Colorado
English Language Assessment Program

2007 Technical Report

Submitted to the
Colorado Department of Education

October 1, 2007

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 Assessment (CELA) was in Spring 2006. That first administration consisted of CTB/McGraw-Hill's *LAS Links, Form A,* which was administered to Colorado English Language Learners in Kindergarten through 12th grade. All five of the *LAS Links* grade spans were administered.

For the 2007 CELA administration, the tests for Grades 6-8 and 9-12 remained unchanged, but the tests for Kindergarten through Grade 5 were reconfigured to conform to the Colorado grade spans. The grade spans used in 2006 and 2007 are shown in Table 1, below:

Table 1. Comparison of 2006 and 2007 CELA Grade Spans

Grade	CELA Grade Spans	
	2006	2007
K	**1** (K-1)	**1** (K-2)
1		
2	**2** (2-3)	
3		**2** (3-5)
4	**3** (4-5)	
5		
6	**4** (6-8)	**3** (6-8)
7		
8		
9	**5** (9-12)	**4** (9-12)
10		
11		
12		

The reconfigured CELA forms for the 2007 Grade Spans 1 and 2 were created using items from *LAS Links Form A*, Levels 1, 2, and 3 (K-1, 2-3, and 4-5). These reconfigured tests, like the *LAS Links* assessments upon which they were based, were built upon standards and blueprints that are aligned to the Teachers of English to Speakers of Other Languages (TESOL) standards and to several states' language proficiency standards.

The *LAS Links*/CELA tests include multiple-choice and performance-based questions to assess students' English language proficiency in Reading, Writing, Listening, and Speaking. Items from these four skill areas are also combined to yield a Comprehension score (from selected Reading and Listening items) and an Oral score (from Listening and Speaking items). A total score is also reported. The total score is computed by averaging each student's scale scores in Reading, Writing, Listening and Speaking.

The 2007 CELA tests for Grade Spans 3 and 4 are identical to the *LAS Links Form A* tests for the corresponding grades. A description of the *LAS Links* development, scaling, alignment, and standardization is available in the *LAS Links Technical Manual* (CTB, 2006). All items were field tested and reviewed for fairness and bias. Bias and sensitivity reviews were conducted via committee reviews and bias was further statistically investigated via differential item functioning (DIF) analyses.

Comprehensive technical documentation for *LAS Links* is available in the *LAS Links Technical Manual* (CTB, 2006). The present document is intended to serve as a supplement to that manual, and contains information specific to the 2007 CELA test administration.   This report describes the reconfiguration of tests for the lower grade spans, the alignment of CELA to the Colorado Learning Standards, the methods and procedures employed in the scoring of the Colorado tests, the item and test statistics for Colorado students, and the performance of Colorado students on each of the tests.

Part 1: Standards

The Colorado English Language Assessment (CELA) is the language proficiency assessment used for classifying and monitoring the progress of Colorado English Language Learners (ELLs) in the acquisition of English. This assessment measures the competencies necessary for successful social and academic language use in four major modalities: Listening, Speaking, Reading, and Writing – along a continuum of five proficiency levels: Beginning, Early Intermediate, Intermediate, Proficient, and Advanced. The assessment takes into account the students' maturation and cognitive skills by providing age appropriate tests covering four grade spans: K-2, 3-5, 6-8, and 9-12.

A combination of item types -- Dichotomous Constructed-Response (Correct or Incorrect), Constructed-Response (CR), and Selected-Response (SR) items -- provide a variety of ways for students to demonstrate proficiency and to maintain reasonable testing times. Constructed-Response (CR) items assess the productive domains of Speaking and Writing, whereas the Selected-Response (SR) 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 the interpretation of the results in multiple ways.

Alignment Studies

An important indicator of the validity of a standardized 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. Alternatively, in the case of an already existing test, an alignment study (a research report that gives the detailed alignment tables showing the standards and matching test items) may be provided to a customer who requests one in order to determine whether to adopt a particular test and/or as evidence of accountability that may be submitted for NCLB reporting requirements.

The following discussion of the CELA alignment to Colorado Standards is intended to provide a brief overview of the alignment process and results. A more comprehensive description of the alignment study procedures and results has been provided to CDE in a separate document.

Colorado has four general standards for English language learners (ELL), 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.

Within each of the four general standards listed above, Colorado specifies subskills for each grade level, which are similar (but not identical) to the *LAS Links* standards. The matching of *LAS Links* items to Colorado standards was done by two expert raters with backgrounds in K-12 teaching and English language acquisition. In the first step, similar subskills were grouped across proficiency levels into an easily readable horizontal grid. Then, the *LAS Links* items were matched as closely as possible to individual Colorado standards. In performing the match, the raters took into account two factors: (1) the wording of the standards so that the skills mentioned in the standards could be matched with the skills and competencies of the item, and (2) the estimated proficiency level of the item (whether beginning, intermediate, or advanced students would be expected to get the item correct).

The protocol for conducting alignments has been established in a number of studies (LaMarca, 2001; Sato et al, 2005; Webb, 1997, 1999); CTB uses an adaptation of these protocols in aligning its assessment instruments to state standards. The CTB operational definition of an alignment is expressed as the percentage of assessable standards for which there are matching test items.

An initial alignment was done between the CELA standards and LAS Links, to determine where new test items were needed to match specific standards. After multiple reviews, CTB found that the alignment could be increased with a fewer amount of items.

The method below describes the process used in developing the customized version of the CELA tests in order to achieve the maximum possible alignment to the Colorado ELD standards.

Method

The following procedures were followed in conducting the alignment:

a. Proficiency levels, modalities (reading, writing, speaking, listening), and grade spans of state standards and the assessment instrument were compared to determine the degree of correspondence.

b. Standards were rearranged into tables (representing a general model of academic language competence) that shows the progression of skills from one proficiency level to the next.

c. A determination was made as to which standards are not assessable in a standardized testing format for various reasons.

d. Test items were matched to standards by determining:

    i. what skills are represented in a test item;
    ii. the proficiency level of the item;
    iii. partial or indirect matching of items to standards;
    iv. whether items tested in one modality of the test match standards at another modality (e.g., a listening standard may specify sound discrimination, which is actually tested in the reading portion of a given test).

e. The test item numbers aligned to one standard.

f. For rubric-scored writing and speaking items, each item was repeated in each cell at each proficiency level.

g. Ratings were compared with other rater(s) and a consensus was reached when there were differences in the item to standard matching.

h. The number of standards for which there are items listed in the cell was divided by the number of total number of *assessable* standards (not the total standards, which include non-assessable standards) for each modality to get an alignment percentage.

