Colorado
English Language Acquisition Assessment Program

2012 Technical Report

FINAL

Submitted to the
Colorado Department of Education

July 31, 2012

CTB McGraw-Hill

Developed and published under contract with Colorado Department of Education by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703.

Table of Contents

Appendices

List of Figures

List of Tables

Overview

The first administration of the Colorado English Language Acquisition Assessments (CELApro) occurred in Spring 2006. At that time, the assessments were identical to CTB's *LAS Links* ® (Form A), except for customized Colorado test book covers and answer sheets.

LAS Links (Form A) continues to provide a solid foundation for all the CELApro tests. The *LAS Links* assessments were developed from a framework that reflects sound principles of second-language acquisition (Schmidt, 2001; Savignon, 1972, 1997; Bachman and Palmer, 1996; O'Malley and Valdez-Pierce, 1996; Chamot and O'Malley, 1994; Bachman, 1990). Each*LAS Links* test consists of four separately scored sections (Speaking, Listening, Reading, and Writing). In addition to these four component scores, all of the Speaking and Listening items are combined to produce an Oral score, and selected Listening and Reading items are combined to yield a Comprehension score.

Approximately 30,000 students participated in the field test, *LAS Links* (Form A), which was calibrated and scaled using item response theory and a common-item equating design to place all grade levels on a common scale and to ensure that skill area scores have the same meaning across forms, grades, and years.

The *LAS Links* tests are aligned to CTB/McGraw-Hill's English Language Proficiency Assessment Standards (ELPAS), which were developed to include key standards from the national English as a Second Language (ESL) and Teachers of English to Speakers of Other Languages(TESOL) standards and from several state ESL standards. In order to match the Colorado standards, the five *LAS Links* grade spans were modified and reduced to four grade spans for the 2007 CELApro tests. In 2008, the K–2 grade span was divided into three separate grades, resulting in the current six CELApro grade spans: K, 1, 2, 3–5, 6–8, and 9–12. In 2008, additional items were added to the Reading section in Grades K, 1, and 2 and to the Listening and Writing sections in all six grade spans. Table 1 shows a comparison of grade spans by year.

Table 1. Comparison of *LAS Links* and CELApro Grade Spans

<table>
<tr><th colspan="3">Grade Spans</th></tr>
<tr><th>LAS Links</th><th>CELApro 2007</th><th>CELApro 2008–2012</th></tr>
<tr><td rowspan="2">K–1</td><td rowspan="3">K–2</td><td>K</td></tr>
<tr><td>1</td></tr>
<tr><td rowspan="2">2–3</td><td>2</td></tr>
<tr><td rowspan="2">3–5</td><td rowspan="2">3–5</td></tr>
<tr><td>4–5</td></tr>
<tr><td>6–8</td><td>6–8</td><td>6–8</td></tr>
<tr><td>9–12</td><td>9–12</td><td>9–12</td></tr>
</table>

The 2012 CELApro tests are identical to those administered in 2008 through 2011. Grades K, 1, 2, and 3–5 have a scannable test book; the other grade spans have a reusable test book and a scannable answer book. The Speaking items and the Writing constructed-response (CR) items appear only in the answer book for Grade Spans 6–8 and 9–12.

Part 1: Standards

The Colorado English Language Acquisition Assessment (CELApro) is the language proficiency assessment used for classifying and monitoring the progress of English Language Learners (ELLs) in Colorado in the acquisition of English. *LAS Links* (Form A) assessments form the core of the CELApro tests.

The CELApro assessments measure the competencies necessary for successful social and academic language use in four major modalities— Speaking, Listening, Reading, and Writing—along a continuum of five proficiency levels: Beginning, Early Intermediate, Intermediate, Proficient, and Advanced. The assessments take into account the students' maturation and cognitive skills by providing age-appropriate tests covering six grade spans: K, 1, 2, 3–5, 6–8, and 9–12.

A combination of item types—constructed-response (CR) and multiple-choice (MC)—provide a variety of ways for students to demonstrate proficiency and to maintain reasonable testing times. CR items assess the productive domains of Speaking and Writing, whereas the MC items assess the receptive domains of Listening, Reading, and Writing (grammar). The variety of item types ensures measurement of the full spectrum of possible tasks required for each language subskill and allows for multiple ways to interpret results.

Alignment Studies

An important indicator of the validity of a standardized English language test is the degree of *alignment* (i.e., the match) between the state English language development (ELD) standards and the test content. In developing standardized tests, test items are written to cover as many standards as possible.

Colorado has four general standards for English language learners, organized by modality (Listening, Speaking, Reading, and Writing) and applicable at all grade levels. The standards specify general skills in social and academic language.

- **Standard 1:** English Language Learners listen for information and understanding, using a variety of sources, for academic and social purposes.
- **Standard 2:** English Language Learners speak to convey information and understanding, using a variety of sources, for academic and social purposes.
- **Standard 3:** English Language Learners read for information and understanding, using a variety of sources, for academic and social purposes.
- **Standard 4:** English Language Learners write to convey information and understanding, using a variety of sources, for academic and social purposes.

A detailed description of the standards, by grade and proficiency level, is provided in Appendix F.

In order to increase the alignment of CELApro to the Colorado ELD standards, additional test items were written for the 2008 tests to assess individual standards that were not already assessed by *LAS Links* items.

CTB conducted an alignment analysis of the 2008 CELApro assessments to evaluate the match between the test and the standards. In performing an alignment analysis, it is sometimes necessary to eliminate some standards because they cannot be easily assessed by a standardized test. For example, a standard may require an extended process outside of the test situation, such as the steps for writing a research paper, specify instructional strategies rather than student skills, or specify parameters outside of the testing situation, such as "participate in group discussions." Of the 397 Colorado ELD standards, 104 were eliminated as nonassessable.

In performing the CELApro alignment, the raters independently matched items to all of the assessable standards on the basis of direct, indirect, or partial alignment. The test item numbers were then entered into the cells of the matching standards. The degree of alignment was calculated by totalling the number of assessable standards that were measured by at least one CELApro item. All of the standards are assessed by at least one test item.

This alignment analysis was reviewed by CTB and the Colorado Department of Education (CDE) at a meeting in April 2007. CTB then conducted a final review with a committee of English language acquisition experts, finalizing the alignment, which is shown in Table 2. This table also reflects the current alignment because the CELApro tests were unchanged from 2008 through 2012.

Table 2. Item Alignment Percentages by Grade Span

	K–2		**3–5**		**6–8**		**9–12**	
Listening	14/14	100	15/15	100	14/14	100	15/15	100
Beginning	4/4		5/5		5/5		5/5	
Intermediate	5/5		5/5		5/5		5/5	
Advanced	5/5		5/5		4/4		5/5	
Speaking	14/14	100	13/13	100	12/12	100	11/11	100
Beginning	4/4		4/4		5/5		4/4	
Intermediate	5/5		4/4		4/4		4/4	
Advanced	5/5		5/5		3/3		3/3	
Reading	16/16	100	15/15	100	14/14	100	12/12	100
Beginning	5/5		4/4		3/3		3/3	
Intermediate	7/7		6/6		5/5		5/5	
Advanced	4/4		5/5		6/6		4/4	

(continued on the next page)

Table 2. Item Alignment Percentages by Grade Span (continued)

	K–2		**3–5**		**6–8**		**9–12**	
Writing	13/13	100	14/14	100	9/9	100	7/7	100
Beginning	2/2		4/4		3/3		2/2	
Intermediate	5/5		5/5		3/3		3/3	
Advanced	6/6		5/5		3/3		2/2	

Part 2: Test Development

The 2012 CELApro tests are identical to the 2008, 2009, 2010, and 2011 tests and consist of both *LAS Links* items and items owned by the Colorado Department of Education. For Grade Spans 6–8 and 9–12, the organization of the CELApro assessments is identical to the corresponding *LAS Links* assessments. The reconfigured tests for Grade Spans K–2 and 3–5 were created using selected items from the *LAS Links* assessments for the appropriate grades. The lowest grade span was also broken into separate tests for Kindergarten, Grade 1, and Grade 2. All K–2 students take the same Speaking and Listening items but some different Reading and Writing items. All of these items were written by writers with experience or training in the areas being tested. Before writing items, all writers went through extensive training and were instructed to:

- Study each standard to be assessed.
- Decide what is important for the student to know and do in order to demonstrate mastery of the standard. Avoid the trivial.
- Write the item so that it focuses on the particular content or skill to be assessed.
- Develop answer choices that relate logically to the stem and standard. The correct response should be clear to students who have mastered the concept or skill. The distractors should be clearly wrong to students who have mastered the content or skill. Test items should not be "tricky" or contain information unfamiliar to most students.
- Provide documentation from source material (e.g., photocopies of encyclopedia entries and other reliable reference materials) to verify that all information included in the stimulus and item is correct. All factual statements in stimuli, stems, and correct responses must be checked against reliable sources. Also, distractors should be verified as incorrect.
- Use appropriate subject matter. Refrain from explicit references to or descriptions of alcohol or drug abuse, sex, or vulgar language. Exercise caution when developing religious, political, social, or philosophical issues as subject matter. Individual beliefs should not influence content.
- Avoid using very controversial material. Large-scale (national, state, or district) assessments are administered to student populations with different experiences and beliefs.
- Verify that the item is free of content that could be offensive, insensitive, stereotypical, or that introduces other types of bias.
- Check that the content of the stimulus and/or the item is developmentally and age appropriate for the students being tested.
- Write a range of items representing all levels of proficiency in English within a specific standard.

The tests have been structured to comprehensively assess the four language skills of Speaking, Listening, Reading, and Writing. Comprehension is assessed using selected Listening and Reading items. A combination of CR, dichotomous CR (correct or incorrect), and MC items is used to provide diverse opportunities for students to demonstrate proficiency and to maintain reasonable testing times. CR items are used to assess the productive domains of Speaking and Writing, whereas the MC items are used to assess the receptive domains of Listening and Reading and of the Writing Use Conventions subtest. The structure of the 2012 CELApro is shown in Table 3.

Table 3. 2012 CELApro Test Structure

Content	Grade Span	Sub-Content	Item Type	Items	Score Points	CR/DCR Items Scored By	Administration
Speaking 20 items, 41 pts	4 grade spans: K–2, 3–5, 6–8, 9–12	Speak Words	DCR	10	10	Local Test Administrator	Individual
		Sentences	CR	5	15		
		Conversation	CR	4	12		
		Tell a Story	CR	1	4		
Listening K=21 items, 20 pts* 1–2=21 items, 21 pts 3–12=23 items, 23 pts	K	Listen for Information	MC	11	11	Not Applicable	Individual/Group
		Listen in the Classroom	MC	5	5		
		Listen & Comprehend	MC	4	4		
	1–2	Listen for Information	MC	11	11		
		Listen in the Classroom	MC	6	6		
		Listen & Comprehend	MC	4	4		
	3–12	Listen for Information	MC	10	10		
		Listen in the Classroom	MC	9	9		
		Listen & Comprehend	MC	4	4		
Reading K=31 items, 31 pts 1–2=36 items, 36 pts 3–12=35 items, 35 pts	K	Analyze Words	MC	11	11	Not Applicable	Individual
		Read Words	MC	10	10		
		Understanding	MC	10	10		
	1–2	Analyze Words	MC	11	11		
		Read Words	MC	10	10		
		Understanding	MC	15	15		
	3–12	Analyze Words	MC	10	10		Group
		Read Words	MC	10	10		
		Understanding	MC	15	15		
Writing K–1=25 items, 35 pts 3–5=26 items, 37 pts 2, 6–12=25 items, 36 pts	K–1	Conventions	MC	20	20	CTB Handscoring	Group (or Individual for K)
		Write About	CR	2	6		
		Write Why	CR	3	9		
	3–5	Conventions	MC	21	21	CTB Handscoring	Group
		Write About	CR	2	6		
		Write Why	CR	2	6		
		Write in Detail	CR	1	4		
	2, 6–12	Conventions	MC	20	20	CTB Handscoring	Group
		Write About	CR	2	6		
		Write Why	CR	2	6		
		Write in Detail	CR	1	4		
Oral K–2=40 items, 61 pts 1–2=41 items, 62 pts 3–12=43 items, 64 pts	K–2	Listening and Speaking	MC	20	20	Local Test Administrator	Not Applicable
			SCR	10	31		
			ECR	10	10		
	1–2	Listening and Speaking	MC	21	21		
			SCR	10	31		
			ECR	10	10		
	3–12	Listening and Speaking	MC	23	23		
			SCR	9	27		
			ECR	1	4		
			CR	10	10		
Comprehension K=39 items, 39 pts 1–2=45 items, 45 pts 3–5=48 items, 48 pts 6–12=50 items, 50 pts	K	Listening and Reading	MC	39	39	Not Applicable	Not Applicable
	1–2	Listening and Reading	MC	45	45		
	3–5	Listening and Reading	MC	48	48		
	6–12	Listening and Reading	MC	50	50		

KEY: DCR=Dichotomous CR; SCR=Short CR; ECR=Extended CR

*There were 21 items in the Kindergarten Listening test, but one item was suppressed prior to scoring.

Item Review and Test Fairness

All items are expected to be fair for all examinees. Various procedures are employed to review item bias. Once the items are developed, they must go through a series of content and bias reviews and analyses prior to being selected as part of the item pool. A content and bias review has two purposes: to ensure that the items are grade-level appropriate and to ensure that any sensitivity issues are identified and addressed. Grade-level appropriateness is evaluated by grade-level teachers who possess the on-the-ground knowledge of how content is taught in the classroom. Sensitivity reviews ensure that items are free of offensive, disturbing, or inappropriate language or content.

Content reviews and sensitivity and bias reviews were conducted on all operational items. The item review committees reviewed all operational items before the operational test administration.

Item Selection

In selecting items for the reconfigured CELApro tests in Grades K–2 and 3–5, the primary criterion was to meet the content specifications represented by test blueprints, while at the same time maintaining the desired statistical properties of *LAS Links*. This involved an iterative process in which test characteristic curves and standard errors of measurement were examined after each preliminary item selection. Selections were revised as necessary in order to obtain an acceptable match to the statistical properties of the previous *LAS Links* assessments at each grade level.

Minimizing Test Bias

The position of CTB/McGraw-Hill concerning test bias is based on two general propositions. First, students may differ in their background knowledge, cognitive and academic skills, language, attitudes, and values. To the degree that these differences are large, no single curriculum and no single set of instructional materials will be equally suitable for all. Therefore, no single test will be equally appropriate for all. Furthermore, it is difficult to specify what amount of difference can be called large and to determine how these differences will affect the outcome of a particular test.

Second, schools have been assigned the tasks of developing certain basic cognitive skills and supporting English language proficiency (ELP) among all students. Therefore, there is a need for ELP tests that measure the common skills and bodies of knowledge that are common to English language learners. The test publisher's task is to develop assessments that measure English language proficiency without introducing extraneous or construct-irrelevant elements in the performances on which the measurement is based. If these tests require that students have culture-specific knowledge and skills not taught in school, differences in performance among students can occur because of differences in student background and out-of-school learning. Such tests are measuring different things for different groups and can be called biased (Camilli and Shepard, 1994; Green, 1975). In order to lessen this bias, CTB/McGraw-Hill strives to minimize the role of the extraneous elements, thereby increasing the number of students for whom the test is appropriate. Careful attention is given during the test construction process to lessen the influence of these elements for large numbers of students. Unfortunately, in some cases, these elements may continue to play a substantial role.

Four measures were taken to minimize bias in the *LAS Links* assessments. The first was based

on the premise that careful editorial attention to validity is an essential step in keeping bias to a minimum. Bias can occur only if the test is measuring different things for different groups. If the test entails irrelevant skills or knowledge, however common, the possibility of bias is increased. Thus, careful attention was given to content validity during the item-writing and item-selection process.

The second way bias was minimized was by following the McGraw-Hill guidelines designed to reduce or eliminate bias. Item writers were directed to the following published guidelines: *Guidelines for Bias-Free Publishing* (MacMillan/McGraw-Hill, 1993a) and *Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals* (Macmillan/McGraw-Hill, 1993b). Developers reviewed *LAS Links* assessment materials with these considerations in mind. Such internal editorial reviews were conducted by at least four different people: a content editor, who directly supervised the item writers; the project director; a style editor; and a proofreader. The final test built from the tryout materials was again reviewed by at least these same people.

In the third effort to minimize bias, educational community professionals who represent various ethnic groups reviewed all *LAS Links* tryout materials. They were asked to consider and comment on the appropriateness of language, subject matter, and representation of groups of people.

It is believed that these three procedures both improve the quality of an assessment and reduce item and test bias. However, current evidence suggests that expertise in this area is no substitute for data. Reviewers are often wrong about which items perform differently between specific subgroups of students, apparently because some of their ideas about how students will react to items may be inaccurate (Camilli and Shepard, 1994; Sandoval and Mille, 1979; Scheuneman, 1984). Thus, a fourth method for minimizing bias, an empirical approach, was also used to identify potential sources of item bias. For language tests, these are differential item functioning (DIF) studies, since criterion-related validities are essentially unobtainable for such tests. DIF studies include a systematic item analysis to determine whether examinees with the same underlying level of ability have the same probability of getting the item correct. Items identified with DIF are then examined to determine whether item performance differences between identifiable subgroups of the population are due to extraneous or construct-irrelevant information, making the items unfairly difficult. The inclusion of these items is minimized in the test development process. DIF studies have been routinely done for all major test batteries published by CTB/McGraw-Hill after 1970. DIF of the *LAS Links* assessment tryout items was assessed for students identified as males and females at each grade level in which the items were administered. In most cases, each item was administered at two grade spans.

