Colorado Spanish Language Arts

Technical Report

2017

Table of Contents

PART I: HISTORICAL OVERVIEW AND SUMMARY OF PROCESSES

CHAPTER 1: INTRODUCTION AND BACKGROUND

All public school students enrolled in Colorado are required by state law to take a standards-based 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 2017 CSLA assessments, including content, assessment development, administration, scoring, and technical attributes.

The Spring 2017 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 2017 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 learners (ELs) 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 ELs 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 EL 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: https://www.cde.state.co.us/assessment/csla_flowchart_17-18

Description of CSLA

CSLA is 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 ELs and educators who are familiar with the instruction and needs of the students in an English language development program. In addition to 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 Partner, The Colorado Education Initiative
- Dr. Jonathan Dings, Executive Director of Student Assessment and Program Evaluation, Boulder Valley School District
- Dr. Lisa Escarcega, Executive Director, Colorado Association of School Executives
- Dr. Michael Kolen, 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.

CSLA passage and item development is 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 includes the following steps:

Passage Development

Using the CMAS: PARCC ELA Passage Selection Guidelines, Tri-Lin Spanish language arts content specialists and assessment developers are trained to develop appropriate passages that meet 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 applies 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 conduct 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 are analyzed and rated for text complexity prior to item writing as readily accessible, moderately complex, or very complex.

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

After the CSLA passages are approved by CDE, Tri-Lin begins the item development process.

Item Development

After the passages are approved by CDE, Spanish item writers are trained and begin developing items. The CMAS: PARCC Item Guidelines for ELA/Literacy Summative Assessment and the Cognitive Complexity Framework guide item development to ensure that text complexity and item/task complexity interact to determine the overall complexity of a task.

Three main sources of item complexity are 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 are written, Tri-Lin's team of content specialists, assessment developers, and editors conduct 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 ensure 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 conducts 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 are completed, the items are reviewed and approved for presentation to the CDE by the lead assessment specialist. Prior to the educator committee reviews, CDE reviews and approves the CSLA items.

Educator Content and Bias Review Meetings

CDE experts, Colorado educators, and postsecondary faculty with diverse backgrounds from across the state conduct 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 are conducted either in person or virtually and includes group training on the expectations and processes of each meeting, followed by breakout groupings into grade/subject working committees where additional training is provided.

The committee members are trained and instructed to verify that each passage and item:

- Uses clear, unambiguous, and grade-level appropriate language
- Avoids complex sentence structure
- Has one correct answer
- Contains plausible distractors
- Represents the range of cognitive complexities and include challenging items for students performing at all levels

- Is appropriate for students in the assigned grade in terms of reading level, vocabulary, interest, and experience
- Has scoring guidelines that capture exemplar responses at each score point
- Includes appropriate and clear graphics that are relevant
- Is free of ethnic, gender, political, and religious bias

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

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

Data Review

After development of the CSLA items, selected items are placed on the operational test in embedded field-test positions. 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 are experts in bilingual instruction and ELs at grades 3 and 4 are convened to review the newly developed items along with the student performance data. The data review committee members are provided passages, item images, and content metadata along with classical statistics and Differential Item Functioning (DIF) statistics to review.

The classical statistics include 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 are used to classify items as having either negligible, moderate, or large DIF. Items that are classified as having moderate or large DIF are reviewed by the data review committee.

During the data review meeting, educators are trained to interpret the statistical information, and while the committee use the data as a tool to inform their judgments, the committee is instructed not to base their final assessment of the appropriateness or fairness of items solely on these data. Committee members review each item and make a recommendation as to whether to "accept" or "reject" the item.

Following the data review meeting, a Data Review Reconciliation meeting is held which includes CDE, Tri-Lin, and Pearson staff. At the reconciliation meeting, the assessment specialists and psychometricians discuss the committee comments from the data review meeting as well as any concerns they have about the items. After the item outcomes are finalized during reconciliation, field test items that are accepted are re-classified in the item bank as available for

use on future operational assessments. Items that are rejected are re-classified to eliminate them from use on a test. These items may be modified and field tested again on future test forms.

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 CMAS 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 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 educators with relevant grade level and 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 2017 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 ELs 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 development program. In addition to 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. The full CSLA standard setting report can be found in the *Spring 2016 CSLA Technical Report*.

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 2017 Score Interpretive Guide at http://www.cde.state.co.us/assessment/2017cmascoaltinterpretiveguide.

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 D.

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:

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

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. The Rasch scale can then be transformed to create scale scores to allow for the interpretation of test scores. The RPCM and Winsteps (Linacre, 2011) were used for all equating analyses.

Equating and Scaling

The fixed common items approach was used to equate the Spring 2017 CSLA assessments to the Spring 2016 operational scales. The operational items used to equate the 2017 assessments to the 2016 scales are called anchor items. The anchor items are a set of common items that are placed on forms from adjacent administrations. This set of items represents the CSLA blueprint in terms of content and item types and represents approximately 30% of a full form. To obtain equated Rasch parameter estimates for the Spring 2017 assessments, anchor item parameter estimates were fixed to their 2016 parameter estimates before calibrating the remaining non-anchor operational items. This method placed the non-anchor operational items on the same scale as the anchor items.

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

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
  o 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
  o 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 2017 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 2,500 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. Fifty-five 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		**1701**	**737**	**24.5**	**656**	**848**	**0.90**
Gender	Female	871	740	22.3	667	848	0.89
	Male	830	733	24.1	656	805	0.91
Ethnicity	American Indian	2	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	0	-	-	-	-	-
	Hispanic/Latino	1695	737	24.0	656	848	0.90
	White	3	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	1	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	120	741	24.5	675	798	0.90
	Economically Disadvantaged	1581	736	23.4	656	848	0.90
Students With Disabilities	504	8	-	-	-	-	-
	IEP	126	714	20.7	656	779	0.85
	No and Missing	1567	738	22.7	667	848	0.89
Reading Score		**1701**	**44**	**8.5**	**19**	**84**	**0.89**
Gender	Female	871	45	8.2	19	84	0.88
	Male	830	43	8.8	19	71	0.90
Ethnicity	American Indian	1	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	0	-	-	-	-	-
	Hispanic/Latino	1695	44	8.6	19	84	0.89
	White	3	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	1	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	120	45	8.5	26	67	0.89
	Economically Disadvantaged	1581	44	8.6	19	84	0.89
Students with Disabilities	504	8	-	-	-	-	-
	IEP	126	37	6.8	19	71	0.80
	No and Missing	1567	45	8.4	19	84	0.89

Group Type	Subgroup	N	Mean	SD	Min	Max	Alpha
Writing Score		**1701**	**32**	**8.7**	**10**	**60**	**0.76**
Gender	Female	871	33	8.1	10	60	0.73
	Male	830	30	9.1	10	57	0.79
Ethnicity	American Indian	2	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	0	-	-	-	-	-
	Hispanic/Latino	1695	32	8.7	10	60	0.76
	White	3	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	1	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	120	33	9.5	10	57	0.79
	Economically Disadvantaged	1581	31	8.7	10	60	0.76
Students with Disabilities	504	8	-	-	-	-	-
	IEP	126	23	9.7	10	43	0.77
	No and Missing	1235	32	8.2	10	60	0.75