In alignments in general, it is often possible to find a direct one-to-one match of an item to a single standard. In other cases, an item may align to several standards, may indirectly or partially align to a standard, or may align to the wording of a standard at a particular proficiency level even though the item difficulty is at a higher or lower level of difficulty. On the other hand, some standards may not be efficiently or directly testable on a standardized test. These standards have been eliminated from the final alignment because they cannot or should not be tested due to one of the following reasons:

(1) assessment of the standard is not feasible within the time constraints of the test,

(2) assessment of the standard would require students to provide personal information that would not pass the content and bias review,

(3) the standard specifies parameters or situations outside of a standardized testing situation, such as:

(a) participate in class or group discussions,

(b) proofread or correct own work – writing or read aloud,

(c) produce humor, idiomatic or figurative expressions, which are difficult to elicit in a test situation,

(d) specify how words were taught to the students (word which have been taught in context, for example, cannot be determined on a test),

(e) require other materials at the test site, e.g., books or electronic media or various resources,

(f) require assistance or support from the tester/teacher,

(g) specify prior knowledge or background knowledge as part of the standard,

(h) require extensive outside preparation (formal reports and presentations),

(i) specify multiple steps or strategies ("plan, draft, revise" or "organize classroom procedure"),

(j) combine modalities (e.g., in the listening section, “demonstrate comprehension by explaining, paraphrasing, giving opinions,” requires the student to speak or write an answer),

(k) require use of student’s native language.

In performing the alignment, the raters independently matched items to all possible alignable standards on the basis of direct, indirect, or partial alignment. The test item numbers were then entered into the cell of the matching standard. A detailed description of the standards, by grade and proficiency level, is provided in Appendix F.

The final step of the alignment involved the calculation of an alignment percentage, i.e., the percentage of alignable standards that are covered by items in the CELA test. The numbers of alignable standards having matched items from CELA were divided by the total number of alignable standards, as shown in Table 2.

Table 2. Item Alignment Percentages by Grade Span

	Grades K-2									
	Standard 1 Listening		Standard 2 Speaking		Standard 3 Reading		Standard 4 Writing		Total	
LEVEL	# Aligned/ Alignable Stds (Total Stds)	Align %	# Aligned/ Alignable Stds (Total Stds)	Align %	# Aligned/ Alignable Stds (Total Stds)	Align %	# Aligned/ Alignable Stds (Total Stds)	Align %	# Aligned/ Alignable Stds (Total Stds)	Align %
Beginning	8/8 (8)*	100%	5//5 (7)	100%	5/6 (11)	83%	6/6 (10)	100%	24/25 (36)	96%
Intermediate	5/5 (5)	100%	7/7 (7)	100%	8/9 (12)	89%	8/9 (11)	89%	28/30 (35)	93%
Advanced	4/5 (6)	80%	5/5 (5)	100%	7/7 (9)	100%	7/8 (9)	88%	23/2 (29)	92%
Total	17/18 (19)	94%	17/17 (19)	100%	20/22 (32)	91%	21/23 (30)	91%	75/80 (100)	94%
Grades 3-5										
Beginning	7/8 (8)	88%	5/5 (7)	100%	6/7 (8)	86%	4/5 (9)	80%	22/25 (32)	88%
Intermediate	5/5 (5)	100%	7/8 (9)	88%	9/9 (10)	100 %	9/9 (11)	100%	30/31 (35)	97%
Advanced	4/5 (6)	80%	5/6 (7)	83%	7/8 (9)	88 %	7/9 (10)	78%	23/28 (32)	82%
Total	16/18 (19)	89%	17/19 (23)	89%	22/24 (27)	92%	20/23 (30)	87%	75/84 (99)	89%
Grades 6-8										
Beginning	5/8 (8)	62%	6/7 (8)	86%	2/3 (6)	67%	6/6 (7)	100%	19/24 (29)	79%
Intermediate	5/5 (5)	100%	6/8(10)	75%	8/8 (8)	100%	5/8 (11)	62%	24/29 (34)	83%
Advanced	6/6 (6)	100%	4/5 (7)	80%	6/8 (9)	75%	5/6 (11)	83%	21/25 (33)	84%
Total	16/19 (19)	84%	16/20 (25)	80%	16/19 (23)	84%	16/20 (29)	80%	64/78 (96)	82%
Grades 9-12										
Beginning	6/8 (8)	75%	6/8 (8)	75%	4/5 (6)	80%	4/7 (8)	57%	20/28 (32)	71%
Intermediate	4/5 (5)	80%	8/8 (10)	100%	6/7 (9)	86%	5/10 (13)	50%	23/29 (37)	76%
Advanced	6/6 (6)	100%	4/5 (7)	80%	5/6 (6)	83%	7/9 (13)	78%	22/26 (32)	85%
Total	16/19 (19)	84%	18/21 (25)	86%	15/18 (21)	83%	15/26 (34)	58%	64/84 (99)	76%

Notes:
1. Indirectly tested standards are included in the alignment percentages
2. Numbers in parentheses indicate the total number of standards

Part 2: Test Development

The 2007 CELA tests consist entirely of *LAS Links* items. For the two upper grade spans (Grades 6-8 and 9-12) the CELA tests are identical to the corresponding *LAS Links* assessments. The reconfigured tests for the two lower grade spans (Grades K-2 and 3-5) were created using selected items from the *LAS Links* assessments for the appropriate grades. 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 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. Distractors also 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 Dichotomous Constructed-Response (Correct or Incorrect), Constructed- Response, and Multiple-Choice items are used to provide diverse opportunities for students to demonstrate proficiency and to maintain reasonable testing times. Constructed-Response items are used to assess the productive domains of Speaking and Writing, whereas the Multiple-Choice items are used to assess the receptive domains of Listening, Reading, and the Writing Use Conventions subtest. The structure of the 2007 CELA is shown in Table 3.

