	-0.3400	0.3400		0.93	0.92
12	EBSR	-0.5867	0	0.6919	-0.6919		0.80	0.75
13	EBSR	-0.0485	0	0.6277	-0.6277		0.92	0.90
14	EBSR	0.3229	0	0.1189	-0.1189		1.23	1.37
15	EBSR	-0.1010	0	-0.1372	0.1372		1.05	1.10
16	EBSR	-0.1605	0	0.8103	-0.8103		0.95	0.98
17	EBSR	-0.3812	0	0.2267	-0.2267		1.00	1.01
18	EBSR	-0.2812	0	-0.4568	0.4568		0.91	0.92
19	EBSR	0.1757	0	0.6654	-0.6654		1.02	1.08
20	EBSR	-0.0611	0	-0.0725	0.0725		1.07	1.09
21	EBSR	-0.0496	0	-0.6647	0.6647		0.97	0.98
22	EBSR	0.0421	0	0.0228	-0.0228		1.25	1.33
23	EBSR	0.4364	0	0.7369	-0.7369		1.16	1.37
24	EBSR	-0.3893	0	0.7830	-0.7830		0.98	0.99
25	EBSR	0.0371	0	0.7356	-0.7356		1.11	1.18
26	EBSR	0.3778	0	0.3868	-0.3868		1.20	1.35
27	CR	1.0080	0	-1.2804	-0.7347	2.0151	1.04	1.00
28	CR	0.4441	0	-1.0648	-0.0554	1.1202	0.91	0.89

ITEM	TYPE	B	D1	D2	D3	D4	INFIT	OUTFIT
29	CR	1.1312	0	-2.3586	-0.0945	2.4531	0.92	0.91
30	CR	0.5540	0	-1.6385	-0.5060	2.1445	0.91	0.89
31	CR	0.3699	0	-1.3685	-0.7700	2.1385	0.94	0.93
32	CR	0.5107	0	-1.6573	-0.6031	2.2604	0.93	0.91

Table 8. Grade 4 Item Parameter Estimates

ITEM	TYPE	B	D1	D2	D3	D4	D5	INFIT	OUTFIT
1	EBSR	-0.4769	0	0.3637	-0.3637			0.94	0.96
2	EBSR	-0.3241	0	0.1086	-0.1086			0.95	0.93
3	EBSR	-0.6048	0	-0.0390	0.0390			0.90	0.89
4	EBSR	-0.328	0	1.5702	-1.5702			0.84	0.82
5	EBSR	0.4510	0	0.4403	-0.4403			1.04	1.24
6	EBSR	-0.4277	0	-0.2772	0.2772			0.97	0.98
7	EBSR	0.5687	0	1.4197	-1.4197			1.29	1.72
8	EBSR	0.0246	0	-0.2140	0.2140			1.07	1.06
9	EBSR	-0.2051	0	-0.2037	0.2037			0.99	0.98
10	EBSR	0.4387	0	0.3590	-0.3590			1.09	1.18
11	EBSR	0.1537	0	0.3764	-0.3764			0.98	0.99
12	EBSR	0.1504	0	1.1685	-1.1685			1.08	1.18
13	EBSR	-0.1743	0	0.0735	-0.0735			1.12	1.08
14	EBSR	-0.2744	0	0.3309	-0.3309			0.94	0.94
15	EBSR	0.0126	0	0.3194	-0.3194			1.06	1.15
16	EBSR	0.3902	0	0.2923	-0.2923			1.13	1.18
17	EBSR	-0.3301	0	0.3118	-0.3118			0.96	0.99
18	EBSR	0.5356	0	0.1951	-0.1951			1.15	1.38
19	EBSR	-0.8693	0	0.6812	-0.6812			0.85	0.82
20	EBSR	-0.5510	0	0.6280	-0.6280			1.03	1.02
21	EBSR	-0.1550	0	-0.0719	0.0719			1.23	1.37

ITEM	TYPE	B	D1	D2	D3	D4	D5	INFIT	OUTFIT
22	EBSR	-0.6247	0	0.2803	-0.2803			0.82	0.79
23	EBSR	0.0942	0	0.2213	-0.2213			0.99	0.98
24	EBSR	0.2217	0	0.5489	-0.5489			1.19	1.26
25	EBSR	-0.1400	0	0.5859	-0.5859			1.14	1.19
26	EBSR	0.4829	0	-0.0098	0.0098			0.98	1.05
27	EBSR	0.3039	0	0.5475	-0.5475			1.11	1.33
28	EBSR	0.1006	0	0.1604	-0.1604			1.05	1.14
29	CR	0.8175	0	-1.9137	-1.2490	0.6524	2.5103	0.77	0.76
30	CR	1.0247	0	-1.3408	-1.0827	0.1330	2.2904	0.95	0.89
31	CR	-0.3505	0	0.1007	-0.6091	0.5084		0.84	0.84
32	CR	0.1263	0	-1.0661	-0.7611	1.8272		0.87	0.85
33	CR	0.2222	0	-1.0080	-1.0273	2.0352		0.86	0.86
34	CR	-0.2838	0	-0.9948	-0.7638	1.7586		0.85	0.86

Table 9. Cut Scores and Students in Each Performance Level

	Cut Scores				Performance Levels											
Grade	Level 2	Level 3	Level 4	Level 5	Level 1		Level 2		Level 3		Level 4		Level 5		Levels 4 and 5 Combined	
					N	%	N	%	N	%	N	%	N	%	N	%
3	11	26	48	70	134	9	472	31	572	38	299	20	35	2	334	22
4	12	28	53	73	51	10	181	34	222	42	66	13	7	1	73	14

Table 10. Scale Score Ranges for Each Performance Level

Grade	Level 1	Level 2	Level 3	Level 4	Level 5
3	650–699	700–724	725–749	750–778	779–850
4	650–699	700–724	725–749	750–771	772–850

Table 11. Students in Each Subclaim Performance Category

Grade	Subclaims	Category	N	%	Grade	Subclaims	Category	N	%
3	Reading Literary	1	327	22	4	Reading Literary	1	92	17
		2	508	34			2	180	34
		3	677	45			3	255	48
	Reading Informational	1	322	21		Reading Informational	1	61	12
		2	545	36			2	266	50
		3	645	43			3	200	38
	Reading Vocabulary	1	394	26		Reading Vocabulary	1	93	18
		2	484	32			2	163	31
		3	634	42			3	271	51
	Written Expression	1	417	28		Written Expression	1	89	17
		2	255	17			2	205	39
		3	840	56			3	233	44
	Writing Knowledge and Language Conventions	1	439	29		Writing Knowledge and Language Conventions	1	123	23
		2	487	32			2	143	27
		3	586	39			3	261	50

Note. Category 1=Met or Exceeded Expectations; Category 2=Approached Expectations; Category 3=Did Not Yet Meet or Partially Met Expectations.

Table 12. Grade 3 Scale Score Frequency Distributions

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
650	1	0.07	1	0.07
653	1	0.07	2	0.13
664	3	0.20	5	0.33
672	5	0.33	10	0.66
678	10	0.66	20	1.32
683	21	1.39	41	2.71
687	23	1.52	64	4.23
691	19	1.26	83	5.49
694	24	1.59	107	7.08
697	27	1.79	134	8.86
700	26	1.72	160	10.58
702	25	1.65	185	12.24
705	36	2.38	221	14.62
707	34	2.25	255	16.87
709	40	2.65	295	19.51
711	27	1.79	322	21.30
712	29	1.92	351	23.21
714	40	2.65	391	25.86
716	28	1.85	419	27.71
717	31	2.05	450	29.76
719	23	1.52	473	31.28
720	34	2.25	507	33.53
722	39	2.58	546	36.11
723	38	2.51	584	38.62
724	22	1.46	606	40.08
725	36	2.38	642	42.46
727	37	2.45	679	44.91
728	26	1.72	705	46.63
729	32	2.12	737	48.74
730	20	1.32	757	50.07
732	19	1.26	776	51.32
733	34	2.25	810	53.57
734	35	2.31	845	55.89
735	21	1.39	866	57.28
736	24	1.59	890	58.86
737	30	1.98	920	60.85
738	20	1.32	940	62.17
739	25	1.65	965	63.82
740	16	1.06	981	64.88
741	31	2.05	1012	66.93
742	28	1.85	1040	68.78
744	29	1.92	1069	70.70
745	24	1.59	1093	72.29
746	27	1.79	1120	74.07

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
747	21	1.39	1141	75.46
748	23	1.52	1164	76.98
749	14	0.93	1178	77.91
750	25	1.65	1203	79.56
751	17	1.12	1220	80.69
752	19	1.26	1239	81.94
753	15	0.99	1254	82.94
754	22	1.46	1276	84.39
756	16	1.06	1292	85.45
757	16	1.06	1308	86.51
758	11	0.73	1319	87.24
759	11	0.73	1330	87.96
760	19	1.26	1349	89.22
762	18	1.19	1367	90.41
763	17	1.12	1384	91.53
764	15	0.99	1399	92.53
765	16	1.06	1415	93.58
767	14	0.93	1429	94.51
768	9	0.60	1438	95.11
770	5	0.33	1443	95.44
771	11	0.73	1454	96.16
773	6	0.40	1460	96.56
774	9	0.60	1469	97.16
776	8	0.53	1477	97.69
779	1	0.07	1478	97.75
781	5	0.33	1483	98.08
783	7	0.46	1490	98.54
785	6	0.40	1496	98.94
787	3	0.20	1499	99.14
790	3	0.20	1502	99.34
792	4	0.26	1506	99.60
795	1	0.07	1507	99.67
797	1	0.07	1508	99.74
800	1	0.07	1509	99.80
804	1	0.07	1510	99.87
824	1	0.07	1511	99.93
850	1	0.07	1512	100.00

Table 13. Grade 3 Reading Scale Score Frequency Distributions

Reading Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10	1	0.07	1	0.07
17	3	0.20	4	0.26
21	5	0.33	9	0.60
24	7	0.46	16	1.06
27	29	1.92	45	2.98
28	36	2.38	81	5.36
30	43	2.84	124	8.20
31	49	3.24	173	11.44
33	55	3.64	228	15.08
34	67	4.43	295	19.51
35	66	4.37	361	23.88
36	104	6.88	465	30.75
37	56	3.70	521	34.46
38	45	2.98	566	37.43
39	54	3.57	620	41.01
40	89	5.89	709	46.89
41	43	2.84	752	49.74
42	84	5.56	836	55.29
43	56	3.70	892	58.99
44	30	1.98	922	60.98
45	64	4.23	986	65.21
46	57	3.77	1043	68.98
47	26	1.72	1069	70.70
48	64	4.23	1133	74.93
49	70	4.63	1203	79.56
50	23	1.52	1226	81.08
51	46	3.04	1272	84.13
52	19	1.26	1291	85.38
53	56	3.70	1347	89.09
54	21	1.39	1368	90.48
55	43	2.84	1411	93.32
56	16	1.06	1427	94.38
57	19	1.26	1446	95.63
58	15	0.99	1461	96.63
59	10	0.66	1471	97.29
60	8	0.53	1479	97.82
61	9	0.60	1488	98.41
62	7	0.46	1495	98.88
64	4	0.26	1499	99.14
65	4	0.26	1503	99.40
67	5	0.33	1508	99.74
69	2	0.13	1510	99.87
75	1	0.07	1511	99.93
80	1	0.07	1512	100.00

Table 14. Grade 3 Writing Scale Score Frequency Distributions

Writing Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10	111	7.34	111	7.34
14	51	3.37	162	10.71
19	36	2.38	198	13.10
21	59	3.90	257	17.00
23	63	4.17	320	21.16
25	61	4.03	381	25.20
26	92	6.08	473	31.28
28	82	5.42	555	36.71
29	69	4.56	624	41.27
30	106	7.01	730	48.28
31	81	5.36	811	53.64
32	154	10.19	965	63.82
33	59	3.90	1024	67.72
34	73	4.83	1097	72.55
35	61	4.03	1158	76.59
36	45	2.98	1203	79.56
37	49	3.24	1252	82.80
38	78	5.16	1330	87.96
39	29	1.92	1359	89.88
40	52	3.44	1411	93.32
41	16	1.06	1427	94.38
42	8	0.53	1435	94.91
44	24	1.59	1459	96.49
45	18	1.19	1477	97.69
46	6	0.40	1483	98.08
47	8	0.53	1491	98.61
49	10	0.66	1501	99.27
50	6	0.40	1507	99.67
52	1	0.07	1508	99.74
56	2	0.13	1510	99.87
58	1	0.07	1511	99.93
60	1	0.07	1512	100.00

Table 15. Grade 4 Scale Score Frequency Distributions

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
651	2	0.38	2	0.38
663	1	0.19	3	0.57
670	2	0.38	5	0.95
676	2	0.38	7	1.33
681	3	0.57	10	1.90
685	6	1.14	16	3.04
689	8	1.52	24	4.55
692	5	0.95	29	5.50
695	11	2.09	40	7.59
698	11	2.09	51	9.68
700	11	2.09	62	11.76
702	15	2.85	77	14.61
704	11	2.09	88	16.70
706	9	1.71	97	18.41
708	7	1.33	104	19.73
710	8	1.52	112	21.25
711	7	1.33	119	22.58
713	12	2.28	131	24.86
714	17	3.23	148	28.08
716	17	3.23	165	31.31
717	19	3.61	184	34.91
719	6	1.14	190	36.05
720	11	2.09	201	38.14
721	12	2.28	213	40.42
722	8	1.52	221	41.94
724	11	2.09	232	44.02
725	10	1.90	242	45.92
726	14	2.66	256	48.58
727	8	1.52	264	50.09
728	11	2.09	275	52.18
729	5	0.95	280	53.13
730	9	1.71	289	54.84
731	16	3.04	305	57.87
732	9	1.71	314	59.58
733	13	2.47	327	62.05
734	10	1.90	337	63.95
735	14	2.66	351	66.60
736	11	2.09	362	68.69
737	3	0.57	365	69.26
738	8	1.52	373	70.78
739	9	1.71	382	72.49
740	6	1.14	388	73.62
741	8	1.52	396	75.14
742	7	1.33	403	76.47

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
743	9	1.71	412	78.18
744	7	1.33	419	79.51
745	9	1.71	428	81.21
746	8	1.52	436	82.73
747	3	0.57	439	83.30
748	8	1.52	447	84.82
749	7	1.33	454	86.15
750	6	1.14	460	87.29
751	8	1.52	468	88.80
752	3	0.57	471	89.37
753	1	0.19	472	89.56
754	3	0.57	475	90.13
755	5	0.95	480	91.08
756	4	0.76	484	91.84
757	7	1.33	491	93.17
758	4	0.76	495	93.93
759	4	0.76	499	94.69
761	4	0.76	503	95.45
762	2	0.38	505	95.83
763	4	0.76	509	96.58
765	4	0.76	513	97.34
766	1	0.19	514	97.53
767	1	0.19	515	97.72
768	1	0.19	516	97.91
769	3	0.57	519	98.48
771	1	0.19	520	98.67
772	3	0.57	523	99.24
773	2	0.38	525	99.62
775	1	0.19	526	99.81
779	1	0.19	527	100.00

Table 16. Grade 4 Reading Scale Score Frequency Distributions

Reading Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
17	3	0.57	3	0.57
21	1	0.19	4	0.76
24	3	0.57	7	1.33
26	6	1.14	13	2.47
28	6	1.14	19	3.61
30	13	2.47	32	6.07
31	12	2.28	44	8.35
33	20	3.80	64	12.14
34	29	5.50	93	17.65
35	32	6.07	125	23.72
36	22	4.17	147	27.89
37	25	4.74	172	32.64
38	21	3.98	193	36.62
39	56	10.63	249	47.25
40	17	3.23	266	50.47
41	36	6.83	302	57.31
42	15	2.85	317	60.15
43	33	6.26	350	66.41
44	15	2.85	365	69.26
45	19	3.61	384	72.87
46	10	1.90	394	74.76
47	29	5.50	423	80.27
48	22	4.17	445	84.44
49	13	2.47	458	86.91
50	6	1.14	464	88.05
51	14	2.66	478	90.70
52	13	2.47	491	93.17
53	10	1.90	501	95.07
54	9	1.71	510	96.77
55	7	1.33	517	98.10
56	1	0.19	518	98.29
57	2	0.38	520	98.67
58	4	0.76	524	99.43
59	3	0.57	527	100.00