Table 2. Grade 4 Performance by Subgroups

Group Type	Subgroup	N	Mean	SD	Min	Max	Alpha
Total Score		**835**	**726**	**21.2**	**668**	**794**	**0.87**
Gender	Female	400	732	20.8	674	794	0.86
	Male	435	721	20.1	6668	779	0.86
Ethnicity	American Indian	1	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	0	-	-	-	-	-
	Hispanic/Latino	827	726	21.2	668	794	0.87
	White	5	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	0	-	-	-	-	-
	Not Indicated	2	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	66	725	19.7	679	764	0.86

Group Type	Subgroup	N	Mean	SD	Min	Max	Alpha
	Economically Disadvantaged	769	726	21.4	668	794	0.87
Students With Disabilities	504	7	-	-	-	-	-
	IEP	75	704	16.7	668	744	0.74
	No and Missing	753	728	20.4	674	794	0.86
Reading Score		**835**	**41**	**7.7**	**21**	**66**	**0.85**
Gender	Female	400	42	7.6	24	66	0.85
	Male	435	39	7.4	21	63	0.84
Ethnicity	American Indian	1	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	0	-	-	-	-	-
	Hispanic/Latino	827	41	7.7	21	66	0.85
	White	5	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	0	-	-	-	-	-
	Not Indicated	2	-	-	-	-	-
Economic Status	Not Economically Disadvantaged	66	41	7.4	26	60	0.85
	Economically Disadvantaged	769	41	7.7	21	66	0.85
Students with Disabilities	504	7	-	-	-	-	-
	IEP	75	33	5.4	21	47	0.62
	No and Missing	753	41	7.5	21	66	0.85
Writing Score		**835**	**28**	**8.4**	**10**	**52**	**0.72**
Gender	Female	400	30	7.8	10	52	0.71
	Male	435	26	8.3	10	47	0.68
Ethnicity	American Indian	1	-	-	-	-	-
	Asian	0	-	-	-	-	-
	Black or African American	0	-	-	-	-	-
	Hispanic/Latino	827	28	8.4	10	52	0.72
	White	5	-	-	-	-	-
	Native Hawaiian or other Pacific Islander	0	-	-	-	-	-
	Two or More Races	0	-	-	-	-	-
	Not Indicated	2	-	-	-	-	-

Group Type	Subgroup	N	Mean	SD	Min	Max	Alpha
Economic Status	Not Economically Disadvantaged	66	27	8.3	10	44	0.70
	Economically Disadvantaged	769	28	8.4	10	52	0.72
Students with Disabilities	504	7	-	-	-	-	-
	IEP	75	20	8.7	10	37	0.61
	No and Missing	753	29	8.0	10	52	0.71

Table 3. SEMs

Grade	Total Score SEM	Reading SEM	Writing SEM
3	7.6	2.9	4.2
4	7.8	3.0	4.4

Table 4. Subclaim Performance Summary

Grade	Subclaim	Max Possible Score	Mean	SD	Min	Max	Alpha
3	Reading Literary	25	10	5.7	0	25	0.78
	Reading Informational	21	7	4.6	0	21	0.75
	Reading Vocabulary	12	6	3.0	0	12	0.61
	Written Expression	27	10	5.8	0	27	0.65
	Writing Knowledge and Language Conventions	9	4	2.7	0	9	0.79
4	Reading Literary	26	10	5.1	0	25	0.75
	Reading Informational	26	7	4.3	0	24	0.65
	Reading Vocabulary	12	5	2.8	0	12	0.54
	Written Expression	33	9	6.4	0	30	0.61
	Writing Knowledge and Language Conventions	9	3	2.9	0	9	0.79

Table 5. Grade 3 Classical Statistics

ITEM	TYPE	% 0	% 1	% 2	% 3	% Omit	MEAN SCORE	ITEM-TOTAL CORR
1	EBSR	34.3	8.8	56.6		0.3	1.220	0.496
2	EBSR	38.7	43.9	17.2		0.2	0.784	0.374
3	EBSR	37.3	15.6	46.7		0.4	1.091	0.558
4	EBSR	37.6	15.9	46.1		0.3	1.082	0.485
5	EBSR	68.2	10.9	20.3		0.6	0.516	0.349
6	EBSR	38.8	22.7	37.9		0.6	0.985	0.506
7	CR	33.5	35.0	27.7	1.7	2.2	0.955	0.622
8	CR	25.0	41.2	17.1	14.5	2.2	1.188	0.569
9	EBSR	42.9	22.1	31.7		3.4	0.855	0.524
10	EBSR	44.3	10.9	40.7		4.2	0.922	0.598
11	EBSR	43.6	12.3	39.4		4.6	0.912	0.554
12	EBSR	46.7	14.9	33.3		5.1	0.816	0.524
13	EBSR	22.5	31.5	45.9		0.2	1.232	0.452
14	EBSR	37.6	15.9	46.1		0.4	1.081	0.614
15	EBSR	54.0	15.4	30.3		0.3	0.760	0.496
16	EBSR	58.1	20.7	20.9		0.4	0.624	0.294
17	EBSR	41.9	28.7	28.8		0.6	0.864	0.433
18	EBSR	50.7	14.8	34.0		0.5	0.828	0.510
19	CR	26.1	32.2	38.9	1.6	1.2	1.148	0.699
20	CR	25.1	36.7	16.9	20.0	1.2	1.307	0.612
21	EBSR	50.0	14.8	34.7		0.5	0.843	0.396
22	EBSR	43.6	13.1	43.2		0.2	0.994	0.522
23	EBSR	60.0	18.8	21.0		0.2	0.608	0.373
24	EBSR	54.3	25.1	20.2		0.4	0.655	0.375
25	CR	27.4	38.7	29.4	3.6	0.8	1.085	0.661
26	CR	24.1	36.0	11.6	27.4	0.8	1.415	0.603
27	EBSR	56.3	18.5	23.8		1.4	0.660	0.416
28	EBSR	63.7	16.2	18.5		1.7	0.531	0.391

ITEM	TYPE	% 0	% 1	% 2	% 3	% Omit	MEAN SCORE	ITEM-TOTAL CORR
29	EBSR	66.0	17.5	14.9		1.5	0.474	0.341
30	EBSR	48.6	26.9	22.6		1.8	0.722	0.479
31	EBSR	73.5	7.6	16.9		1.9	0.415	0.391
32	EBSR	55.6	26.2	16.3		1.9	0.588	0.375