Table 3. 2007 CELA Test Structure

Content	Gradespan	Sub-Content	Item Type	Items	Score Points	CR/DCR Items Scored By	Administration
Speaking 20 items, 41 pts	4 gradespans: 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 20 items. 20 pts	4 gradespans: K-2, 3-5, 6-8, 9-12	Listen for Information	MC	10	10	Not Applicable	Individual
		Listen in the Classroom	MC	6	6		
		Listen & Comprehend	MC	4	4		
Reading K=25 items, 25 pts 1-12=35 items.35pts	K Only	Analyze Words	MC	10	10	Not Applicable	Individual
		Read Words	MC	10	10		
		Understanding	MC	5	5		
	4 gradespans: 1-2, 3-5, 6-8, 9-12	Analyze Words	MC	10	10	Not Applicable	Group
		Read Words	MC	10	10		
		Understanding	MC	15	15		
Writing K-1=25 items, 35pts 2-12=25 items,36pts	K-1	Conventions	MC	20	20	CTB Handscoring	Group (or Individual for K)
		Write About	CR	2	6		
		Write Why	CR	3	9		
	4 gradespans: 2, 3-5, 6-8, 9-12	Conventions	MC	20	20	CTB Handscoring	Group
		Write About	CR	2	6		
		Write Why	CR	2	6		
		Write in Detail	CR	1	4		
Oral 40 items, 61 pts	4 gradespans: K-2, 3-5, 6-8, 9-12	Listening & Speaking	MC	20	20	Local Test Administrator	N/A
			DCR	10	10		
			CR	10	31		
Comprehension K = 33 items, 33 pts 1-2 = 43 items, 43 pts 3-5 = 45 items, 45 pts 6-12=47 items, 47 pts	K	Listening & Reading	MC	33	33	Not Applicable	N/A
	1-2	Listening & Reading	MC	43	43		
	3-5	Listening & Reading	MC	45	45		
	6-8 & 9-12	Listening & Reading	MC	47	47		

Item Review and Test Fairness

All items are expected to be fair for all examinees. Various procedures are employed to review item fairness, also referred to as bias. Once the items are developed, they must go through a series of 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 CELA 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 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 one curriculum and no one set of instructional materials will be equally suitable for all. Therefore, no one 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 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 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 cultural 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 & 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 taken in 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 was 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 paid 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 separate 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 & Shepard, 1994; Sandoval & 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 if 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 if 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. Differential item functioning 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, & Wardrop, 1981; Camilli & Shepard, 1994). However, these procedures require a minimum of 800 to 1000 cases in each group of comparison to produce reliable and consistent results. In contrast, the Linn-Harnisch procedure (Linn & 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 85,997 students participated in the 2007 CELA testing. Students in kindergarten and first grade formed the largest groups of examinees (12,366 and 11,956, respectively), with numbers generally decreasing at successive grade levels. The number of male examinees was slightly greater than the number of females at each grade level. The examinee counts by grade and gender are shown in Table 4, below. Note that not all students completed all four of the CELA 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	6,027	6,338	1	12,366
1	5,902	6,051	3	11,956
2	4,772	5,313	1	10,086
3	4,538	4,802	5	9,345
4	3,793	4,026	1	7,820
5	3,276	3,583	2	6,861
6	2,469	3,011	0	5,480
7	2,305	2,601	0	4,906
8	1,990	2,402	0	4,392
9	2,028	2,470	1	4,499
10	1,706	1,894	1	3,601
11	1,238	1,380	0	2,618
12	973	1,093	1	2,067
Total	41,017	44,964	16	85,997

Student ethnicity and home language is summarized by grade span in Tables 5 and 6.

Table 5. Ethnicity by Grade Span

Ethnicity	Grade Span								Total	
	Grades K-2		Grades 3-5		Grades 6-8		Grades 9-12			
	N	%	N	%	N	%	N	%	N	%
AmericanIndian/ Alaska Native	246	0.7%	232	1.0%	184	1.2%	107	0.8%	769	0.9%
Asian/Pacific Islander	2,278	6.6%	1,483	6.2%	895	6.1%	959	7.5%	5,615	6.5%
Black	628	1.8%	397	1.7%	278	1.9%	500	3.9%	1,803	2.1%
Hispanic	29,552	85.9%	21,016	87.5%	12,881	87.2%	10,610	83.0%	74,059	86.1%
White	1,700	4.9%	892	3.7%	535	3.6%	609	4.8%	3,736	4.3%
Not Specified	4	0.0%	6	0.0%	5	0.0%	0	0.0%	15	0.0%
TOTAL	34,408	100%	24,026	100%	14,778	100%	12,785	100%	85,997	100%

Table 6. Home Language (191 Languages Represented).