Because *LAS Links* was built using item response theory, DIF analyses that capitalized on the information and item statistics provided by this theory were implemented. There are several IRT-based DIF procedures, including those that assess the equality of item parameters across groups (Lord, 1980) and those that assess area differences between item characteristic curves (Linn, Levine, Hastings, and Wardrop, 1981; Camilli and Shepard, 1994). However, these procedures require a minimum of 800 to 1,000 cases in each group of comparison to produce reliable and consistent results. In contrast, the Linn-Harnisch procedure (Linn and Harnisch, 1981) utilizes the information provided by the three-parameter IRT model but requires fewer cases. This was the procedure used to complete the gender DIF studies for the *LAS Links* field test data.

Part 3: Tested Population

A total of 102,461 students participated in the 2012 CELApro testing. Students in Kindergarten and Grade 1 formed the largest groups of examinees (13,074 and 13,127, respectively), with numbers generally decreasing at successive grade levels. The number of male examinees was slightly greater than the number of female students at each grade level. The examinee counts by grade and gender are shown in Table 4. Note that not all students completed all four of the CELApro content areas, so these numbers differ from those that appear in some of the subsequent tables within this report.

Table 4. Examinee Counts by Grade and Gender

Grade	Number of Examinees			Total
	Females	Males	Not Specified	
Kindergarten	6328	6746	0	13074
1	6399	6727	1	13127
2	5867	6378	0	12245
3	5619	5973	0	11592
4	4903	5362	0	10265
5	4252	4633	0	8885
6	3430	3877	0	7307
7	2654	3204	0	5858
8	2364	2749	0	5113
9	2009	2449	0	4458
10	1819	2165	0	3984
11	1521	1805	0	3326
12	1546	1681	0	3227
Total	48711	53749	1	102461

Student ethnicity, race, and home language are summarized by grade span in Tables 5 through 8. It should be noted that the results for ethnicity and race are not strictly comparable to those from prior years because the procedure for collecting and reporting this information was changed in 2011. Instead of obtaining this information from a single ethnicity variable, as was done in the past, the information was collected using a two-part question. The first part is an ethnicity variable, classifying students as Hispanic/Latino or non-Hispanic/non-Latino. The second part is a race variable, which allows for the classification of each student into one or more racial groups. The responses to these two parts are then combined to place each student into one and only one of seven distinct Federal Reporting Categories. Students who are identified in the first part as Hispanic/Latino are classified as Hispanic/Latino. Students who are identified as non-Hispanic/non-Latino are classified into the six remaining categories on the basis of the race information indicated in the second part of the question. The resulting percentages of students in the new Federal Reporting Categories are very similar to the percentages that resulted from the prior years' single-variable classification. Table 5 shows the distribution of students in the new Federal Reporting Categories. Tables 6 and 7 summarize the responses to the two-part question about ethnicity and race. Table 6 shows the numbers and percentages of students whose ethnicity was identified as Hispanic/Latino and

non-Hispanic/non-Latino. Table 7 shows the numbers and percentages of students in each racial group.

Table 5. Examinee Counts and Percents by Federal Ethnicity Reporting Category and Grade Span

Ethnicity	Grade Span							
	Grades K–2		Grades 3–5		Grades 6–8		Grades 9–12	
	N	%	N	%	N	%	N	%
Hispanic or Latino	31871	31.11	26341	25.71	15954	15.57	12199	11.91
American Indian/ Alaska Native*	152	0.15	138	0.13	82	0.08	83	0.08
Asian*	2922	2.85	2062	2.01	1078	1.05	1361	1.33
Black or African American*	1258	1.23	862	0.84	531	0.52	700	0.68
White*	1926	1.88	1168	1.14	547	0.53	563	0.55
Native Hawaiian or Other Pacific Islander*	90	0.09	60	0.06	43	0.04	43	0.04
Two or more races*	226	0.22	111	0.11	43	0.04	46	0.04
Missing	1	0	0	0	0	0	0	0
Total	38446	37.52	30742	30	18278	17.84	14995	14.63

* Note that these Federal Reporting Categories exclude students who were classified as Hispanic/Latino.

Table 6. Examinee Counts and Percents by Reported Hispanic/Latino Ethnicity and Grade Span

Hispanic/Latino	Grade Span							
	Grades K–2		Grades 3–5		Grades 6–8		Grades 9–12	
	N	%	N	%	N	%	N	%
Hispanic/Latino	31871	31.11	26341	25.71	15954	15.57	12199	11.91
Not Hispanic/Latino	6575	6.42	4401	4.30	2324	2.27	2796	2.73
Total	38446	37.52	30742	30	18278	17.84	14995	14.63

Table 7. Examinee Counts and Percents for Racial Classifications by Grade Span and Reported Hispanic/Latino Ethnicity*.

Racial Classifications	Ethnicity: Hispanic/Latino							
	Grades K-2 (N=32516)		Grades 3-5 (N=26910)		Grades 6-8 (N=16297)		Grades 9-12 (N=12512)	
	N	%	N	%	N	%	N	%
American Indian/Alaska Native	8083	7.72	6930	6.61	4060	3.88	3237	3.09
Asian	120	0.11	82	0.08	49	0.05	35	0.03
Black/ African American	177	0.17	136	0.13	69	0.07	64	0.06
White	23975	22.88	19597	18.71	12046	11.5	9121	8.71
Native Hawaiian or Other Pacific Islander	161	0.15	165	0.16	73	0.07	55	0.05

Racial Classifications	Ethnicity: Not Hispanic/Latino							
	Grades K-2 (N=6804)		Grades 3-5 (N=4516)		Grades 6-8 (N=2369)		Grades 9-12 (N=2843)	
	N	%	N	%	N	%	N	%
American Indian/Alaska Native	175	0.17	155	0.15	92	0.09	88	0.08
Asian	3100	2.96	2143	2.05	1108	1.06	1401	1.34
Black/ African American	1305	1.25	884	0.84	541	0.52	703	0.67
White	2105	2.01	1241	1.18	576	0.55	585	0.56
Native Hawaiian or Other Pacific Islander	119	0.11	93	0.09	52	0.05	66	0.06

*The percentages in Table 7 sum to more than 100% because some students were identified as belonging to more than one racial group. This table does not include students for whom Hispanic/Non-Hispanic ethnicity information was missing.

Table 8 shows the distribution of students by home language. As in previous years, the home language was indicated as Spanish for the overwhelming majority of students (84% in Grades K–2, 86% in Grades 3–5, 87% in Grades 6–8, and 82% in Grades 9–12).

Table 8. Home Language (204 Languages Represented)

Language	Grade Span							
	Grades K–2		Grades 3–5		Grades 6–8		Grades 9–12	
	N	%	N	%	N	%	N	%
Abkhaz	3	0.00	1	0.00	0	0.00	0	0.00
Afrikaans	12	0.01	5	0.00	3	0.00	7	0.01
Akan	16	0.02	5	0.00	2	0.00	5	0.00
Albanian	14	0.01	6	0.01	5	0.00	2	0.00
Algonquin	2	0.00	0	0.00	0	0.00	0	0.00
Amharic	279	0.27	168	0.16	70	0.07	96	0.09
Anuak	1	0.00	2	0.00	0	0.00	1	0.00
Apache	1	0.00	0	0.00	0	0.00	0	0.00
Arabic	590	0.58	375	0.37	169	0.16	201	0.20
Arapaho	0	0.00	0	0.00	1	0.00	0	0.00
Armenian	7	0.01	14	0.01	2	0.00	3	0.00
Assamese	2	0.00	6	0.01	7	0.01	2	0.00
Assyrian	1	0.00	1	0.00	0	0.00	0	0.00
Awadhi	0	0.00	0	0.00	1	0.00	0	0.00
Azerbaijani	1	0.00	0	0.00	0	0.00	0	0.00
Bambara	5	0.00	5	0.00	0	0.00	2	0.00
Bashkir	0	0.00	0	0.00	1	0.00	2	0.00
Bassa	3	0.00	0	0.00	1	0.00	0	0.00
Bengali	29	0.03	14	0.01	8	0.01	9	0.01
Berber	1	0.00	1	0.00	1	0.00	0	0.00
Bhojpuri	2	0.00	0	0.00	0	0.00	0	0.00
Bislama	0	0.00	0	0.00	0	0.00	1	0.00
Bosnian	48	0.05	29	0.03	18	0.02	8	0.01
Bulgarian	26	0.03	13	0.01	1	0.00	1	0.00
Burmese	95	0.09	94	0.09	72	0.07	129	0.13
Caddo	1	0.00	1	0.00	0	0.00	1	0.00
Catalan	1	0.00	0	0.00	0	0.00	1	0.00
Cebuano	3	0.00	6	0.01	3	0.00	1	0.00
Chamorro	4	0.00	4	0.00	3	0.00	4	0.00
Chinese, Cantonese	138	0.13	95	0.09	25	0.02	44	0.04
Chinese, Hakka	3	0.00	4	0.00	2	0.00	5	0.00
Chinese, Mandarin	402	0.39	210	0.20	91	0.09	136	0.13
Chinese, Min Nan	1	0.00	0	0.00	1	0.00	0	0.00
Chinese, Wu	5	0.00	0	0.00	1	0.00	0	0.00
Chinook, Upper	1	0.00	0	0.00	0	0.00	0	0.00
Chinyanja	3	0.00	2	0.00	0	0.00	0	0.00

Table 8. Home Language (continued)

Language	Grade Span							
	Grades K–2		Grades 3–5		Grades 6–8		Grades 9–12	
	N	%	N	%	N	%	N	%
Choctaw	0	0.00	0	0.00	0	0.00	1	0.00
Chuuk	1	0.00	0	0.00	0	0.00	0	0.00
Chuukese	17	0.02	9	0.01	8	0.01	6	0.01
Chuvash	0	0.00	1	0.00	2	0.00	0	0.00
Cora	32	0.03	13	0.01	17	0.02	22	0.02
Croatian	6	0.01	6	0.01	1	0.00	1	0.00
Czech	24	0.02	5	0.00	6	0.01	5	0.00
Dakota	0	0.00	1	0.00	0	0.00	0	0.00
Danish	12	0.01	8	0.01	2	0.00	0	0.00
Dari	3	0.00	6	0.01	1	0.00	1	0.00
Deccan	1	0.00	2	0.00	0	0.00	0	0.00
Dinka	11	0.01	7	0.01	3	0.00	0	0.00
Dutch	19	0.02	13	0.01	2	0.00	2	0.00
English	12	0.01	7	0.01	9	0.01	18	0.02
Estonian	1	0.00	0	0.00	0	0.00	0	0.00
Ewe	13	0.01	4	0.00	3	0.00	2	0.00
Eyak	0	0.00	0	0.00	0	0.00	1	0.00
Faroese	0	0.00	1	0.00	1	0.00	1	0.00
Farsi, Eastern	74	0.07	50	0.05	14	0.01	17	0.02
Farsi, Western	28	0.03	16	0.02	6	0.01	9	0.01
Fijian	0	0.00	0	0.00	1	0.00	0	0.00
Finnish	7	0.01	6	0.01	0	0.00	2	0.00
French	151	0.15	123	0.12	58	0.06	102	0.10
French Creole	17	0.02	22	0.02	10	0.01	9	0.01
Fulfulde, Nigerian	12	0.01	6	0.01	6	0.01	5	0.00
Ga	4	0.00	1	0.00	1	0.00	0	0.00
Ganda	4	0.00	2	0.00	1	0.00	2	0.00
Georgian	3	0.00	0	0.00	0	0.00	2	0.00
German	137	0.13	64	0.06	23	0.02	30	0.03
Gokana	0	0.00	0	0.00	2	0.00	0	0.00
Gola	0	0.00	0	0.00	1	0.00	2	0.00
Grebo	0	0.00	0	0.00	1	0.00	0	0.00
Greek	7	0.01	6	0.01	1	0.00	1	0.00
Guarani	0	0.00	1	0.00	0	0.00	0	0.00
Gujarati	21	0.02	9	0.01	3	0.00	0	0.00
Haitian, Creole French	7	0.01	5	0.00	5	0.00	3	0.00
Hausa	1	0.00	1	0.00	0	0.00	2	0.00

Table 8. Home Language (continued)

Language	Grade Span							
	Grades K–2		Grades 3–5		Grades 6–8		Grades 9–12	
	N	%	N	%	N	%	N	%
Hawaiian	1	0.00	0	0.00	0	0.00	2	0.00
Hebrew	10	0.01	6	0.01	3	0.00	2	0.00
Hindi	102	0.10	36	0.04	14	0.01	20	0.02
Hmong	176	0.17	138	0.13	75	0.07	59	0.06
Hungarian	17	0.02	9	0.01	4	0.00	2	0.00
Icelandic	0	0.00	1	0.00	0	0.00	0	0.00
Igbo	24	0.02	10	0.01	6	0.01	1	0.00
Ilocano	0	0.00	1	0.00	2	0.00	1	0.00
Indonesian	59	0.06	32	0.03	16	0.02	8	0.01
Italian	16	0.02	10	0.01	6	0.01	7	0.01
Iu Mien	2	0.00	0	0.00	0	0.00	0	0.00
Japanese	96	0.09	61	0.06	22	0.02	15	0.01
Kanjobal	31	0.03	18	0.02	16	0.02	5	0.00
Kannada	11	0.01	2	0.00	1	0.00	0	0.00
Karelian	1	0.00	0	0.00	2	0.00	1	0.00
Karen	92	0.09	112	0.11	91	0.09	140	0.14
Kazakh	0	0.00	1	0.00	1	0.00	0	0.00
Keres, Eastern	3	0.00	0	0.00	2	0.00	0	0.00
Khasi	1	0.00	0	0.00	0	0.00	0	0.00
Khmer	61	0.06	68	0.07	29	0.03	33	0.03
Kikuyu	1	0.00	0	0.00	1	0.00	0	0.00
Kinyarwanda	5	0.00	4	0.00	5	0.00	7	0.01
Konkani	1	0.00	1	0.00	0	0.00	0	0.00
Korean	277	0.27	215	0.21	99	0.10	96	0.09
Kosraen	5	0.00	3	0.00	0	0.00	0	0.00
Kpelle	0	0.00	0	0.00	1	0.00	0	0.00
Krahn	0	0.00	1	0.00	2	0.00	4	0.00
Krio	0	0.00	4	0.00	3	0.00	1	0.00
Kru Languages	0	0.00	1	0.00	1	0.00	2	0.00
Kru Pidgin English	0	0.00	0	0.00	0	0.00	1	0.00
Kurdi/Kurdish Bandinani	13	0.01	15	0.01	8	0.01	12	0.01
Lakota	0	0.00	0	0.00	0	0.00	2	0.00
Lao	62	0.06	61	0.06	33	0.03	20	0.02
Latin	1	0.00	0	0.00	0	0.00	0	0.00
Latvian	0	0.00	3	0.00	0	0.00	3	0.00
Liberian English	3	0.00	7	0.01	4	0.00	3	0.00
Lingala	2	0.00	4	0.00	1	0.00	2	0.00

Table 8. Home Language (continued)

Language	Grade Span							
	Grades K–2		Grades 3–5		Grades 6–8		Grades 9–12	
	N	%	N	%	N	%	N	%
Lithuanian	6	0.01	5	0.00	4	0.00	2	0.00
Maay	12	0.01	15	0.01	5	0.00	8	0.01
Macedonian	0	0.00	1	0.00	0	0.00	0	0.00
Malagasy	0	0.00	1	0.00	0	0.00	0	0.00
Malay	9	0.01	6	0.01	5	0.00	2	0.00
Malayalam	23	0.02	13	0.01	4	0.00	4	0.00
Mandan	1	0.00	0	0.00	0	0.00	0	0.00
Mandinka	16	0.02	8	0.01	6	0.01	6	0.01
Maniinkakan, Western	0	0.00	0	0.00	0	0.00	1	0.00
Marathi	20	0.02	5	0.00	0	0.00	0	0.00
Marshallese	19	0.02	15	0.01	10	0.01	11	0.01
Marwari	2	0.00	0	0.00	0	0.00	0	0.00
Maya	0	0.00	2	0.00	1	0.00	4	0.00
Mende	1	0.00	0	0.00	0	0.00	2	0.00
Mongolian	39	0.04	24	0.02	10	0.01	21	0.02
Mono	0	0.00	1	0.00	0	0.00	0	0.00
Navajo	49	0.05	58	0.06	33	0.03	37	0.04
Ndebele	1	0.00	0	0.00	0	0.00	0	0.00
Nepali	159	0.16	138	0.13	141	0.14	281	0.27
Newari	1	0.00	0	0.00	1	0.00	0	0.00
Norwegian	0	0.00	3	0.00	1	0.00	1	0.00
Nuer	5	0.00	6	0.01	0	0.00	1	0.00
Nyanja	0	0.00	2	0.00	1	0.00	0	0.00
Omaha	0	0.00	1	0.00	1	0.00	0	0.00
Oriya	4	0.00	3	0.00	0	0.00	0	0.00
Oromo, West-Central	55	0.05	16	0.02	12	0.01	19	0.02
Palauan	2	0.00	1	0.00	1	0.00	3	0.00
Pampangan	2	0.00	2	0.00	1	0.00	0	0.00
Panjabi, Eastern	33	0.03	9	0.01	4	0.00	9	0.01
Panjabi, Western	5	0.00	3	0.00	0	0.00	0	0.00
Papago	1	0.00	0	0.00	0	0.00	0	0.00
Pashto, Central	2	0.00	2	0.00	3	0.00	5	0.00
Pashto, Northern	4	0.00	3	0.00	1	0.00	1	0.00
Pashto, Southern	5	0.00	6	0.01	1	0.00	0	0.00
Pohnpeian	13	0.01	5	0.00	1	0.00	4	0.00
Polish	92	0.09	38	0.04	16	0.02	10	0.01
Portuguese	49	0.05	34	0.03	13	0.01	23	0.02