Table 6. Grade 4 Classical Statistics

ITEM	TYPE	% 0	% 1	% 2	% 3	% 4	% Omit	MEAN SCORE	ITEM-TOTAL CORR
1	EBSR	44.7	21.3	33.9			0.1	0.891	0.443
2	EBSR	46.7	29.5	23.7			0.1	0.769	0.488
3	EBSR	34.9	27.8	37.2			0.1	1.023	0.451
4	EBSR	54.5	12.8	32.7			0	0.782	0.431
5	EBSR	73.9	16.5	9.3			0.2	0.352	0.340
6	EBSR	43.1	32.2	24.3			0.4	0.808	0.467
7	CR	29.5	31.5	29.2	4.8	1.3	3.7	1.096	0.691
8	CR	45.9	20.0	12.8	17.6		3.7	0.984	0.580
9	EBSR	47.7	12.9	34.9			4.6	0.826	0.354
10	EBSR	50.7	23.8	20.0			5.5	0.638	0.341
11	EBSR	47.3	30.8	15.9			6.0	0.626	0.331
12	EBSR	62.6	18.8	12.5			6.1	0.437	0.288
13	EBSR	69.2	17.7	13.1			0	0.438	0.329
14	EBSR	69.6	10.5	19.8			0.1	0.501	0.272
15	EBSR	55.2	24.9	19.6			0.2	0.642	0.287
16	EBSR	50.7	26.1	22.9			0.4	0.719	0.444
17	EBSR	61.9	20.0	17.2			0.8	0.545	0.261
18	EBSR	68.4	17.6	13.5			0.5	0.447	0.148
19	EBSR	51.5	21.2	26.9			0.4	0.751	0.420
20	EBSR	70.9	20.0	8.3			0.8	0.365	0.231

ITEM	TYPE	% 0	% 1	% 2	% 3	% 4	% Omit	MEAN SCORE	ITEM-TOTAL CORR
21	CR	29.1	26.1	29.9	10.8	1.6	2.5	1.246	0.678
22	CR	46.0	17.2	16.4	17.8		2.5	1.036	0.577
23	EBSR	31.5	18.7	49.5			0.4	1.176	0.430
24	EBSR	30.8	23.2	45.6			0.4	1.145	0.496
25	EBSR	35.8	21.1	42.8			0.4	1.066	0.508
26	EBSR	30.8	29.6	39.4			0.2	1.084	0.474
27	CR	57.4	25.5	13.5	2.0		1.6	0.587	0.602
28	CR	51.6	23.5	9.9	13.4		1.6	0.836	0.499
29	EBSR	43.1	38.3	17.7			0.8	0.738	0.465
30	EBSR	65.0	21.0	12.9			1.1	0.468	0.387
31	EBSR	59.9	24.0	14.9			1.3	0.537	0.263
32	EBSR	41.3	23.6	33.1			2.0	0.897	0.485
33	EBSR	41.8	31.1	25.3			1.8	0.817	0.497
34	EBSR	65.7	21.9	10.9			1.4	0.437	0.254

Table 7. Grade 3 Item Parameter Estimates

ITEM	TYPE	B	D1	D2	D3	D4	INFIT	OUTFIT
1	EBSR	-0.7121	0	1.4121	-1.4121		0.98	1.01
2	EBSR	0.1525	0	-0.7696	0.7696		1.07	1.07
3	EBSR	-0.5164	0	0.7692	-0.7692		0.91	0.88
4	EBSR	-0.5746	0	0.3322	-0.3322		1.09	1.08
5	EBSR	0.3508	0	1.0613	-1.0613		1.11	1.32
6	EBSR	-0.3640	0	0.3838	-0.3838		0.98	0.99
7	EBSR	-0.1053	0	0.3271	-0.3271		0.95	0.93
8	EBSR	-0.2850	0	0.9059	-0.9059		0.88	0.83
9	EBSR	-0.1933	0	1.0029	-1.0029		0.93	0.93
10	EBSR	-0.0608	0	0.7906	-0.7906		0.95	0.97
11	EBSR	-0.8181	0	-0.1996	0.1996		0.98	0.99
12	EBSR	-0.5867	0	0.6919	-0.6919		0.83	0.77
13	EBSR	0.0419	0	0.7515	-0.7515		0.99	1.02
14	EBSR	0.3229	0	0.1189	-0.1189		1.24	1.31
15	EBSR	-0.1010	0	-0.1372	0.1372		1.08	1.11
16	EBSR	-0.1605	0	0.8103	-0.8103		0.96	0.94
17	EBSR	-0.1072	0	0.8167	-0.8167		1.16	1.23
18	EBSR	-0.3544	0	0.9800	-0.9800		0.98	0.94
19	EBSR	0.3520	0	0.4025	-0.4025		1.12	1.22
20	EBSR	0.3013	0	0.0462	-0.0462		1.12	1.21
21	EBSR	0.2376	0	0.4636	-0.4636		1.08	1.16
22	EBSR	0.4992	0	0.5268	-0.5268		1.06	1.15
23	EBSR	0.6500	0	0.3623	-0.3623		1.11	1.27
24	EBSR	0.1711	0	-0.0085	0.0085		0.98	1.01
25	EBSR	0.6704	0	1.3082	-1.3082		1.02	1.35
26	EBSR	0.4680	0	-0.0840	0.0840		1.10	1.18
27	CR	1.0080	0	-1.2804	-0.7347	2.0151	0.99	0.99
28	CR	0.7583	0	-1.5765	-1.1516	2.7280	0.83	0.82

ITEM	TYPE	B	D1	D2	D3	D4	INFIT	OUTFIT
29	CR	0.5189	0	-1.4580	-0.4166	1.8746	0.86	0.85
30	CR	0.5540	0	-1.6385	-0.5060	2.1445	1.40	1.40
31	CR	-0.1902	0	-0.8313	0.6807	0.1507	1.00	0.98
32	CR	-0.3567	0	-0.7213	1.1518	-0.4304	1.02	1.00

Table 8. Grade 4 Item Parameter Estimates

ITEM	TYPE	B	D1	D2	D3	D4	D5	INFIT	OUTFIT
1	EBSR	-0.5860	0	0.4252	-0.4252			0.98	0.98
2	EBSR	-0.3371	0	-0.0538	0.0538			0.92	0.90
3	EBSR	-0.6048	0	-0.0390	0.0390			1.02	1.01
4	EBSR	-0.3280	0	1.5702	-1.5702			1.00	1.10
5	EBSR	0.5246	0	0.2755	-0.2755			1.00	1.08
6	EBSR	-0.4277	0	-0.2772	0.2772			0.96	0.95
7	EBSR	-0.5038	0	1.0251	-1.0251			1.12	1.17
8	EBSR	-0.1130	0	0.1624	-0.1624			1.07	1.08
9	EBSR	-0.2051	0	-0.2037	0.2037			1.02	1.03
10	EBSR	0.4387	0	0.3590	-0.3590			1.20	1.45
11	EBSR	0.2843	0	0.3436	-0.3436			1.03	1.10
12	EBSR	0.1504	0	1.1685	-1.1685			1.23	1.45
13	EBSR	-0.1743	0	0.0735	-0.0735			1.12	1.14
14	EBSR	-0.2744	0	0.3309	-0.3309			0.91	0.92
15	EBSR	0.0126	0	0.3194	-0.3194			1.14	1.21
16	EBSR	0.3902	0	0.2923	-0.2923			1.36	1.59
17	EBSR	-0.3499	0	0.3945	-0.3945			1.02	1.02
18	EBSR	0.5659	0	0.0057	-0.0057			1.10	1.21
19	EBSR	-1.0468	0	0.5863	-0.5863			0.97	1.00
20	EBSR	-1.0071	0	0.3149	-0.3149			0.89	0.88
21	EBSR	-0.8705	0	0.4479	-0.4479			0.90	0.90