	Test Level									
	Grades K-2		Grades 3-5		Grades 6-8		Grades 9-12		TOTAL	
	N	%	N	%	N	%	N	%	N	%
Afrikaans	9	0%	4	0%	7	0%	9	0%	29	0%
Akan	12	0%	4	0%	1	0%	6	0%	23	0%
Albanian	16	0%	5	0%	5	0%	7	0%	33	0%
Amharic	110	0%	65	0%	52	0%	98	1%	325	0%
Anuak	0	0%	0	0%	0	0%	1	0%	1	0%
Apache	3	0%	0	0%	0	0%	0	0%	3	0%
Arabic	257	1%	130	1%	64	0%	82	1%	533	1%
Arapaho	1	0%	0	0%	0	0%	0	0%	1	0%
Armenian	9	0%	4	0%	1	0%	4	0%	18	0%
Assamese	11	0%	7	0%	5	0%	8	0%	31	0%
Azerbaijani	0	0%	1	0%	0	0%	0	0%	1	0%
Bambara	2	0%	1	0%	0	0%	1	0%	4	0%
Bangla	0	0%	0	0%	0	0%	2	0%	2	0%
Bashkir	1	0%	2	0%	1	0%	1	0%	5	0%
Bassa	0	0%	0	0%	0	0%	1	0%	1	0%
Bemba	1	0%	1	0%	0	0%	0	0%	2	0%
Bengali	8	0%	6	0%	3	0%	4	0%	21	0%
Bihari	0	0%	1	0%	0	0%	0	0%	1	0%
Bosnian	46	0%	31	0%	20	0%	15	0%	112	0%
Bulgarian	20	0%	9	0%	12	0%	10	0%	51	0%
Burmese	12	0%	7	0%	7	0%	2	0%	28	0%
Cakchiquel, E.	0	0%	1	0%	0	0%	0	0%	1	0%
Catalan	0	0%	1	0%	0	0%	1	0%	2	0%
Chamorro	1	0%	2	0%	1	0%	2	0%	6	0%
Cheyenne	3	0%	0	0%	1	0%	0	0%	4	0%
Cantonese	113	0%	70	0%	51	0%	67	1%	301	0%
Chinese Hakka	34	0%	27	0%	15	0%	18	0%	94	0%
Chinese Mandarin	188	1%	112	1%	61	0%	90	1%	451	1%
Chinese Min Nan	1	0%	2	0%	0	0%	2	0%	5	0%
Chinese Wu	0	0%	0	0%	1	0%	0	0%	1	0%
Choctaw	0	0%	1	0%	0	0%	0	0%	1	0%
Chuukese	4	0%	4	0%	5	0%	3	0%	16	0%
Cora	25	0%	27	0%	17	0%	5	0%	74	0%
Cree	1	0%	0	0%	0	0%	0	0%	1	0%
Creole	18	0%	11	0%	11	0%	12	0%	52	0%
Croatian	5	0%	4	0%	2	0%	4	0%	15	0%
Crow	1	0%	2	0%	0	0%	0	0%	3	0%
Czech	12	0%	8	0%	6	0%	4	0%	30	0%
Danish	8	0%	8	0%	3	0%	2	0%	21	0%
Dari	3	0%	4	0%	3	0%	9	0%	19	0%

	Test Level									
	Grades K-2		Grades 3-5		Grades 6-8		Grades 9-12		TOTAL	
	N	%	N	%	N	%	N	%	N	%
Deccan	4	0%	3	0%	0	0%	1	0%	8	0%
Dinka	9	0%	4	0%	4	0%	3	0%	20	0%
Dutch	13	0%	2	0%	2	0%	3	0%	20	0%
Eleme	0	0%	1	0%	0	0%	0	0%	1	0%
English	20	0%	15	0%	12	0%	13	0%	60	0%
Eskimo	0	0%	2	0%	0	0%	0	0%	2	0%
Estonian	0	0%	0	0%	1	0%	0	0%	1	0%
Ewe	6	0%	1	0%	1	0%	2	0%	10	0%
Fante	0	0%	1	0%	1	0%	2	0%	4	0%
Farsi, Eastern	43	0%	18	0%	17	0%	23	0%	101	0%
Farsi, Western	42	0%	17	0%	8	0%	21	0%	88	0%
Filip-Taga	1	0%	2	0%	3	0%	3	0%	9	0%
Finnish	8	0%	2	0%	0	0%	3	0%	13	0%
Fon	0	0%	0	0%	0	0%	2	0%	2	0%
French	110	0%	72	0%	33	0%	56	0%	271	0%
French Cree	0	0%	1	0%	0	0%	0	0%	1	0%
Fulani	2	0%	0	0%	0	0%	2	0%	4	0%
Ga	1	0%	1	0%	0	0%	2	0%	4	0%
Gaelic	1	0%	0	0%	0	0%	0	0%	1	0%
Ganda	0	0%	1	0%	0	0%	0	0%	1	0%
German	96	0%	54	0%	22	0%	37	0%	209	0%
Grebo	0	0%	0	0%	0	0%	2	0%	2	0%
Greek	5	0%	1	0%	3	0%	2	0%	11	0%
Gujarati	9	0%	3	0%	1	0%	0	0%	13	0%
Gujari	1	0%	0	0%	0	0%	0	0%	1	0%
Haitian, Creole Fr	1	0%	1	0%	0	0%	0	0%	2	0%
Han Chinese	0	0%	0	0%	0	0%	1	0%	1	0%
Hausa	0	0%	3	0%	1	0%	4	0%	8	0%
Hawaiian	4	0%	1	0%	2	0%	0	0%	7	0%
Hebrew	15	0%	12	0%	11	0%	13	0%	51	0%
Hindi	45	0%	16	0%	8	0%	11	0%	80	0%
Hmong	164	1%	155	1%	112	1%	115	1%	546	1%
Hopi	0	0%	1	0%	1	0%	0	0%	2	0%
Hungarian	9	0%	3	0%	0	0%	5	0%	17	0%
Ibo	1	0%	0	0%	0	0%	2	0%	3	0%
Icelandic	0	0%	0	0%	0	0%	1	0%	1	0%
Igbo	7	0%	6	0%	2	0%	4	0%	19	0%
Ilocano	5	0%	3	0%	0	0%	0	0%	8	0%
Indonesian	44	0%	31	0%	13	0%	14	0%	102	0%
Italian	17	0%	9	0%	8	0%	9	0%	43	0%
Iu Mien	1	0%	0	0%	1	0%	2	0%	4	0%
Japanese	72	0%	25	0%	19	0%	13	0%	129	0%