Table 8. Home Language (continued)

Language	Grade Span							
	Grades K–2		Grades 3–5		Grades 6–8		Grades 9–12	
	N	%	N	%	N	%	N	%
Pulaar	2	0.00	0	0.00	4	0.00	1	0.00
Quechua, Huanuco, Panao	1	0.00	0	0.00	0	0.00	0	0.00
Quiche, Central	0	0.00	0	0.00	0	0.00	3	0.00
Romani	2	0.00	0	0.00	0	0.00	0	0.00
Romanian	35	0.03	14	0.01	7	0.01	8	0.01
Ruanda	0	0.00	1	0.00	1	0.00	0	0.00
Rundi	20	0.02	16	0.02	12	0.01	13	0.01
Russian	399	0.39	261	0.25	142	0.14	156	0.15
Rwanda	0	0.00	1	0.00	3	0.00	5	0.00
Samoan	18	0.02	10	0.01	7	0.01	4	0.00
Sango	3	0.00	1	0.00	2	0.00	0	0.00
Serbian	10	0.01	4	0.00	1	0.00	6	0.01
Serbo-Croatian	4	0.00	5	0.00	7	0.01	6	0.01
Sesotho	2	0.00	0	0.00	0	0.00	0	0.00
Shona	7	0.01	3	0.00	1	0.00	0	0.00
Sindhi	3	0.00	0	0.00	1	0.00	0	0.00
Sinhala	1	0.00	2	0.00	0	0.00	0	0.00
Sioux	1	0.00	0	0.00	1	0.00	0	0.00
Slovak	6	0.01	5	0.00	3	0.00	3	0.00
Slovenian	3	0.00	0	0.00	1	0.00	0	0.00
Somali	267	0.26	186	0.18	147	0.14	161	0.16
Spanish	32145	31.37	26533	25.90	16024	15.64	12236	11.94
Spokane	2	0.00	0	0.00	0	0.00	0	0.00
Sundanese	1	0.00	0	0.00	1	0.00	2	0.00
Susu	0	0.00	0	0.00	0	0.00	1	0.00
Swahili	39	0.04	47	0.05	45	0.04	75	0.07
Swedish	34	0.03	8	0.01	2	0.00	7	0.01
Sylhetti	0	0.00	0	0.00	0	0.00	1	0.00
Tagalog	103	0.10	100	0.10	53	0.05	52	0.05
Tahitian	0	0.00	1	0.00	0	0.00	0	0.00
Tajik	3	0.00	1	0.00	2	0.00	1	0.00
Tamil	54	0.05	14	0.01	3	0.00	1	0.00
Telugu	95	0.09	27	0.03	8	0.01	2	0.00
Thai	32	0.03	24	0.02	19	0.02	23	0.02
Tibetan	8	0.01	3	0.00	2	0.00	3	0.00
Tigrigna	105	0.10	49	0.05	31	0.03	53	0.05

Table 8. Home Language (continued)

Language	Grade Span							
	Grades K–2		Grades 3–5		Grades 6–8		Grades 9–12	
	N	%	N	%	N	%	N	%
Tiwa, Northern	0	0.00	0	0.00	1	0.00	0	0.00
Tonga	6	0.01	2	0.00	2	0.00	4	0.00
Tongan	4	0.00	3	0.00	1	0.00	0	0.00
Tonkawa	0	0.00	0	0.00	0	0.00	1	0.00
Tsonga	2	0.00	1	0.00	0	0.00	1	0.00
Tswana	2	0.00	1	0.00	0	0.00	0	0.00
Turkish	44	0.04	27	0.03	21	0.02	14	0.01
Turkmen	1	0.00	0	0.00	0	0.00	1	0.00
Twi	54	0.05	47	0.05	22	0.02	31	0.03
Ukrainian	29	0.03	30	0.03	27	0.03	18	0.02
Urdu	73	0.07	36	0.04	9	0.01	7	0.01
Ute	2	0.00	11	0.01	8	0.01	16	0.02
Uzbek	10	0.01	3	0.00	3	0.00	5	0.00
Vengo	1	0.00	0	0.00	0	0.00	0	0.00
Vietnamese	738	0.72	532	0.52	235	0.23	237	0.23
Welsh	1	0.00	0	0.00	0	0.00	0	0.00
Wolof	3	0.00	6	0.01	2	0.00	1	0.00
Yapese	0	0.00	2	0.00	0	0.00	0	0.00
Yoruba	10	0.01	6	0.01	8	0.01	3	0.00
Zarma	1	0.00	0	0.00	0	0.00	0	0.00
Zulu	0	0.00	2	0.00	0	0.00	0	0.00
Not Specified	32	0.02	11	0.01	9	0.00	22	0.01
Total	38446	37.52	30742	30.00	18278	17.84	14995	14.63

Because some students required accommodations in order to access the items, the following accommodations were available:

- Braille
- Large Print
- Use of a Scribe to Record Responses
- Signing
- Use of Assistive Communicative Devices
- Use of Approved Nonstandard Accommodations
- Oral Presentation

These accommodations are summarized by content area and grade in Tables 9 through 12.

Table 9. Speaking Accommodations by Grade

	Grade													Total
	K	1	2	3	4	5	6	7	8	9	10	11	12	
One Accommodation														
None	12859	12921	12038	11387	10082	8743	7127	5688	4980	4343	3879	3216	3132	100395
Braille	0	0	0	1	1	3	0	1	1	1	0	0	2	10
Large-Print	0	1	3	4	5	5	0	0	0	0	0	0	0	18
Signing	3	2	1	1	1	3	2	0	3	2	1	3	2	24
Assistive-Tech	3	1	4	2	2	5	1	1	1	0	0	1	1	22
Non-standard	0	2	0	0	1	0	0	0	0	0	1	0	1	5
Two Accommodations														
None & Braille	1	2	0	3	0	2	0	0	2	0	0	0	0	10
None & Large-Print	1	0	0	2	1	0	0	0	1	0	0	0	0	5
None & Signing	0	1	0	0	0	0	0	0	0	2	1	0	2	6
None & Assistive-Tech	2	0	1	1	1	0	0	0	1	0	0	0	0	6
None & Non-standard	1	5	1	1	0	1	2	2	2	0	0	0	0	15
Signing & Assistive-Tech	3	2	2	3	3	1	0	0	1	0	0	0	0	15
Assistive-Tech & Non-standard	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Four Accommodations														
None & Braille & Large-Print & Non-stand	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Not Specified	201	189	195	187	167	122	175	166	121	110	102	106	87	1928
Total	13074	13127	12245	11592	10265	8885	7307	5858	5113	4458	3984	3326	3227	102461

Table 10. Listening Accommodations by Grade

	Grade													Total
	K	1	2	3	4	5	6	7	8	9	10	11	12	
One Accommodation														
None	12851	12914	12027	11388	10075	8743	7130	5687	4985	4343	3874	3213	3132	100362
Braille	0	0	0	2	1	3	0	1	1	1	0	0	2	11
Large-Print	1	1	3	5	6	5	1	0	0	0	1	0	0	23
Signing	3	2	0	0	1	3	2	0	3	2	1	3	4	24
Assistive-Tech	3	1	4	1	2	5	1	1	1	0	0	1	1	21
Non-standard	0	2	0	1	1	0	1	0	0	0	0	0	1	6
Two Accommodations														
None & Braille	0	0	0	1	0	0	0	0	1	0	0	0	0	2
None & Large-Print	0	0	1	1	0	0	0	0	1	0	0	0	0	3
None & Signing	0	0	0	0	0	0	0	0	0	2	1	1	0	4
None & Assistive-Tech	2	0	1	1	0	0	0	0	0	0	0	0	0	4
None & Non-standard	2	5	2	0	1	1	0	1	0	0	1	0	0	13
Signing & Assistive-Tech	3	2	2	3	3	1	0	0	1	0	0	0	0	15
Assistive-Tech & Non-standard	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Three Accommodations														
None & Braille & Large-Print	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Not Specified	209	200	205	189	174	124	172	167	120	110	106	108	87	1971
Total	13074	13127	12245	11592	10265	8885	7307	5858	5113	4458	3984	3326	3227	102461

Table 11. Reading Accommodations by Grade

	Grade													Total
	K	1	2	3	4	5	6	7	8	9	10	11	12	
One Accommodation														
None	12791	12873	11985	11367	10028	8715	7106	5666	4967	4329	3852	3196	3108	99983
Braille	0	0	0	3	1	3	0	1	1	1	0	0	1	11
Large-Print	2	1	3	5	5	5	1	0	0	0	2	0	1	25
Scribe	2	4	4	0	2	2	4	4	1	1	1	1	1	27
Signing	2	2	0	0	1	2	2	0	3	2	1	3	4	22
Assistive-Tech	3	0	2	1	2	6	1	1	0	0	0	1	0	17
Non-standard	1	2	0	1	0	0	0	0	0	0	0	0	1	5
Two Accommodations														
None & Braille	0	1	0	0	2	0	0	0	0	0	0	0	0	3
None & Large-Print	0	0	0	0	0	0	0	0	1	0	0	0	0	1
None & Scribe	1	1	2	2	0	0	0	0	0	1	0	0	2	9
None & Signing	1	1	0	0	0	0	0	0	0	1	1	1	0	5
None & Assistive-Tech	2	0	1	1	0	0	0	0	0	0	0	0	0	4
None & Non-standard	0	5	1	0	0	1	1	2	0	0	0	0	0	10
Braille & Scribe	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Large-Print & Scribe	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Scribe & Signing	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Scribe & Assistive-Tech	0	0	0	0	0	0	1	0	1	0	0	0	0	2
Signing & Assistive-Tech	3	2	2	3	3	1	0	0	1	0	0	0	0	15

Table 11. Reading Accommodations by Grade (continued)

	Grade													
	K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
Three Accommodations														
Scribe & Assistive-Tech & Non-standard	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Not Specified	266	235	245	209	219	150	191	183	138	123	127	124	108	2318
Total	13074	13127	12245	11592	10265	8885	7307	5858	5113	4458	3984	3326	3227	102461

Table 12. Writing Accommodations by Grade

	Grade													Total
	K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
One Accommodation														
None	12784	12870	11981	11358	10026	8706	7097	5665	4963	4327	3847	3194	3110	99928
Braille	0	0	0	3	1	3	0	1	1	1	0	0	1	11
Large-Print	2	1	3	4	5	3	1	0	1	0	2	0	1	23
Scribe	2	4	5	4	2	4	5	3	2	1	1	1	1	35
Signing	2	1	0	0	1	1	2	0	3	2	1	3	4	20
Assistive-Tech	3	0	3	1	3	8	4	2	0	0	0	2	0	26
Non-standard	1	2	0	1	0	0	0	0	0	0	0	0	1	5

Table 12. Writing Accommodations by Grade (continued)

	Grade													
	K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
Two Accommodations														
None & Scribe	1	0	1	1	0	1	0	1	0	0	0	0	2	7
None & Signing	0	0	0	0	0	0	0	0	0	1	1	1	0	3
None & Assistive-Tech	2	0	1	1	0	1	0	0	0	0	0	0	0	5
None & Non-standard	0	1	1	0	0	1	1	1	0	0	0	0	0	5
Braille & Scribe	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Large-Print & Scribe	0	1	0	0	1	2	0	0	0	0	0	0	0	4
Large-Print & Assistive-Tech	0	0	0	1	0	0	0	0	0	0	0	0	0	1
Scribe & Signing	0	1	0	0	0	1	0	1	0	0	0	0	0	3
Scribe & Assistive-Tech	0	1	0	0	0	0	1	0	1	0	0	0	0	3
Scribe & Non-standard	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Signing & Assistive-Tech	3	2	2	3	3	1	0	0	1	0	0	0	0	15
Three Accommodations														
None & Braille & Large-Print	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Scribe & Assistive-Tech & Non-standard	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Not Specified	274	243	248	215	221	153	196	184	140	126	132	125	106	2363
Total	13074	13127	12245	11592	10265	8885	7307	5858	5113	4458	3984	3326	3227	102461

Part 4: Test Administration

The Colorado English Language Assessment was first administered in Spring 2006. In 2007, the administration was moved to winter, and in January 2012, the CELApro was administered to 102,461students. This test consists of four separately administered sections assessing Speaking, Listening, Reading, and Writing proficiency.
The CELApro Speaking section is individually administered. The Listening, Reading, and Writing sections may be administered to a group or individually administered, depending upon the needs of the particular examinees being tested.

CELApro test examiners must be proficient English speakers who are able to model clear pronunciation of English phonemes. For group-administered K–2 Reading and Writing sections, students must be grouped by grade. Students in Grades 3 and above may be grouped either by grade or by grade span for all of the group-administered sections. Examiners are also instructed to group students by English proficiency in different rooms or at different times, if possible.

All sections of the test are untimed in order to give students every opportunity to demonstrate their proficiency in English. The estimated administration times and administration modes are shown in Table 13. Actual times may vary.

Table 13. Estimated Administration Time and Administration Mode by Skill Area

Skill Area	Estimated Administration Time (all tests are untimed)	Administration Mode
Speaking	10 Minutes – All Grades	Individual
Listening	20 Minutes – All Grades	Group or Individual
Reading	35 Minutes – Kindergarten 45 Minutes – Grades 1–12	Group or Individual
Writing	35 Minutes – Grades K–1 45 Minutes – Grades 2–12	Group or Individual

All test examiners, school assessment coordinators (SACs), and district assessment coordinators (DACs) were instructed in standardized test administration and scoring procedures prior to the test administration.

The Speaking Subtests

The Speaking test is individually administered by a fluent English speaker who reads the test questions aloud while pointing to illustrations. All items are in CR format and scored with performance-based rubrics that direct the attention of the rater (generally the examiner) to the student's use of vocabulary, social and academic language, complex and grammatically correct verbal expressions, and length of responses. The Speaking test takes approximately 10 minutes per student to administer and consists of four subtests as follows:

Speak in Words
In *Speak in Words*, the examiner points to objects depicted in cue pictures and asks questions such as, "What is this?" and "What is it used for?" Students respond with single words and short phrases to identify the objects and answer questions related to those objects. Student responses are scored as correct (C), incorrect (I), or no response (NR).

Speak in Sentences
In *Speak in Sentences*, students respond in complete sentences to describe activities or actions. The examiner points to each cue picture and directs the student to respond to prompts such as, "Tell me what is happening in the picture," "Tell me exactly where the book is located," and "Please give me clear directions on how to go from Place A to Place B." Student responses are scored with a 0–3 point rubric.

Make Conversation
Students also respond in complete sentences in *Make Conversation*. However, instead of describing cue pictures, students respond to the examiner's prompts, such as, "Tell someone to do something," "Ask someone for something," "Describe how to do something," and "Explain why we do something." Student responses are scored with a 0–3 point rubric.

Tell a Story
In *Tell a Story*, students produce multiple sentences explaining what is happening in a series of four cue pictures. The pictures illustrate a story with a beginning, a middle, and an end. Pointing to the series of four pictures, the examiner begins the story by reading a story starter to contextualize the pictures without giving away vocabulary or key content. Student responses are scored with a 0–4 point rubric.

The Listening Subtests

The Listening test is administered to a group of students by a fluent English speaker who reads from the Examiner's Guide and uses the audio CD. All Listening items are in multiple-choice format and measure general comprehension and inferential and critical thinking skills at a discourse level that integrates academic language. Students listen to classroom English to demonstrate language proficiency levels within each grade span. The Listening test takes approximately 20 minutes per group to administer and consists of three subtests as follows:

Listen for Information
In *Listen for Information*, students hear instructions typical of those provided by a classroom teacher. Instructions vary in length from one to three sentences and must be played from the audio CD. The examiner then asks students which of three answer choices restates the instructions they heard. Instructions and answers may contain idioms and different syntactical structures.

Listen in the Classroom
Listen in the Classroom assesses comprehension of academic language, where students hear two short exchanges typical of classroom discussions. The listening passages, questions, and text answer choices must be played from the audio CD. After

listening, students respond to three questions about what they heard. Each question has three answer choices. For each Grade Span, the passage length is as follows:

Grade Span	Passage Length
K–2	50–60 words
3–5	60–90 words
6–8	60–100 words
9–12	90–130 words

Listen and Comprehend

A longer listening passage included in *Listen and Comprehend* assesses comprehension of narratives. Questions ask about main ideas, details, inferences, and idioms. The listening passages, questions, and text answer choices must be played from the audio CD. Students are asked four questions about the passage. Each question has three answer choices.