ITEM	TYPE	B	D1	D2	D3	D4	D5	INFIT	OUTFIT
22	EBSR	-0.9121	0	-0.0102	0.0102			0.92	0.93
23	EBSR	-0.2142	0	-0.4970	0.4970			0.92	0.92
24	EBSR	0.2548	0	0.1482	-0.1482			0.99	1.03
25	EBSR	0.1183	0	0.0411	-0.0411			1.13	1.18
26	EBSR	-0.5937	0	0.3017	-0.3017			0.92	0.90
27	EBSR	-0.4266	0	-0.1116	0.1116			0.90	0.89
28	EBSR	0.3628	0	0.0236	-0.0236			1.10	1.34
29	CR	0.8175	0	-1.9137	-1.2490	0.6524	2.5103	0.99	0.98
30	CR	0.2415	0	-1.0429	-1.0598	0.4280	1.6748	0.94	0.93
31	CR	0.6911	0	-0.6477	-0.4830	1.1307		0.89	0.81
32	CR	0.1263	0	-1.0661	-0.7611	1.8272		1.59	1.54
33	CR	-0.2825	0	0.3952	-0.2861	-0.1091		1.02	0.98
34	CR	-0.0848	0	0.0405	0.4004	-0.4409		1.09	1.02

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	25	48	71	97	6	425	25	679	40	445	26	55	3	500	29
4	14	31	55	75	103	12	292	35	317	38	111	13	12	1	123	15

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	497	29	4	Reading Literary	1	151	18
		2	635	37			2	283	34
		3	569	33			3	401	48
	Reading Informational	1	415	24		Reading Informational	1	118	14
		2	748	44			2	303	36
		3	538	32			3	414	50
	Reading Vocabulary	1	498	29		Reading Vocabulary	1	148	18
		2	539	32			2	245	29
		3	664	39			3	442	53
	Written Expression	1	759	45		Written Expression	1	193	23
		2	311	18			2	261	31
		3	631	37			3	381	47
	Writing Knowledge and Language Conventions	1	655	39		Writing Knowledge and Language Conventions	1	180	22
		2	455	27			2	198	24
		3	591	35			3	457	55

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
656	1	0.06	1	0.06
667	4	0.24	5	0.29
675	8	0.47	13	0.76
681	6	0.35	19	1.12
685	11	0.65	30	1.76
690	17	1.00	47	2.76
693	12	0.71	59	3.47
696	14	0.82	73	4.29
699	24	1.41	97	5.70
700	23	1.35	120	7.05
704	30	1.76	150	8.82
706	30	1.76	180	10.58
708	25	1.47	205	12.05
710	33	1.94	238	13.99
712	28	1.65	266	15.64
714	29	1.70	295	17.34
715	36	2.12	331	19.46
717	26	1.53	357	20.99
718	34	2.00	391	22.99
720	37	2.18	428	25.16
721	31	1.82	459	26.98
723	35	2.06	494	29.04
724	28	1.65	522	30.69
725	23	1.35	545	32.04
726	26	1.53	571	33.57
728	28	1.65	599	35.21
729	33	1.94	632	37.15
730	27	1.59	659	38.74
731	34	2.00	693	40.74
732	24	1.41	717	42.15
733	34	2.00	751	44.15
734	33	1.94	784	46.09
735	31	1.82	815	47.91
737	41	2.41	856	50.32
738	35	2.06	891	52.38
739	35	2.06	926	54.44
740	25	1.47	951	55.91
741	28	1.65	979	57.55
742	37	2.18	1016	59.73
743	26	1.53	1042	61.26

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
744	22	1.29	1064	62.55
745	20	1.18	1084	63.73
746	26	1.53	1110	65.26
747	31	1.82	1141	67.08
748	28	1.65	1169	68.72
749	32	1.88	1201	70.61
750	25	1.47	1226	72.08
751	20	1.18	1246	73.25
752	30	1.76	1276	75.01
753	28	1.65	1304	76.66
754	22	1.29	1326	77.95
755	23	1.35	1349	79.31
756	24	1.41	1373	80.72
758	36	2.12	1409	82.83
759	25	1.47	1434	84.30
760	12	0.71	1446	85.01
761	14	0.82	1460	85.83
762	22	1.29	1482	87.13
763	19	1.12	1501	88.24
765	24	1.41	1525	89.65
766	15	0.88	1540	90.53
767	13	0.76	1553	91.30
769	18	1.06	1571	92.36
770	15	0.88	1586	93.24
771	12	0.71	1598	93.94
773	13	0.76	1611	94.71
775	11	0.65	1622	95.36
776	11	0.65	1633	96.00
778	13	0.76	1646	96.77
779	11	0.65	1657	97.41
782	5	0.29	1662	97.71
784	8	0.47	1670	98.18
786	5	0.29	1675	98.47
788	4	0.24	1679	98.71
790	3	0.18	1682	98.88
793	1	0.06	1683	98.94
795	5	0.29	1688	99.24
798	4	0.24	1692	99.47
805	5	0.29	1697	99.76
809	1	0.06	1698	99.82
818	2	0.12	1700	99.94
848	1	0.06	1701	100.00

Table 13. Grade 3 Reading Scale Score Frequency Distributions

Reading Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
19	2	0.12	2	0.12
23	7	0.41	9	0.53
26	10	0.59	19	1.12
28	19	1.12	38	2.23
30	22	1.29	60	3.53
31	40	2.35	100	5.88
33	37	2.18	137	8.05
34	58	3.41	195	11.46
35	53	3.12	248	14.58
36	64	3.76	312	18.34
37	70	4.12	382	22.46
38	134	7.88	516	30.34
39	55	3.23	571	33.57
40	59	3.47	630	37.04
41	94	5.53	724	42.56
42	48	2.82	772	45.39
43	87	5.11	859	50.50
44	34	2.00	893	52.50
45	99	5.82	992	58.32
46	77	4.53	1069	62.85
47	33	1.94	1102	64.79
48	80	4.70	1182	69.49
49	69	4.06	1251	73.54
50	37	2.18	1288	75.72
51	66	3.88	1354	79.60
52	60	3.53	1414	83.13
53	30	1.76	1444	84.89
54	60	3.53	1504	88.42
55	23	1.35	1527	89.77
56	24	1.41	1551	91.18
57	29	1.70	1580	92.89
58	34	2.00	1614	94.89
59	18	1.06	1632	95.94
60	12	0.71	1644	96.65
61	13	0.76	1657	97.41
62	9	0.53	1666	97.94
64	15	0.88	1681	98.82
65	3	0.18	1684	99.00
67	10	0.59	1694	99.59
69	3	0.18	1697	99.76