Table 17. Grade 4 Writing Scale Score Frequency Distributions

Writing Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10	43	8.16	43	8.16
15	17	3.23	60	11.39
19	12	2.28	72	13.66
21	16	3.04	88	16.70
23	21	3.98	109	20.68
24	13	2.47	122	23.15
25	20	3.80	142	26.94
26	14	2.66	156	29.60
27	18	3.42	174	33.02
28	55	10.44	229	43.45
29	23	4.36	252	47.82
30	43	8.16	295	55.98
31	38	7.21	333	63.19
32	17	3.23	350	66.41
33	36	6.83	386	73.24
34	31	5.88	417	79.13
35	23	4.36	440	83.49
36	18	3.42	458	86.91
37	19	3.61	477	90.51
38	8	1.52	485	92.03
39	2	0.38	487	92.41
40	25	4.74	512	97.15
41	4	0.76	516	97.91
43	2	0.38	518	98.29
44	2	0.38	520	98.67
45	1	0.19	521	98.86
46	1	0.19	522	99.05
48	2	0.38	524	99.43
49	2	0.38	526	99.81
51	1	0.19	527	100.00

Table 19. Grade 3 Scale Scores and Conditional Standard Error of Measurement (CSEM)

Raw Score	Scale Score	CSEM
0	650	15
1	650	15
2	653	15
3	664	15
4	672	14
5	678	12
6	683	11
7	687	10
8	691	10
9	694	9
10	697	9
11	700	8
12	702	8
13	705	8
14	707	8
15	709	7
16	711	7
17	712	7
18	714	7
19	716	7
20	717	7
21	719	6
22	720	6
23	722	6
24	723	6
25	724	6
26	725	6
27	727	6
28	728	6
29	729	6
30	730	6
31	732	6
32	733	6
33	734	6
34	735	6
35	736	6
36	737	6
37	738	6
38	739	5
39	740	5

Raw Score	Scale Score	CSEM
40	741	5
41	742	5
42	744	5
43	745	5
44	746	5
45	747	5
46	748	5
47	749	5
48	750	6
49	751	6
50	752	6
51	753	6
52	754	6
53	756	6
54	757	6
55	758	6
56	759	6
57	760	6
58	762	6
59	763	6
60	764	6
61	765	6
62	767	6
63	768	6
64	770	6
65	771	6
66	773	7
67	774	7
68	776	7
69	778	7
70	779	7
71	781	7
72	783	7
73	785	8
74	787	8
75	790	8
76	792	8
77	795	9
78	797	9
79	800	9
80	804	10
81	807	10

Raw Score	Scale Score	CSEM
82	811	10
83	815	11
84	819	11
85	824	12
86	829	12
87	834	13
88	841	14
89	848	15
90	850	15
91	850	15
92	850	15
93	850	15
94	850	15

Table 20. Grade 3 Reading Scale Scores and Conditional Standard Error of Measurement (CSEM)

Raw Score	Scale Score	CSEM
0	10	8
1	10	8
2	17	8
3	21	6
4	24	5
5	27	5
6	28	4
7	30	4
8	31	4
9	33	4
10	34	3
11	35	3
12	36	3
13	36	3
14	37	3
15	38	3
16	39	3
17	40	3
18	40	3
19	41	3
20	42	3
21	42	3
22	43	3

Raw Score	Scale Score	CSEM
23	43	3
24	44	3
25	45	3
26	45	3
27	46	3
28	46	3
29	47	3
30	48	3
31	48	3
32	49	3
33	49	3
34	50	3
35	51	3
36	51	3
37	52	3
38	53	3
39	53	3
40	54	3
41	55	3
42	55	3
43	56	3
44	57	3
45	58	3
46	59	3
47	60	3
48	61	4
49	62	4
50	64	4
51	65	4
52	67	5
53	69	5
54	72	6
55	75	7
56	80	8
57	89	12
58	90	12

Table 21. Grade 3 Writing Scale Scores and Conditional Standard Error of Measurement (CSEM)

Raw Score	Scale Score	CSEM
0	10	6
1	14	6
2	19	4
3	21	4
4	23	3
5	25	3
6	26	3
7	28	3
8	29	3
9	30	2
10	31	2
11	32	2
12	32	2
13	33	2
14	34	2
15	35	2
16	36	2
17	37	2
18	38	2
19	38	2
20	39	2
21	40	2
22	41	2
23	42	3
24	44	3
25	45	3
26	46	3
27	47	3
28	49	3
29	50	3
30	52	3
31	54	3
32	56	4
33	58	4
34	60	4
35	60	4
36	60	4

Table 22. Grade 4 Scale Scores and Conditional Standard Error of Measurement (CSEM)

Raw Score	Scale Score	CSEM
0	650	15
1	650	15
2	651	15
3	663	15
4	670	14
5	676	13
6	681	11
7	685	11
8	689	10
9	692	9
10	695	9
11	698	8
12	700	8
13	702	8
14	704	8
15	706	7
16	708	7
17	710	7
18	711	7
19	713	7
20	714	7
21	716	6
22	717	6
23	719	6
24	720	6
25	721	6
26	722	6
27	724	6
28	725	6
29	726	6
30	727	6
31	728	6
32	729	6
33	730	6
34	731	6
35	732	5
36	733	5
37	734	5
38	735	5
39	736	5

Raw Score	Scale Score	CSEM
40	737	5
41	738	5
42	739	5
43	740	5
44	741	5
45	742	5
46	743	5
47	744	5
48	745	5
49	746	5
50	747	5
51	748	5
52	749	5
53	750	5
54	751	5
55	752	5
56	753	5
57	754	5
58	755	5
59	756	5
60	757	5
61	758	6
62	759	6
63	760	6
64	761	6
65	762	6
66	763	6
67	765	6
68	766	6
69	767	6
70	768	6
71	769	6
72	771	6
73	772	6
74	773	6
75	775	6
76	776	6
77	778	7
78	779	7
79	781	7
80	782	7
81	784	7

Raw Score	Scale Score	CSEM
82	786	7
83	788	7
84	790	8
85	792	8
86	794	8
87	796	8
88	799	9
89	801	9
90	804	9
91	807	9
92	810	10
93	813	10
94	817	11
95	821	11
96	825	11
97	830	12
98	835	12
99	841	13
100	847	14
101	850	15
102	850	15
103	850	15
104	850	15
105	850	15
106	850	15

Table 23. Grade 4 Reading Scale Scores and Conditional Standard Error of Measurement (CSEM)

Raw Score	Scale Score	CSEM
0	10	8
1	10	8
2	17	8
3	21	6
4	24	6
5	26	5
6	28	5
7	30	4
8	31	4
9	33	4
10	34	4

Raw Score	Scale Score	CSEM
11	35	3
12	36	3
13	37	3
14	38	3
15	39	3
16	39	3
17	40	3
18	41	3
19	41	3
20	42	3
21	43	3
22	43	3
23	44	3
24	45	3
25	45	3
26	46	3
27	47	3
28	47	3
29	48	3
30	48	3
31	49	3
32	49	3
33	50	3
34	51	3
35	51	3
36	52	3
37	52	3
38	53	3
39	54	3
40	54	3
41	55	3
42	55	3
43	56	3
44	57	3
45	58	3
46	58	3
47	59	3
48	60	3
49	61	3
50	62	3
51	63	3
52	64	4

Raw Score	Scale Score	CSEM
53	65	4
54	66	4
55	67	4
56	69	4
57	71	5
58	73	5
59	75	6
60	79	7
61	83	8
62	89	9
63	90	9
64	90	9

Table 24. Grade 4 Writing Scale Scores and Conditional Standard Error of Measurement (CSEM)

Raw Score	Scale Score	CSEM
0	10	6
1	15	6
2	19	4
3	21	3
4	23	3
5	24	3
6	25	2
7	26	2
8	27	2
9	28	2
10	28	2
11	29	2
12	30	2
13	30	2
14	31	2
15	31	2
16	32	2
17	33	2
18	33	2
19	34	2
20	34	2
21	35	2
22	36	2
23	36	2

Raw Score	Scale Score	CSEM
24	37	2
25	38	2
26	39	2
27	40	2
28	40	2
29	41	2
30	43	3
31	44	3
32	45	3
33	46	3
34	48	3
35	49	3
36	51	3
37	52	3
38	54	4
39	57	4
40	60	4
41	60	4
42	60	4

Table 25. Correlations between Claims and Subclaims

Grade		RD	RL	RI	RV	WR	WE	WKLC		Grade		RD	RL	RI	RV	WR	WE	WKLC
3	RD	1	0.94	0.91	0.84	0.67	0.65	0.56		4	RD	1	0.90	0.84	0.79	0.69	0.69	0.53
	RL		1	0.76	0.70	0.64	0.62	0.52			RL		1	0.60	0.60	0.64	0.64	0.49
	RI			1	0.67	0.62	0.60	0.53			RI			1	0.54	0.59	0.59	0.47
	RV				1	0.51	0.49	0.44			RV				1	0.50	0.50	0.38
	WR					1	0.98	0.83			WR					1	0.98	0.83
	WE						1	0.69			WE						1	0.71
	WKLC							1			WKLC							1

Note: RD=Reading, RL=Reading Literary, RI=Reading Informational, RV=Reading Vocabulary, WR=Writing, WE=Written Expression, and WKLC=Writing Knowledge of Language and Conventions

Table 26. Classification Accuracy and Consistency

	Accuracy	Consistency		
Grade	Prob of Accurate Classification (PA)	Prob of Consistent Classification (PC)	Prob of Consistent Classification by Chance (Chance)	Kappa
3	0.76	0.66	0.29	0.52
4	0.74	0.64	0.32	0.47

Table 27. Spring 2016 Rater Agreement Statistics

Grade	Item	N	Exact	Adjacent	Non-Adjacent
3	1	152	78.9	19.8	1.4
	2	152	84.9	13.2	2.0
	3	152	80.9	18.4	0.6
	4	152	77.0	21.8	1.4
	5	152	77.6	20.4	2.0
	6	152	82.9	15.2	2.0
4	1	53	83.0	15.0	1.8
	2	53	81.1	18.8	0.0
	3	53	75.5	22.6	1.8
	4	53	79.2	20.8	0.0
	5	53	96.2	3.8	0.0
	6	53	88.7	11.4	0.0

Table 28. Spring 2016 Items Field Tested and Data Review Outcomes

	Grade 3	Grade 4
Number of test forms	2	2
Number of items field tested	24	24
Number of items flagged and reviewed	14	15
Number of accepted items	11	14

APPENDICES

APPENDIX A: CSLA ELIGIBILITY FLOWCHART

Colorado Spanish Language Arts Decision Making Flowchart

Grade 3 and 4

NEP

Non-English Proficient

LEP

Limited-English Proficient

FEP

Fully English Proficient M1, M2 or Exited

FELL

Former English Language Learner

Has the student received instruction in Spanish language arts within the last nine months? *

Yes

No

Note: 5 years does not include Pre-School or Kindergarten.

How long has the student had instruction in an English Language Development program?

5 Years or less

More than 5 years

Eligible for the Spanish Language Arts Assessment (or locally translated CoAlt: DLM ELA)

Districts must determine if the grades 3/4 Spanish Language Arts Assessment is the best choice for the student or if the CMAS: PARCC ELA with linguistic accommodations or translated CoAlt: DLM ELA will better help the student access the content on the assessment.

Student MUST participate in CMAS: PARCC ELA or CoAlt: DLM ELA
(with or without accommodations)

Standard and Linguistic accommodations are available on CMAS: PARCC and CoAlt: DLM to provide access to content on the assessment. Appropriate accommodations should be used.

* District assessment leadership should collaborate with EL staff to evaluate appropriateness and eligibility of students to take CSLA.

September 2016

APPENDIX B: CSLA TEST BLUEPRINTS

Grade 3 Blueprint

Unit	Task/Item Set	Passages	Claims/Sub-Claims	Item Types		CR Points
				EBSR Items (Points)	CR Items	
Unit 1	Literary Analysis Task	2	Reading Literary Text	4 (8)	1	3
			Reading Vocabulary	2 (4)		0
			Written Expression	0		9
			Writing Knowledge of Language and Conventions	0		3
	Literary short passage set	1	Reading Literary Text	3 (6)		N/A
			Reading Vocabulary	1 (2)		
Unit 2	Research Simulation Task	2	Reading Informational Text	4 (8)	1	3
			Reading Vocabulary	2 (4)		0
			Written Expression	0		9
			Writing Knowledge of Language and Conventions	0		3
Unit 3	Narrative Writing Task	1	Reading Literary Text	4 (8)	1	0
			Written Expression	0		9
			Writing Knowledge of Language and Conventions	0		3
	Informational long passage set	1	Reading Informational Text	5 (10)		N/A
			Reading Vocabulary	1 (2)		
	Totals			52 Reading		6 Reading 32 Writing

Grade 4 Blueprint

Unit	Task/Item Set	Passages	Claims/Sub-Claims	Item Types		CR Points
				EBSR Items (Points)	CR Items	
Unit 1	Literary Analysis Task	2	Reading Literary Text	4 (8)	1	4
			Reading Vocabulary	2 (4)		0
			Written Expression	0		12
			Writing Knowledge of Language and Conventions	0		3
	Literary short passage set	1	Reading Literary Text	3 (6)		N/A
			Reading Vocabulary	1 (2)		
Unit 2	Research Simulation Task	3	Reading Informational Text	6 (12)	1	4
			Reading Vocabulary	2 (4)		0
			Written Expression	0		12
			Writing Knowledge of Language and Conventions	0		3
Unit 3	Narrative Writing Task	1	Reading Literary Text	4 (8)	1	0
			Written Expression	0		9
			Writing Knowledge of Language and Conventions	0		3
	Informational long or paired passage set	1 or 2	Reading Informational Text	5 (10)		N/A
			Reading Vocabulary	1 (2)		
	Totals			56 Reading		8 Reading 42 Writing

APPENDIX C: CSLA CONSTRUCTED RESPONSE RUBRICS

GRADE 3 (August 2015)
SCORING RUBRIC FOR PROSE CONSTRUCTED RESPONSE ITEMS

Research Simulation Task (RST) and Literary Analysis Task (LAT)

Construct Measured	Score Point 3	Score Point 2	Score Point 1	Score Point 0
Reading Comprehension and Written Expression	The student response • demonstrates **full comprehension** by providing an **accurate** explanation/description/ comparison; • addresses the prompt and provides **effective** development of the topic that is **consistently appropriate** to task, purpose, and audience; • uses **clear** reasoning supported by **relevant**, text-based evidence in the development of the topic; • is **effectively** organized with **clear and coherent** writing**;** • uses language **effectively** to clarify ideas.	The student response • demonstrates **comprehensio**n by providing a **mostly accurate** explanation/ description/comparison; • addresses the prompt and provides **some** development of the topic that is **generally** appropriate to task, purpose, and audience; • uses reasoning and **relevant**, text-based evidence in the development of the topic; • is organized with **mostly clear and coherent** writing**;** • uses language in a way that is **mostly effective** to clarify ideas.	The student response • demonstrates **limited comprehension**; • addresses the prompt and provides **minimal** development of the topic that is **limited in its appropriateness** to task, purpose, and audience • uses **limited** reasoning and text-based evidence; • demonstrates **limited** organization and coherence; • uses language to express ideas with **limited** clarity.	The student response • does not demonstrate comprehension; • is **undeveloped and/or inappropriate** to the task, purpose, and audience; • includes ***little to no*** text-based evidence; • **lacks** organization and coherence; • **does not** use language to express ideas with clarity.
Knowledge of Language and Conventions	The student response to the prompt demonstrates **full command** of the conventions of standard Spanish at an appropriate level of complexity. There **may** be a **few minor errors** in mechanics, grammar, and usage, but **meaning is clea**r.	The student response to the prompt demonstrates **some command** of the conventions of standard Spanish at an appropriate level of complexity. There **may** be errors in mechanics, grammar, and usage that **occasionally impede understanding**, but the **meaning is generally clear**.	The student response to the prompt demonstrates **limited command** of the conventions of standard Spanish at an appropriate level of complexity. There **may** be errors in mechanics, grammar, and usage that **often impede understanding.**	The student response to the prompt **does not demonstrate command** of the conventions of standard Spanish at the appropriate level of complexity. **Frequent and varied errors** in mechanics, grammar, and usage **impede understanding**.