Grade Span	Passage Length	Genre	Percentage
K–2	150–200 words	Fiction	83%
		Nonfiction	17%
3–5	200–250 words	Fiction	50%
		Nonfiction	50%
6–8	200–250 words	Fiction	33%
		Nonfiction	67%
9–12	225–325 words	Fiction	33%
		Nonfiction	67%

The Reading Subtests

The Reading test is usually administered to a group by a fluent English speaker who reads from the Examiner's Guide. All Reading items are in MC format. Some items evaluate phonemic awareness as the basis for recognizing words and developing vocabulary. In other items, students read literary and informational grade-appropriate texts to demonstrate sentence-level and discourse-level reading ability as well as inferential skills. The Reading test takes approximately 35–45 minutes to administer and consists of three subtests as follows:

Analyze Words

In *Analyze Words*, students respond to discrete items in a variety of formats addressing four word analysis tasks: identifying rhyming words, applying letter-sound relationships to read English words, applying letter-sound relationships to read English phonemes, and applying knowledge of morphemes and syntax to word meaning. Each question has three answer choices.

Read Words

In *Read Words*, for Grades K–2, students demonstrate vocabulary by classifying words, selecting written words to match those spoken by the examiner, and matching pictures of objects to their written descriptions. In all other grade levels, students demonstrate vocabulary by choosing synonyms or antonyms of a given word and/or choosing words that correctly complete sentences. Additionally, students in Grades 6–12 are tested on idiomatic expressions. Each question has three answer choices.

Read for Understanding

Higher-level reading skills are evaluated in *Read for Understanding*, in which students respond to passages representing various literary genres (e.g., fiction, nonfiction, and poetry). Questions address three tasks: demonstrating reading comprehension, identifying important literary features of text, and applying learning strategies to interpretation. Students in Kindergarten read along as the examiner reads passages aloud; then students identify one of three picture choices that correspond with the reading passage. Students in Grades 1 and 2 read two additional passages independently. Students in Grades 3–5 read passages without assistance and choose corresponding pictures or text. Students in higher grades read passages without assistance and select from four written answer choices.

Grade Span	Passage Length	Genre	Percentage
K	50–100 words	Fiction	100%
		Nonfiction	0%
1–2	100–150 words	Fiction	100%
		Nonfiction	0%
3–5	175–275 words	Fiction	50%
		Nonfiction	50%
6–8	250–350 words	Fiction (Poetry)	50%
		Nonfiction	50%
9–12	250–450 words	Fiction (Poetry)	50%
		Nonfiction	50%

The Writing Subtests

The Writing test is usually administered to a group by a fluent English speaker who reads from the Examiner's Guide. The test includes both MC and CR items that assess both receptive and productive domains. In the first section, MC items engage students to identify appropriate grammar, mechanics, and syntax. In the second section, students respond to prompts in the form of phrases, sentences, and paragraphs.

Responses to CR items are evaluated with performance-based rubrics (on a 0–3 or 0–4 point scale depending on the item) that direct the rater's attention to the student's use of English grammar and the appropriate use of discourse. The test takes approximately 35–45 minutes to administer and consists of four subtests as follows, with the exception that students in Grades K–1 do not take *Write in Detail*:

Use Conventions
Discrete point items in *Use Conventions* assess whether students can identify correct uses of grammar, capitalization, punctuation, and sentence structure. Each item has three answer choices.

Write About
In *Write About*, students in Grades K–1 write one sentence and students in Grades 2–12 write two sentences to describe a picture. Responses are scored with a 0–3 point rubric.

Write Why
In *Write Why*, students make a choice between two alternatives and write to explain the reason for the choice they make. In Grades K–1, students write one reason; in Grades 2–12, students write two reasons. Responses are scored with a 0–3 point rubric.

Write in Detail
Prompts in *Write in Detail* elicit longer responses. Students in Grade 2 write to describe what is happening in a sequence of four pictures. Students in Grades 3–12 organize their ideas and write paragraphs or essays responding to a written prompt. Responses are scored with a 0–4 point rubric. Students in Grades K–1 do not take *Write in Detail*.

Administration Training

The Administration Training Workshops for 2012 were conducted by the Colorado Department of Education. Training took place in 4 locations in, Limon, Pueblo, , Denver, Grand Junction, and online. These locations were selected to cover the state's training needs geographically as well as in terms of district size. A total of 139 participants attended the CELApro workshops. Table 14 shows the breakdown of attendees by workshop location.

Table 14. Number of Attendees at Pre-Administration Training Workshops

Location and Date	Number of Attendees
Limon	22
Pueblo	33
Denver	22
Grand Junction	18
Online	44
Total Attendees	139

Workshop Setup

The environment of the Pre-Administration Workshop is friendly and facilitates small-group discussion. Participants' seats were not assigned. CELApro training was first, followed by the CSAPA Administration Training.

Training Materials Development

The training materials were developed to reduce complexity, mirror the trainer's script, and ensure clarity in the use of the contents within the Training Folder and Training DVD throughout the training. Following are the details of the purpose of each component.

Training Folder
The CELApro Administration Training Folder contained a copy of a PowerPoint presentation, a list of acronyms, Speaking Practice Scoring Sheets, and Speaking Rubrics. The purpose of these materials was to allow for easy navigation. Navigation through the training materials is key when training a large number of participants, which in turn facilitates the learning process and helps participants gain the understanding needed to conduct their own trainings.

Training DVD
Another important part of the training materials is the coordination between the DVD and the Training Folder. The Training DVD gives an overview of each subtest for all grade spans: Speaking, Listening, Reading, and Writing. Because the Speaking test is scored by Test Examiners during test administration, the DVD component is critical for training. Teachers and CTB experts scored all Speaking samples. Participants used Speaking Practice Scoring Sheets as part of the scoring calibration exercises.

Part 5: Scoring

The 2012 CELApro tests were scored and processed by CTB's scoring team using the standardized methods and procedures previously developed for the *LAS Links* program. The CELApro scoring team consists of trained technical specialists who are responsible for coordinating all scoring and reporting activities related to the processing of CELApro test documents. Document preparation, interdepartmental coordination and communication, processing specifications, and problem resolution are performed by a designated Scoring Project Manager from this team. The scoring team works closely with all CTB departments to ensure successful scoring and reporting.

CTB maintains a professional staff of specialized data processing technicians to lead the verification process and ensure the integrity of the student response data at both group and individual levels. Raw scoring and editing of scanned data is performed in a client/server system (WinScore), where a sophisticated system of edits are invoked to review the integrity of each batch scanned and to produce a list of error suspects. While the editors can view data from any document online, the error suspect list concentrates on the most likely problems based on predefined guidelines. This system reduces editing time and provides a high degree of quality control. CTB continues to enhance the capability of editing software to simplify the detection and correction of errors. Online editing screens allow an editor to focus on potential problems and then they provide related information. The actual scanned documents are always available to the editor, and the software supports the review and correction of any field in the scanned record. Entry and verification of the necessary corrections are enhanced to ensure that each error is actually corrected. As batches are extracted for scoring, a final edit is performed to ensure that all requirements for scoring are met. This automated final edit flags a batch for further editing if any error is still detected. A batch containing errors cannot be extracted for reporting. This ensures a high level of accuracy of the scored data.

When the editing process is completed, documents are moved to a staging area to be prepared for retention. Bundles are caged, warehoused in a recoverable location, and retained for possible retrieval during the specified retention period. Once this period is over, documents are destroyed according to procedures that ensure security is maintained.

Handscoring Process

For the CELApro assessments, CTB's imaging handscoring system presents images of scanned test books to trained readers who assign scores for CR items. Scanned images are viewed on high-quality, 19-inch workstation monitors. Images of each student's responses are automatically routed to two or more readers when required, and images of specific subsets of test items are routed to designated groups of readers trained to score these items. CTB is committed to using the finest imaging equipment, software presentation system, data management system, and quality control to provide valid, reliable, cost-efficient scoring.

Readers

In order to work as a handscoring reader at CTB, one must possess and show evidence of either a BA or BS degree. The evaluator staff is comprised of individuals from many walks of life, from retired or current educators to engineers, all possessing BA degrees to PhDs.

Many CTB readers also have a great deal of classroom teaching experience. Our reader pool includes editors, published authors, and a number of individuals with advanced degrees. The minimum qualification for all Scoring Center readers is a Bachelor's degree.

Team Leaders
Scoring team leaders are selected on the basis of having demonstrated a high degree of scoring accuracy and consistency, often across multiple subjects and grades. They must also possess good interpersonal and leadership skills in order to be effective when training and counseling readers. The ratio of readers to team leaders is no more than 10 to 1. While it is possible to conduct handscoring with more readers per team leader, it has been CTB's experience that inter-rater reliability and production goals are jeopardized unless a trained leader can frequently monitor all readers.

Scoring Supervisors
Scoring Supervisors are the core group at CTB scoring centers. They direct and organize the assessment process and train team leaders and readers. Scoring Supervisors have extensive experience as team leaders prior to their qualification and selection. The Scoring Supervisors are subject-area experts in the content(s) that they supervise and train.

Anchor and Training Papers
Prior to the actual scoring, the CTB Scoring Center creates training materials. The process includes several presorting steps and subsequent iterative/consensus processes in order to achieve ever-increasing agreement and precision through a kind of "round robin" scoring, followed by discussion and selection. When all papers for a form are selected and assigned a status as good anchors, training, qualifying, or check-set papers, they are consolidated into training formats. Scoring Guides (consisting of rubrics, anchors, and annotations) serve as a constant, setting the course for all subsequent training and scoring.

Rater Training and Validation
Validation is a critical task in the assessment training process. It is the final determinant of reader readiness. All readers, including team leaders, must achieve at least 80% exact agreement on the qualifying round following training. Those readers not validating on the first attempt receive further training prior to taking an additional qualifying round. Only those who successfully validate are qualified as readers and allowed to score tests. Team leaders are required to complete two validation rounds with at least 80% exact agreement in each round.

Intra-Rater Reliability

Throughout the course of the handscoring process, calibration sets of pre-scored papers (check-sets) are administered daily to the team leaders as well as to the readers to monitor scoring accuracy and to maintain a consistent focus on the established rubric and guidelines. Imaging permits this monitoring without reader knowledge of when a check-set is administered. Readers whose check-set scores fall below the qualifying level are removed from live scoring and are given additional training and another qualifying (validation) round. Readers unable to qualify are dismissed.

The "read-behind" is another valuable intra-rater reliability monitoring technique. On a daily basis, each team leader reads a random selection of each reader's scored items. The scores are compared, and if they agree, the team leader is able to offer feedback, which enhances the reader's confidence and ability to score quickly and accurately. However, if an individual is

straying from the standard established in the training and validation samples, the aberrant scoring is detected, and the team leader is able to offer the guidance necessary to refocus the reader's effort. Team leaders conduct read-behinds more frequently for any reader whose scoring is inconsistent. Thus, any scoring variation is corrected.

Inter-Rater Reliability

Intraclass correlation coefficients and weighted Kappa coefficients were calculated to measure reader agreement (Fleiss and Cohen, 1973) for each of the handscored CELApro items[1, 2] using scores assigned to all item responses that received second reads. The intraclass correlation coefficients ranged from 0.83 to 0.96, with 89 percent of the coefficients greater than or equal to 0.90. The weighted Kappa values also were high[3] for all items, ranging from 0.66 to 0.92, indicating good agreement between the first and second readers. Inter-rater agreement statistics for all of the handscored items are shown in Table 15.

The percentages of discrepant ratings were higher in Kindergarten than in any of the higher grades, ranging from 9 to 11 percent. The percentage of discrepant ratings was 4 percent or less for all items in Grades 1 through 12.

[1] If agreement is perfect, both the intraclass correlation coefficient and Kappa will be equal to +1. If agreement is at chance levels, then both coefficients will be equal to zero.
[2] The intraclass correlation does not consider chance agreement between two raters, but the weighted Kappa does take into account chance agreement. Therefore, in general, weighted Kappa will have values equal to or less than the intraclass correlations.
[3] Kappa values between 0.75 and 0.91 represent good agreement beyond chance, and values below 0.40 indicate poor agreement.

Table 15. Inter-Rater Agreement for CELApro Writing Responses

Grade Span		Item	Max Score	% Perfect Agreement	% Adjacent Scores	% Special Codes	% Discrepant (>1 point)	Intraclass Correlation	Wtd Kappa
K–2	K	21	3	0.49	0.07	0.35	0.09	0.93	0.85
		22	3	0.47	0.06	0.36	0.11	0.93	0.86
		23	3	0.52	0.08	0.31	0.10	0.94	0.89
		24	3	0.51	0.08	0.31	0.10	0.93	0.86
		25	3	0.51	0.07	0.31	0.11	0.94	0.87
	1	21	3	0.74	0.17	0.07	0.02	0.92	0.84
		22	3	0.76	0.15	0.08	0.02	0.94	0.87
		23	3	0.72	0.18	0.07	0.03	0.92	0.85
		24	3	0.71	0.17	0.08	0.03	0.93	0.86
		25	3	0.73	0.16	0.08	0.03	0.93	0.87
	2	21	3	0.76	0.17	0.05	0.03	0.94	0.87
		22	3	0.77	0.15	0.06	0.03	0.94	0.87
		23	3	0.76	0.17	0.05	0.03	0.92	0.84
		24	3	0.71	0.19	0.06	0.04	0.92	0.83
		25	4	0.60	0.29	0.07	0.04	0.92	0.83
3–5		22	3	0.73	0.22	0.03	0.03	0.88	0.75
		23	3	0.77	0.19	0.03	0.02	0.90	0.80
		24	3	0.72	0.22	0.03	0.02	0.88	0.76
		25	3	0.74	0.20	0.04	0.02	0.90	0.79
		26	4	0.66	0.24	0.07	0.03	0.91	0.82
6–8		21	3	0.81	0.16	0.02	0.01	0.90	0.80
		22	3	0.83	0.14	0.02	0.01	0.90	0.80
		23	3	0.69	0.28	0.02	0.01	0.83	0.66
		24	3	0.71	0.25	0.02	0.02	0.84	0.69
		25	4	0.62	0.33	0.04	0.01	0.88	0.75
9–12		21	3	0.80	0.13	0.06	0.01	0.95	0.90
		22	3	0.82	0.11	0.07	0.01	0.96	0.92
		23	3	0.79	0.12	0.07	0.01	0.95	0.90
		24	3	0.75	0.17	0.07	0.01	0.93	0.86
		25	4	0.72	0.19	0.08	0.01	0.94	0.88

Scoring and Technology Quality Control Procedures

The Technology and Scoring Departments at CTB both have quality assurance groups specifically charged with reviewing scoring data and reports during all stages of the process. The Technology quality assurance team verifies the accuracy of all reporting programs before they become operational. The Scoring quality assurance team verifies the accuracy of report information during the scoring process. After all data are entered into the scoring system and all reporting programs are completed, a sample of reports are printed and submitted to the Scoring quality assurance group, which reviews the sample reports to verify the accuracy and correct presentation of all data.

Numerous quality assurance checks are in place throughout the scoring process to ensure the accuracy of reports. Prior to delivering any electronic files or hard-copy score reports, all reports are given a final, extensive quality check, known as a “Red Team Review.” Red Teams are comprised of individuals from every CTB department coming together to form an interdisciplinary team. Samples of each type of report are printed from the active scoring system, and the Red Team carefully reviews these samples for accuracy and correct format. Student-level information is compared by hand with student rosters and other documentation. Reports are not sent out until all necessary corrections determined by the Red Team are resolved.

Part 6: Data Analysis and Results

This section of the technical report contains a description of the 2009 calibration and equating and differential item functioning (DIF) procedures and results, along with details of the 2012 classical item analysis that was conducted for each test. Because the 2012 CELApro tests were identical to those administered in 2009, no new calibration, equating, or DIF analyses were conducted this year. The 2012 CELApro tests were scored using the raw-to-scale score tables that were produced from the results of the 2009 calibration/equating/scaling analyses. This section also includes a subsection describing student performance on the 2012 tests, along with comparisons with the 2011,2010, and 2009 results.

IRT Item Calibration

Student item responses on each of the CELApro assessments were calibrated in 2009 using the three-parameter logistic (3PL) model to scale the MC items and the two-parameter partial credit (2PPC) model to scale the CR items. A brief explanation of the models is provided below.

The 3PL model (Lord and Novick, 1968; Lord, 1980) defines a MC item in terms of three item parameters: (a) item discrimination, (b) item difficulty or location, and (c) probability of a student with very low ability answering the item correctly (i.e., a guessing parameter). In this model, the probability that a student with scale score θ will respond correctly to item *j* is defined as

$$p_j(\theta) = c_j + \frac{(1 - c_j)}{1 + \exp[-1.7a_j(\theta - b_j)]},$$

where a_j is the item discrimination,
b_j is the item difficulty, and
c_j is the probability of a correct response by a very low ability student.