	Test Level									
	Grades K-2		Grades 3-5		Grades 6-8		Grades 9-12		TOTAL	
	N	%	N	%	N	%	N	%	N	%
Javanese	2	0%	0	0%	0	0%	0	0%	2	0%
Kanjobal	16	0%	12	0%	7	0%	9	0%	44	0%
Kannada	5	0%	1	0%	0	0%	1	0%	7	0%
Kawaiisu	0	0%	0	0%	1	0%	0	0%	1	0%
Keres, Eastern	1	0%	0	0%	0	0%	0	0%	1	0%
Keres, Western	1	0%	1	0%	0	0%	0	0%	2	0%
Khmer	72	0%	56	0%	31	0%	26	0%	185	0%
Kikuyu	0	0%	2	0%	0	0%	1	0%	3	0%
Kinyarwanda	1	0%	1	0%	1	0%	1	0%	4	0%
Kirundi	1	0%	2	0%	0	0%	0	0%	3	0%
Korean	288	1%	214	1%	142	1%	158	1%	802	1%
Kosraen	0	0%	0	0%	1	0%	0	0%	1	0%
Kpelle	1	0%	0	0%	0	0%	0	0%	1	0%
Krahn	5	0%	4	0%	3	0%	3	0%	15	0%
Krio	5	0%	0	0%	1	0%	7	0%	13	0%
Kru	2	0%	2	0%	1	0%	2	0%	7	0%
Kurdi/Kurdish	15	0%	11	0%	3	0%	3	0%	32	0%
Lakota	5	0%	1	0%	1	0%	0	0%	7	0%
Lao	71	0%	55	0%	21	0%	19	0%	166	0%
Latvian	2	0%	2	0%	0	0%	0	0%	4	0%
Lebanese	4	0%	0	0%	0	0%	0	0%	4	0%
Liberian Eng.	7	0%	3	0%	3	0%	7	0%	20	0%
Lingala	0	0%	2	0%	0	0%	2	0%	4	0%
Lithuanian	7	0%	9	0%	3	0%	3	0%	22	0%
Luganda	4	0%	2	0%	3	0%	11	0%	20	0%
Lwo	0	0%	0	0%	0	0%	1	0%	1	0%
Maay	16	0%	7	0%	4	0%	1	0%	28	0%
Macedonian	0	0%	2	0%	0	0%	0	0%	2	0%
Magyar	0	0%	0	0%	0	0%	1	0%	1	0%
Makah	0	0%	0	0%	1	0%	0	0%	1	0%
Malay	3	0%	2	0%	0	0%	1	0%	6	0%
Malayalam	9	0%	2	0%	3	0%	4	0%	18	0%
Malinke	0	0%	1	0%	0	0%	0	0%	1	0%
Mandinka	9	0%	3	0%	1	0%	2	0%	15	0%
Marathi	9	0%	1	0%	0	0%	0	0%	10	0%
Marshallese	14	0%	10	0%	2	0%	6	0%	32	0%
Maya	1	0%	0	0%	1	0%	1	0%	3	0%
Mende	1	0%	0	0%	0	0%	1	0%	2	0%
Mongolian	33	0%	39	0%	32	0%	38	0%	142	0%
Navajo	119	0%	114	1%	90	1%	59	1%	382	0%
Nepali	41	0%	42	0%	28	0%	32	0%	143	0%
Norwegian	4	0%	1	0%	2	0%	1	0%	8	0%

	Test Level									
	Grades K-2		Grades 3-5		Grades 6-8		Grades 9-12		TOTAL	
	N	%	N	%	N	%	N	%	N	%
Nuer	1	0%	2	0%	1	0%	2	0%	6	0%
Nyanja	1	0%	2	0%	0	0%	0	0%	3	0%
Oriya	0	0%	1	0%	0	0%	0	0%	1	0%
Oromo, W.Ctr.	13	0%	11	0%	9	0%	30	0%	63	0%
Palauan	0	0%	3	0%	2	0%	2	0%	7	0%
Pampangan	1	0%	0	0%	0	0%	0	0%	1	0%
Panjabi, Eastern	23	0%	14	0%	12	0%	10	0%	59	0%
Panjabi, Western	2	0%	1	0%	0	0%	2	0%	5	0%
Pashto, Central	6	0%	5	0%	3	0%	5	0%	19	0%
Pashto, Northern	3	0%	0	0%	0	0%	0	0%	3	0%
Pashto, Southern	2	0%	1	0%	0	0%	3	0%	6	0%
Patois	3	0%	1	0%	1	0%	0	0%	5	0%
Phonpeian	1	0%	1	0%	0	0%	0	0%	2	0%
Polish	74	0%	27	0%	11	0%	15	0%	127	0%
Portuguese	36	0%	17	0%	14	0%	23	0%	90	0%
Pulaar	1	0%	3	0%	0	0%	2	0%	6	0%
Quechua, Ch.	1	0%	0	0%	0	0%	0	0%	1	0%
Romanian	14	0%	11	0%	5	0%	3	0%	33	0%
Russian	325	1%	219	1%	145	1%	165	1%	854	1%
Rwanda	1	0%	2	0%	1	0%	0	0%	4	0%
Samoan	8	0%	7	0%	3	0%	5	0%	23	0%
Saraiki	0	0%	1	0%	0	0%	0	0%	1	0%
Seminole	0	0%	0	0%	1	0%	0	0%	1	0%
Serbian	9	0%	0	0%	1	0%	0	0%	10	0%
Serbo-Croatian	12	0%	16	0%	9	0%	16	0%	53	0%
Sesotho	0	0%	1	0%	1	0%	0	0%	2	0%
Setswana	0	0%	1	0%	0	0%	0	0%	1	0%
Shona	1	0%	1	0%	1	0%	0	0%	3	0%
Sibo	2	0%	0	0%	1	0%	0	0%	3	0%
Sinhala	2	0%	0	0%	0	0%	0	0%	2	0%
Sioux	1	0%	1	0%	0	0%	0	0%	2	0%
Slovak	5	0%	0	0%	0	0%	2	0%	7	0%
Slovenian	2	0%	1	0%	1	0%	0	0%	4	0%
Somali	83	0%	68	0%	39	0%	72	1%	262	0%
Soninke	1	0%	0	0%	1	0%	0	0%	2	0%
Spanish	29,995	87%	21,149	88%	12,994	88%	10,702	84%	74,840	87%
Spokane	1	0%	2	0%	2	0%	0	0%	5	0%
Sundanese	1	0%	1	0%	1	0%	2	0%	5	0%
Susu	1	0%	2	0%	0	0%	2	0%	5	0%
Swahili	15	0%	8	0%	21	0%	20	0%	64	0%
Swedish	11	0%	8	0%	3	0%	1	0%	23	0%
Tagalog	70	0%	64	0%	37	0%	45	0%	216	0%