GRADE 3 (August 2015)
SCORING RUBRIC FOR PROSE CONSTRUCTED RESPONSE ITEMS

Narrative Task (NT)

Construct Measured	Score Point 3	Score Point 2	Score Point 1	Score Point 0
Written Expression	The student response • **is effectively** developed with narrative elements and is **consistently appropriate** to the task; • is ***effectively*** organized with ***clear and coherent*** writing • uses language **effectively** to clarify ideas.	The student response • is developed with **some** narrative elements and is **generally appropriate** to the task; • is organized with ***mostly coherent*** writing; • uses language in a way that is ***mostly effective*** to clarify ideas.	The student response • is **minimally** developed with **few** narrative elements and is **limited in its appropriateness** to the task; • demonstrates ***limited*** organization and coherence; • uses language to express ideas with **limited** clarity.	The student response • is **undeveloped** and/or **inappropriate** to the task**;** • **lacks** organization and coherence; • **does not** use language to express ideas with clarity.
Knowledge of Language and Conventions	The student response to the prompt demonstrates **full command** of the conventions of standard Spanish at an appropriate level of complexity. There may be a **few minor errors** in mechanics, grammar, and usage, but **meaning is clea**r.	The student response to the prompt demonstrates **some command** of the conventions of standard Spanish at an appropriate level of complexity. There **may** be errors in mechanics, grammar, and usage that **occasionally impede understanding**, but the **meaning is generally clear**.	The student response to the prompt demonstrates **limited command** of the conventions of standard Spanish at an appropriate level of complexity. There **may** be errors in mechanics, grammar, and usage that **often impede understanding.**	The student response to the prompt **does not demonstrate command** of the conventions of standard Spanish at the appropriate level of complexity. **Frequent and varied errors** in mechanics, grammar, and usage **impede understanding**.

NOTE:

- The reading dimension is not scored for elicited narrative stories.
- Per the CCSS, narrative elements in grades 3-5 may include: establishing a situation, organizing a logical event sequence, describing scenes, objects or people, developing characters' personalities, and using dialogue as appropriate.
- The elements of organization to be assessed are expressed in the grade-level standards W1-W3.

A response is considered unscoreable if it cannot be assigned a score based on the rubric criteria. For unscoreable student responses, one of the following condition codes will be applied.

Coded Responses:

A=No response
B=Response is unintelligible or undecipherable
C=Response is not written in Spanish
D=Off-topic
E=Refusal to respond
F=Don't understand/know

GRADE 4 (August 2015)
SCORING RUBRIC FOR PROSE CONSTRUCTED RESPONSE ITEMS

Research Simulation Task and Literary Analysis Task

Construct Measured	Score Point 4	Score Point 3	Score Point 2	Score Point 1	Score Point 0
Reading Comprehension and Written Expression	The student response • demonstrates ***full comprehension*** of ideas stated explicitly and/or inferentially by providing an ***accurate*** analysis; • addresses the prompt and provides ***effective*** development of the topic that is ***consistently appropriate*** to task, purpose, and audience; • uses ***clear*** reasoning supported by ***relevant***, text-based evidence in the development of the topic; • is ***effectively*** organized with ***clear and coherent*** writing; • uses language ***effectively*** to clarify ideas.	The student response • demonstrates ***comprehension*** of ideas stated explicitly and/or inferentially by providing a ***mostly accurate*** analysis; • addresses the prompt and provides ***mostly effective*** development of the topic that is ***appropriate*** to task, purpose, and audience; • uses **mostly clear** reasoning supported by ***relevant*** text-based evidence in the development of the topic; • is organized with ***mostly clear and coherent*** writing • uses language that is ***mostly effective*** to clarify ideas.	The student response • demonstrates ***basic comprehension*** of ideas stated explicitly and/or inferentially by providing a ***generally accurate*** analysis; • addresses the prompt and provides ***some*** development of the topic that is ***somewhat appropriate*** to task, purpose, and audience; • uses ***some*** reasoning and text-based evidence in the development of the topic; • demonstrates ***some*** organization with ***somewhat*** coherent writing; • uses language to express ideas with ***some*** clarity.	The student response • demonstrates ***limited comprehension*** of ideas by providing a ***minimally accurate*** analysis; • addresses the prompt and provides ***minimal*** development of the topic that is ***limited in its appropriateness*** to task, purpose, and audience • uses ***limited*** reasoning and text-based evidence; • demonstrates ***limited*** organization and coherence; • uses language to express ideas with ***limited*** clarity.	The student response • demonstrates ***no comprehension*** of ideas by providing an ***inaccurate or no*** analysis. • is ***undeveloped and/or inappropriate*** to the task, purpose, and audience; • includes ***little to no*** text-based evidence; • ***lacks*** organization and coherence; • ***does not*** use language to express ideas with clarity.
Knowledge of Language and Conventions		The student response to the prompt demonstrates **full command** of the conventions of standard Spanish at an appropriate level of complexity. There may be a **few minor errors** in mechanics, grammar, and usage, but **meaning is clea**r.	The student response to the prompt demonstrates **some command** of the conventions of standard Spanish at an appropriate level of complexity. There **may** be errors in mechanics, grammar, and usage that **occasionally impede understanding**, but the **meaning is generally clear**.	The student response to the prompt demonstrates **limited command** of the conventions of standard Spanish at an appropriate level of complexity. There **may** be errors in mechanics, grammar, and usage that **often impede understanding.**	The student response to the prompt **does not demonstrate command** of the conventions of standard Spanish at the appropriate level of complexity. **Frequent and varied errors** in mechanics, grammar, and usage **impede understanding**.

GRADE 4 (August 2015)
SCORING RUBRIC FOR PROSE CONSTRUCTED RESPONSE ITEMS

Narrative Task (NT)

Construct Measured	Score Point 3	Score Point 2	Score Point 1	Score Point 0
Written Expression	The student response • is ***effectively*** developed with narrative elements and is ***consistently appropriate*** to the task; • is ***effectively*** organized with ***clear and coherent*** writing • uses language **effectively** to clarify ideas.	The student response • is developed with ***some*** narrative elements and is ***generally appropriate*** to the task; • is organized with ***mostly coherent*** writing; • uses language that is ***mostly effective*** to clarify ideas.	The student response • is ***minimally*** developed with ***few*** narrative elements and is ***limited in its appropriateness*** to the task; • demonstrates ***limited*** organization and coherence; • uses language to express ideas with ***limited*** clarity.	The student response • is ***undeveloped*** and/or ***inappropriate*** to the task; • **lacks** organization and coherence; • ***does not*** use language to express ideas with clarity.
Knowledge of Language and Conventions	The student response to the prompt demonstrates **full command** of the conventions of standard Spanish at an appropriate level of complexity. There may be a **few minor errors** in mechanics, grammar, and usage, but **meaning is clea**r.	The student response to the prompt demonstrates **some command** of the conventions of standard Spanish at an appropriate level of complexity. There **may** be errors in mechanics, grammar, and usage that **occasionally impede understanding**, but the **meaning is generally clear**.	The student response to the prompt demonstrates **limited command** of the conventions of standard Spanish at an appropriate level of complexity. There **may** be errors in mechanics, grammar, and usage that **often impede understanding.**	The student response to the prompt **does not demonstrate command** of the conventions of standard Spanish at the appropriate level of complexity. **Frequent and varied errors** in mechanics, grammar, and usage **impede understanding**.

NOTE:

- The reading dimension is not scored for elicited narrative stories.
- Per the CCSS, narrative elements in grades 3-5 may include: establishing a situation, organizing a logical event sequence, describing scenes, objects or people, developing characters' personalities, and using dialogue as appropriate.
- The elements of organization to be assessed are expressed in the grade-level standards W1-W3.

A response is considered unscoreable if it cannot be assigned a score based on the rubric criteria. For unscoreable student responses, one of the following condition codes will be applied.

Coded Responses:

A=No response
B=Response is unintelligible or undecipherable
C=Response is not written in Spanish
D=Off-topic
E=Refusal to respond
F=Don't understand/know

APPENDIX D: CSLA STANDARD SETTING REPORT

Colorado Spanish Language Arts (CSLA) Standard Setting Report

&

July 22, 2016

Table of Contents

OVERVIEW **4**

PREPARATION FOR STANDARD SETTING **5**

PANELIST SELECTION AND COMPOSITION **5**

DEVELOPMENT OF PLDS **5**

CREATION OF MATERIALS **6**

Agendas *6*

Slides *6*

CSLA Test Book *6*

Description of the Item Scoring *6*

Sample Rubric Responses *7*

Item Mean and Score Point Distribution Reports *7*

External Data *7*

Forms *7*

FACILITATOR AND DATA ANALYST TRAINING **7**

STANDARD SETTING MEETING ACTIVITIES **8**

INTRODUCTORY SESSION **8**

THE STANDARD SETTING PROCESS **8**

Experience the Assessment *9*

Review and Discuss PLDs *9*

Development of Borderline Student Descriptors *9*

Standard Setting Training and Practice Round *10*

Readiness Survey *11*

Round 1 ..... *11*

Round 1 Feedback *12*

Round 2 ..... *12*

Round 2 Feedback *12*

Round 3 ..... *13*

Round 3 Feedback *13*

ROUND 3 RECOMMENDED CUT SCORES **14**

VERTICAL ARTICULATION **15**

Evaluation *15*

VERTICAL ARTICULATION RECOMMENDED CUT SCORES **16**

REFERENCES **18**

APPENDIX A: PANEL COMPOSITION **19**

APPENDIX B: PERFORMANCE LEVEL DESCRIPTORS **20**

APPENDIX C: GENERAL AGENDA..... 31

APPENDIX D: EXAMPLE PRACTICE EXERCISE RATING SHEET..... 32

APPENDIX E: EXAMPLE ROUND READINESS SURVEY..... 33

APPENDIX F: AN EXAMPLE RATINGS FORM..... 34

APPENDIX G: STANDARD SETTING EVALUATION..... 36

OVERVIEW

Standard setting is the process of determining cut scores, or performance standards. In June 2016, the Colorado Department of Education (CDE) convened a standard setting committee to recommend cut scores for the new Colorado Spanish Language Arts (CSLA) assessments. The standard setting process integrated content knowledge, educator classroom experience and expert judgment, Performance Level Descriptors (PLDs), and empirical data to recommend four cut scores that distinguished five performance levels on each CSLA assessment. The performance levels for the assessments are as follows:

- Level 5: Exceeded Expectations
- Level 4: Met Expectations
- Level 3: Approached Expectations
- Level 2: Partially Met Expectations
- Level 1: Did Not Yet Meet Expectations

Beginning with the Spring 2016 administration, third and fourth grade students who met eligibility criteria participated in CSLA instead of CMAS: PARCC English Language Arts (ELA). Colorado School Law §22-7-409 (3.5) (a) and (b) requires a Spanish Language Arts assessment for third and fourth grade. The previous Spanish Language Arts assessments, Lectura and Escritura, were aligned to the Colorado Model Content Standards. The new CSLA assessments, however, are aligned to the skills and concepts in the Colorado Academic Standards. These paper-based assessments have been created using blueprints that mirror the CMAS: PARCC ELA assessments and are intended to be an accommodated version of CMAS: PARCC ELA.

The CSLA assessments contain two item types: Evidence-Based Selected Response (EBSR) items and Constructed Response (CR) items. The EBSR items are machine-scored items and scored on a 0−2 point scale. The CR items are hand-scored items and can be categorized as Prose Constructed Response (PCR) items or Narrative Prose Constructed Response (NPCR) items. Both the PCR and NPCR items have two trait dimensions. The PCR traits are 1) Reading Comprehension and Written Expression (RCWE) and 2) Writing Knowledge of Language and Conventions (WKLC). The NPCR traits are 1) Written Expression (WE) and 2) Writing Knowledge of Language and Conventions (WKLC). For the PCR literary analysis tasks and research simulation tasks, the RCWE trait is worth 0−3 points for grade 3 and 0−4 points for grade 4. The PCR trait of WKLC is worth 0−3 points for both grades 3 and 4. For the NPCR items, all traits are worth 0−3 points. Weighting is also applied to the RCWE and WE traits as part of the test design.

Performance standards were set on the full test score for each CSLA assessment. Each assessment also reports out subscores for various content domains (e.g. a reading subscore and a writing subscore); however, classification into performance levels was based on the full test score.

The Modified Extended Angoff approach (Cizek, 2012; Cizek, Bunch, & Koons, 2004; Hambleton & Plake, 1995) was used to set performance standards on the CSLA assessments.

With this methodology, standard setting panelists review the content of each test item, and considering the content the item is measuring and the content knowledge of the students at the cut scores (i.e., borderline students), the panelists make a judgment about what score a borderline student would receive on the item to be considered "just-barely" in a performance level. Panelists use the PLDs to conceptualize "borderline" students (those students just barely in a particular performance level) in order to determine the score the borderline student would obtain on each item. The individual item-level cut scores for each particular performance level are then summed for each panelist to obtain the recommended test-level cut scores that are used to define the performance levels.

PREPARATION FOR STANDARD SETTING

Preparation for standard setting began months before the actual meeting. This section provides details about the selection of panelists, the development of the PLDs, the various materials that were gathered or created for the meeting, and the training for those who facilitated the meeting and analyzed the data.

PANELIST SELECTION AND COMPOSITION

One committee was convened to recommend performance standards for both grades 3 and 4. The final CSLA standard setting committee consisted of ten panelists. Twelve panelists were in attendance at the start of the meeting; however, due to family circumstances, two panelists were unable to attend the full three-day meeting. Panelists were grouped into tables of three with three to four panelists per table. The CSLA panelists included educators who teach English language learners at grades 3 and 4, are content experts with knowledge of the subject-area curriculum, and are familiar with the instruction and specific needs of the students in an English language proficiency program. In addition to classroom teachers, educators in higher education and school administrators and/or directors who are familiar with instruction in classrooms where the Spanish language is used also participated in the meeting. Appendix A describes the panel composition of the committee.

DEVELOPMENT OF PLDS

PLDs are an important tool for recommending cut scores. PLDs outline the expectations of student performance at each performance level of a test. As a component of the standard-setting process, PLDs serve to anchor training activities and guide committee members by establishing a common understanding of expected performance on each CSLA assessment. In addition to their use in standard setting, PLDs have been published to serve as a tool for classroom instruction and to help educators interpret student performance on the assessments. PLDs can also enhance parents' understanding of their child's academic strengths and weaknesses and can help them understand the test scores and the level of performance required of students on the test.

Each CSLA assessment has five performance levels. The performance levels range from Level 1 (Did Not Meet Expectations), representing the lowest level of student performance, to Level 5 (Exceeded Expectations), representing the highest level of student performance. As mentioned previously, the CSLA assessment was designed to mirror the CMAS: PARCC ELA assessment and is considered an accommodated version of CMAS: PARCC ELA. As a result, the CSLA PLDs were based on the CMAS: PARCC ELA PLDs. However, the CMAS: PARCC ELA

PLDs were updated as needed for CSLA to take into account classroom instruction given in Spanish and the population of students taking the CSLA assessment. The CSLA PLDs were developed in English and then translated into Spanish for standard setting and for later posting to the CDE website. After the CSLA PLDs were created, both language versions were posted online for public review. CDE then reviewed the feedback and incorporated feedback into the PLDs as appropriate. For the standard setting meeting, panelists were provided the CSLA PLDs in both English and Spanish. During the standard setting meeting, panelists were also offered the opportunity to provide additional feedback regarding the clarity of the PLDs; however, they were not able to change the content of the PLDs during the meeting. Following the standard setting meetings, CDE reviewed panelists' suggestions on how to clarify the PLDs and incorporated feedback into the PLDs where appropriate. Appendix B contains the final CSLA PLDs.

CREATION OF MATERIALS

A standard setting meeting requires a myriad of materials. Documents were obtained from several different sources for the meeting. Some documents were uniquely created for panelists to use during the meeting and other documents were obtained from the materials distributed during the Spring 2016 CSLA administration. CDE reviewed, edited, and approved all documents prior to the standard setting meeting. This section outlines the primary materials that were used for the meeting.

Agendas

A general agenda, which contained an outline of the standard setting tasks that all the panelists would be completing during the meeting, was created and provided to the panelists at the beginning of the meeting. A specific agenda was also created and provided to CDE and other standard setting staff. This agenda outlined the same tasks listed in the general agenda, but with more detail regarding each task and the specific times each task was to begin and end.