The 2PPC model defines a CR item in terms of item discrimination as well as location parameter for each score point. The 2PPC model is a special case of Bock's (1972) nominal model. Bock's model states that the probability of an examinee with ability θ having a score at the *k*th level of the *j*th item is

$$P_{jk}(\theta) = P(x_j = k - 1|\theta) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j} \exp Z_{ji}}, k = 1,..., m_{j,}$$

where m_j is the number of score levels, and

$Z_{jk} = A_{jk}\theta + C_{jk}$,
$A_{jk} = \alpha_j(k-1)$, k = 1, 2,…m_j, and
$C_{jk} = -\sum_{i=0}^{k-1}\gamma_{ji}$, where $\gamma_{j0} = 0$,

where A_{jk} is the discrimination parameter of the *k*th category of item *j*, C_{jk} is the intercept parameter of the nonlinear response function associated with the *k*th category of item *j*, and α_j and γ_{ji} are the parameters to be estimated from the data.

For each item, there are m_j –1 independent γ_{ji} parameters and one α_j parameter; a total of m_j independent item parameters is estimated.

All of the CELApro assessments were calibrated in 2009 using the 3PL/2PPC models described above. Separate calibrations were conducted for Speaking, Listening, Reading, Writing, Comprehension, and Oral scales in each grade span.

Equating and Scaling

The calibrated tests were placed on the existing CELApro/*LAS Links* scale in 2009 through a Stocking and Lord (1983) characteristic curve equating procedure. The 2008 operational item parameters for almost all of the test items were used as equating anchors in this procedure.

The 2009 M1 and M2 conversion parameters were computed as

$M1_{New} = A^* M1_{Old}$, and

$M2_{New} = A^* M2_{Old} + B$,

where $M1_{New}$ and $M2_{New}$ are the new transformation constants calculated to place the 2009 test items onto the existing scale, and $M1_{Old}$ and $M2_{Old}$ are the transformation constants from the anchor set.

The *A* and *B* values are derivatives of the input (initial) and estimated (final) values for the anchor set and are computed as

$$A = \frac{SD_{New}}{SD_{Old}}$$

$$B = (Mean_{New} - \frac{SD_{New}}{SD_{Old}} Mean_{Old})$$

where

SD_{New} is the standard deviation of anchor estimates in scale score metric,

SD_{Old} is the standard deviation of anchor input values in scale score metric,

$Mean_{New}$ is the mean of anchor estimates in scale score metric, and

$Mean_{Old}$ is the mean of anchor input in scale score metric.

This equating procedure was performed in 2009 for each of the grade spans (K, 1, 2, 3–5, 6–8, and 9–12). Consequently, the equated results were used to create raw-to-scale score tables for each of the six content areas (Speaking, Listening, Reading, Writing, Oral, and Comprehension). Because the total score is computed as the unweighted mean of the scale scores on Reading, Writing, Listening, and Speaking, no separate calibration, equating, scaling, or scoring table was required for the total score.

The resulting scoring tables for all grade spans, which were used to score the 2012 CELApro, are included in Appendix E.

Results of Calibration and Equating

Tables 1B through 31B and Figures B1 through B62 in Appendix B show the alignment of the original and equated "A" parameters (using the log of A) and the alignment of the corresponding "B" parameters for Speaking, Listening, Reading, and Writing. In these figures, the original parameters are the 2008 CELApro item parameters, and the equated parameters are the 2009 CELApro parameters. Since no equating was performed, the 2012 CELApro parameters are identical to the 2009 item parameters.

Figures 1C through 12C in Appendix C show the CELApro test characteristic curves (TCCs) and the standard errors of measurement (SEMs) for each grade span and content domain. For a vertically scaled test such as the CELApro/*LAS Links*, we would expect to see a pattern in which the TCCs are arrayed in grade-level sequence from left to right (i.e., tests increasing in difficulty as grade level increases). The TCCs show this expected pattern.

The correlations between the 2009 equated and input anchor item parameters and *p*-values (*P*) are shown in Table 16. For MC scales, these represent the correlations of the *A* and *B* parameters. For CR items, the correlations of item parameters represent the alpha and gamma correlations, respectively.

Table 16. Stocking and Lord Parameter Correlations

Grade Span K–2			
	P	Discrimination	Location
Speaking	1.00	0.99	1.00
Listening	1.00	0.97	0.93
Reading	1.00	0.98	1.00
Writing	0.99	0.83	0.97
Comprehension	1.00	0.98	1.00
Oral	1.00	0.93	0.95

(continued on next page)

Table 16. Stocking and Lord Parameter Correlations (continued)

Grade Span 3–5			
	P	Discrimination	Location
Speaking	1.00	0.98	0.97
Listening	1.00	0.98	1.00
Reading	1.00	0.98	0.99
Writing	0.97	0.99	1.00
Comprehension	1.00	0.98	1.00
Oral	1.00	0.96	0.99
Grade Span 6–8			
	P	Discrimination	Location
Speaking	1.00	0.98	0.98
Listening	1.00	0.94	0.98
Reading	1.00	0.98	0.98
Writing	1.00	0.96	0.98
Comprehension	1.00	0.98	0.99
Oral	1.00	0.90	0.98
Grade Span 9–12			
	P	Discrimination	Location
Speaking	1.00	0.98	0.98
Listening	1.00	0.96	0.97
Reading	1.00	0.97	0.99
Writing	1.00	0.97	0.99
Comprehension	1.00	0.97	0.99
Oral	1.00	0.98	0.99

For all contents and grade spans, the *p*-value correlations are all greater than 0.95.

For each of the six content domains, Appendix D contains the test characteristic curves for the anchor item input parameters, the equated anchor item estimated parameters, and the equated total test. As shown in these plots, the total test and the anchor test are closely aligned to each other.

Item Analysis

Classical item analysis statistics were computed for the 2012 CELApro administration for each content domain at each grade span. The tables in Appendix A present item-level descriptive statistics for each grade span and content domain. These tables contain the following information: item number, item type, *p*-value, item correlation with the total test score, correlation between each item choice and the total test score, and percent omit. The *p*-value for an MC item represents the proportion of students who answered the item correctly. The *p*-value for a CR item represents the mean raw score for the item divided by the maximum possible score for that item.

The point biserial correlation between the item score and the total score on the test was also computed for each of the MC items. For each CR item, the Pearson product-moment correlation between the item score and the total score on the test was computed. For these correlations, the studied item was excluded from the computation of the total score so as not to artificially inflate the correlation.

Item Difficulty Statistics (*p*-Values)

The 2012 statistics for individual items at each grade span are provided in the item analysis tables in Appendix A. In these tables, item difficulty is expressed in terms of *p*-values. For MC items, the *p*-value is the proportion of students answering the item correctly. For CR items, the *p*-value is the mean item score expressed as a proportion of the total score points possible on that item (i.e., each raw item score is divided by the maximum possible score on the item). The item-level results, overall, are consistent with the results obtained in 2011.

The *p*-values in Appendix A are at or above 0.20 except for four Kindergarten Writing items, and two Kindergarten Speaking items and two Kindergarten Oral items; most are in the desired difficulty range between 0.30 and 0.90. The range of *p*-values varies by grade span and content domain. Across grade spans, the *p*-values range from 0.10. to 0.98 for Speaking, 0.24 to 0.99 for Listening, 0.20 to 0.99 for Reading, 0.11 to 0.97 for Writing, 0.24 to 0.99 for Comprehension, and 0.10 to 0.99 for Oral. Within grade spans, *p*-values range from 0.10 to 0.99 in Grade Span K–2, from 0.32 to 0.99 in Grade Span 3–5, from 0.26 to 0.96 in Grade Span 6–8, and from 0.32 to 0.94 in Grade Span 9–12.

Average item difficulty for each content area, grade, and grade span is summarized in Table 17.

Table 17. Mean *p*-Values by Grade Span and Grade

Grade	Speaking	Listening	Reading	Writing	Comprehension	Oral
Grade Span 1	0.69	0.71	0.70	0.52	0.69	0.70
K	0.55	0.52	0.57	0.31	0.55	0.54
1	0.72	0.74	0.68	0.55	0.69	0.73
2	0.81	0.85	0.82	0.71	0.83	0.83
Grade Span 2	0.82	0.73	0.70	0.78	0.71	0.77
3	0.78	0.67	0.61	0.71	0.63	0.72
4	0.83	0.74	0.71	0.79	0.73	0.78
5	0.86	0.79	0.78	0.83	0.78	0.82
Grade Span 3	0.84	0.80	0.71	0.79	0.75	0.82
6	0.82	0.77	0.67	0.78	0.72	0.79
7	0.84	0.80	0.71	0.79	0.76	0.82
8	0.85	0.82	0.75	0.81	0.79	0.83
Grade Span 4	0.83	0.77	0.65	0.77	0.71	0.8
9	0.82	0.75	0.61	0.76	0.68	0.79
10	0.83	0.77	0.64	0.77	0.71	0.8
11	0.84	0.79	0.67	0.78	0.73	0.81
12	0.83	0.78	0.67	0.76	0.73	0.81

Item-Total Correlations

An important indicator of item quality is the correlation of scores on that item with scores on the total test. The 2012 item-total correlations (point biserial correlation coefficients for MC items and Pearson product-moment correlations for CR items) are summarized in Table 18. To compute these correlations, the "total" score was defined as the total score on the specific content domain. To avoid artificially inflating the correlation coefficients, the contribution of the item in question was removed from the total when calculating each of the correlations. Thus, performance on each Speaking item was correlated with the total Speaking score minus the score on the item in question and so on for the Listening, Reading, Writing, Comprehension, and Oral scales.

Individual item-total correlations for each content area and grade span are provided in the item analysis tables in Appendix A. Across Grades 1–12, item-total correlations range from 0.22 to 0.84 for Speaking, from 0.20 to 0.56 for Listening, from 0.06 to 0.58 for Reading, and from 0.03 to 0.77 for Writing. Comprehension item-total correlations range from 0.02 to 0.59, and Oral item-total correlations range from 0.11 to 0.81. Item-total correlations for Kindergarten were generally lower than the other grades, ranging from 0.33 to 0.81 for Speaking, from 0.27 to 0.56 for Listening, from 0.06 to 0.53 for Reading, from 0.20 to 0.68 for Writing, from 0.02 to 0.51 for Comprehension, and from 0.15 to 0.78 for Oral.

The average (mean) item-total correlation coefficients for each content area, grade span, and grade are shown in Table 18. The average item-total correlation coefficients ranged from 0.53 to 0.68 for Speaking, from 0.35 to 0.44 for Listening, from 0.37 to 0.45 for Reading, from 0.37 to 0.51 for Writing, from 0.36 to 0.43 for Comprehension, and from 0.38 to 0.50 for Oral.

Table 18. Average Item-Total Correlations by Grade Span and Grade

Grade	Speaking	Listening	Reading	Writing	Comprehension	Oral
Grade Span 1	0.59	0.41	0.39	0.44	0.37	0.44
K	0.63	0.42	0.40	0.37	0.37	0.46
1	0.59	0.42	0.37	0.45	0.36	0.44
2	0.56	0.40	0.40	0.50	0.37	0.42
Grade Span 2	0.54	0.37	0.44	0.50	0.38	0.40
3	0.53	0.35	0.42	0.49	0.36	0.38
4	0.54	0.37	0.44	0.50	0.38	0.40
5	0.55	0.38	0.45	0.50	0.40	0.42
Grade Span 3	0.60	0.40	0.41	0.44	0.38	0.44
6	0.58	0.38	0.39	0.43	0.37	0.42
7	0.60	0.40	0.40	0.43	0.38	0.44
8	0.63	0.41	0.42	0.44	0.40	0.47
Grade Span 4	0.68	0.42	0.41	0.51	0.41	0.49
9	0.68	0.40	0.39	0.50	0.38	0.48
10	0.67	0.43	0.41	0.51	0.41	0.50
11	0.68	0.43	0.43	0.51	0.42	0.50
12	0.68	0.44	0.43	0.51	0.43	0.50

Item Omit Rates

The item analysis tables in Appendix A also show the rate at which students omitted items. Omit rates are often useful in determining whether testing times are sufficient, particularly if there is a high rate of items omitted at the end of a test section. In cases where speededness is not an issue, high item omit rates may often indicate ambiguity or extreme item difficulty.

Omit rates were generally low for students in Grades 3 through 12. Omit rates for Grade Spans 6–8 and 9–12 were below 5 percent for all of the items in all content areas. For Grade Span 3–5, there were two items in Reading with omit rates of 21.81 percent and 22.66; these same two items in Comprehension had omit rates of 21.82 percent and 22.66 percent. The same omit rate or slight difference in omit rates across the two content domains reflect the fact that omit rates for each domain are based on the responses of students with valid scores on that domain.

Omit rates were generally higher for Grade Span K–2. Omit rates were between 0.04 and 7.59 percent for all of the Speaking items, with 2 items above 5 percent. For the Listening items, omit

rates were above 5 percent for 4 items, all of them in Kindergarten. Grades K–2 in Reading had 24 items above 5 percent—13 in Kindergarten, 10 in Grade 1, and 1 in Grade 2. The highest omit rates were for the Kindergarten Writing items, with omit rates ranging from 3.77 to 22.16 percent and all but 1 item above 5 percent. Only 1 item in Grade 1 was slightly above 5 percent, and no items were above 5 percent for students in Grade 2.

Differential Item Functioning (DIF) Statistics

In addition to the analyses that were conducted as part of the *LAS Links* development process, Linn-Harnisch (1981) gender DIF analyses were conducted on data from the 2009 CELApro administration. The procedures employed for CELApro DIF analyses apply IRT-based models to analyze the item calibration data. Because no new calibrations have been conducted since 2009, no new DIF analyses were conducted. The DIF analyses and results in this technical report are based on the 2009 item parameters and thus are identical to those that were reported in the last two years.

For the 2009 CELApro analyses, a separate IRT calibration and a separate DIF analysis were conducted for each grade span and content domain (Speaking, Listening, Reading, Writing, Comprehension, and Oral). To calculate DIF for the CELApro assessments, the IRT parameters for each item (a_i, b_i, c_i) and the trait or ability estimate (θ_i) for each examinee were estimated for the three-parameter logistic model as

$$P_{ij} = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7a_i\left(\theta_j - b_i\right)\right]},$$

where P_{ij} is the probability that examinee j will pass item i. The total population is then divided into two groups by gender, and the members in each group are sorted into ten equal score categories (deciles) based on their location on the scale score (θ_i) scale. The expected proportion correct for each group based on the model prediction is compared to the observed (actual) proportion correct obtained by the group. The proportion of examinees in decile g who are expected to answer item i correctly is

$$P_{ig} = \frac{1}{n_g}\sum_{j \in g} P_{ij},$$

where n_g is the number of examinees in decile g. The proportion of examinees expected to answer item i correctly (over all deciles) for a group (e.g., female) is

$$P_i = \frac{\sum_{g=1}^{10} n_g P_{ig}}{\sum_{g=1}^{10} n_g}.$$

The corresponding observed proportion correct for examinees in a decile (O_{ig}) is defined as the number of examinees in decile *g* who answered item *i* correctly divided by the total number of examinees in the decile (n_g). That is,

$$O_{ig} = \frac{\sum_{j \varepsilon g} u_{ij}}{n_g},$$

where u_{ij} is the dichotomous score for item *i* for examinee *j*.

The corresponding formula to compute the observed proportion answering each item correctly (over all deciles) for a complete gender or ethnic target group is given by

$$O_i = \frac{\sum_{g=1}^{10} n_g O_{ig}}{\sum_{g=1}^{10} n_g}.$$

After the values are calculated for these variables, the difference between the observed proportion correct and the expected proportion correct can be computed. The decile group difference (D_{ig}) for observed and expected proportion correctly answering item *i* in decile *g* is

$$D_{ig} = O_{ig} - P_{ig},$$

and the overall group difference (D_i) between observed and expected proportion correct for item *i* in the complete group (over all deciles) is

$$D_i = O_i - P_i.$$

DIF is defined in terms of the decile group and total target subsample differences, the D_{i-} (sum of the negative group differences) and D_{i+} (sum of the positive group differences) values, and the corresponding standardized difference (Z_i) for the subsample (see Linn and Harnisch, 1981, p. 112). Items for which $|D_i| \geq 0.10$ and $|Z_i| \geq 2.58$ are flagged as DIF items. If D_i is positive, the item favors the target subsample. If D_i is negative, the item favors the standard sample.

These indices are indicators of the degree to which members of a target group perform better or worse than expected on each item based on the parameter estimates from all subsamples. Differences for decile groups provide an index for each of the ten regions on the scale score (θ) scale. The decile group difference (D_{ig}) can be either positive or negative. Use of the decile group differences as well as the overall group difference allows one to detect items that give a large positive difference in one range of θ and a large negative difference in another range of θ

yet have a small overall difference. A generalization of the Linn and Harnisch (1981) procedure was used to measure DIF for CR items.

The results of the 2009 DIF analyses are shown in Table 19. Again, since calibrations did not occur, the results are the same as for the last three years. Overall, no items were flagged for DIF against males or females, and very few items exhibited DIF by ethnicity. Across all grades and content areas, 24 items (3.02%) were flagged in favor of and 29 items (3.65%) were flagged against American Indian/Alaska Native examinees, 11 items (1.39%) were flagged in favor of and 13 items (1.64%) were flagged against Asian/Pacific Islander examinees, 27 items (3.40%) were flagged in favor of and 23 items (2.90%) were flagged against Black examinees, and 2 items (0.25%) were flagged against White examinees. DIF was not performed on Hispanics because they constitute an overwhelming majority of the respondents.