	Test Level									
	Grades K-2		Grades 3-5		Grades 6-8		Grades 9-12		TOTAL	
	N	%	N	%	N	%	N	%	N	%
Tahitian	1	0%	1	0%	0	0%	0	0%	2	0%
Tamil	20	0%	5	0%	2	0%	1	0%	28	0%
Telugu	32	0%	6	0%	3	0%	0	0%	41	0%
Thai	29	0%	15	0%	14	0%	8	0%	66	0%
Tibetan	5	0%	2	0%	3	0%	2	0%	12	0%
Tigrigna	25	0%	21	0%	16	0%	33	0%	95	0%
Tiwa, Northern	1	0%	0	0%	0	0%	0	0%	1	0%
Tonga	4	0%	7	0%	3	0%	2	0%	16	0%
Tongan	0	0%	0	0%	0	0%	1	0%	1	0%
Trukese	1	0%	4	0%	0	0%	2	0%	7	0%
Tsonga	0	0%	0	0%	0	0%	1	0%	1	0%
Turkish	35	0%	20	0%	12	0%	26	0%	93	0%
Twi	24	0%	12	0%	13	0%	21	0%	70	0%
Ukrainian	45	0%	35	0%	29	0%	25	0%	134	0%
Urdu	53	0%	22	0%	9	0%	16	0%	100	0%
Ute	49	0%	66	0%	58	0%	33	0%	206	0%
Uzbek	1	0%	1	0%	1	0%	1	0%	4	0%
Vengo	2	0%	1	0%	1	0%	0	0%	4	0%
Vietnamese	735	2%	412	2%	212	1%	198	2%	1,557	2%
Visayan	0	0%	1	0%	2	0%	2	0%	5	0%
Welsh	1	0%	0	0%	0	0%	0	0%	1	0%
Wolof	1	0%	1	0%	0	0%	1	0%	3	0%
Yapese	0	0%	0	0%	0	0%	1	0%	1	0%
Yoruba	8	0%	8	0%	3	0%	5	0%	24	0%
Zuni	0	0%	0	0%	1	0%	0	0%	1	0%
NonValid Codes	45	0%	20	0%	14	0%	14	0%	93	0%
Not Specified	98	0%	81	0%	35	0%	21	0%	235	0%
TOTAL	**34,408**	**100%**	**24,026**	**100%**	**14,778**	**100%**	**12,785**	**100%**	**85,997**	**100%**

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
- Oral Presentation

These accommodations are summarized, by content area and grade, in Tables 7 to 10.

Table 7. Speaking Accommodations by Grade

Grade	Speaking Accommodations Provided					Total
	None	Braille	Large Print	Signing	Assistive Com. Device	
KG	12,338	0	1	0	2	12,341
1	11,925	1	0	2	4	11,932
2	10,067	0	0	1	1	10,069
3	9,314	0	2	3	4	9,323
4	7,803	1	2	2	0	7,808
5	6,831	0	2	2	3	6,838
6	5,461	0	2	0	0	5,463
7	4,879	1	0	0	0	4,880
8	4,371	0	0	3	0	4,374
9	4,487	1	0	3	0	4,491
10	3,591	0	1	0	0	3,592
11	2,606	0	1	1	0	2,608
12	2,059	0	0	0	0	2,059
TOTAL	85,732	4	11	17	14	85,778

Table 8. Listening Accommodations by Grade

Grade	Listening Accommodations Provided					Total
	None	Braille	Large Print	Signing	Assistive Com. Device	
KG	12,334	0	1	0	2	12,337
1	11,925	1	0	2	4	11,932
2	10,068	0	0	1	0	10,069
3	9,311	0	2	4	4	9,321
4	7,794	2	2	4	1	7,803
5	6,829	0	2	2	3	6,836
6	5,460	0	3	0	0	5,463
7	4,878	1	0	0	1	4,880
8	4,369	0	0	3	0	4,372
9	4,488	1	0	3	0	4,492
10	3,591	0	1	0	0	3,592
11	2,606	0	1	1	0	2,608
12	2,059	0	0	0	0	2,059
TOTAL	85,712	5	12	20	15	85,764

Table 9. Reading Accommodations by Grade

Grade	Reading Accommodations Provided						Total
	None	Braille	Large Print	Scribe	Signing	Assistive Com. Device	
KG	12,338	0	1	2	2	1	12,344
1	11,920	0	0	9	3	4	11,936
2	10,052	0	0	15	2	0	10,069
3	9,310	0	2	7	5	4	9,328
4	7,791	1	2	10	4	0	7,808
5	6,828	0	2	7	2	3	6,842
6	5,457	0	2	3	2	0	5,464
7	4,880	1	0	4	1	0	4,886
8	4,367	0	0	1	4	0	4,372
9	4,493	1	0	1	1	0	4,496
10	3,593	0	0	1	0	0	3,594
11	2,608	0	1	0	1	0	2,610
12	2,061	0	0	0	0	0	2,061
TOTAL	85,698	3	10	60	27	12	85,810

Table 10. Writing Accommodations by Grade

Grade	Writing Accommodations Provided							Total
	None	Braille	Large Print	Scribe	Signing	Assistive Com. Device	Oral Presentation	
KG	11,007	0	1	4	2	2	1,328	12,344
1	10,772	0	0	7	2	3	1,151	11,935
2	9,743	0	0	13	2	0	310	10,068
3	9,054	0	2	17	5	6	245	9,329
4	7,580	1	2	10	4	1	208	7,806
5	6,662	0	2	10	2	3	162	6,841
6	5,340	0	1	6	2	0	116	5,465
7	4,825	0	0	7	1	0	52	4,885
8	4,310	0	0	0	4	0	55	4,369
9	4,477	1	0	2	1	0	14	4,495
10	3,577	0	0	1	0	0	16	3,594
11	2,607	0	1	0	1	0	1	2,610
12	2,056	0	0	0	0	0	5	2,061
TOTAL	82,010	2	9	77	26	15	3,663	85,802

Part 4: Test Administration

The Colorado English Language Assessment was first administered in Spring 2006. In 2007 the administration was moved to winter, and the CELA was administered to 85,997 students in January and February 2007. This test consists of four separately administered sections assessing speaking, listening, reading, and writing proficiency.