Slides

For the introductory session, a PowerPoint presentation was created to provide a general overview of the CSLA assessment and the standard setting process. For the remainder of the meeting, an additional PowerPoint presentation was developed and presented to panelists.

CSLA Test Book

To allow the panelists the opportunity to become familiar with the items and the scoring of the CSLA assessment, the Spring 2016 CSLA test book corresponding to each grade was provided to panelists as part of the standard setting process. All operational items and field-test items appeared in the test book, but only the operational items were reviewed as part of the standard setting process. In addition to the test book, the answer key and content standards for the operational items were also provided to the panelists.

Description of the Item Scoring

Reference material describing how the EBSR and CR items are scored was also provided to the panelists. Having the documents easily accessible allowed panelists the opportunity to refer to the scoring rules as often as they are needed during the standard setting process.

Sample Rubric Responses

A set of sample student responses for the operational CR items were obtained from the Spring 2016 student data and distributed to panelists during the meeting. The panelists used the sample responses to review the type of student work that would earn a specific rubric score.

Item Mean and Score Point Distribution Reports

Item means and score point distributions were provided to panelists as part of the feedback provided after Round 1 recommendations. The item mean is the average score students earned on an EBSR item or a trait dimension of a CR item. The score point distribution is the percentage of students who earned each score point on an EBSR item or a trait dimension of a CR item.

External Data

As part of the feedback provided to panelists after Round 2 recommendations, external data were shared with panelists to provide a point of reference for their CSLA results. The CMAS: PARCC ELA performance data corresponding to each grade level were provided to panelists to review.

Forms

Various forms were created for panelists to complete during the meeting and include the following:

- Panelist Information Form: While some demographic information is included in the database of Colorado educators, the panelist information sheet was used to collect additional demographic information from the panelists.

- Readiness Survey: A brief questionnaire was provided to panelists before each round of the standard setting process in which panelists were asked to verify that they understand the task at hand and are ready to move forward with providing their cut score recommendations.

- Ratings Recommendation Forms: The ratings forms were used to collect panelists' item ratings for each round.

- Standard Setting Evaluation: An evaluation was provided to panelists after the standard setting to gather information on panelists' perceptions of the meeting.

FACILITATOR AND DATA ANALYST TRAINING

Because only one committee was convened, only one facilitator and one data analyst was required for the meeting. For the facilitator training, an overview of the new CSLA assessment was provided and the meeting slides were also reviewed and discussed in detail to ensure that the facilitator understood how to lead the panelists through the standard setting process and the logistics of the meeting. In addition to reviewing the slides and the meeting logistics, the facilitator also reviewed his/her facilitator materials and the materials to be distributed to the panelists. As part of the training, content specialists who were to attend the standard setting meeting also met with the lead facilitator to discuss the standard setting process and meeting

logistics. The content specialists were available throughout the standard setting meeting to answer any content-related questions posed by panelists. For the data analyst, it was important that the analysis code and spreadsheets be set up properly to ensure accurate and rapid analysis of panelists' recommendations. All the analysis code and spreadsheets created for the meetings were tested and verified before the meetings.

STANDARD SETTING MEETING ACTIVITIES

The CSLA standard setting was held on June 27–29, 2016. During the three-day meeting, panelists received training on the assessment and the standard setting process, reviewed the grade-level PLDs, reviewed the Spring 2016 operational items, reviewed the borderline student descriptors, and apply the Modified Extended Angoff method to establish cut score recommendations across three rounds of rating. During the process of establishing cut score recommendations, panelists also reviewed the content standards assessed by the CSLA items, reviewed CMAS: PARCC ELA external data, engaged in table level and whole group discussions, and considered the impact of their cut scores on student performance when making their CSLA cut score recommendations. The specific procedures involved in the CSLA standard setting are described in the sections that follow.

INTRODUCTORY SESSION

The standard setting meeting began with an introductory session in which panelists received a general overview of the meeting and the directions regarding their tasks during the meeting. To begin the introductory session, a representative from CDE welcomed the panelists to the meeting and provided the context for the meeting by presenting details describing the purpose of the CSLA assessment. This information was provided to help the panelists understand what standard setting is and the reason they were asked to be part of a standard setting committee. Then there were formal introductions of all panelists participating in the meeting as well as the standard setting staff. After introductions were complete, CDE continued by presenting details describing the CSLA item and test development process, the test design, and the importance of standard setting in the test development process.

Next, a member of the Pearson Psychometric Services staff provided an overview of the standard setting process and a description of the Modified Extended Angoff method, including the rationale behind the procedure and the types of decisions panelists would be asked to make during the meeting. A high-level agenda containing the tasks the panelists were to complete over the three-day meeting was also provided to the panelists (see Appendix C for an example). Once the introductory session was completed, the panelists began the specific standard setting tasks.

THE STANDARD SETTING PROCESS

The standard setting specific tasks were completed over the course of three days as outlined in this section. The committee recommended performance standards for the grade 3 assessment first and then repeated the same process for the grade 4 assessment.

Experience the Assessment

After the introductory session was completed, the Pearson facilitator followed with a review of the meeting agenda, housekeeping tasks (i.e., non-disclosure agreements and security protocols), and answered any panelist's questions regarding meeting logistics and the standard setting process.

After the general housekeeping tasks were completed, panelists were given the grade 3 CSLA assessment as part of their first standard setting task. To become more familiar with the test for which they would be setting performance standards, the panelists were asked to experience the CSLA assessment. Panelists took the CSLA assessment and were encouraged to think about the test from a student's perspective. For CR items, they were not required to draft a full response but were encouraged to think about strategies that they would use to engage with the items. After taking the assessment, the panelists were trained on how the CSLA items are scored and given an opportunity to apply the test's answer key to score their own responses to the test questions. As part of this task, the panelists were also given sample student responses for the CR items to understand the type of student response that earns a specific rubric score. After taking the assessment and scoring the items, panelists were then asked to discuss their test-taking experience, the types of knowledge and skills the students are asked to demonstrate for each item, and the difficulty of the test.

Review and Discuss PLDs

After experiencing the assessment, panelists reviewed and discussed the PLDs. Panelists used the PLDs to obtain a common understanding of the knowledge, skills, and abilities possessed by a typical student in each performance level for each grade. After being given a copy of the PLDs for the specific grade-level assessment, panelists were asked to review the performance labels and the PLDs and write down any comments they may have regarding the PLDs. These comments were later given to CDE to review and apply to the PLDs as appropriate. After the panelists provided comments regarding the PLDs, the meeting facilitator then led the panelists in a discussion of the characteristics that most differentiate the adjacent performance levels. The panelists were instructed to refer to these characteristics as they moved through the standard setting process.

Development of Borderline Student Descriptors

With this task, the panelists were introduced to the differences between a *typical* student and a *borderline* student within a performance level. Panelists were reminded that the main purpose for reviewing and discussing the PLDs is to operationalize the performance levels to *support the standard setting task*. The focus of this activity was on the borderline students—those students who "just barely" make it into a particular performance level (i.e., those students who are minimally qualified to be classified within a particular performance level). These students are the focus of standard setting because it is these students the panelists must consider when recommending the cut scores that define the five performance levels. The goal of this activity was to have the panelists develop borderline student descriptors as a whole group to gain a common understanding of these students so that when panelists were asked to think about a borderline student, they were all in agreement regarding what such a student knows and can do. Because the CSLA assessment has five performance levels, panelists were asked to develop borderline descriptors for four groups of students:

- Level 2 borderline students (cut score between Level 1 and Level 2)
- Level 3 borderline students (cut score between Level 2 and Level 3)
- Level 4 borderline students (cut score between Level 3 and Level 4)
- Level 5 borderline students (cut score between Level 4 and Level 5)

Prior to the standard setting meeting, a set of "draft" borderline descriptors for each assessment was created by a set of content experts familiar with the PLDs for the assessment. These draft descriptors provided panelists a starting point for their work and contained a unique list of knowledge, skills, and abilities that a borderline student would be expected to demonstrate. Panelists worked with these draft descriptors in a 3-step process as outlined below to create a final set of borderline descriptors that they used when making their cut score judgments.

- **Step 1**: A representative from each table was assigned a borderline level (2, 3, 4, or 5). Borderline level groups met first to discuss the draft descriptors for their assigned level.
    - How well do these describe the borderline student as we envision them?
- **Step 2**: Table groups reconvened and discussed what they learned about each borderline student group and then work as a table group to edit assigned descriptors.
    - Are there any concepts and skills that you would modify/revise in the descriptors to better reflect the limited capabilities of the “just-barely” student?
- **Step 3**: The facilitator reviewed the compiled edited descriptors from each group with the whole group and additional edits or clarifications to the descriptors were made as needed. Once finalized, the borderline descriptors were printed for each panelist to use throughout the standard setting activity.

Standard Setting Training and Practice Round

After the development of the borderline student descriptors, panelists were introduced to the Modified Extended Angoff standard-setting method. The meeting facilitator introduced the method to the panelists and then explained the steps the panelists needed to complete as part of the method. Following the training session, panelists engaged in a practice round of standard setting using a small set of items. The purpose of this exercise was to have the panelists practice evaluating and rating items to make sure they were comfortable with the task.

For the practice exercise, a set of six items and a Practice Exercise Ratings sheet were presented to the panelists. The practice item set included EBSR items and one PCR item. For each practice item, the panelists were asked to review the item, the answer key, the PLDs, the scoring rules and rubrics, and the borderline descriptors. Based on the review of the item, the borderline descriptors, and other related materials, the panelists were asked to answer the following question:

"How many points would a borderline performance level student likely earn if they answer this item?"

Likely was defined for panelists as 2 out of 3 times. The panelists were instructed that the item judgments should be made for all performance levels for one item before moving on to the next item, starting with performance level 2 and moving up to performance level 5. The following outlines the specific steps that were followed for the Level 2 cut.

1. Review the items listed on the rating sheet.
2. Identify the skills required for the item.
    - Think about the content the item is measuring.
3. Decide: How would performance appear for the borderline students?
    - Think about the content of the item, the difficulty of the item, how the item is scored, and should the borderline student be able to demonstrate the skills assessed by the item.
4. Decide: How many points would a borderline performance level student likely earn if they answer this item?
5. On the ratings sheet, indicate the item-level score you feel describes what a borderline student should be able to obtain.

The same steps were repeated for the "Level 3", "Level 4", and "Level 5" cuts. Panelists were asked to complete their judgments independently and without discussion from other panelists. Before beginning their practice ratings, panelists were also asked to complete a practice round readiness form to indicate they understood the steps of the process and were ready to provide the item-level cut scores for each performance level (see Appendix D for an example). After the panelists provided their ratings on their Practice Exercise Ratings sheet, the facilitator asked the panelists to share their rating results with the whole group, leading to a group discussion where panelists discuss their ratings and the general process employed. Based on the panelists' discussion, the facilitator provided additional instructions and guidance as needed. After the practice exercise discussion, the panelists were ready for the rounds of rating.

Readiness Survey

To evaluate whether the panelists understand their task for each round of rating, a readiness survey was completed by each panelist prior to beginning each round. The readiness survey asked panelists to report if they understood the task asked of them as well as any feedback data provided before making subsequent ratings (see Appendix E for an example). Results of the readiness survey indicated that panelists understood their tasks for each round and understood the data presented.

Round 1

After completing the readiness survey, the panelists were ready to begin Round 1. Prior to beginning Round 1, panelists were reminded to think about the content each item measures, the answer key, the PLDs, the scoring rules and rubrics, and the borderline student descriptors. During Round 1, panelists received a round readiness form and a Round 1 Ratings form to complete (see Appendix F for an example). Within the round, panelists were asked to consider each operational item in the test form, starting with first operational item to the last operational item. Panelists worked independently to make their item-level cut score ratings for each performance level. Panelists worked on their Round 1 ratings until they were dismissed at the end of day 1. At the conclusion of day 1, the meeting facilitator collected all ratings forms and secure materials before the panelists were dismissed. The ratings forms and secure materials were then redistributed at the beginning of day 2 so the panelists could complete their Round 1 ratings. When the panelists were finished providing their ratings, the meeting facilitator collected each panelist's ratings form and reviewed the form to ensure all the ratings for each item were present. The ratings sheets were then submitted for analysis.

Round 1 Feedback

After the Round 1 data were analyzed, panelists received their feedback results which included several pieces of data. With each piece of data, the panelists were reminded that the data are intended to inform their decisions, but not to dictate them.

Panelists were presented with feedback showing their individual, table, and committee-level test-level cut scores. The whole group feedback included the minimum, maximum, mean, and median test-level cut scores for Level 2, Level 3, Level 4, and Level 5, as well as a bar chart reflecting the panelists' cut score agreement for the performance levels. The table-level feedback included the same type of statistics shown for the committee-level cut scores. Panelists also received a summary of the frequency distribution of item scores for each item at each performance level. The panelists' Round 1 Ratings form was also redistributed with the Round 1 feedback, so the panelists could refer to their initial ratings as they reviewed the summary of the frequency distribution of the item scores as a table group.

Item mean scores and score point distributions were also presented to the panelists. The item means and score point distributions were intended to be used to validate panelists' perceptions of item difficulty.

After receiving the various pieces of data, panelists held table-level discussions regarding the results. During the table-level discussions, panelists were instructed to consider how close their recommendations are to those of others in their table group as well as the whole group and discuss why they may have had different ratings for certain items. As part of their discussions, several items with the greatest level of panelist disagreement for each performance level was also discussed to help panelists think about the specific factors that could lead to the different ratings. While the discussion of the Round 1 feedback was conducted to encourage panelists to re-evaluate their cut score recommendations, the main purpose of this activity was to allow panelists the opportunity to think through and discuss the recommendation process; it was not done to have panelists arrive at a consensus.

Round 2

After discussing Round 1 feedback and completing the readiness survey for Round 2, panelists worked independently to re-evaluate their recommendations and decided whether they wanted to revise their ratings. During Round 2, the panelists continued to consider the content each item is measuring, the answer key, the PLDs, the scoring rules and rubrics, and the borderline descriptors before providing their item-level cut score ratings. As before, panelists were reminded that their recommendations should be grounded in the content and what students should know and be able to do, not what they can do or are currently doing. Panelists recorded their Round 2 recommendations on their Round 2 Ratings form and submitted it to the facilitator.

Round 2 Feedback

As done previously, several pieces of feedback data were provided to panelists based on their Round 2 recommendations. Panelists received the same summary statistics as in Round 1, but the statistics were based on their Round 2 recommendations. Table-level and committee-level discussions were also conducted for these data.

For this round, impact data based on the Spring 2016 operational administration was also provided. Based on Round 2 recommendations, bar graphs indicating the percentage of students who would be in each of the performance levels were displayed. The impact data for the overall test was based on the median test-level cut scores and was also shown disaggregated by gender and economic disadvantage. After reviewing the impact data for the total group and by gender and by economic disadvantage, panelists were then asked to discuss whether the percentage of students in each performance level met their expectations given what they know about the population of students tested and the test content.

Impact data for the corresponding CMAS: PARCC ELA grade-level assessment was also shown to panelists during this round. Before being shown the CMAS: PARCC ELA impact data, panelists were asked their expectations regarding the CMAS: PARCC ELA impact data in relation to the CSLA impact data. Both sets of impact data were intended to provide a reasonableness check, but panelists were reminded that any modifications to their cut score recommendations should be based on content and not driven by impact data alone.

Round 3

After discussing Round 2 feedback and completing the readiness survey for Round 3, panelists worked independently to again re-evaluate their recommendations. Panelists completed their round readiness form for this last round and then recorded their final item-level ratings on the Round 3 Ratings form and submit their completed ratings sheet to the facilitator.

Round 3 Feedback

After completing their Round 3 ratings, panelists were shown their Round 3 feedback. They reviewed the committee-level cut score recommendations for each performance level and panelist agreement data. Impact data based on their Round 3 ratings was also shown to the panelists which was based on the median test-level cut scores. Panelists were also given an opportunity to discuss the Round 3 feedback data. This discussion and the Round 3 results were the primary inputs to the vertical articulation process.