All items flagged for DIF are carefully reviewed by CTB's content development experts to try to determine whether race, native language, or another characteristic might have caused the DIF. If that review suggests that the DIF statistics are likely to reflect racial bias rather than only meaningful language differences, the items will be replaced in revised future forms whenever suitable replacement items are available.

Table 19. Number of Items Exhibiting Differential Item Functioning

Subject	Grade Span	American Indian/Alaska Native		Asian/Pacific Islander		Black		White	
		For	Against	For	Against	For	Against	For	Against
Speaking	K–2	0	2	1	0	0	0	0	0
	3–5	1	1	0	0	0	1	0	0
	6–8	2	3	0	1	3	2	0	0
	9–12	0	2	1	0	3	4	0	0
Listening	K–2	2	1	0	0	0	0	0	0
	3–5	0	1	0	0	0	0	0	0
	6–8	0	0	0	0	0	0	0	0
	9–12	0	0	0	0	1	1	0	0
Reading	K–2	0	2	0	0	0	0	0	0
	3–5	1	2	1	0	0	0	0	0
	6–8	2	1	0	0	1	1	0	0
	9–12	1	2	1	4	3	2	0	1
Writing	K–2	0	0	1	0	0	0	0	0
	3–5	0	0	0	1	0	0	0	0
	6–8	1	2	1	0	0	0	0	0
	9–12	1	0	3	2	0	2	0	0
Comprehension	K–2	2	1	0	0	0	0	0	0
	3–5	0	2	0	0	0	0	0	0
	6–8	0	0	0	0	1	0	0	0
	9–12	0	2	1	4	4	3	0	1
Oral	K–2	4	1	1	0	2	0	0	0
	3–5	1	2	0	0	1	1	0	0
	6–8	2	1	0	1	5	2	0	0
	9–12	4	1	0	0	3	4	0	0

Student Performance on the 2012 CELApro

This section of the report summarizes student performance on the 2012 CELApro. Results are presented for the total population and for various subgroups of interest. In addition, results are compared with performance on the 2011 CELApro. To facilitate interpretation of the score distributions provided in this report, the lowest obtainable scale scores (LOSS) and the highest obtainable scale scores (HOSS) on the CELApro are provided in Table 20. These values do not change from year to year.

Table 20. CELApro Lowest and Highest Obtainable Scale Scores

		Speaking	Listening	Reading	Writing	Comp (R+L)	Oral (L+S)	Total
Grade K	LOSS	300	300	240	200	270	280	260
	HOSS	580	560	570	630	570	620	585
Grade 1	LOSS	300	300	240	200	270	280	260
	HOSS	580	560	590	630	590	620	590
Grade 2	LOSS	300	300	240	200	270	280	260
	HOSS	580	560	590	640	590	620	592
Grades 3–5	LOSS	310	310	300	270	320	290	297
	HOSS	635	630	660	680	660	680	651
Grades 6–8	LOSS	325	360	380	300	360	310	341
	HOSS	645	640	690	690	680	700	666
Grades 9–12	LOSS	330	370	390	310	380	320	350
	HOSS	650	650	700	700	700	710	675

Note: LOSS = Lowest Obtainable Scale Score; HOSS = Highest Obtainable Scale Score

Table 21 shows the 2012 total scale score means and standard deviations (SD) by grade span, and Table 22 shows the results for each individual grade in 2010, 2011, and 2012.

Table 21. 2012 Total Scale Score Means and Standard Deviations by Grade Span

	N	Mean	SD
Grade Span 1	37538	446.32	53.25
Grade Span 2	30231	529.49	43.44
Grade Span 3	17864	560.02	41.51
Grade Span 4	14103	551.63	44.72

Table 22. 2010, 2011, and 2012 Total Scale Score Means and Standard Deviations by Grade

	2010			2011			2012		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
K	12073	388.87	38.52	12587	393.20	37.86	12666	396.47	37.57
1	12539	447.63	39.64	12612	451.28	39.45	12874	454.78	37.44
2	11859	486.48	37.60	12091	488.38	37.14	11998	489.86	35.81
3	10926	508.02	41.09	11308	510.56	41.21	11372	510.93	40.18
4	9473	529.19	41.50	9851	532.08	41.12	10114	533.38	40.42
5	7424	545.08	41.97	8484	547.48	42.04	8745	549.12	41.06
6	6198	546.65	40.61	6504	549.58	40.22	7135	551.84	39.61
7	5100	555.04	42.49	5650	558.59	42.68	5729	561.97	40.48
8	4400	561.50	44.74	4608	565.18	44.09	5000	569.46	43.01
9	4056	541.33	44.27	4223	543.03	43.72	4317	546.16	42.33
10	3174	547.03	46.06	3700	550.24	45.87	3819	551.32	44.51
11	2737	554.90	44.45	2884	553.91	45.98	3128	556.93	45.14
12	2311	555.08	46.23	2707	555.39	47.14	2839	554.52	47.10

The 2012 total scale scores were higher than the 2010 and 2011 scores in grades K–11. The greatest increase in scores occurred in grades 6–9.

The 2012 performance on the six component scales of Speaking, Listening, Reading, Writing, Comprehension, and Oral proficiency is summarized by grade and grade span in Table 23 and by grade and gender in Table 24. Note that because not all students had valid scores on all four components, the component scale score means may be based upon larger numbers of students than the total scale score means.

Overall, female students tended to score somewhat higher than male students. The greatest gender differences were observed in Reading, Writing, Comprehension, and Oral. Female students scored higher than male students on the Writing test at all grade levels. Differences in the mean Writing scores were most evident in the Kindergarten through Grade 7 where the female score advantage ranged from 11 points to more than 16 points, with smaller differences observed at higher grade levels. Male students, on the other hand, tended to score slightly higher than female students on the Speaking and Listening tests. The difference in mean Speaking scores was highest in Grade 11, where the mean score for male students was 15 points higher than the mean for female students. These results are displayed graphically in Figures 1 through 7.

Table 23. CELApro Scale Score Means and Standard Deviations: Component Scales

	Speaking			Listening			Reading			Writing			Comprehension			Oral		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Grade Span 1	38139	488.77	41.80	38050	467.14	45.82	37972	418.36	63.80	37827	411.06	92.68	37900	448.14	52.16	37908	479.57	38.53
K	12979	465.63	41.25	12904	430.69	36.25	12876	363.34	46.11	12746	325.75	76.74	12840	403.18	39.34	12881	453.73	37.87
1	13025	492.79	36.13	13007	472.48	36.46	12979	425.18	45.03	12976	429.01	65.26	12959	453.35	37.36	12955	484.01	30.20
2	12135	509.21	35.54	12139	500.15	35.03	12117	469.52	49.40	12105	481.66	55.51	12101	490.25	37.92	12072	502.37	29.99
Grade Span 2	30511	540.42	46.11	30447	521.52	47.91	30402	521.37	55.69	30367	535.19	61.10	30379	520.40	46.90	30373	534.47	42.92
3	11498	527.36	42.82	11484	502.28	43.59	11472	498.62	52.45	11425	515.95	59.51	11460	500.20	42.38	11452	518.35	36.44
4	10194	542.89	45.01	10158	525.63	44.88	10153	526.29	52.34	10150	539.55	58.23	10146	524.48	43.60	10137	537.19	40.77
5	8819	554.60	46.83	8805	541.87	47.12	8777	545.42	51.91	8792	555.15	58.94	8773	542.07	45.29	8784	552.35	45.30
Grade Span 3	18056	564.59	52.03	18008	558.96	55.95	17999	551.72	46.98	17968	565.37	51.98	17954	546.80	43.18	17938	558.86	48.61
6	7224	557.99	49.59	7190	548.40	53.70	7200	541.27	44.37	7179	560.30	52.08	7169	537.21	40.43	7165	550.22	44.13
7	5786	566.34	51.11	5782	561.79	55.51	5774	553.67	45.74	5765	566.49	51.74	5766	548.77	42.20	5756	560.86	47.91
8	5046	572.03	55.22	5036	570.77	56.86	5025	564.47	48.56	5024	571.33	51.41	5019	558.23	44.98	5017	568.91	53.14
Grade Span 4	14286	554.44	60.06	14253	549.75	58.53	14245	555.58	44.84	14222	547.68	52.07	14209	553.77	49.81	14156	550.81	55.92
9	4355	550.96	57.70	4343	541.53	55.09	4341	547.49	42.05	4340	545.56	51.31	4334	545.00	46.08	4328	544.25	50.78
10	3853	553.52	59.42	3844	549.60	58.74	3843	554.91	44.77	3840	548.24	51.42	3833	553.25	49.41	3833	550.59	56.34
11	3161	559.24	60.99	3163	556.22	59.31	3162	562.22	45.11	3152	551.25	51.58	3155	560.82	50.91	3137	556.64	57.28
12	2917	555.66	62.91	2903	555.21	60.86	2899	561.33	46.68	2890	546.23	54.35	2887	559.94	52.39	2858	554.65	60.07

Table 24. CELApro Scale Score Means and Standard Deviations by Grade and Gender

		Speaking			Listening			Reading			Writing			Comprehension			Oral		
		N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
K	F	6278	467.40	42.32	6252	432.75	35.82	6231	366.50	45.13	6175	331.42	77.45	6215	406.14	38.65	6238	455.45	38.45
	M	6701	463.97	40.15	6652	428.76	36.55	6645	360.38	46.83	6571	320.42	75.69	6625	400.40	39.78	6643	452.12	37.25
1	F	6357	495.13	36.77	6350	474.09	35.54	6337	428.13	43.95	6343	436.89	61.23	6328	455.72	36.27	6328	485.87	29.74
	M	6667	490.56	35.37	6656	470.95	37.26	6641	422.36	45.86	6632	421.48	68.06	6630	451.10	38.24	6626	482.24	30.54
2	F	5812	511.15	35.96	5814	500.73	34.33	5810	471.60	47.34	5808	487.06	52.00	5801	491.90	36.42	5782	503.72	30.17
	M	6323	507.41	35.06	6325	499.61	35.65	6307	467.60	51.15	6297	476.67	58.13	6300	488.73	39.19	6290	501.14	29.76
3	F	5578	527.76	43.26	5570	501.25	42.29	5567	502.05	51.89	5544	524.39	57.93	5562	501.76	41.54	5557	518.00	36.08
	M	5920	526.98	42.40	5914	503.26	44.76	5905	495.38	52.77	5881	507.99	59.89	5898	498.73	43.11	5895	518.68	36.78
4	F	4879	543.96	44.76	4864	524.50	44.21	4861	530.04	50.97	4860	547.37	57.11	4858	526.24	42.72	4853	537.20	40.05
	M	5315	541.90	45.22	5294	526.68	45.46	5292	522.84	53.34	5290	532.36	58.33	5288	522.86	44.32	5284	537.18	41.43
5	F	4220	555.87	46.96	4220	540.56	45.94	4208	548.16	50.26	4213	563.66	57.79	4206	543.16	43.62	4208	552.42	45.03
	M	4599	553.43	46.69	4585	543.07	48.15	4569	542.91	53.27	4579	547.32	58.92	4567	541.06	46.76	4576	552.28	45.56
6	F	3398	557.81	49.07	3382	552.57	52.95	3389	544.93	42.89	3380	566.93	49.36	3375	540.39	39.53	3373	551.43	44.39
	M	3826	558.16	50.05	3808	544.70	54.10	3811	538.01	45.41	3799	554.40	53.71	3794	534.39	41.02	3792	549.14	43.87
7	F	2622	563.87	50.77	2624	564.66	55.35	2621	556.94	44.76	2615	572.52	51.55	2618	551.21	42.36	2608	560.30	49.16
	M	3164	568.39	51.32	3158	559.39	55.55	3153	550.94	46.37	3150	561.49	51.37	3148	546.74	41.96	3148	561.32	46.85
8	F	2334	570.77	54.77	2331	575.06	56.19	2324	567.67	46.47	2326	575.63	50.82	2322	561.26	43.96	2320	570.02	53.55
	M	2712	573.12	55.59	2705	567.07	57.18	2701	561.72	50.14	2698	567.62	51.63	2697	555.62	45.69	2697	567.96	52.78
9	F	1950	543.38	53.14	1950	541.29	52.51	1948	547.32	40.52	1945	548.84	49.81	1946	544.78	44.02	1940	540.23	47.73
	M	2405	557.10	60.47	2393	541.72	57.11	2393	547.63	43.26	2395	542.89	52.35	2388	545.17	47.71	2388	547.51	52.91
10	F	1770	546.71	55.97	1763	548.94	57.50	1762	555.30	42.94	1762	551.82	51.03	1757	552.97	47.88	1759	546.07	52.77
	M	2083	559.31	61.62	2081	550.15	59.77	2081	554.58	46.28	2078	545.20	51.56	2076	553.48	50.68	2074	554.42	58.94
11	F	1448	551.25	59.38	1451	555.66	58.73	1451	562.23	44.18	1449	554.10	51.55	1448	560.67	49.75	1439	551.57	56.56
	M	1713	565.99	61.53	1712	556.68	59.81	1711	562.21	45.89	1703	548.82	51.49	1707	560.94	51.89	1698	560.93	57.56
12	F	1399	550.85	60.95	1394	555.19	60.02	1393	561.14	46.74	1388	548.07	53.35	1389	559.58	51.99	1367	552.29	59.02
	M	1518	560.09	64.37	1509	555.24	61.64	1506	561.51	46.64	1502	544.54	55.22	1498	560.27	52.76	1491	556.83	60.96