Reading Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
71	3	0.18	1700	99.94
84	1	0.06	1701	100.00

Table 14. Grade 3 Writing Scale Score Frequency Distributions

Writing Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10	94	5.53	94	5.53
15	40	2.35	134	7.88
19	23	1.35	157	9.23
21	50	2.94	207	12.17
23	44	2.59	251	14.76
25	53	3.12	304	17.87
26	74	4.35	378	22.22
27	61	3.59	439	25.81
28	59	3.47	498	29.28
29	74	4.35	572	33.63
30	57	3.35	629	36.98
31	175	10.29	804	47.27
32	63	3.70	867	50.97
33	72	4.23	939	55.20
34	153	8.99	1092	64.20
35	58	3.41	1150	67.61
36	72	4.23	1222	71.84
37	109	6.41	1331	78.25
38	56	3.29	1387	81.54
39	59	3.47	1446	85.01
40	58	3.41	1504	88.42
41	49	2.88	1553	91.30
43	29	1.70	1582	93.00
44	44	2.59	1626	95.59
45	26	1.53	1652	97.12
47	4	0.24	1656	97.35
49	15	0.88	1671	98.24
51	20	1.18	1691	99.41
53	1	0.06	1692	99.47
55	3	0.18	1695	99.65
57	4	0.24	1699	99.88
60	2	0.12	1701	100.00

Table 15. Grade 4 Scale Score Frequency Distributions

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
668	1	0.12	1	0.12
674	4	0.48	5	0.60
679	4	0.48	9	1.08
683	8	0.96	17	2.04
686	13	1.56	30	3.59
689	10	1.20	40	4.79
692	17	2.04	57	6.83
695	12	1.44	69	8.26
697	16	1.92	85	10.18
699	18	2.16	103	12.34
700	16	1.92	119	14.25
703	11	1.32	130	15.57
705	28	3.35	158	18.92
707	13	1.56	171	20.48
709	19	2.28	190	22.75
710	14	1.68	204	24.43
712	15	1.80	219	26.23
713	16	1.92	235	28.14
715	21	2.51	256	30.66
716	16	1.92	272	32.57
717	8	0.96	280	33.53
719	20	2.40	300	35.93
720	16	1.92	316	37.84
721	19	2.28	335	40.12
722	25	2.99	360	43.11
723	20	2.40	380	45.51
724	15	1.80	395	47.31
725	18	2.16	413	49.46
727	18	2.16	431	51.62
728	11	1.32	442	52.93
729	22	2.63	464	55.57
730	17	2.04	481	57.60
731	18	2.16	499	59.76
732	14	1.68	513	61.44
733	17	2.04	530	63.47
734	16	1.92	546	65.39
735	14	1.68	560	67.07
736	13	1.56	573	68.62
737	15	1.80	588	70.42
738	11	1.32	599	71.74
739	8	0.96	607	72.69
740	6	0.72	613	73.41

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
741	19	2.28	632	75.69
742	7	0.84	639	76.53
743	15	1.80	654	78.32
744	12	1.44	666	79.76
745	9	1.08	675	80.84
746	13	1.56	688	82.40
747	8	0.96	696	83.35
748	2	0.24	698	83.59
749	14	1.68	712	85.27
750	10	1.20	722	86.47
751	7	0.84	729	87.31
752	10	1.20	739	88.50
753	11	1.32	750	89.82
754	11	1.32	761	91.14
755	7	0.84	768	91.98
756	4	0.48	772	92.46
757	6	0.72	778	93.17
758	6	0.72	784	93.89
759	5	0.60	789	94.49
760	7	0.84	796	95.33
761	11	1.32	807	96.65
762	1	0.12	808	96.77
763	1	0.12	809	96.89
764	4	0.48	813	97.37
766	3	0.36	816	97.72
768	3	0.36	819	98.08
769	2	0.24	821	98.32
771	2	0.24	823	98.56
772	3	0.36	826	98.92
773	1	0.12	827	99.04
779	2	0.24	829	99.28
781	2	0.24	831	99.52
784	1	0.12	832	99.64
786	2	0.24	834	99.88
794	1	0.12	835	100.00

Table 16. Grade 4 Reading Scale Score Frequency Distributions

Reading Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
21	2	0.24	2	0.24
24	6	0.72	8	0.96
26	10	1.20	18	2.16
27	12	1.44	30	3.59
29	24	2.87	54	6.47
30	19	2.28	73	8.74
31	28	3.35	101	12.10
32	28	3.35	129	15.45
33	33	3.95	162	19.40
34	40	4.79	202	24.19
35	38	4.55	240	28.74
36	26	3.11	266	31.86
37	30	3.59	296	35.45
38	53	6.35	349	41.80
39	32	3.83	381	45.63
40	71	8.50	452	54.13
41	37	4.43	489	58.56
42	47	5.63	536	64.19
43	16	1.92	552	66.11
44	40	4.79	592	70.90
45	17	2.04	609	72.93
46	39	4.67	648	77.60
47	24	2.87	672	80.48
48	13	1.56	685	82.04
49	31	3.71	716	85.75
50	27	3.23	743	88.98
51	10	1.20	753	90.18
52	25	2.99	778	93.17
53	7	0.84	785	94.01
54	26	3.11	811	97.13
55	6	0.72	817	97.84
56	2	0.24	819	98.08
57	3	0.36	822	98.44
58	3	0.36	825	98.80
59	2	0.24	827	99.04
60	2	0.24	829	99.28
61	2	0.24	831	99.52
62	1	0.12	832	99.64
63	2	0.24	834	99.88
66	1	0.12	835	100.00

Table 17. Grade 4 Writing Scale Score Frequency Distributions

Writing Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10	94	11.26	94	11.26
15	19	2.28	113	13.53
19	1	0.12	114	13.65
22	49	5.87	163	19.52
23	33	3.95	196	23.47
25	17	2.04	213	25.51
26	59	7.07	272	32.57
27	64	7.66	336	40.24
28	48	5.75	384	45.99
29	26	3.11	410	49.10
30	65	7.78	475	56.89
31	65	7.78	540	64.67
32	38	4.55	578	69.22
33	59	7.07	637	76.29
34	43	5.15	680	81.44
35	35	4.19	715	85.63
36	21	2.51	736	88.14
37	29	3.47	765	91.62
38	16	1.92	781	93.53
39	20	2.40	801	95.93
40	8	0.96	809	96.89
41	3	0.36	812	97.25
42	3	0.36	815	97.60
43	3	0.36	818	97.96
44	4	0.48	822	98.44
45	10	1.20	832	99.64
46	1	0.12	833	99.76
47	1	0.12	834	99.88
52	1	0.12	835	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	656	15
3	667	15
4	675	14
5	681	12
6	685	11
7	690	10
8	693	10
9	696	9
10	699	9
11	700	8
12	704	8
13	706	8
14	708	7
15	710	7
16	712	7
17	714	7
18	715	7
19	717	7
20	718	6
21	720	6
22	721	6
23	723	6
24	724	6
25	725	6
26	726	6
27	728	6
28	729	6
29	730	6
30	731	6
31	732	6
32	733	6
33	734	6
34	735	5
35	737	5
36	738	5
37	739	5