After the panelists completed their final round of recommendations for grade 3, they completed the same steps of the standard setting process for the grade 4 assessment. They began their tasks for the grade 4 assessment during the afternoon of day 2 and completed the grade 4 standard setting tasks during the afternoon of the final day of the meeting.

ROUND 3 RECOMMENDED CUT SCORES

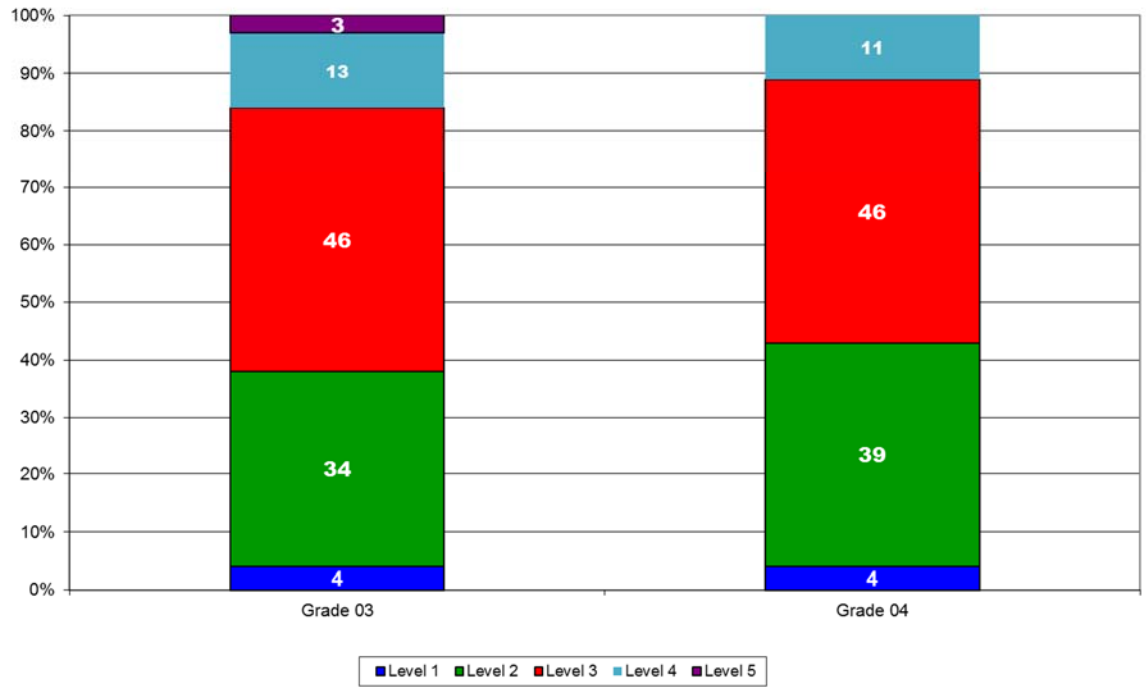
This section provides results from the standard setting meeting. Table 1 shows the median of panelists' recommendations by round.

Table 1. Panelists Recommendations by Round

		Level 2: Partially Met Expectations	Level 3: Approached Expectations	Level 4: Met Expectations	Level 5: Exceeded Expectations
Grade 3	**Round 1**	14	35	61	78
	Round 2	11	33	58	76
	Round 3	8	25	53	70
Grade 4	**Round 1**	9	30	59	81
	Round 2	8	29	54	76
	Round 3	9	28	55	79

Table 2 shows the percentages of students who would fall into each performance level based on the Round 3 recommendations using the Spring 2016 administration data.

Table 2. Round 3 Estimated Impact Data for CSLA



VERTICAL ARTICULATION

Once the performance standards were recommended for the grade 3 and the grade 4 assessments, the standard setting panelists made cross-grade comparisons during vertical articulation. The purpose of vertical articulation was to review the impact data associated with the recommended cut scores across both grades to determine if the trend of the impact data is reasonable given the PLDs, the test-taking population, and the concepts and skills presented on the assessments.

After a brief introduction to the vertical articulation process, panelists spent some time reviewing the PLDs for both grades. Panelists reviewed the PLDs independently with instructions to look for differences across grades. The panelists then discussed the differences as a whole group. The review of both sets of PLDs helped provide a complete picture of the developmental continuum for the content area.

After reviewing the PLDs, the expectations for impact across the grade levels were discussed as a group. Both the CSLA and CMAS: PARCC ELA expectations were discussed. The following questions were posed to the group:

- What are your expectations of the student performance data progression across the grades for CSLA?
    - Do you expect similar percentages of students in performance levels across grades? Why or Why not?
        - Is there a progression of skills in the PLDs that suggest differential impact from grade to grade?
        - Do populations differ significantly as you move from grade to grade?
    - What other trends might you expect to see and why?
- Do you expect similar percentages of students in performance levels across grades between CSLA and CMAS: PARCC ELA? Why or why not?

After discussing expectations, the impact data associated with the Round 3 recommended cuts for each grade was shown in a side-by-side chart. Panelists were encouraged to discuss how/if cut scores should be changed to be consistent with impact expectations. These changes were made directly at the test level and did not involve item-level judgments.

Throughout this discussion, it was stressed to panelists that the intent is not to undo all that was done in the standard setting meeting. Rather, the goal was to provide reasonable cut score recommendations to policy makers that consider both the content-based recommendations and the expectations about how students should perform across performance levels. However, any desire to change the cuts needed to be justified based on the PLDs and the assessment items. Once the group reached a shared recommendation, final results were displayed.

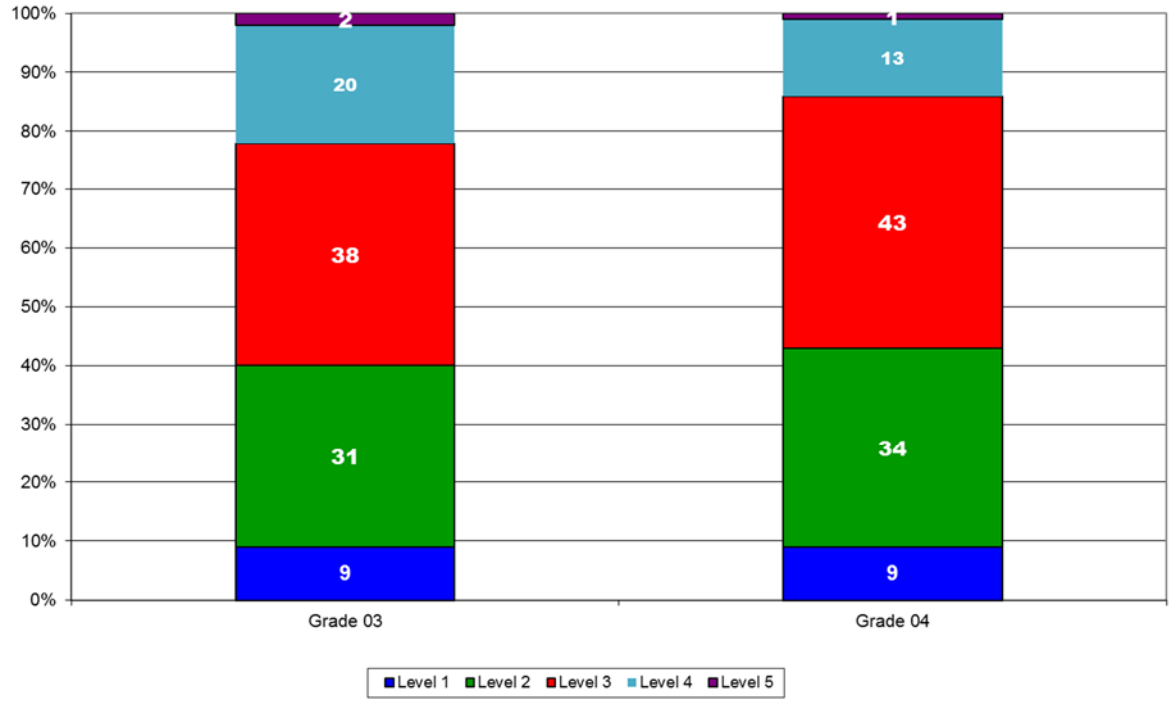
Evaluation

To end the meeting, panelists were asked to complete a brief evaluation. The evaluation asked panelists about their level of comfort with the standard setting procedure, their understanding of the performance levels, and their satisfaction with the final recommended cut scores. The standard setting evaluation and results can be found in Appendix G. Upon completing the evaluations, panelists were thanked for their time and participation and dismissed.

VERTICAL ARTICULATION RECOMMENDED CUT SCORES

The updated impact data for grades 3 and 4 are reflected in Table 3. The panelists recommended adjustments to the cut scores for each grade during vertical articulation. Discussions around adjusting the cut scores across both grades focused on the content assessed on the assessments, the characteristics of the student populations, and the rigor of the cut scores.

Table 3. Post-Vertical Articulation Impact Data for CSLA



At the completion of vertical articulation, the cut score recommendations were then reviewed by CDE to ensure that the performance standards contributed to a well-articulated and coherent assessment program. Table 4 shows the final recommended cut scores, and Table 5 shows the scale score ranges resulting from the final recommended cut scores.

Table 4. CSLA Final Recommended Cut Scores

	Level 2: Partially Met Expectations	Level 3: Approached Expectations	Level 4: Met Expectations	Level 5: Exceeded Expectations
Grade 3	11	26	48	70
Grade 4	12	28	53	73

Table 5. CSLA Scale Score Ranges

	Level 1: Did Not Yet Meet Expectations	Level 2: Partially Met Expectations	Level 3: Approached Expectations	Level 4: Met Expectations	Level 5: Exceeded Expectations
Grade 3	650–699	700–724	725–749	750–778	779–850
Grade 4	650–699	700–724	725–749	750–771	772–850

REFERENCES

Cizek, G. J. (2012). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). New York: Routledge.

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

Hambleton, R. K. & Plake, B.S. (1995). Using an Extended Angoff Procedure to Set Standards on Complex Performance Assessments. *Applied Measurement in Education*, 8, 41–56.

APPENDIX A: PANEL COMPOSITION

Panelist Breakdown by Expertise

	Total
ELL Teacher/Specialist	**6**
Administrator	**4**
Higher Ed	**2**
Total	**12**

Panelists Breakdown by School Type

	Total
Charter/Innovation School	**1**
Neither Charter nor Innovation School	**5**
District Level	**4**
Other	**2**
Total	**12**

Panelists Breakdown by Region

	Total
Denver Metro	**8**
North Central	**2**
Northeast	**0**
Northwest	**2**
Pikes Peak	**0**
Southeast	**0**
Southwest	**0**
West Central	**0**
Total	**12**

Panelists Breakdown by Experience with Special Populations

	Total
Students who are English language learners	**12**
Students receiving special education services	**9**
Students of low socioeconomic status	**12**

APPENDIX B: PERFORMANCE LEVEL DESCRIPTORS

Grade 3 Colorado Spanish Language Arts (CSLA) Performance Level Descriptors

Performance level descriptors (PLDs) outline the knowledge, skills, and abilities possessed by a typical student in each performance level. The PLDs for the Colorado Spanish Language Arts (CSLA) assessment are based on the CMAS PARCC Consortium's English Language Arts (ELA) PLDs (©2015 PARCC Inc.). This interpretation of the CMAS PARCC ELA PLDs is designed to help educators and parents understand the CSLA assessment results.

Performance Level	Level of Text Complexity[1]	Range of Accuracy[2]	Quality of Evidence[3]
5	Very Complex Moderately Complex Readily Accessible	Mostly Accurate Mostly Accurate Accurate	Explicit Explicit Explicit
4	Very Complex Moderately Complex Readily Accessible	Generally Accurate Generally accurate Mostly Accurate	Explicit Explicit Explicit
3	Very Complex Moderately Complex Readily Accessible	Minimally Accurate Generally accurate Mostly Accurate	Explicit Explicit Explicit
2	Very Complex Moderately Complex Readily Accessible	Inaccurate Minimally accurate Partially accurate	Explicit Explicit Explicit

1. Text Complexity

The complexity framework reflects the importance of text complexity as it relates to the CCSS, which indicates that 50 percent of an item's complexity is linked to the complexity of the text(s) used as the stimulus for that item. Consequently, to determine students' performance levels, it is critical to identify the pattern of responses when students respond to items linked to passages with distinct text complexities. To this end, CMAS PARCC has developed a clear and consistent model to define text complexity and has determined to use three text complexity levels: readily accessible, moderately complex, or very complex. For more information on text complexity, refer to the CCSS Appendix A (http://www.corestandards.org/ELA-Literacy) and Appendix B (http://www.corestandards.org/ELA-Literacy).

CMAS PARCC and CSLA use two components for determining text complexity for **all** passages:

a. CMAS PARCC uses two quantitative text complexity measures (Reading Maturity Metric and Lexile) to analyze all reading passages to determine **an initial** recommendation for placement of a text into a grade band and subsequently a grade level. At the time initial CSLA passages were being developed, quantitative text complexity measures were not available in Spanish. The initial CSLA passages were analyzed for text complexity by Spanish content experts then evaluated and approved for grade-level placement by Colorado educators proficient in Spanish and bilingual education practices.
b. Text Analysis Worksheets (http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents), one for informational text and one for literary text, are then used to determine qualitative measures. Trained evaluators use these worksheets to determine a recommendation for qualitative text complexity within the grade level, with each text defined as readily accessible, moderately complex, or very complex.

2. Range of Accuracy

There are two types of items on the CSLA assessments: Evidence-Based Selected Response (EBSR) and Prose-Constructed Response (PCR) items. For EBSR items, the design is such that the items help contribute to an understanding of how accurately students comprehend text (demonstrate mastery of CCSS Reading Standards 2-10). These items offer opportunities for students to receive partial credit based on the range of accuracy. For PCR items, CSLA uses the CMAS PARCC scoring rubrics accommodated for Spanish that include a Reading dimension to measure comprehension. Scores on the PCR items contribute to an evaluation of the degree to which a student can accurately comprehend a text.

The CSLA assessment Performance Level Descriptors (PLDs) describe five levels of accuracy at grades 3-4 that are determined using the reading data collected through EBSR and PCR items:

Accurate – The student is able to accurately state both the general ideas expressed in the text(s) and the key and supporting details. The response is complete, and the student demonstrates full understanding.

Mostly accurate – The student is able to accurately state most of the general ideas expressed in the text(s) and the key and supporting details, but the response is incomplete or contains minor inaccuracies. The student demonstrates understanding.

Generally accurate – The student is able to accurately state the gist of the text(s) but fails to accurately state the key and supporting details in the text or to connect such details to the overarching meaning of the text(s). The student demonstrates basic understanding.

Partially accurate – The student is able to accurately state the gist of the text(s) but is unable to state some of the key or supporting details with accuracy. The student is partially able to connect the specific details of the text to the overarching meaning(s) of the text. The student demonstrates partial understanding**.**

Minimally accurate – The student is unable to accurately state the gist of the text(s) but is able to minimally state some of the key or supporting details with accuracy. The student does not connect the specific details of the text to the overarching meaning(s) of the text. The student demonstrates minimal understanding.

Inaccurate – The student is unable to accurately state either the gist of the text or the key and supporting details evident in the text. The student demonstrates limited understanding.

3. Quality of Evidence

All items are designed to contribute to an understanding of how students "read closely to determine what the text says explicitly and to make logical inferences from it" and "cite specific textual evidence when writing or speaking to support conclusions drawn from the text" (CCSS Anchor Reading Standard 1). Some items offer opportunities for students to receive partial credit based on the quality of evidence provided. Students support their comprehension with explicit and/or inferential evidence:

Explicit evidence – Students show how the explicit words and phrases (details) from the text support statements made about the meaning of the text.

Inferential evidence – Students show how inferences drawn from the text support statements made about the meaning of the text.