The CELA speaking section is individually administered. The listening, reading, and writing sections may be group administered or individually administered, depending upon the needs of the particular examinees being tested.

CELA 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. For all of the group-administered sections, students in the upper grades (grades 3-12) may be grouped either by grade or by grade span. 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 11, below. Actual times may vary.

Table 11. 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	15 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 while pointing to illustrations. All items are in constructed-response format, scored with performance-based rubrics that direct the rater's (generally the examiner) attention to the student's use of vocabulary, social and academic language, complex 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 rubric.

Make Conversation

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

Tell a Story

In *Tell a Story*, students produce multiple sentences explaining what is happening in a series of four 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 on a 0–4 rubric.

The Listening Subtests

The Listening test is usually administered to a group of students who listen to audio prompts. All Listening items are multiple-choice in format and measure general comprehension as well as inferential and critical thinking skills based on academic discourse. Students listen to a variety of classroom passages with a range of difficulty levels. The Listening test takes approximately 15 minutes per group to administer and consists of three subtests:

Listen for Information

In *Listen for Information*, students hear instructions typical of those provided by a classroom teacher. The instructions, read by the examiner, vary in length from one to three sentences. 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. Students hear two short exchanges typical of classroom discussions. The listening passages, questions, and text answer choices are read by the examiner. After listening, students respond to questions about what they heard. Each question has three answer choices.

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 focus on main ideas, details, inferences, and idioms. The listening passages, questions, and text answer choices are read by the examiner. 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%
		Non-Fiction	17%
3–5	200–250 words	Fiction	50%
		Non-Fiction	50%
6–8	200–250 words	Fiction	33%
		Non-Fiction	67%
9–12	225–325 words	Fiction	33%
		Non-Fiction	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 multiple-choice in 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:

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

For Grades K–5, 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 upper grades read passages without assistance and select from four written answer choices.

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

<u>The Writing Subtests</u>

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 multiple-choice and constructed-response items that assess both receptive and productive domains. In the first section, selected response items engage students to identify appropriate grammar, mechanics, and syntax, and in the second section, students respond to prompts in the form of phrases, sentences, and paragraphs.

Responses to constructed-response items are evaluated with performance-based rubrics (on a 0-3 or 0-4 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 (except 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 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 rubric.

Write in Detail

Prompts in *Write in Detail* elicit longer responses. Students in Grades 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 rubric. Students in Grades K–1 do not take *Write in Detail*.

Teacher Training

The Pre-Administration Training Workshops for 2007 were conducted in six locations in Colorado: Durango, Glenwood, Greeley, Pueblo, Aurora, and Golden. These locations were selected to cover the state's training needs geographically as well as in terms of district size. A total of 397 participants attended to the workshops. Table 12 shows the breakdown of attendees per workshop location.

Table 12. Number of Attendees at Pre-Administration Training Workshops.

Durango 11/9	32
Glenwood Springs 11/10	45
Greeley 11/13	84
Pueblo 11/14	115
Aurora 11/15	63
Golden/Lakewood 11/16	58
Total attendees	397

Workshop Set up

The environment of the Pre-Administration Workshop was friendly and facilitated small-group discussion. Participants were assigned tables, and table leaders were chosen according to experience with CELA, job title, and the size/location of their district/school site.

Training Materials Development

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

Training Manual
The CELA Pre-Administration Training Manual consists of a table of contents that corresponds directly to the organization of the materials. This allows for easy navigation through the training manual. 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 Audio CD
Another important part of the training materials is the coordination between the audio component and the training manual. Because the Speaking test is scored by Test Examiners during test administration, the audio component is critical for training. There are two audio CDs that provide student sample responses for all grade spans, organized as follows: K-5 and 6-12. This format allows the trainer to facilitate inter-rater reliability and to give each participant the opportunity to score items in a range of grade spans. All samples were scored by CTB experts and teachers. Participants use Scoring Sheets as part of calibration exercises.

Part 5: Scoring

The 2007 CELA tests were scored and processed by CTB's scoring team using the standardized methods and procedures previously developed for the *LAS Links* program. The CELA scoring team consists of trained technical specialists who are responsible for coordinating all scoring and reporting activities related to the processing of CELA 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 on-line, the error suspect list concentrates on the most likely problems based on pre-defined 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. On-line editing screens focus an editor on potential problems and then 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 each error is actually corrected. As batches are extracted for scoring, a final edit is performed to ensure 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 CELA assessments, CTB's imaging handscoring system presents images of scanned test books to trained readers, who assign scores for constructed response items. Scanned images are viewed on high quality 19″ 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 BAs 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 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 in reader readiness. All readers, including team leaders, must achieve 80 percent 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 training who successfully validate are qualified as readers and could score tests. Team leaders are required to complete two validation rounds with 80 percent 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. Readers whose scoring is inconsistent are read behind more frequently by their team leaders. Thus, any scoring variation is corrected.

Inter-rater Reliability

Intraclass correlation coefficients and weighted Kappa coefficients were calculated to measure reader agreement (Fleiss & Cohen, 1973) for each of the hand-scored CELA items,[1,2] using scores assigned to all item responses that received second reads. The intraclass correlation coefficients were consistently high, ranging from .87 to .99, with 80 percent of the coefficients greater than or equal to .90. The weighted Kappa values also were high[3] for all items, indicating good agreement between the first and second readers. Inter-rater agreement statistics for all of the hand-scored items are shown in Table 13.

The percentage of discrepant scores was 5% or less for all items in the upper three grade spans. Within the first grade span, the percentage of discrepant scores reached a maximum of 7% for those items that were administered to students in Kindergarten and Grade 1, but was less than 5% for all other items.