Raw Score	Scale Score	CSEM
38	740	5
39	741	5
40	742	5
41	743	5
42	744	5
43	745	5
44	746	5
45	747	5
46	748	5
47	749	5
48	750	5
49	751	5
50	752	5
51	753	5
52	754	5
53	755	6
54	756	6
55	758	6
56	759	6
57	760	6
58	761	6
59	762	6
60	763	6
61	765	6
62	766	6
63	767	6
64	769	6
65	770	6
66	771	6
67	773	7
68	775	7
69	776	7
70	778	7
71	779	7
72	782	7
73	784	8
74	786	8
75	788	8
76	790	8
77	793	9
78	795	9
79	798	9

Raw Score	Scale Score	CSEM
80	802	10
81	805	10
82	809	11
83	813	11
84	818	11
85	823	12
86	828	13
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	11
1	12	11
2	19	8
3	23	6
4	26	5
5	28	5
6	30	4
7	31	4
8	33	4
9	34	4
10	35	3
11	36	3
12	37	3
13	38	3
14	38	3
15	39	3
16	40	3
17	41	3
18	41	3
19	42	3
20	43	3
21	43	3
22	44	3
23	45	3
24	45	3
25	46	3
26	46	3
27	47	3
28	48	3
29	48	3
30	49	3
31	49	3
32	50	3
33	51	3
34	51	3
35	52	3
36	52	3
37	53	3
38	54	3

Raw Score	Scale Score	CSEM
39	54	3
40	55	3
41	56	3
42	57	3
43	58	3
44	58	3
45	59	3
46	60	3
47	61	4
48	62	4
49	64	4
50	65	4
51	67	4
52	69	5
53	71	5
54	74	6
55	78	7
56	84	9
57	90	9
58	90	9

Table 21. Grade 3 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	4
4	23	3
5	25	3
6	26	3
7	27	2
8	28	2
9	29	2
10	30	2
11	31	2
12	31	2
13	32	2
14	33	2
15	34	2
16	34	2
17	35	2
18	36	2
19	37	2
20	37	2
21	38	2
22	39	2
23	40	3
24	41	3
25	43	3
26	44	3
27	45	3
28	47	3
29	49	3
30	51	3
31	53	4
32	55	4
33	57	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	650	15
3	660	15
4	668	14
5	674	13
6	679	11
7	683	11
8	686	10
9	689	9
10	692	9
11	695	8
12	697	8
13	699	8
14	700	8
15	703	7
16	705	7
17	707	7
18	709	7
19	710	7
20	712	7
21	713	6
22	715	6
23	716	6
24	717	6
25	719	6
26	720	6
27	721	6
28	722	6
29	723	6
30	724	6
31	725	6
32	727	6
33	728	6
34	729	6
35	730	5
36	731	5
37	732	5
38	733	5
39	734	5

Raw Score	Scale Score	CSEM
40	735	5
41	736	5
42	737	5
43	738	5
44	739	5
45	740	5
46	741	5
47	742	5
48	743	5
49	744	5
50	745	5
51	746	5
52	747	5
53	748	5
54	749	5
55	750	5
56	751	5
57	752	5
58	753	5
59	754	5
60	755	5
61	756	5
62	757	6
63	758	6
64	759	6
65	760	6
66	761	6
67	762	6
68	763	6
69	764	6
70	766	6
71	767	6
72	768	6
73	769	6
74	771	6
75	772	6
76	773	6
77	775	6
78	776	7
79	778	7
80	779	7
81	781	7

Raw Score	Scale Score	CSEM
82	783	7
83	784	7
84	786	7
85	788	8
86	790	8
87	792	8
88	794	8
89	797	8
90	799	9
91	802	9
92	805	9
93	808	10
94	811	10
95	814	10
96	818	11
97	822	11
98	827	12
99	832	13
100	838	13
101	845	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	14	8
3	18	7
4	21	6
5	24	5
6	26	5
7	27	4
8	29	4
9	30	4
10	31	4
11	32	4
12	33	3
13	34	3
14	35	3
15	36	3
16	37	3
17	38	3
18	38	3
19	39	3
20	40	3
21	40	3
22	41	3
23	42	3
24	42	3
25	43	3
26	44	3
27	44	3
28	45	3
29	46	3
30	46	3
31	47	3
32	47	3
33	48	3
34	49	3
35	49	3
36	50	3
37	50	3
38	51	3

Raw Score	Scale Score	CSEM
39	52	3
40	52	3
41	53	3
42	54	3
43	54	3
44	55	3
45	56	3
46	57	3
47	57	3
48	58	3
49	59	3
50	60	3
51	61	3
52	62	4
53	63	4
54	64	4
55	66	4
56	67	4
57	69	5
58	71	5
59	74	6
60	76	6
61	80	7
62	86	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	22	3
4	23	3
5	25	3
6	26	2
7	27	2
8	27	2
9	28	2
10	29	2
11	30	2
12	30	2
13	31	2
14	31	2
15	32	2
16	33	2
17	33	2
18	34	2
19	34	2
20	35	2
21	35	2
22	36	2
23	37	2
24	37	2
25	38	2
26	39	2
27	39	2
28	40	2
29	41	2
30	42	2
31	43	2
32	44	2
33	45	3
34	46	3
35	47	3
36	49	3
37	50	3
38	52	3

Raw Score	Scale Score	CSEM
39	55	4
40	58	5
41	60	5
42	60	5

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	0.93	0.91	0.82	0.68	0.68	0.53	4	RD	1	0.91	0.87	0.80	0.67	0.68	0.45
	RL		1	0.74	0.66	0.65	0.65	0.49		RL		1	0.65	0.62	0.65	0.65	0.45
	RI			1	0.66	0.62	0.62	0.47		RI			1	0.59	0.57	0.59	0.36
	RV				1	0.53	0.52	0.43		RV				1	0.48	0.48	0.35
	WR					1	0.97	0.84		WR					1	0.96	0.79
	WE						1	0.68		WE						1	0.59
	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	0.51
4	0.73	0.63	0.30	0.47

Table 27. Spring 2017 Rater Agreement Statistics

Grade	Item	Exact	Exact+Adjacent
3	1	77.3%	100%
	2	77.9%	99.0%
	3	80.2%	99.0%
	4	77.9%	97.8%
	5	72.6%	97.2%
	6	83.1%	96.6%
4	1	78.5%	97.8%
	2	86.9%	99.0%
	3	78.5%	97.8%
	4	89.2%	99.0%
	5	75.0%	99.0%
	6	78.5%	99.0%