Reading Sub-Claims	**Reading Literature** Students demonstrate comprehension and draw evidence from readings of grade-level, complex literary text.	**Reading Information** Students demonstrate comprehension and draw evidence from readings of grade-level, complex informational text.	**Vocabulary Interpretation and Use** Students use context to determine the meaning of words and phrases.
EVIDENCES: Students are expected to produce responses that demonstrate the skills and content listed in the evidence tables at the accuracy level and with the quality of evidence as described for students at each level.	**See Literary Evidence Table** http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents	**See Informational Evidence Table** http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents	**See Vocabulary Evidence Table** http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents

Level 5	**Level 4**	**Level 3**	**Level 2**
A student who achieves at **Level 5 exceeds expectations** for the assessed standards.	A student who achieves at **Level 4 meets expectations** for the assessed standards.	A student who achieves at **Level 3 approaches expectations** for the assessed standards.	A student who achieves at **Level 2 partially meets expectations** for the assessed standards.
In **reading**, the pattern exhibited by student responses indicates: ● With very complex text, students demonstrate the ability to be mostly accurate when asking and/or answering questions, showing understanding of the text when referring to explicit details and examples in the text. ● With moderately complex text, students demonstrate the ability to be mostly accurate when asking and/or answering questions, showing understanding of the text when referring to explicit details and examples in the text. ● With readily accessible text, students demonstrate the ability to be accurate when asking and/or answering questions, showing full understanding of the text when referring to explicit details and examples in the text.	In **reading**, the pattern exhibited by student responses indicates: ● With very complex text, students demonstrate the ability to be generally accurate when asking and/or answering questions, showing general understanding of the text when referring to explicit details and examples in the text. ● With moderately complex text, students demonstrate the ability to be generally accurate when asking and/or answering questions, showing general understanding of the text when referring to explicit details and examples in the text. ● With readily accessible text, students demonstrate the ability to be mostly accurate when asking and/or answering questions, showing understanding of the text when referring to explicit details and examples in the text.	In **reading**, the pattern exhibited by student responses indicates: ● With very complex text, students demonstrate the ability to be minimally accurate when asking and/or answering questions, showing minimal understanding of the text when referring to explicit details and examples in the text. ● With moderately complex text, students demonstrate the ability to be generally accurate when asking and/or answering questions, showing basic understanding of the text when referring to explicit details and examples in the text. ● With readily accessible text, students demonstrate the ability to be mostly accurate when asking and/or answering questions, showing understanding of the text when referring to explicit details and examples in the text.	In **reading**, the pattern exhibited by student responses indicates: ● With very complex text, students demonstrate the inability to ask or answer questions, showing limited understanding of the text when referring to explicit details and examples in the text. ● With moderately complex text, students demonstrate the ability to be minimally accurate when asking and/or answering questions, showing minimal understanding of the text when referring to explicit details and examples in the text. ● With readily accessible text, students demonstrate the ability to be partially accurate when asking and/or answering questions, showing partial understanding of the text when referring to explicit details and examples in the text.

Writing Sub-Claim for Written Expression: Students produce clear and coherent writing in which the development, organization, and style are appropriate to the task, purpose, and audience.	
EVIDENCES: Students are expected to produce responses that demonstrate the skills and content listed in the evidence tables at the accuracy level and with the quality of evidence as described for students at each level.	**See Writing Evidence Table** http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents

Level 5	Level 4	Level 3	Level 2
A student who achieves at **Level 5 exceeds expectations** for the assessed standards.	A student who achieves at **Level 4 meets expectations** for the assessed standards.	A student who achieves at **Level 3 approaches expectations** for the assessed standards.	A student who achieves at **Level 2 partially meets expectations** for the assessed standards.
In **writing**, students address the prompts and provide effective development of ideas, including when drawing evidence from multiple sources, in the majority of instances demonstrating purposeful and controlled organization. The student: ● Provides effective development of the topic and/or narrative elements, using reasoning, details, text-based evidence, and/or description. ● Develops topic and/or narrative elements in a manner that is appropriate to the task and purpose. ● Demonstrates purposeful organization that includes an introduction and/or conclusion. ● Effectively uses linking words and phrases, descriptive words, and/or temporal words to express ideas with clarity.	In **writing**, students address the prompts and provide development of ideas, including when drawing evidence from multiple sources, while in the majority of instances demonstrating purposeful and mostly controlled organization. The student: ● Develops the topic and/or narrative elements using reasoning, details, text-based evidence, and/or description. ● Develops topic and/or narrative elements in a manner that is mostly appropriate to the task and purpose. ● Demonstrates purposeful organization that is mostly controlled and may include an introduction and/or conclusion. ● Uses linking words and phrases, descriptive words, and/or temporal words to express ideas with clarity.	In **writing**, students address the prompts and provide basic development of ideas, including when drawing evidence from multiple sources, while in the majority of instances demonstrating organization that sometimes is controlled. The student: ● Develops the topic and/or narrative elements using some reasoning, details, text- based evidence, and/or description. ● Demonstrates some organization. ● Includes some linking words and phrases, descriptive words, and/or temporal words, limiting the clarity with which ideas are expressed.	In **writing**, students address the prompts and provide minimal development of ideas, including when drawing evidence from multiple sources, while in the majority of instances demonstrating organization that often is not controlled. The student: ● Provides minimal development of the topic and/or narrative elements and is, therefore, inappropriate to the task and purpose. ● Demonstrates minimal organization. ● Includes minimal linking words and phrases, descriptive words, and/or temporal words, limiting the clarity with which ideas are expressed.

Writing Sub-Claim for Knowledge of Language and Conventions: Students demonstrate knowledge of conventions and other important elements of language.	
EVIDENCES: Students are expected to produce responses that demonstrate the skills and content listed in the evidence tables at the accuracy level and with the quality of evidence as described for students at each level.	**See Writing Evidence Table** http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents

Level 5	Level 4	Level 3	Level 2
A student who achieves at **Level 5 exceeds expectations** for the assessed standards.	A student who achieves at **Level 4 meets expectations** for the assessed standards.	A student who achieves at **Level 3 approaches expectations** for the assessed standards.	A student who achieves at **Level 2 partially meets expectations** for the assessed standards.
In **writing**, students demonstrate full command of the conventions of Standard Spanish consistent with edited writing. There may be some errors in grammar and usage, but overall meaning is clear.	In **writing**, students demonstrate command of the conventions of Standard Spanish consistent with edited writing. There are errors in grammar and usage that may occasionally impede understanding.	In **writing**, students demonstrate basic command of the conventions of Standard Spanish consistent with edited writing. There are few patterns of errors in grammar and usage that impede understanding, demonstrating partial control over language.	In **writing**, students demonstrate minimal command of the conventions of Standard Spanish consistent with edited writing. There are patterns of errors in grammar and usage that impede understanding, demonstrating minimal control over language.

Grade 4 Colorado Spanish Language Arts (CSLA) Performance Level Descriptors

Performance level descriptors (PLDs) outline the knowledge, skills, and abilities possessed by a typical student in each performance level. The PLDs for the Colorado Spanish Language Arts (CSLA) assessment are based on the CMAS PARCC Consortium's English Language Arts (ELA) PLDs (©2015 PARCC Inc.). This interpretation of the CMAS PARCC ELA PLDs is designed to help educators and parents understand the CSLA assessment results.

Performance Level	Level of Text Complexity [1]	Range of Accuracy [2]	Quality of Evidence [3]
5	Very Complex Moderately Complex Readily Accessible	Mostly Accurate Mostly Accurate Accurate	Explicit & Inferential Explicit & Inferential Explicit & Inferential
4	Very Complex Moderately Complex Readily Accessible	Generally Accurate Generally accurate Mostly Accurate	Explicit & Inferential Explicit & Inferential Explicit & Inferential
3	Very Complex Moderately Complex Readily Accessible	Minimally Accurate Generally accurate Mostly Accurate	Explicit & Inferential Explicit & Inferential Explicit & Inferential
2	Very Complex Moderately Complex Readily Accessible	Inaccurate Minimally accurate Partially accurate	Explicit & Inferential Explicit & Inferential Explicit & Inferential

1. Text Complexity

The complexity framework reflects the importance of text complexity as it relates to the CCSS, which indicates that 50 percent of an item's complexity is linked to the complexity of the text(s) used as the stimulus for that item. Consequently, to determine students' performance levels, it is critical to identify the pattern of responses when students respond to items linked to passages with distinct text complexities. To this end, CMAS PARCC has developed a clear and consistent model to define text complexity and has determined to use three text complexity levels: readily accessible, moderately complex, or very complex. For more information on text complexity, refer to the CCSS Appendix A (http://www.corestandards.org/ELA-Literacy) and Appendix B (http://www.corestandards.org/ELA-Literacy).

CMAS PARCC and CSLA use two components for determining text complexity for **all** passages:

a. CMAS PARCC uses two quantitative text complexity measures (Reading Maturity Metric and Lexile) to analyze all reading passages to determine **an initial** recommendation for placement of a text into a grade band and subsequently a grade level. At the time initial CSLA passages were being developed, quantitative text complexity measures were not available in Spanish. The initial CSLA passages were analyzed for text complexity by Spanish content experts then evaluated and approved for grade-level placement by Colorado educators proficient in Spanish and bilingual education practices.
b. Text Analysis Worksheets (http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents), one for informational text and one for literary text, are then used to determine qualitative measures. Trained evaluators use these worksheets to determine a recommendation for qualitative text complexity within the grade level, with each text defined as readily accessible, moderately complex, or very complex.

2. Range of Accuracy

There are two types of items on the CSLA assessments: Evidence-Based Selected Response (EBSR) and Prose-Constructed Response (PCR) items. For EBSR items, the design is such that the items help contribute to an understanding of how accurately students comprehend text (demonstrate mastery of CCSS Reading Standards 2-10). These items offer opportunities for students to receive partial credit based on the range of accuracy. For PCR items, CSLA uses the CMAS PARCC scoring rubrics accommodated for Spanish that include a Reading dimension to measure comprehension. Scores on the PCR items contribute to an evaluation of the degree to which a student can accurately comprehend a text.

The CSLA assessment Performance Level Descriptors (PLDs) describe five levels of accuracy at grades 3-4 that are determined using the reading data collected through EBSR and PCR items:

Accurate – The student is able to accurately state both the general ideas expressed in the text(s) and the key and supporting details. The response is complete, and the student demonstrates full understanding.

Mostly accurate – The student is able to accurately state most of the general ideas expressed in the text(s) and the key and supporting details, but the response is incomplete or contains minor inaccuracies. The student demonstrates understanding.

Generally accurate – The student is able to accurately state the gist of the text(s) but fails to accurately state the key and supporting details in the text or to connect such details to the overarching meaning of the text(s). The student demonstrates basic understanding.

Partially accurate – The student is able to accurately state the gist of the text(s) but is unable to state some of the key or supporting details with accuracy. The student is partially able to connect the specific details of the text to the overarching meaning(s) of the text. The student demonstrates partial understanding**.**

Minimally accurate – The student is unable to accurately state the gist of the text(s) but is able to minimally state some of the key or supporting details with accuracy. The student does not connect the specific details of the text to the overarching meaning(s) of the text. The student demonstrates minimal understanding.

Inaccurate – The student is unable to accurately state either the gist of the text or the key and supporting details evident in the text. The student demonstrates limited understanding.

3. Quality of Evidence

All items are designed to contribute to an understanding of how students "read closely to determine what the text says explicitly and to make logical inferences from it" and "cite specific textual evidence when writing or speaking to support conclusions drawn from the text" (CCSS Anchor Reading Standard 1). Some items offer opportunities for students to receive partial credit based on the quality of evidence provided. Students support their comprehension with explicit and/or inferential evidence:

Explicit evidence – Students show how the explicit words and phrases (details) from the text support statements made about the meaning of the text.

Inferential evidence – Students show how inferences drawn from the text support statements made about the meaning of the text.

Grade 4 Colorado Spanish Language Arts (CSLA) Performance Level Descriptors

Reading Sub-Claims	**Reading Literature** Students demonstrate comprehension and draw evidence from readings of grade-level, complex literary text.	**Reading Information** Students demonstrate comprehension and draw evidence from readings of grade-level, complex informational text.	**Vocabulary Interpretation and Use** Students use context to determine the meaning of words and phrases.
EVIDENCES: Students are expected to produce responses that demonstrate the skills and content listed in the evidence tables at the accuracy level and with the quality of evidence as described for students at each level.	**See Literary Evidence Table** http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents	**See Informational Evidence Table** http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents	**See Vocabulary Evidence Table** http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents

Level 5	**Level 4**	**Level 3**	**Level 2**
A student who achieves at **Level 5 exceeds expectations** for the assessed standards.	A student who achieves at **Level 4 meets expectations** for the assessed standards.	A student who achieves at **Level 3 approaches expectations** for the assessed standards.	A student who achieves at **Level 2 partially meets expectations** for the assessed standards.
In **reading**, the pattern exhibited by student responses indicates: ● With very complex text, students demonstrate the ability to be mostly accurate when asking and/or answering questions, showing understanding of the text when referring to explicit details and examples in the text and when explaining inferences drawn from the text. ● With moderately complex text, students demonstrate the ability to be mostly accurate when asking and/or answering questions, showing understanding of the text when referring to explicit details and examples in the text and when explaining inferences drawn from the text. ● With readily accessible text, students demonstrate the ability to be accurate when asking and/or answering questions, showing full understanding of the text when referring to explicit details and examples in the text and when explaining inferences drawn from the text.	In **reading**, the pattern exhibited by student responses indicates: ● With very complex text, students demonstrate the ability to be generally accurate when asking and/or answering questions, showing general understanding of the text when referring to explicit details and examples in the text and when explaining inferences drawn from the text. ● With moderately complex text, students demonstrate the ability to be generally accurate when asking and/or answering questions, showing general understanding of the text when referring to explicit details and examples in the text and when explaining inferences drawn from the text. ● With readily accessible text, students demonstrate the ability to be mostly accurate when asking and/or answering questions, showing understanding of the text when referring to explicit details and examples in the text and when explaining inferences drawn from the text.	In **reading**, the pattern exhibited by student responses indicates: ● With very complex text, students demonstrate the ability to ask and/or answer questions with minimal accuracy , showing minimal understanding of the text when referring to explicit details and examples in the text. ● With moderately complex text, students demonstrate the ability to be generally accurate when asking and/or answering questions, showing basic understanding of the text when referring to explicit details and examples in the text. ● With readily accessible text, students demonstrate the ability to be mostly accurate when asking and/or answering questions, showing understanding of the text when referring to explicit details and examples in the text and when explaining inferences drawn from the text.	In **reading**, the pattern exhibited by student responses indicates: ● With very complex text, students demonstrate the inability to be accurate when asking and/or answering questions, showing limited understanding of the text when referring to explicit details and examples in the text. ● With moderately complex text, students demonstrate the ability to ask and/or answer questions with minimal accuracy, showing minimal understanding of the text when referring to explicit details and examples in the text. ● With readily accessible text, students demonstrate the ability to be partially accurate when asking and/or answering questions, showing partial understanding of the text when referring to explicit details and examples in the text and when explaining inferences drawn from the text.

Writing Sub-Claim for Written Expression: Students produce clear and coherent writing in which the development, organization, and style are appropriate to the task, purpose, and audience.	
EVIDENCES: Students are expected to produce responses that demonstrate the skills and content listed in the evidence tables at the accuracy level and with the quality of evidence as described for students at each level.	**See Writing Evidence Table** http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents

Level 5	Level 4	Level 3	Level 2
A student who achieves at **Level 5 exceeds expectations** for the assessed standards.	A student who achieves at **Level 4 meets expectations** for the assessed standards.	A student who achieves at **Level 3 approaches expectations** for the assessed standards.	A student who achieves at **Level 2 partially meets expectations** for the assessed standards.
In **writing**, students address the prompts and provide effective development of ideas, including when drawing evidence from multiple sources, in the majority of instances demonstrating purposeful and controlled organization. The student: ● Provides effective development of the topic and/or narrative elements, using reasoning, details, text-based evidence, and/or description. ● Develops topic and/or narrative elements in a manner that is appropriate to the task and purpose. ● Demonstrates purposeful organization that includes an introduction and/or conclusion. ● Correctly uses linking words and phrases, descriptive words, and/or temporal words to express ideas with clarity.	In **writing**, students address the prompts and provide development of ideas, including when drawing evidence from multiple sources, while in the majority of instances demonstrating purposeful and mostly controlled organization. The student: ● Develops the topic and/or narrative elements using reasoning, details, text- based evidence, and/or description. ● Develops topic and/or narrative elements in a manner that is mostly appropriate to the task and purpose. ● Demonstrates purposeful organization that is mostly controlled and may include an introduction and/or conclusion. ● Uses linking words and phrases, descriptive words, and/or temporal words to express ideas with clarity.	In **writing**, students address the prompts and provide basic development of ideas, including when drawing evidence from multiple sources, while in the majority of instances demonstrating organization that sometimes is controlled. The student: ● Develops the topic and/or narrative elements in a manner that is general in its appropriateness to the task and purpose. ● Demonstrates some organization. ● Includes some linking words and phrases, descriptive words, and/or temporal words, limiting the clarity with which ideas are expressed.	In **writing**, students address the prompts and provide minimal development of ideas, including when drawing evidence from multiple sources, while in the majority of instances demonstrating organization that often is not controlled. The student: ● Provides minimal development of the topic and/or narrative elements and is, therefore, inappropriate to the task and purpose. ● Demonstrates minimal organization. ● Includes minimal linking words and phrases, descriptive words, and/or temporal words, limiting the clarity with which ideas are expressed.