Figure 1. Mean Speaking Scale Scores by Grade and Gender

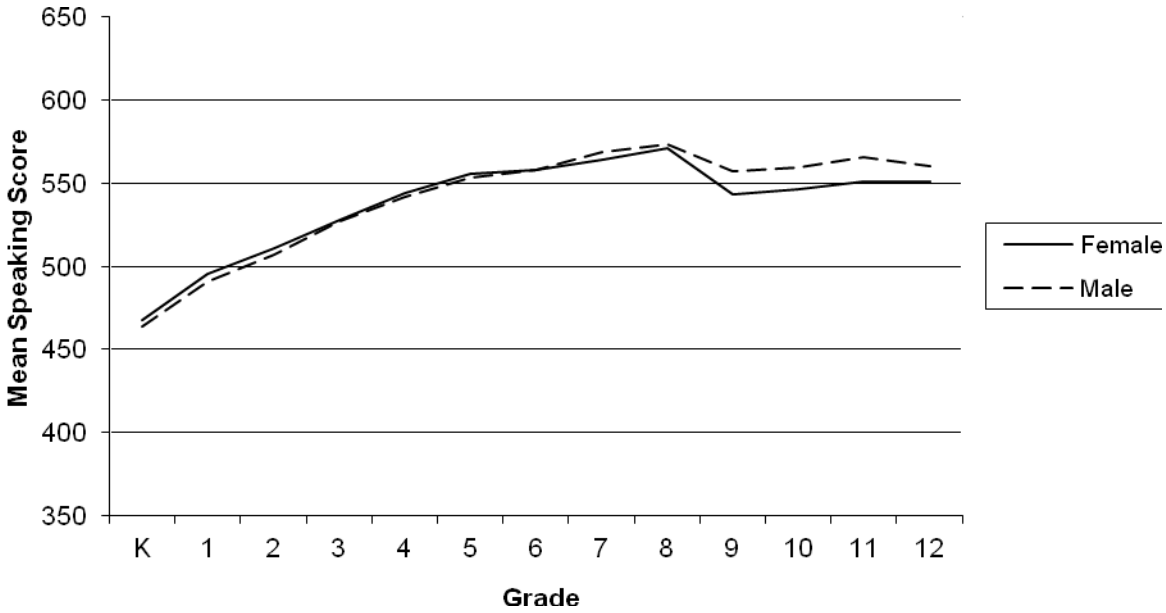


Figure 2. Mean Listening Scale Scores by Grade and Gender

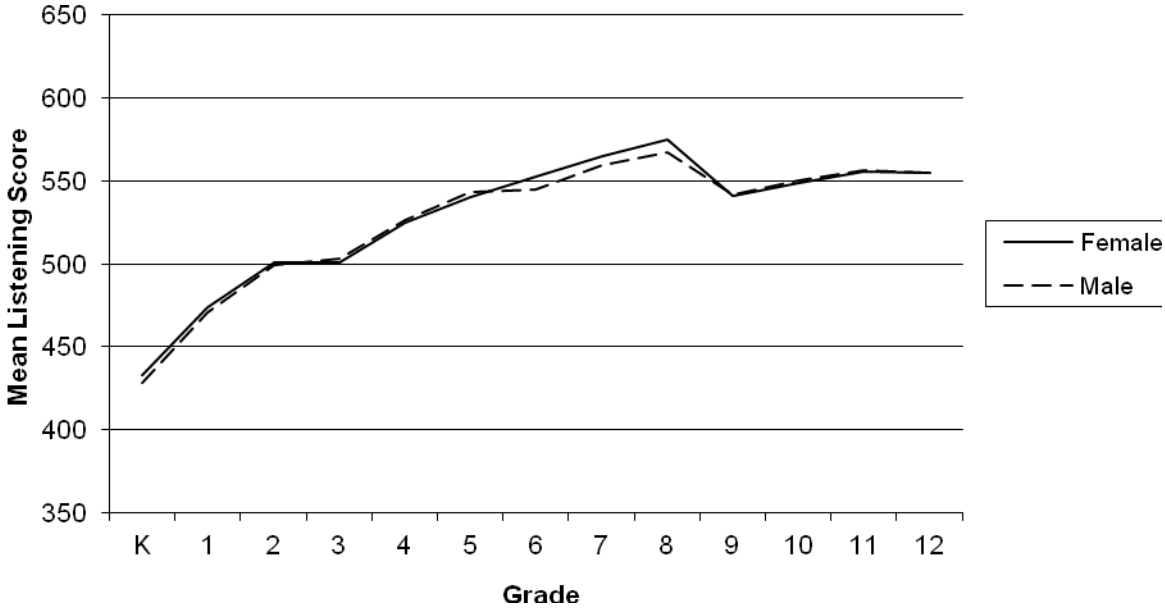


Figure 3. Mean Reading Scale Scores by Grade and Gender

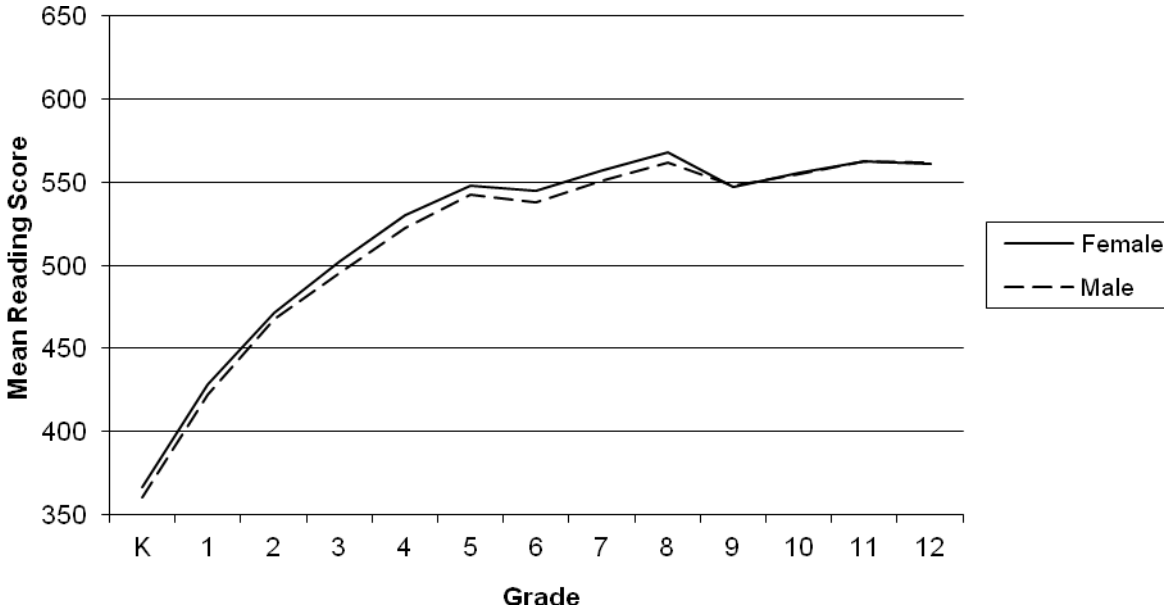


Figure 4. Mean Writing Scale Scores by Grade and Gender

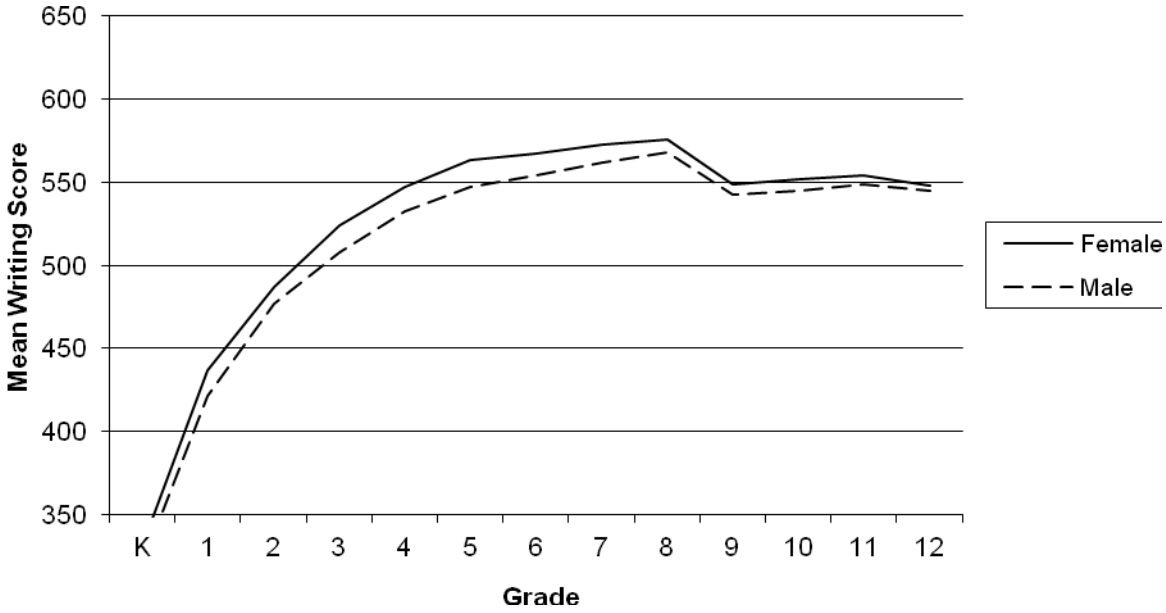


Figure 5. Mean Comprehension Scale Scores by Grade and Gender

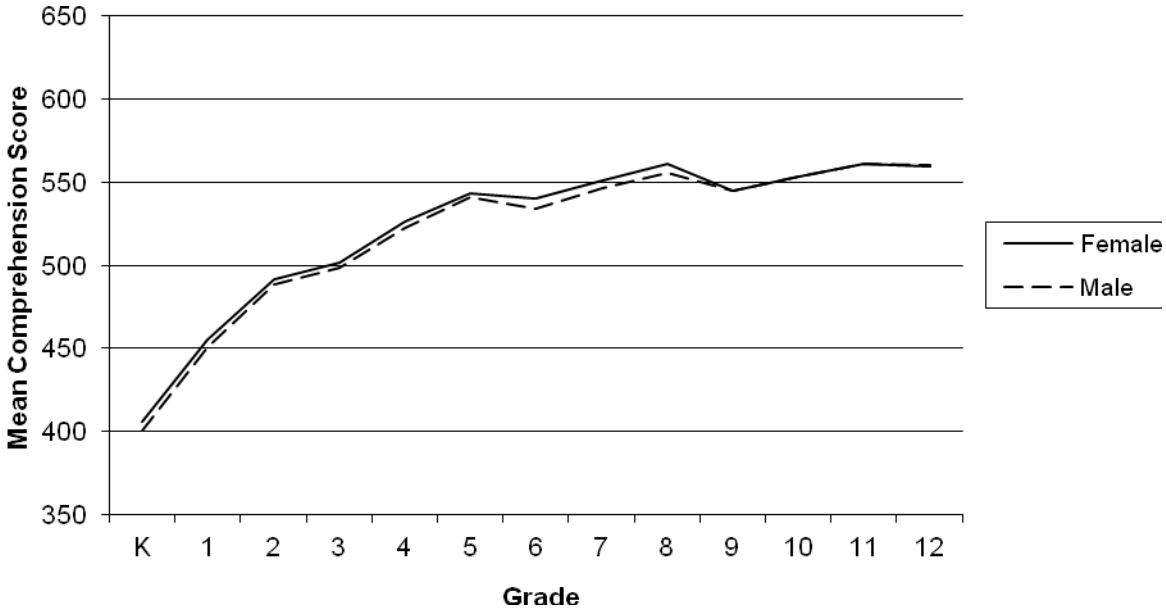


Figure 6. Mean Oral Scale Scores by Grade and Gender

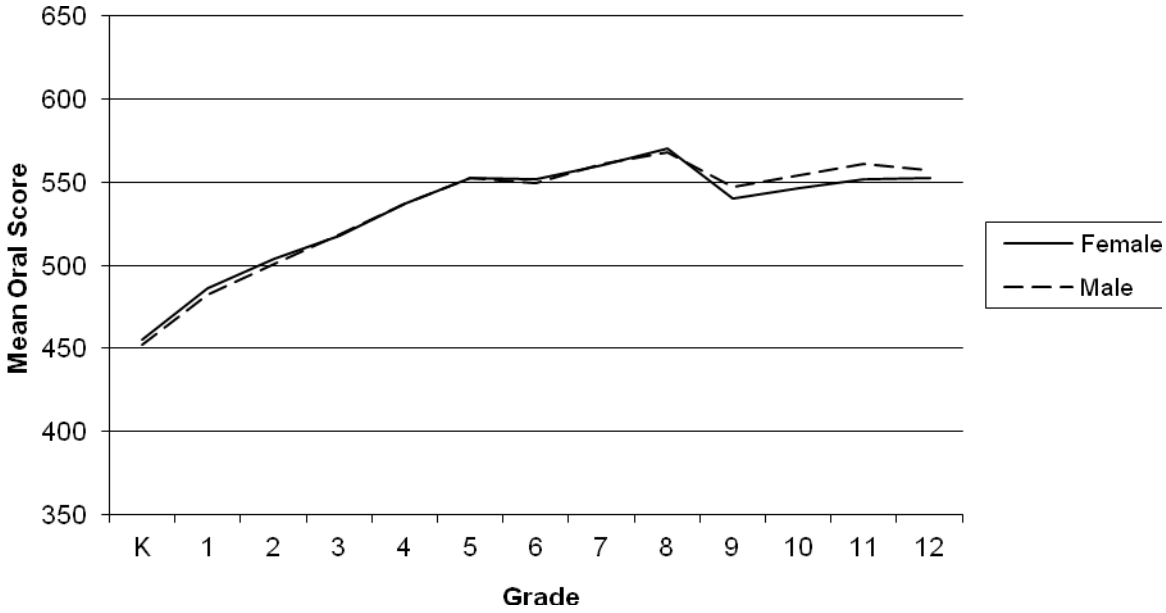
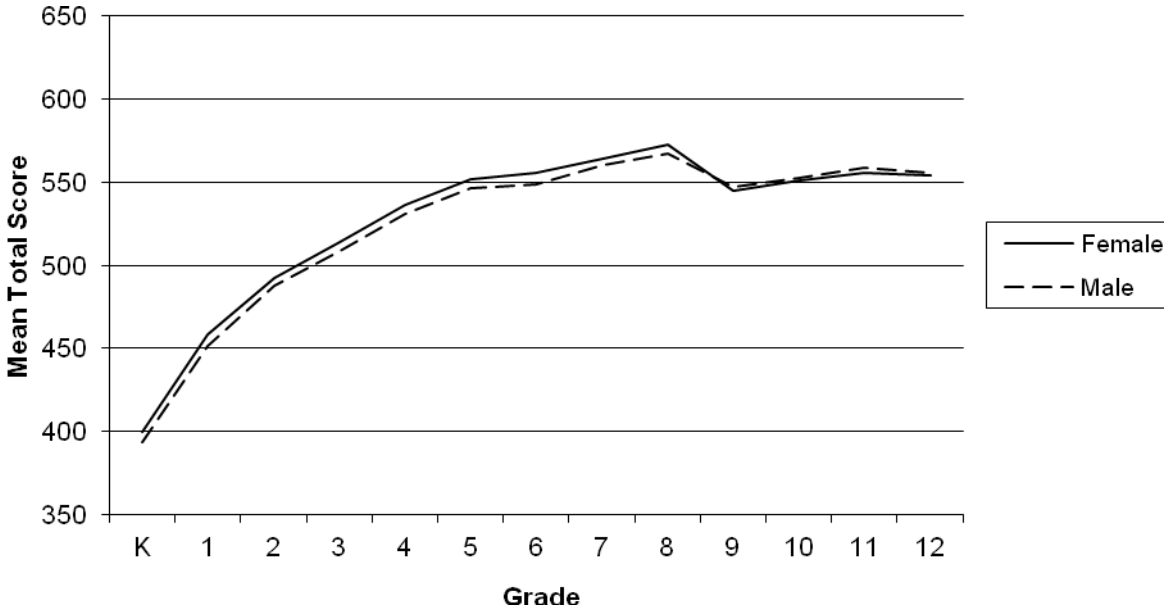


Figure 7. Mean Total Scale Scores by Grade and Gender



The performance of students tested with and without accommodations is provided in Tables 25 and 26. Because the numbers of students receiving accommodations at each grade level are very small, all accommodations for a content domain are combined in these tables.

Table 25. Total Scale Score Means by Grade and Accommodations

Grade	Total Scale Scores								
	Without Accommodations			With Accommodations			Not Specified		
	N	Mean	SD	N	Mean	SD			
K	12395	396.38	37.53	16	377.25	38.47	255	402.01	38.58
1	12611	454.78	37.33	25	410.20	54.68	238	459.26	38.44
2	11745	489.79	35.72	22	449.36	37.69	231	497.37	37.28
3	11144	510.82	40.12	18	487.94	52.80	210	518.76	40.89
4	9892	533.26	40.24	21	491.14	63.11	201	543.37	43.05
5	8576	549.10	40.89	25	501.60	53.40	144	558.61	43.28
6	6931	551.81	39.50	15	474.07	47.81	189	559.26	36.38
7	5541	562.05	40.39	11	512.00	70.75	177	562.66	39.02
8	4850	569.41	42.99	13	536.23	42.39	137	574.59	42.61
9	4198	546.13	42.36	6	487.83	42.02	113	550.25	39.18
10	3693	551.56	44.19	6	460.17	73.61	120	548.55	48.17
11	3019	556.69	45.23	7	534.57	52.86	102	565.73	40.95
12	2755	554.61	46.83	10	468.90	89.22	74	562.70	38.03

Table 26. Component Scale Score Means by Grade and Accommodations

Grade	Speaking Scale Scores								
	No Speaking Accommodations			With Speaking Accommodations			Not Specified		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
K	12775	465.43	41.14	11	401.82	73.23	193	482.06	40.09
1	12828	492.65	36.08	15	478.27	42.65	182	504.03	37.27
2	11931	509.13	35.41	12	471.75	34.70	192	516.34	41.35
3	11308	527.22	42.63	15	503.53	70.69	175	538.34	49.82
4	10025	542.70	44.81	14	495.21	86.96	155	559.61	48.09
5	8686	554.45	46.70	17	529.00	58.12	116	569.47	51.45
6	7054	557.53	49.49	5	461.40	51.13	165	580.58	45.90
7	5623	566.00	51.25	2	473.00	55.15	161	579.31	43.50
8	4921	571.86	55.16	10	544.40	50.77	115	581.92	57.26
9	4250	551.06	57.64	5	524.00	20.10	100	548.13	61.37
10	3760	553.61	59.31	3	500.33	30.75	90	551.64	64.09
11	3072	558.78	60.99	4	522.50	44.78	85	577.47	58.87
12	2853	555.68	62.99	7	517.43	101.44	57	559.42	52.18

Grade	Listening Scale Scores								
	No Listening Accommodations			With Listening Accommodations			Not Specified		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
K	12699	430.58	36.30	11	406.36	35.57	194	439.42	31.45
1	12807	472.47	36.45	12	431.75	24.04	188	475.98	36.27
2	11936	500.17	35.02	12	460.67	34.57	191	501.28	34.61
3	11296	502.14	43.60	13	496.15	54.32	175	511.98	40.81
4	9990	525.55	44.87	13	516.62	39.41	155	531.50	45.74
5	8674	541.91	46.98	15	479.13	63.17	116	547.29	49.38
6	7026	548.47	53.65	5	480.40	91.01	159	547.50	53.76
7	5619	561.92	55.52	3	511.33	125.17	160	557.93	53.56
8	4914	570.97	56.71	8	510.88	46.09	114	566.21	61.71
9	4239	541.46	55.03	5	476.60	46.65	99	547.77	56.31
10	3747	549.72	58.32	4	467.25	84.39	93	548.00	71.28
11	3070	556.22	59.46	5	532.40	70.06	88	557.49	53.56
12	2840	555.24	60.87	7	482.57	99.13	56	562.86	48.74

Table 26. Component Scale Score Means by Grade and Accommodations (continued)

Grade	Reading Scale Scores								
	No Reading Accommodations			With Reading Accommodations			Not Specified		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
K	12614	363.18	46.01	14	328.07	48.29	248	373.61	49.38
1	12744	425.14	44.96	16	353.31	60.80	219	432.46	43.34
2	11875	469.40	49.36	14	418.36	47.10	228	478.71	49.45
3	11264	498.47	52.43	13	464.54	77.69	195	509.58	49.97
4	9939	526.04	52.24	16	491.19	65.00	198	541.95	53.19
5	8620	545.40	51.78	17	477.76	65.16	140	555.42	52.25
6	7011	541.25	44.34	10	463.20	42.28	179	546.36	41.93
7	5593	553.76	45.65	6	468.17	52.75	175	553.69	46.05
8	4886	564.44	48.49	8	529.63	40.58	131	567.55	51.02
9	4225	547.41	42.22	5	502.80	20.14	111	552.61	34.37
10	3724	555.12	44.42	5	490.40	61.50	114	550.70	53.07
11	3055	561.94	45.16	6	538.33	51.38	101	572.18	42.01
12	2817	561.26	46.61	9	507.22	84.57	73	570.82	38.61

Grade	Writing Scale Scores								
	No Writing Accommodations			With Writing Accommodations			Not Specified		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
K	12482	325.72	76.79	13	341.00	67.07	251	326.44	74.92
1	12738	429.06	65.08	12	308.42	99.40	226	432.40	67.24
2	11860	481.56	55.35	15	417.47	85.81	230	491.11	58.63
3	11208	515.96	59.48	16	446.13	76.12	201	520.66	56.83
4	9934	539.43	58.03	16	470.63	83.51	200	550.71	61.64
5	8625	555.20	58.76	24	483.21	70.74	143	564.30	60.13
6	6981	560.34	51.85	14	458.07	75.87	184	566.40	51.08
7	5583	566.60	51.71	7	517.29	119.18	175	565.14	48.20
8	4882	571.20	51.28	10	529.20	50.71	132	579.05	54.56
9	4223	545.52	51.45	4	492.00	43.89	113	548.81	45.12
10	3717	548.51	51.36	5	452.00	100.89	118	543.58	46.82
11	3043	550.95	51.67	7	511.86	66.09	102	562.79	45.55
12	2808	546.25	53.93	9	425.89	113.26	73	560.58	40.98

Part 7: Reliability and Validity Evidence

Reliability statistics and standard error of measurement were computed using the data from the Spring 2012 CELApro administration, along with information obtained in previous years. Overall, the 2011 CELApro analyses yielded results that are consistent with previous CELApro and *LAS Links* results.