Table 28. Spring 2017 Items Field Tested and Data Review Outcomes

	Grade 3	Grade 4
Number of test forms	4	2
Number of items field tested	34	21
Number of items flagged and reviewed	15	17
Number of accepted items	13	13

APPENDICES

APPENDIX A: CSLA ELIGIBILITY FLOWCHART

Colorado Spanish Language Arts (CSLA) Decision Making Flowchart

Grades 3 and 4

NEP

Non-English Proficient

LEP

Limited-English Proficient

FEP

Fully English Proficient M1, M2 or Exited

FELL

Former English Language Learner
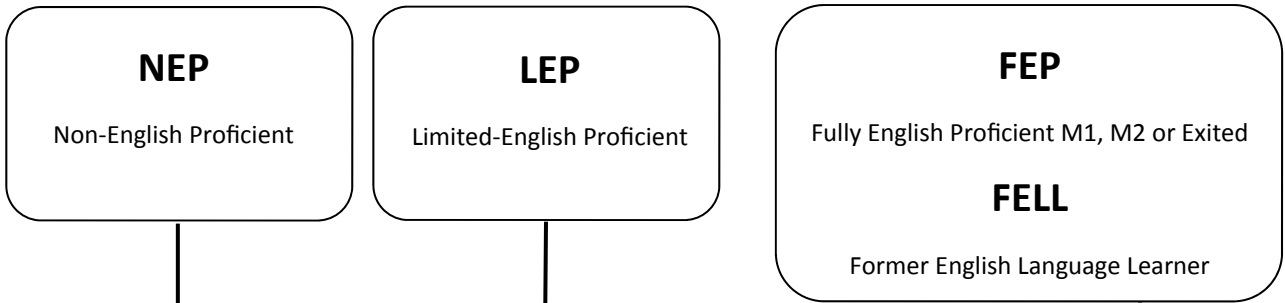
Has the student received instruction in Spanish Language arts within the last nine months?*

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 CSLA Assessment

(or locally translated CoAlt: DLM ELA)

Districts must determine the best fit for the student to access content on the assessment:

- CSLA in grade 3 or 4
- CMAS: ELA with linguistic accommodations
- Locally translated CoAlt: DLM ELA

Student is not eligible for the Spanish version of the assessment.

Student MUST participate in CMAS:ELA or CoAlt: DLM ELA with or without accommodations.

Standard and linguistic accommodations are available on CMAS: ELA and CoAlt: DLM ELA to provide access to content on the assessment.

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

2017/2018

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 **comprehension** 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**.

Colorado Spanish Language Arts

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 clear**.	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**.

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 clear**.	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 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.

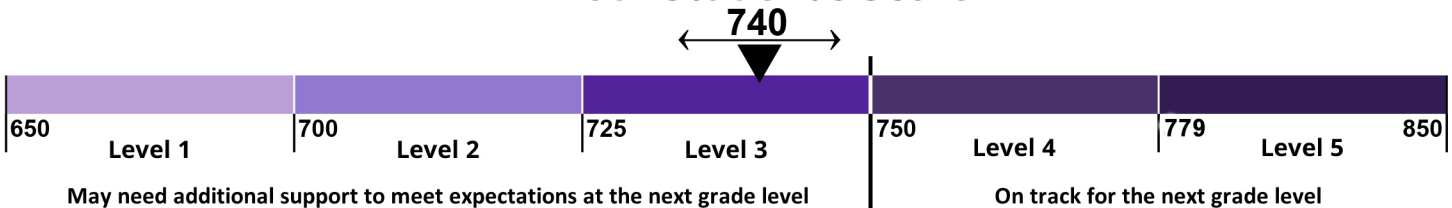## How Did FIRSTNAME Perform Overall?
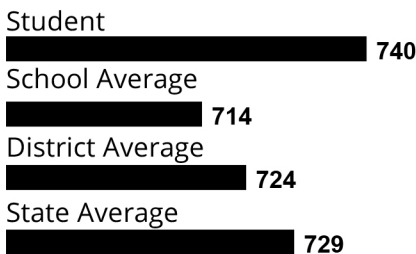
Performance Level 3

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

← **740** →

650	700	725	750	779	850
Level 1	Level 2	Level 3	Level 4	Level 5	

May need additional support to meet expectations at the next grade level

On track for the next grade level

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.

Student
740

School Average
714

District Average
724

State Average
729

650	700	725	750	779	850

How Students in Colorado Performed

15%	25%	35%	15%	10%
Level 1	Level 2	Level 3	Level 4	Level 5
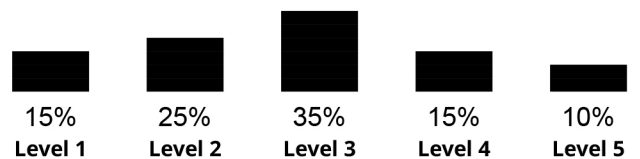
Percentage of students at each performance level

How Did Your Student Perform in Reading and Writing?

READING

Your Student's Score
58

10 ▬▬▬▬▬ **50** ▼ ▬▬▬▬▬ 90

Met expectations

School Average
46

District Average
42

State Average
52

⬆ 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.

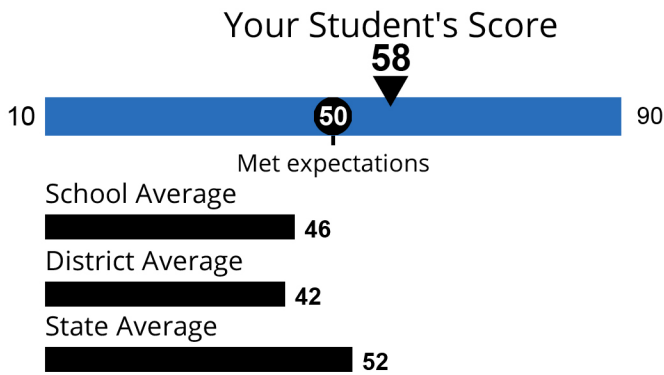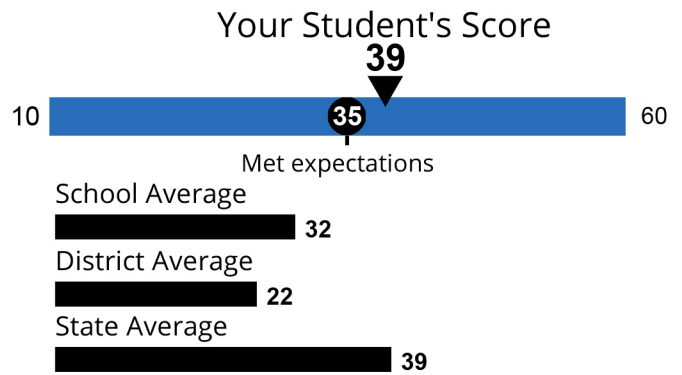### ⬌ INFORMATIONAL TEXT

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

Your Student's Score
39

10 ▬▬▬▬▬ **35** ▼ ▬▬▬▬▬ 60

Met expectations

School Average
32

District Average
22

State Average
39

⬌ 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

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

This report is NOT for public review. Distribution within your school/district must be in accordance with state and federal privacy laws, and local school board policy.