Writing Sub-Claim for Knowledge of Language and Conventions: Students demonstrate knowledge of conventions and other important elements of language.	
EVIDENCES: Students are expected to produce responses that demonstrate the skills and content listed in the evidence tables at the accuracy level and with the quality of evidence as described for students at each level.	**See Writing Evidence Table** http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents

Level 5	**Level 4**	**Level 3**	**Level 2**
A student who achieves at **Level 5 exceeds expectations** for the assessed standards.	A student who achieves at **Level 4 meets expectations** for the assessed standards.	A student who achieves at **Level 3 approaches expectations** for the assessed standards.	A student who achieves at **Level 2 partially meets expectations** for the assessed standards.
In **writing**, students demonstrate full command of the conventions of Standard Spanish consistent with edited writing. There may be some errors in grammar and usage, but overall meaning is clear.	In **writing**, students demonstrate command of the conventions of Standard Spanish consistent with edited writing. There are errors in grammar and usage that may occasionally impede understanding.	In **writing**, students demonstrate basic command of the conventions of Standard Spanish consistent with edited writing. There are few patterns of errors in grammar and usage that impede understanding, demonstrating partial control over language.	In **writing**, students demonstrate minimal command of the conventions of Standard Spanish consistent with edited writing. There are patterns of errors in grammar and usage that impede understanding, demonstrating minimal control over language.

APPENDIX C: GENERAL AGENDA

Colorado Spanish Language Arts (CSLA)

Standard Setting

Grades 3 and 4

June 27–29, 2016
Golden, Colorado

Agenda

Day 1

- Welcome
- Meeting Purpose
- Introductions
- CSLA Overview
- Overview of Standard Setting
- Experience the Test for Grade 3
- Performance Level Descriptors
- Borderline Student Descriptors
- Standard-Setting Training
- Round 1 Recommendations

Day 2

- Round 1Feedback
- Round 2 Recommendations and Feedback
- Round 3 Recommendations and Feedback
- Experience the Test for Grade 4
- Performance Level Descriptors
- Borderline Student Descriptors

Day 3

- Borderline Student Descriptors Cont'd
- Round 1Recommendations and Feedback
- Round 2 Recommendations and Feedback
- Round 3 Recommendations
- Vertical Articulation
- Evaluation

APPENDIX D: EXAMPLE PRACTICE EXERCISE RATING SHEET

Colorado Spanish Language Arts (CSLA)

Assessment: Grade 3

Panelist ID: _____

Table Number: _____

Practice Exercise Ratings

Instructions: For each item, write your item-level cut score recommendation for each performance level in the appropriate box.

When recommending an item rating for Borderline Level 3, 4, or 5, the rating must be the same or higher than the rating for the previous performance level.

				Practice Exercise Ratings			
Item	**Booklet Section**	**Item Type**	**Score Range**	**Borderline Level 2**	**Borderline Level 3**	**Borderline Level 4**	**Borderline Level 5**
Item 1	Unit 1	EBSR	0,1,2				
Item 2	Unit 1	EBSR	0,1,2				
Item 5	Unit 1	EBSR	0,1,2				
Item 6	Unit 1	EBSR	0,1,2				
Item 7.1	Unit 1	PCR (RCWE trait)	0,1,2,3				
Item 7.2	Unit 1	PCR (WKLC trait)	0,1,2,3				

APPENDIX E: EXAMPLE ROUND READINESS SURVEY

Panelist ID: _____

Instructions: Please circle your response to the following questions.

Round 1		
I understand that my task for Round 1 is to use the assessed content, my experience with CSLA students, the scoring rules and rubrics, and the borderline student descriptors to make item-level cut score recommendations. To make my recommendations, I will write my item-level scores on the ratings sheet.	**No**	**Yes**
I am ready to begin Round 1.	**No**	**Yes**

Round 2		
I understand that my task for Round 2 is to use the assessed content, my experience with CSLA students, the scoring rules and rubrics, and the borderline student descriptors to make item-level cut score recommendations. To make my recommendations, I will write my item-level scores on the ratings sheet.	**No**	**Yes**
I understand the panelist feedback data that were presented from Round 1.	**No**	**Yes**
I understand the item mean scores and score point distributions that were provided.	**No**	**Yes**
I am ready to begin Round 2.	**No**	**Yes**

Round 3		
I understand that my task for Round 3 is to use the assessed content, my experience with CSLA students, the scoring rules and rubrics, and the borderline student descriptors to make item-level cut score recommendations. To make my recommendations, I will write my item-level scores on the ratings sheet.	**No**	**Yes**
I understand the impact data that were presented from Round 2.	**No**	**Yes**
I am ready to begin Round 3.	**No**	**Yes**

APPENDIX F: AN EXAMPLE RATINGS FORM

Colorado Spanish Language Arts (CSLA)

Assessment: Grade 3

Panelist ID: _____

Table Number: _____

Round 1 Ratings

Instructions: For each item, write your item-level cut score recommendation for each performance level in the appropriate box.

When recommending an item rating for Borderline Level 3, 4, or 5, the rating must be the same or higher than the rating for the previous performance level.

				Round 1			
Item	**Booklet Section**	**Item Type**	**Score Range**	**Borderline Level 2**	**Borderline Level 3**	**Borderline Level 4**	**Borderline Level 5**
Item 1	Unit 1	EBSR	0,1,2				
Item 2	Unit 1	EBSR	0,1,2				
Item 3	Unit 1	EBSR	0,1,2				
Item 4	Unit 1	EBSR	0,1,2				
Item 5	Unit 1	EBSR	0,1,2				
Item 6	Unit 1	EBSR	0,1,2				
Item 7.1	Unit 1	PCR (RCWE trait)	0,1,2,3				
Item 7.2	Unit 1	PCR (WKLC trait)	0,1,2,3				
Item 8	Unit 1	EBSR	0,1,2				
Item 9	Unit 1	EBSR	0,1,2				
Item 10	Unit 1	EBSR	0,1,2				
Item 11	Unit 1	EBSR	0,1,2				
Item 12	Unit 2	EBSR	0,1,2				
Item 13	Unit 2	EBSR	0,1,2				
Item 14	Unit 2	EBSR	0,1,2				
Item 15	Unit 2	EBSR	0,1,2				
Item 16	Unit 2	EBSR	0,1,2				

Item 17	Unit 2	EBSR	0,1,2				
Item 18.1	Unit 2	PCR (RCWE trait)	0,1,2,3				
Item 18.2	Unit 2	PCR (WKLC trait)	0,1,2,3				
Item 19	Unit 3	EBSR	0,1,2				
Item 20	Unit 3	EBSR	0,1,2				
Item 21	Unit 3	EBSR	0,1,2				
Item 22	Unit 3	EBSR	0,1,2				
Item 23.1	Unit 3	NPCR (WE trait)	0,1,2,3				
Item 23.2	Unit 3	NPCR (WKLC trait)	0,1,2,3				
Item 24	Unit 3	EBSR	0,1,2				
Item 25	Unit 3	EBSR	0,1,2				
Item 26	Unit 3	EBSR	0,1,2				
Item 27	Unit 3	EBSR	0,1,2				
Item 28	Unit 3	EBSR	0,1,2				
Item 29	Unit 3	EBSR	0,1,2				

APPENDIX G: STANDARD SETTING EVALUATION

Colorado Spanish Language Arts (CSLA)
Standard Setting Evaluation Form

The purpose of this evaluation form is to collect information about your experience in recommending performance cut scores for CSLA. Your opinions provide an important part of our evaluation of this meeting. Please do not write your name on this evaluation form as we want your comments to be anonymous. Thank you for your willingness to participate in this survey.

Indicate your response by checking the appropriate box.

	Do not support	**Support with some reservation**	**Moderately support**	**Strongly support**
1. To what degree do you support the recommended cut score for Grade 3 "Level 2: Partially Met Expectations?"	☐	☐	☐	☐
	0%	0%	10%	90%
If you cannot support, please explain why not:				
2. To what degree do you support the recommended cut score for Grade 3 "Level 3: Approached Expectations?"	☐	☐	☐	☐
	0%	0%	10%	90%
If you cannot support, please explain why not:				
3. To what degree do you support the recommended cut score for Grade 3 "Level 4: Met Expectations?"	☐	☐	☐	☐
	0%	0%	30%	70%
If you cannot support, please explain why not:				
4. To what degree do you support the recommended cut score for Grade 3 "Level 5: Exceeded Expectations?"	☐	☐	☐	☐
	0%	0%	20%	80%
If you cannot support, please explain why not:				
5. To what degree do you support the recommended cut score for Grade 4 "Level 2: Partially Met Expectations?"	☐	☐	☐	☐
	0%	0%	20%	80%
If you cannot support, please explain why not:				
6. To what degree do you support the recommended cut score for Grade 4 "Level 3: Approached Expectations?"	☐	☐	☐	☐
	0%	0%	10%	90%

If you cannot support, please explain why not:

7. To what degree do you support the recommended cut score for Grade 4 "Level 4: Met Expectations?"	☐	☐	☐	☐
	0%	0%	20%	80%

If you cannot support, please explain why not:

8. To what degree do you support the recommended cut score for Grade 4 "Level 5: Exceeded Expectations?"	☐	☐	☐	☐
	0%	0%	20%	80%

If you cannot support, please explain why not:

	Way too low	**A bit low**	**Appropriate**	**A bit high**	**Way too high**
9. The recommended cut score for Grade 3 "Level 2" is:	☐	☐	☐	☐	☐
	0%	0%	100%	0%	0%
10. The recommended cut score for Grade 3 "Level 3" is:	☐	☐	☐	☐	☐
	0%	0%	100%	0%	0%
11. The recommended cut score for Grade 3 "Level 4" is:	☐	☐	☐	☐	☐
	0%	10%	90%	0%	0%
12. The recommended cut score for Grade 3 "Level 5" is:	☐	☐	☐	☐	☐
	0%	0%	100%	0%	0%
13. The recommended cut score for Grade 4 "Level 2" is:	☐	☐	☐	☐	☐
	0%	0%	100%	0%	0%
14. The recommended cut score for Grade 4 "Level 3" is:	☐	☐	☐	☐	☐
	0%	0%	100%	0%	0%
15. The recommended cut score for Grade 4 "Level 4" is:	☐	☐	☐	☐	☐
	0%	10%	90%	0%	0%
16. The recommended cut score for Grade 4 "Level 5" is:	☐	☐	☐	☐	☐
	0%	10%	90%	0%	0%

	Strongly Disagree	Disagree	Agree	Strongly Agree
17. The Modified Extended Angoff Method was explained clearly by the group facilitator.	☐	☐	☐	☐
	0%	40%	20%	40%
18. I had a solid understanding of what the test was intended to measure.	☐	☐	☐	☐
	0%	0%	40%	60%
19. I could clearly distinguish between performance levels.	☐	☐	☐	☐
	0%	0%	60%	40%
20. After the first round of recommendations, I felt comfortable with the standard setting procedure.	☐	☐	☐	☐
	0%	20%	40%	40%
21. I found the feedback on the comparison of all panelists' recommendations to be useful in standard setting.	☐	☐	☐	☐
	0%	10%	40%	50%
22. I found the item mean score information to be useful in standard setting.	☐	☐	☐	☐
	0%	0%	40%	60%
23. I found the score point distribution information to be useful in standard setting.	☐	☐	☐	☐
	0%	0%	50%	50%
24. I found the feedback on the percentage of the students tested that would be classified at each performance level to be useful in standard setting.	☐	☐	☐	☐
	0%	0%	10%	90%
25. Table and group discussions were open and honest.	☐	☐	☐	☐
	0%	0%	20%	80%
26. I believe that my opinions were considered and valued by my group.	☐	☐	☐	☐
	0%	0%	10%	90%
27. The facilitator led the group through the standard setting process without imposing ideas about where cut scores should be.	☐	☐	☐	☐
	0%	10%	40%	50%
28. The facilitator led the group through the vertical articulation process without imposing ideas about where cut scores should be.	☐	☐	☐	☐
	0%	0%	40%	60%

29. I am confident that the final cut score recommendations reflect the performance level descriptors associated with CSLA.	☐ 0%	☐ 0%	☐ 40%	☐ 60%
30. I am confident that the final cut score recommendations reflect high expectations consistent with the Colorado Academic Standards.	☐ 0%	☐ 0%	☐ 30%	☐ 70%

Please use the back of this page to provide any additional comments.

APPENDIX E: CSLA SAMPLE SCORE REPORTS

Colorado Measures of Academic Success

Spring 2016

FIRSTNAME M. LASTNAME

ID: 5200154001 **Grade: 3**

SAMPLE DISTRICT NAME

SAMPLE SCHOOL NAME

GRADE 3 CSLA

Colorado Spanish Language Arts Assessment Report

This score report provides information about your student's performance on the Colorado Spanish Language Arts Assessment (CSLA).

- Your student's performance is represented by a scale score and a performance level so that you can see your student's achievement of the grade-level or course-level Colorado Academic Standards at the end of the year.
- School, district, and state information is provided so that you can compare your student's performance to the performance of others.
- Page 2 of the report provides a breakdown of your student's performance on specific skill sets so you can see where your student is excelling or may need improvement. Arrows are included that compare your student's performance to the performance of other students.

How Did FIRSTNAME Perform Overall?

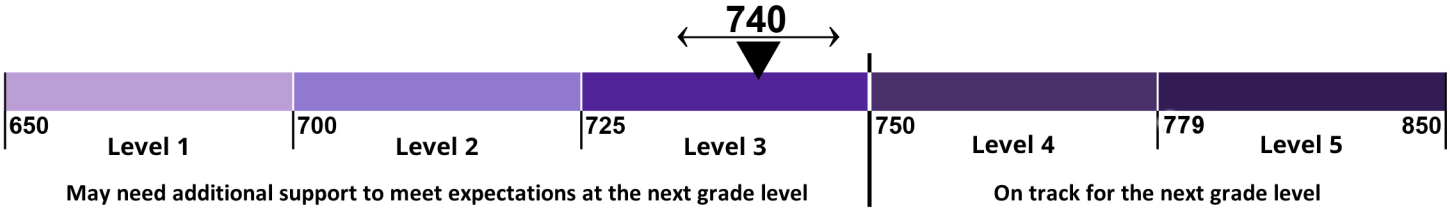
Performance Level 3

Score: **740**

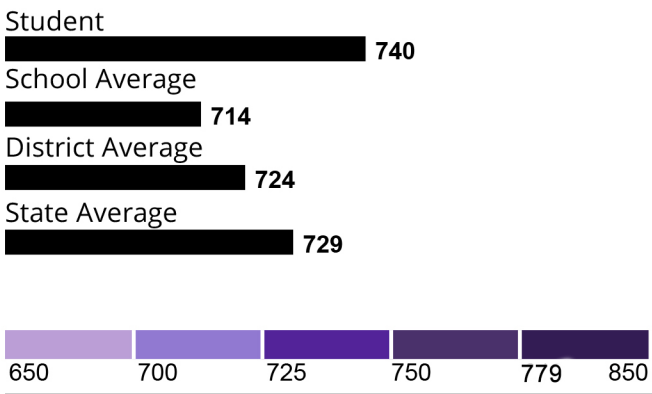
CO Percentile Rank: **75th**

Level 5 Exceeded Expectations
Level 4 Met Expectations
Level 3 Approached Expectations
Level 2 Partially Met Expectations
Level 1 Did Not Yet Meet Expectations

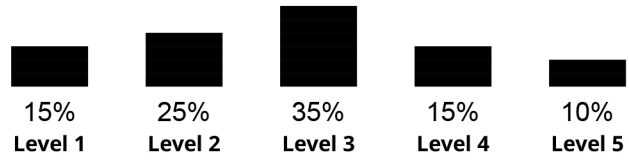
Your Student's Score



The probable range in the student's overall score on this test is plus or minus 7.3 points. This is the amount of change that would be expected in your student's score if he/she were to take the test many times. Arrows beneath your student's score represent the probable range.