Test validation is an ongoing process of gathering evidence from many sources to evaluate the soundness of the desired score interpretation or use. This evidence is acquired from studies of the content of the test as well as from studies involving scores produced by the test. Additionally, reliability is a necessary element for validity. A test cannot be valid if it is not also reliable. All test scores contain some measurement error. Test score reliability refers to the degree to which scores on a particular assessment are free of the kinds of measurement errors that introduce variability in a student's scores. Thus, the reliability coefficient quantifies the expected consistency of student performance across multiple test forms or multiple testing occasions.

Internal Consistency Reliability

Total-test reliability measures, such as Cronbach's (1951) coefficient alpha and standard error of measurement, consider the consistency (reliability) of performance over all test questions in a given form; the results of which imply how well the questions measure the content domain and could continue to do so over repeated administrations. Total-test reliability coefficients, such as coefficient alpha, may range from 0.00 to 1.00, where 1.00 refers to a perfectly consistent test.

The internal consistency reliability of the CELApro Speaking, Listening, Reading, Writing, Comprehension, and Oral scales was evaluated using Cronbach's coefficient alpha, computed with the standard formula

$$C_{\alpha} = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_X^2}\right],$$

where

n = the number of items,

$\sigma^2{}_i$ = the raw item variance, and

$\sigma^2{}_X$ = the raw score variance for each scale.

Because the CELApro total scale score is a composite (the unweighted mean of the four component scale scores on Speaking, Listening, Reading, and Writing), the internal consistency

reliability of the total score was computed using the formula for the reliability of battery composites,

$$\rho_{ZZ'} = 1 - \frac{\sum_{j=1}^{k} \sigma^2{}_{x_j}(1 - \rho_{x_j x'_j})}{k^2 \sigma_Z^2},$$

where

k = the number of component scales (for CELApro, k = 4),

$\rho_{x_j x'_j}$ = reliability of each of the component scales,

$\sigma^2{}_{x_j}$ = scale score variance of each of the component scales, and

$\sigma^2{}_Z$ = variance of the total (mean) scale score.

The internal consistency reliability coefficients for the 2012 CELApro tests are shown in Table 27. Achievement tests are typically considered to be of sound reliability when their reliability coefficients are in the range of 0.80 and above. All of the reliability coefficients for Speaking, Reading, Comprehension, and Oral meet or exceed this criterion. In Writing, all grades except Kindergarten also exceeded this criterion. However, the reliability coefficients for the Listening scale are at or below 0.80 for every grade. This is consistent with the results observed in prior years and may be a consequence of the very small number of items and score points on the Listening test. Because the Listening scores account for one fourth of the total composite, their lower reliability serves to lower the total score reliability as well. In spite of this, the total score reliability coefficients exceed 0.90 for every grade and grade span.

Table 27. Internal Consistency Reliability Coefficients by Grade Span and Grade

	Speaking	Listening	Reading	Writing	Comprehension	Oral	Total Score
Grade Span 1	0.92	0.85	0.87	0.81	0.90	0.93	0.94
K	0.92	0.75	0.82	0.78	0.83	0.92	0.90
1	0.90	0.75	0.82	0.85	0.84	0.90	0.94
2	0.88	0.69	0.85	0.88	0.85	0.89	0.94
Grade Span 2	0.88	0.71	0.89	0.87	0.88	0.88	0.94
3	0.87	0.65	0.86	0.87	0.85	0.87	0.93
4	0.87	0.67	0.87	0.87	0.87	0.87	0.93
5	0.88	0.69	0.88	0.86	0.87	0.88	0.93
Grade Span 3	0.91	0.73	0.84	0.81	0.87	0.90	0.93
6	0.89	0.71	0.83	0.81	0.86	0.89	0.92
7	0.91	0.73	0.84	0.81	0.87	0.90	0.93
8	0.92	0.76	0.85	0.81	0.88	0.92	0.94
Grade Span 4	0.93	0.78	0.85	0.87	0.89	0.93	0.95
9	0.93	0.75	0.83	0.86	0.87	0.92	0.94
10	0.93	0.78	0.85	0.87	0.89	0.93	0.95
11	0.94	0.78	0.86	0.87	0.90	0.93	0.95
12	0.94	0.80	0.86	0.88	0.90	0.93	0.95

Standard Errors of Measurement

Another measure of reliability is a direct estimate of the degree of measurement error in students' reported scores on a test. This second measure of reliability is called the standard error of measurement (SEM) and represents the number of score points about which a given score is expected to vary. The SEM of the CELApro Speaking, Listening, Reading, Writing, Comprehension, and Oral scales was computed with the standard formula

$$SEM = SD * \sqrt{1 - alpha},$$

where *SD* is standard deviation of scale score
alpha is reliability coefficient.

Because the SEM is a function of the test reliability and the standard deviation, there is no prescribed minimum value for this statistic. Rather, the SEM is useful primarily in evaluating the range over which individual scores may be expected to vary. We can expect that two thirds of the obtained scores will be within one SEM of the individuals' true scores.

The SEMs for the Spring 2012 CELApro assessments are shown in Table 28.

Table 28. Standard Errors of Measurement by Grade Span and Grade

	Speaking	Listening	Reading	Writing	Comprehension	Oral	Total Score
Grade Span 1	11.54	17.74	22.93	40.54	16.58	9.95	12.79
K	11.46	18.03	19.70	36.18	16.03	10.99	11.60
1	11.38	18.06	19.20	25.00	15.09	9.39	9.52
2	12.13	19.37	19.19	19.51	14.64	10.13	8.91
Grade Span 2	16.28	25.98	18.88	21.64	16.17	14.82	10.50
3	15.51	25.68	19.46	21.60	16.35	13.33	10.44
4	16.27	25.68	18.67	21.19	15.93	14.73	10.37
5	16.43	26.18	18.31	21.94	16.08	15.71	10.52
Grade Span 3	15.98	28.83	18.60	22.57	15.28	15.05	11.02
6	16.12	28.94	18.27	22.57	15.07	14.47	11.01
7	15.71	28.75	18.46	22.72	15.18	14.92	10.98
8	15.69	28.12	18.84	22.30	15.31	15.28	10.87
Grade Span 4	15.33	27.69	17.33	18.93	16.36	14.94	10.19
9	14.73	27.53	17.51	18.98	16.40	13.97	10.13
10	15.34	27.62	17.36	18.71	16.27	15.01	10.15
11	15.42	27.77	16.96	18.79	16.26	15.10	10.15
12	15.91	27.36	17.33	19.21	16.25	15.64	10.22

Validity Evidence

The purpose of test validation is to validate interpretations of the test scores for particular purposes or uses. Test validation is an ongoing process, beginning at initial conceptualization and continuing throughout the lifetime of an assessment. Every aspect of an assessment provides evidence in support of its validity (or evidence to the contrary), including design, content requirements, item development, and psychometric quality.

The *LAS Links* and CELApro tests were designed and developed to provide English language proficiency scores that are valid for most types of educational decision making. The primary inferences from the test results include measurement of the proficiency of individual students relative to an international sample and relative program effectiveness based on the results of groups of students. Progress can be tracked over years and grades. The results can be used in a norm- and/or criterion-referenced manner to analyze the strengths and weaknesses of a student's growth in each skill area, to plan for further instruction and curriculum development, and to report progress to parents. The results can also be used as one factor in making administrative decisions about program effectiveness, class grouping, needs assessment, and

placement in English Language Development ELD programs.

The *LAS Links* program was developed in accordance with the criteria for test development, administration, and use described in the *Standards for Educational and Psychological Testing* (1999) adopted by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME).

Content Validity

Content-related validity for language proficiency tests is evidenced by a correspondence between test content and instructional content. To ensure such correspondence, developers conducted a comprehensive curriculum review and met with educational experts to determine common educational goals and the knowledge and skills emphasized in curricula across the country. This information guided all phases of the design and development of the *LAS Links* suite of assessments.

As described in Part 1 of this report and summarized in Table 2, a study of the alignment of the CELApro assessments to the Colorado standards was also conducted, and a high level of agreement has been found. This alignment is expected to become even stronger as the CELApro assessments are further customized in future years.

Construct Validity

Construct validity—what test scores mean and what kinds of inferences they support—is the central concept underlying the *LAS Links* test validation process. Evidence for construct validity is comprehensive and integrates evidence from both content- and criterion-related validity. To establish meaningfulness, *LAS Links* should correlate highly with independent measures of achievement and cognitive ability.

Convergent and discriminate validity evidence can also be established through a pattern of high correlations among scales that purport to measure domains that are known to be closely related and lower correlations among scales that purport to measure dissimilar domains. This kind of pattern provides evidence that the scales are actually measuring the constructs that they purport to measure. While we have no external measures available at present to correlate with the CELApro scale scores, the pattern of correlations within CELApro provides preliminary validity evidence. The 2011 intercorrelations among the CELApro scales for each grade and grade span are shown in Tables 29 through 32.

Table 29. CELApro Scale Score Correlations, Grade Span K–2

		Listening	Reading	Writing	Comprehension	Oral	Total
K	Speaking	0.49	0.53	0.26	0.57	0.92	0.68
	Listening		0.63	0.29	0.88	0.69	0.71
	Reading			0.44	0.87	0.60	0.82
	Writing				0.37	0.29	0.78
	Comprehension					0.72	0.81
	Oral						0.74
1	Speaking	0.51	0.52	0.48	0.56	0.92	0.72
	Listening		0.60	0.52	0.82	0.74	0.76
	Reading			0.71	0.91	0.61	0.87
	Writing				0.68	0.55	0.89
	Comprehension					0.73	0.90
	Oral						0.81
2	Speaking	0.45	0.50	0.51	0.54	0.91	0.73
	Listening		0.52	0.50	0.77	0.69	0.73
	Reading			0.72	0.89	0.56	0.87
	Writing				0.71	0.56	0.89
	Comprehension					0.69	0.90
	Oral						0.80

Table 30. CELApro Scale Score Correlations, Grade Span 3–5

		Listening	Reading	Writing	Comprehension	Oral	Total
3	Speaking	0.48	0.47	0.46	0.52	0.91	0.71
	Listening		0.58	0.54	0.82	0.77	0.78
	Reading			0.73	0.89	0.58	0.87
	Writing				0.72	0.55	0.87
	Comprehension					0.72	0.91
	Oral						0.84
4	Speaking	0.46	0.47	0.44	0.51	0.89	0.71
	Listening		0.60	0.52	0.84	0.77	0.78
	Reading			0.71	0.90	0.59	0.87
	Writing				0.70	0.53	0.85
	Comprehension					0.73	0.91
	Oral						0.84
5	Speaking	0.47	0.48	0.44	0.51	0.88	0.72
	Listening		0.61	0.51	0.85	0.78	0.79
	Reading			0.67	0.89	0.58	0.86
	Writing				0.67	0.51	0.84
	Comprehension					0.73	0.90
	Oral						0.84

Table 31. CELApro Scale Score Correlations, Grade Span 6–8

		Listening	Reading	Writing	Comprehension	Oral	Total
6	Speaking	0.46	0.46	0.47	0.52	0.89	0.74
	Listening		0.56	0.50	0.78	0.77	0.80
	Reading			0.66	0.92	0.57	0.83
	Writing				0.68	0.54	0.82
	Comprehension					0.71	0.90
	Oral						0.87
7	Speaking	0.47	0.46	0.47	0.52	0.88	0.75
	Listening		0.56	0.52	0.78	0.78	0.81
	Reading			0.64	0.92	0.56	0.82
	Writing				0.67	0.55	0.82
	Comprehension					0.71	0.89
	Oral						0.87
8	Speaking	0.51	0.50	0.51	0.55	0.89	0.78
	Listening		0.60	0.56	0.80	0.78	0.83
	Reading			0.65	0.92	0.59	0.83
	Writing				0.67	0.58	0.83
	Comprehension					0.73	0.90
	Oral						0.88

Table 32. CELApro Scale Score Correlations, Grade Span 9–12

		Listening	Reading	Writing	Comprehension	Oral	Total
9	Speaking	0.51	0.49	0.56	0.55	0.87	0.80
	Listening		0.63	0.60	0.82	0.82	0.83
	Reading			0.67	0.92	0.62	0.82
	Writing				0.71	0.64	0.85
	Comprehension					0.76	0.90
	Oral						0.91
10	Speaking	0.56	0.52	0.56	0.59	0.88	0.81
	Listening		0.65	0.60	0.83	0.84	0.85
	Reading			0.68	0.93	0.64	0.84
	Writing				0.72	0.62	0.84
	Comprehension					0.76	0.91
	Oral						0.90
11	Speaking	0.54	0.52	0.56	0.57	0.87	0.80
	Listening		0.68	0.60	0.84	0.83	0.85
	Reading			0.67	0.93	0.65	0.84
	Writing				0.70	0.62	0.84
	Comprehension					0.76	0.90
	Oral						0.91
12	Speaking	0.56	0.54	0.57	0.59	0.88	0.81
	Listening		0.68	0.63	0.85	0.83	0.86
	Reading			0.70	0.93	0.64	0.85
	Writing				0.73	0.64	0.85
	Comprehension					0.76	0.91
	Oral						0.90

Overall, the patterns of correlations among the four content domains of Speaking, Listening, Reading, and Writing are similar to the patterns observed in the 2011 data and are consistent with theoretical expectations for the CELApro language constructs. For example, the correlations support the distinction between the receptive language skills (Listening and Reading) and the productive language skills (Speaking and Writing). At almost all grade levels, the component exhibiting the highest correlation with the Listening scale is the Reading scale. The relationship between the productive skills of Speaking and Writing is less clear.

Consistent with 2011, the highest single correlation coefficient among the four domains at every grade except Kindergarten is the correlation between the two orthographic domains of Reading and Writing.

Part 8. Special Studies

No special studies were conducted in the 2012 administration.

References

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443–459.

Burket, G. R. (2002). *PARDUX* [Computer program]. Monterey, CA: CTB/McGraw-Hill.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks: Sage.

Chamot, A. U., & O'Malley, J.M. (1994). *The CALLA handbook*. Reading, MA: Addison-Wesley.

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

CTB/McGraw-Hill. (2007). *LAS links technical manual*. Monterey, CA: Author.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*,613–619.

Green, D. R. (1975). *What does it mean to say a test is biased?* Paper presented at American Educational Research Association, Washington, D.C.

Linn, R. L., & Harnisch, D. (1981). Interactions between item content and group membership in achievement test items. *Journal of Educational Measurement, 18*,109–118.

Linn, R. L., Levine, M.V., Hastings, C.N., &Wardrop, J.L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*,159–173.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacMillan/McGraw-Hill. (1993). *Guidelines for bias-free publishing.* New York: Author.

MacMillan/McGraw-Hill. (1993). *Reflecting diversity: Multicultural guidelines for educational publishing professionals*. New York: Author.

O'Malley, J. M., &Valdez Pierce, L. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. Boston: Addison-Wesley.

Sandoval, J. H., & Mille, M.P. (1979). *Accuracy of judgments of WISC-R item difficulty for minority groups*. Paper presented at the annual meeting of the American Psychological Association, New York.

Savignon, S. (1972). *Communicative competence: An experiment in foreign language teaching*. Philadelphia: The Center for Curriculum Development.

Savignon, S. (1997). *Communicative competence: Theory and classroom practice* (2nd ed.). New York: McGraw-Hill.

Scheuneman, J. D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. *Educational Psychology, 19* (4),219–225.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (3–32). Cambridge, UK: Cambridge University Press.

Stocking, M. L., &Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*,201–210.

Subkoviak, M. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25* (1), 47–55.

.