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 | SCORE | READING* | | | SCORE | WRITING* | |
			LITERARY	INFORMATION	VOCABULARY		EXPRESSION	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

This report is NOT for public review. Distribution within your school/district must be in accordance with state and federal privacy laws, and local school board policy.

mmddyyyy-Batch-1234-5678-1234567

Colorado Measures of Academic Success
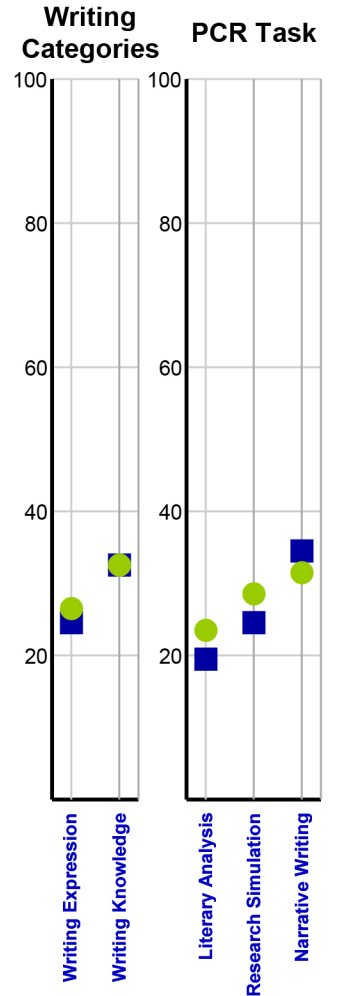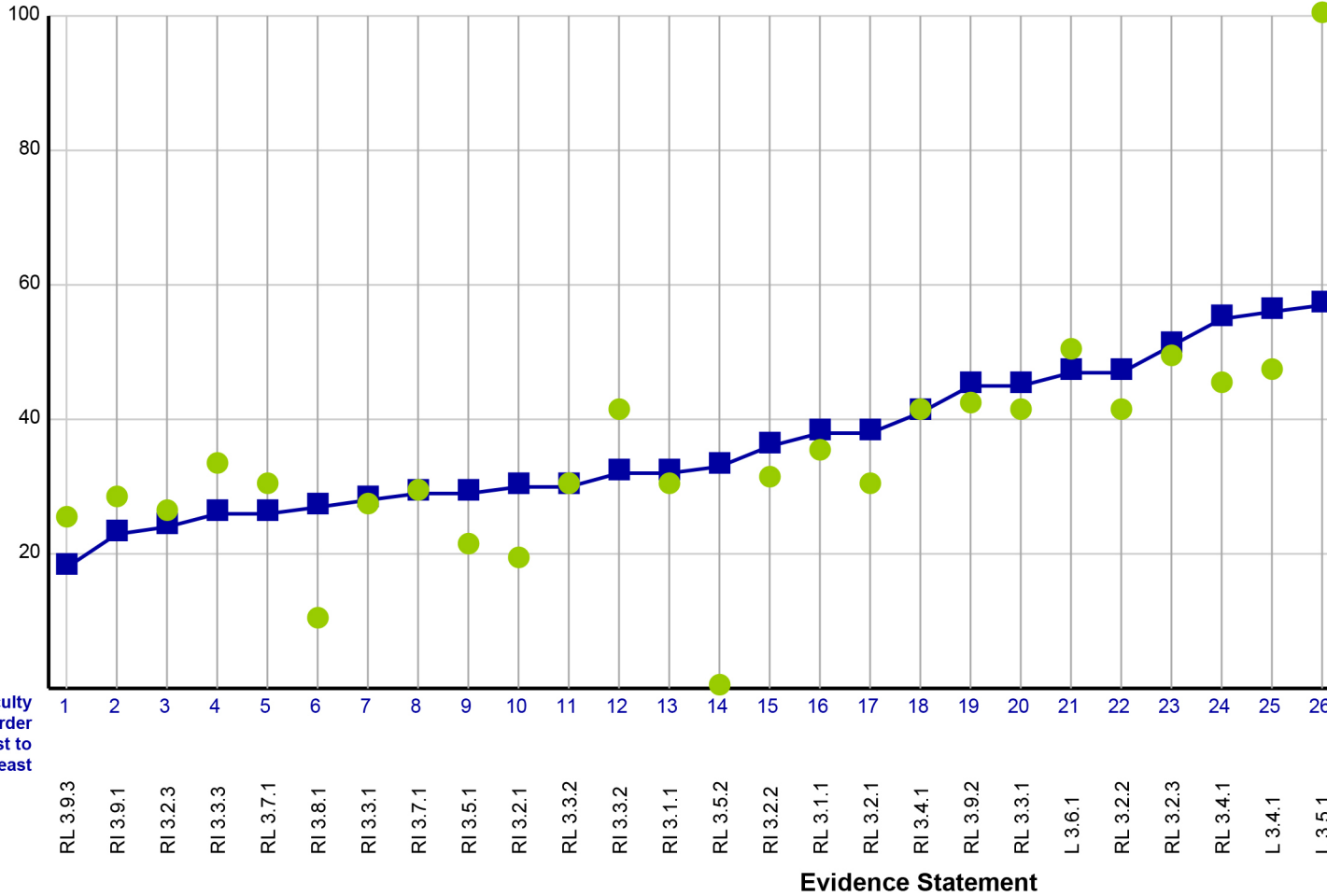## District Evidence Statement Analysis
Spring 2016

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.

Writing Categories — Writing Expression, Writing Knowledge

PCR Task — Literary Analysis, Research Simulation, Narrative Writing

Evidence Statement (Difficulty Order Most to Least):
1 RL 3.9.3, 2 RI 3.9.1, 3 RI 3.2.3, 4 RL 3.3.3, 5 RL 3.7.1, 6 RI 3.8.1, 7 RI 3.3.1, 8 RI 3.7.1, 9 RI 3.5.1, 10 RI 3.2.1, 11 RL 3.3.2, 12 RI 3.3.2, 13 RI 3.1.1, 14 RL 3.5.2, 15 RI 3.2.2, 16 RI 3.1.1, 17 RI 3.2.1, 18 RI 3.4.1, 19 RL 3.9.2, 20 RL 3.3.1, 21 L 3.6.1, 22 RI 3.2.2, 23 RL 3.2.3, 24 RL 3.4.1, 25 L 3.4.1, 26 L 3.5.1

Difficulty level is determined at the State level for all reports.
This report is NOT for public review. Distribution within your school/district must be in accordance with state and federal privacy laws, and local school board policy.

07202016-Sp-Z9999999-5555-0000- 0000000

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