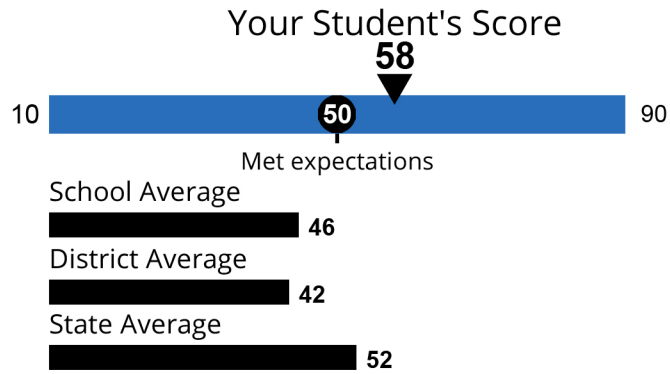
How Students in Colorado Performed



Percentage of students at each performance level

How Did Your Student Perform in Reading and Writing?

READING



LITERARY TEXT

Your student performed about the same as students who **met or exceeded expectations.** Students meet expectations by showing they can read and analyze fiction, drama, and poetry.

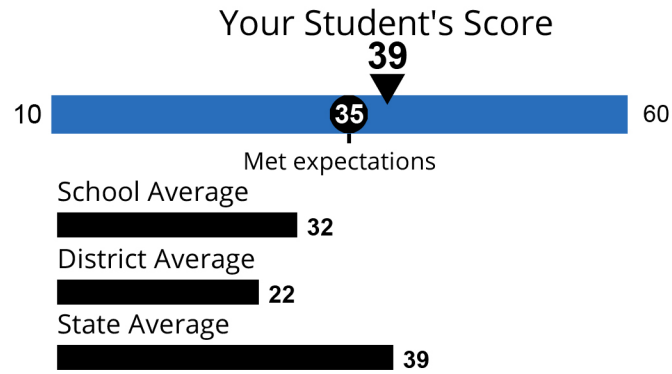
INFORMATIONAL TEXT

Your student performed about the same as students who **approached expectations.** Students meet expectations by showing they can read and analyze nonfiction, history, science, and the arts.

VOCABULARY

Your student performed about the same as students who **did not yet meet or partially met expectations.** Students meet expectations by showing they can use context to determine what words and phrases mean.

WRITING



WRITING EXPRESSION

Your student performed about the same as students who **approached expectations.** Students meet expectations by showing they can compose well-developed writing, using details from what they have read.

KNOWLEDGE AND USE OF LANGUAGE CONVENTIONS

Your student performed about the same as students who **met or exceeded expectations.** Students meet expectations by showing they can compose writing using rules of standard Spanish.

LEGEND

Your student performed about the same as students who:

- Met or Exceeded Expectations
- Approached Expectations
- Did Not Yet Meet or Partially Met Expectations

The Colorado Measures of Academic Success, or CMAS, is a series of state tests administered to students in the content areas of English language arts, math, science, and social studies. Spanish-speaking students in grades 3 and 4 who meet established eligibility criteria may take the CSLA in place of the English language arts assessment. These tests are aligned to the Colorado Academic Standards, which set high expectations for all students in Colorado to help ensure readiness for college or careers after high school graduation.

This test was designed to measure complex skills, like critical-thinking and problem solving. It allows parents and teachers to see how well their students are doing compared to other students in the state, and in some cases, at the school and district level.

Colorado Measures of Academic Success
DISTRICT PERFORMANCE LEVEL SUMMARY
Spring 2016

CONFIDENTIAL - DO NOT DISTRIBUTE

SAMPLE DISTRICT NAME

COLORADO SPANISH LANGUAGE ARTS
Grade 3 Assessment

Purpose: This report describes group achievement in terms of performance levels and average scale scores.

	Number of Valid Scores	Average Scale Score	Performance Levels										≥ Level 4 Met or Exceeded Expectations		No Scores Reported	Total Number of Students
			Level 1 Did Not Yet Meet Expectations		Level 2 Partially Met Expectations		Level 3 Approached Expectations		Level 4 Met Expectations		Level 5 Exceeded Expectations					
			#	%	#	%	#	%	#	%	#	%	#	%	#	#
State	**63,246**	**606**	**18,891**	**29.9%**	**15,001**	**23.7%**	**15,564**	**24.6%**	**11,496**	**18.2%**	**2,294**	**3.6%**	**13,790**	**21.8%**	**2,221**	**65,467**
District	44	616	7	15.9%	12	27.3%	16	36.7%	9	20.5%	0	0.0%	9	20.5%	3	47
Gender																
Female	21	632	3	14.3%	4	19.0%	8	38.1%	6	28.6%	0	0.0%	6	28.6%	1	22
Male	23	301	4	17.4%	7	30.4%	9	39.1%	3	13.0%	0	0.0%	3	13.0%	2	25
Ethnicity/Race																
Hispanic or Latino	12	648	0	0.0%	3	25.0%	5	41.7%	4	33.3%	0	0.0%	4	33.3%	2	14
American Indian or Alaska Native	0	0	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0
Asian	1	705	0	0.0%	0	0.0%	0	0.0%	1	100.0%	0	0.0%	1	100.0%	0	1
Black or African-American	8	557	2	25.0%	3	37.5%	3	37.5%	0	0.0%	0	0.0%	0	0.0%	1	9
Native Hawaiian or Other Pacific Islander	1	775	0	0.0%	0	0.0%	0	0.0%	1	100.0%	0	0.0%	1	100.0%	0	1
White	17	597	4	23.5%	6	35.3%	5	29.4%	2	11.8%	0	0.0%	2	11.8%	3	20
Two or more races	5	646	0	0.0%	0	0.0%	4	80.0%	1	20.0%	0	0.0%	1	20.0%	0	5
Not Indicated	0	0	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0
Economic Disadvantage																
Free/Reduced Lunch Eligible	23	301	4	17.4%	7	30.4%	9	39.1%	3	13.0%	0	0.0%	3	13.0%	2	25
Not Eligible for Free/Reduced Lunch	21	632	3	14.3%	4	19.0%	8	38.1%	6	28.6%	0	0.0%	6	28.6%	3	24

mmddyyyy-Batch-1234-5678-1234567

Colorado Measures of Academic Success
STUDENT ROSTER
Spring 2016

SAMPLE SCHOOL NAME
SAMPLE DISTRICT NAME

COLORADO SPANISH LANGUAGE ARTS
Grade 3 Assessment

STUDENT	CSLA OVERALL SCORE	READING* SCORE	READING* LITERARY	READING* INFORMATION	READING* VOCABULARY	WRITING* SCORE	WRITING* EXPRESSION	WRITING* CONVENTIONS
STATE AVERAGE	746	43	36 21 43	24 63 13	33 21 46	51	38 40 22	51 19 30
DISTRICT AVERAGE	750	37	13 58 71	24 20 56	35 35 30	47	36 17 48	25 38 37
SCHOOL AVERAGE	734	43	34 42 24	46 37 17	29 60 11	51	30 40 30	45 42 13
ALASTNAME, FIRSTNAME M.	751	28	↑	↔	↑	69	↔	↑
BLASTNAME, FIRSTNAME M.	720	28	↑	↔	↑	69	↔	↑
CLASTNAME, FIRSTNAME M.	746	44	↓	↔	↔	55	↑	↔
DLASTNAME, FIRSTNAME M.	713	37	↑	↑	↑	62	↑	↑
ELASTNAME, FIRSTNAME M.	794	28	↓	↑	↔	69	↓	↔
FLASTNAME, FIRSTNAME M.	698	44	↓	↓	↔	55	↔	↓
GLASTNAME, FIRSTNAME M.	724	37	↓	↓	↓	62	↓	↔
HLASTNAME, FIRSTNAME M.	N/A							
ILASTNAME, FIRSTNAME M.	830	28	↑	↔	↑	69	↑	↑
JLASTNAME, FIRSTNAME M.	661	44	↓	↔	↔	55	↓	↓
KLASTNAME, FIRSTNAME M.	726	28	↓	↓	↓	69	↓	↓

1 Did Not Yet Meet Expectations (650-699)	2 Partially Met Expectations (700-724)	3 Approached Expectations (725-749)	4 Met Expectations (750-778)	5 Exceeded Expectations (779-850)

↑ Met or Exceeded Expectations
↔ Approached Expectations
↓ Did Not Yet Meet or Partially Met Expectations

* Numbers are percentages

Colorado Measures of Academic Success
District Evidence Statement Analysis
Spring 2016

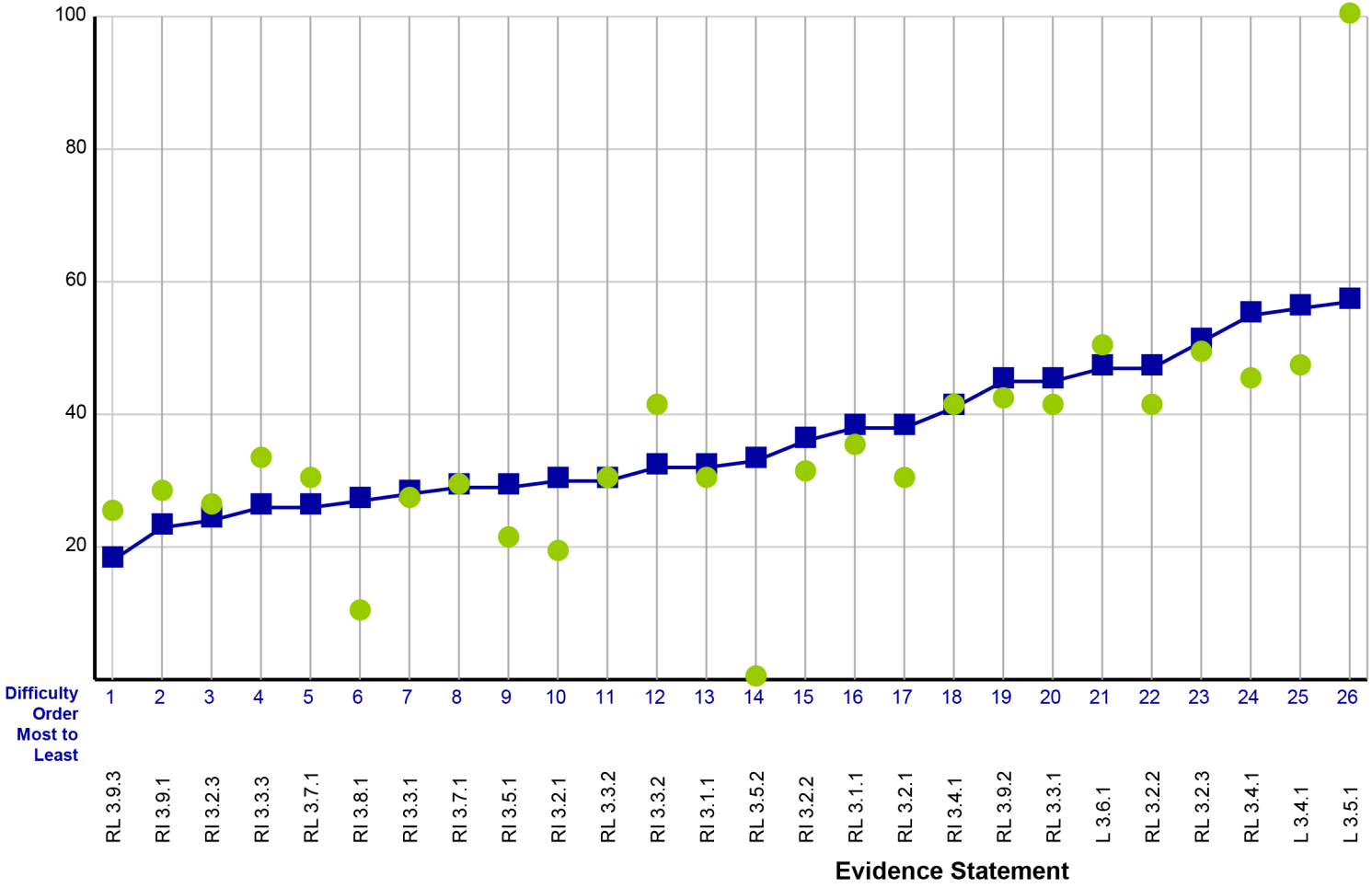
SAMPLE DISTRICT NAME

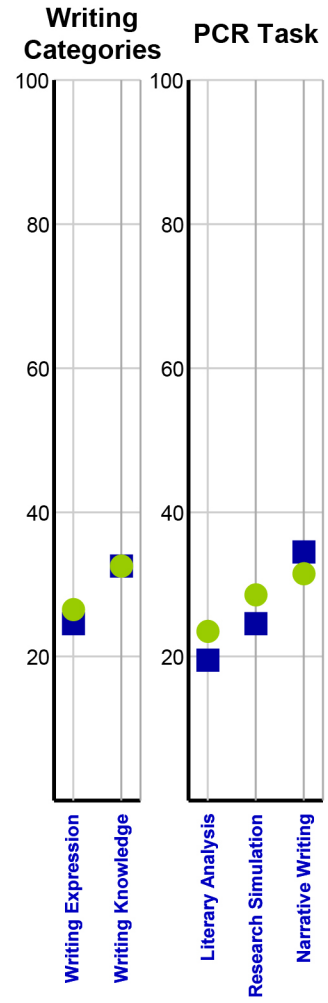
COLORADO SPANISH LANGUAGE ARTS
Grade 3 Assessment

State
District

Students with Valid Scores (104)

Purpose: This report presents the average percent correct by Evidence Statement for district and state.





Difficulty level is determined at the State level for all reports.

Evidence Statement Analysis
Spring 2016

This report shows the operational Evidence Statements for the given grade and subject sorted by difficulty

COLORADO SPANISH LANGUAGE ARTS

Grade 3 Assessment

Difficulty Order Most to Least	Evidence Statement	Colorado Academic Standard(s)	Domain
1	RL 3.9.3	3.2.1.c.iii	Reading: Literature
2	RI 3.9.1	3.2.2.c.iii	Reading: Informational Text
3	RI 3.2.3	3.2.2.a.ii	Reading: Informational Text
4	RI 3.3.3	3.2.2.a.iii	Reading: Informational Text
5	RL 3.7.1	3.2.1.c.i	Reading: Literature
6	RI 3.8.1	3.2.2.c.ii	Reading: Informational Text
7	RI 3.3.1	3.2.2.a.iii	Reading: Informational Text
8	RI 3.7.1	3.2.2.c.i	Reading: Informational Text
9	RI 3.5.1	3.2.2.b.ii	Reading: Informational Text
10	RI 3.2.1	3.2.2.a.ii	Reading: Informational Text
11	RL 3.3.2	3.2.1.a.v	Reading: Literature
12	RI 3.3.2	3.2.2.a.iii	Reading: Informational Text
13	RI 3.1.1	3.2.2.a.i	Reading: Informational Text
14	RL 3.5.2	3.2.1.b.iii	Reading: Literature
15	RI 3.2.2	3.2.2.a.ii	Reading: Informational Text
16	RL 3.1.1	3.2.1.a.i	Reading: Literature
17	RL 3.2.1	3.2.1.a.iii	Reading: Literature
18	RI 3.4.1	3.2.2.b.i	Reading: Informational Text
19	RL 3.9.2	3.2.1.c.iii	Reading: Literature
20	RL 3.3.1	3.2.1.a.v	Reading: Literature
21	L 3.6.1	3.2.3.e	Language
22	RL 3.2.2	3.2.1.a.iii	Reading: Literature
23	RL 3.2.3	3.2.1.a.iii	Reading: Literature
24	RL 3.4.1	3.2.1.b.i	Reading: Literature
25	L 3.4.1	3.2.3.c	Language
26	L 3.5.1	3.2.3.d	Language

Evidence Statements: http://www.parcconline.org/assessments/test-design/ela-literacy/test-specifications-documents

Colorado Academic Standards: http://www.cde.state.co.us/standardsandinstruction/standardsresourcesk12