It should be noted that the percentages of agreement and of discrepant scores are somewhat misleading, especially for the ten items in the K-2 grade span. This is because omitted items are treated as "special codes." Within this grade span, items 21-25 were administered only to students in Kindergarten and Grade 1 and omitted by students in Grade 2, while items 26-30 were administered only to students in Grade 2 and omitted by students in Kindergarten and Grade 1.

Therefore, the adjusted percentage of perfect + adjacent agreement is also provided in Table 13. This percentage reflects the agreement for all responses that were not assigned special codes, and is therefore a more accurate reflection of the actual agreement among ratings of scorable responses.

[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 smaller than the intraclass correlations.

[3] Kappa values between 0.40 and 0.74 represent good agreement beyond chance, and values below 0.40 indicate poor agreement.

Table 13. Inter-Rater Agreement for CELA Writing Responses.

Span	Item	Max Score	% Perfect Agreemnt	% Adjacent Scores	% Special Codes	% Discrepnt (>1 point)	Adjusted* % Perfect +Adjacent	Intraclass Correlation	Wtd Kappa
K-2	21	3	34%	7%	51%	7%	**86%**	**0.95**	**0.90**
	22	3	33%	6%	54%	6%	**87%**	**0.96**	**0.92**
	23	3	34%	7%	52%	7%	**85%**	**0.96**	**0.91**
	24	3	33%	6%	53%	6%	**87%**	**0.96**	**0.91**
	25	3	32%	6%	55%	6%	**87%**	**0.96**	**0.92**
	26	3	21%	5%	73%	1%	**96%**	**0.98**	**0.97**
	27	3	21%	5%	73%	1%	**96%**	**0.98**	**0.97**
	28	3	20%	6%	73%	2%	**93%**	**0.98**	**0.95**
	29	3	17%	7%	73%	3%	**89%**	**0.96**	**0.93**
	30	4	19%	7%	73%	1%	**96%**	**0.99**	**0.97**
3-5	21	3	71%	22%	3%	3%	**97%**	**0.90**	**0.80**
	22	3	75%	19%	4%	2%	**98%**	**0.91**	**0.82**
	23	3	67%	24%	4%	4%	**96%**	**0.89**	**0.77**
	24	3	67%	23%	5%	5%	**95%**	**0.89**	**0.79**
	25	4	56%	33%	7%	5%	**95%**	**0.89**	**0.78**
6-8	21	3	74%	19%	4%	2%	**98%**	**0.91**	**0.81**
	22	3	77%	16%	5%	2%	**98%**	**0.92**	**0.83**
	23	3	65%	26%	5%	4%	**96%**	**0.87**	**0.75**
	24	3	65%	26%	5%	4%	**96%**	**0.88**	**0.77**
	25	4	60%	30%	7%	3%	**97%**	**0.91**	**0.82**
9-12	21	3	76%	15%	8%	1%	**99%**	**0.94**	**0.89**
	22	3	76%	14%	8%	2%	**98%**	**0.94**	**0.89**
	23	3	73%	17%	10%	1%	**99%**	**0.94**	**0.88**
	24	3	74%	15%	10%	1%	**99%**	**0.95**	**0.89**
	25	4	71%	19%	10%	<1%	**>99%**	**0.95**	**0.89**

* Adjusted % Perfect+Adjacent Agreement is computed after excluding responses that were assigned special codes.

Scoring and Technology Quality Control Procedures

The Technology and Scoring Departments at CTB both have quality assurance sections 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 undergo 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

As noted previously in this report, the CELA test forms for grade spans 6-8 and 9-12 are identical to the *LAS Links Form A* tests. Therefore, these tests were scored using the standard LAS Links scoring tables. For the new grade spans K-2 and 3-5, items were recalibrated, and new scoring tables were created using the methods and procedures discussed later in this section of the report.

Test and item information computed from the LAS Links standardization sample is available in the *LAS Links Technical Manual* (CTB, 2006). To supplement that information, CTB conducted new test and item analyses for the Colorado examinee population for these forms as well as for the reconfigured tests that were administered to students in grades K-2 and 3-5.

This section of the technical report contains a description of the calibration and equating procedures and results, along with details of the classical item analysis and differential item functioning analysis that was conducted for each test. This section also includes a subsection describing student performance on the 2007, along with comparisons of the 2007 and 2006 results.

IRT Item Calibration

As noted previously, the reconfiguration of the tests for grades K-2 and 3-5 made it necessary to recalibrate and equate these new test forms. The calibration and equating followed the procedures described below. In addition, the tests for grade spans 6-8 and 9-12 were scored using the standard *LAS Links* scoring tables, these tests were also recalibrated and equated for comparative purposes only.

Student item responses on each of the CELA assessments were calibrated using the three-parameter logistic model (3PL) to scale the selected response (SR) items, and the two-parameter partial credit (2PPC) model to scale the constructed response (CR) items. A brief explanation of the models is provided below.

The 3PL model (Lord & Novick, 1968; Lord, 1980) defines a selected response 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-scoring student.

The 2PPC model defines a constructed response 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

$$Z_{jk} = A_{jk}\,\theta + C_{jk},$$

$$A_{jk} = \alpha_j(k-1),\ k = 1, 2,\ldots m_j \text{, and}$$

$$C_{jk} = -\sum_{i=0}^{k-1} \gamma_{ji}\text{, where } \gamma_{j0} = 0,$$

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*, 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 are estimated.

All of the 2007 CELA assessments were recalibrated using the 3PL/2PPC models described above. Separate calibrations were conducted for Listening, Speaking, Reading, Writing, Comprehension, and Oral scales in each grade span.

Equating and Scaling

The recalibrated tests for the new grade spans K-2 and 3-5 were placed on the existing CELA/*LAS Links* scale through a Stocking and Lord (1983) characteristic curve equating procedure. The original *LAS Links* item parameters were used as equating anchors in this procedure.

The new M1 and M2 conversion parameters were computed as follows:

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

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

where

$M1_{New}$ and $M2_{New}$ are the new transformation constants calculated to place the new field test items onto the *LAS Links* scale,

$M1_{Old}$ 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 follows:

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

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

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 repeated for the 6-8 and 9-12 grade spans for research purposes only, as these grade spans were scored using the original *LAS Links* scoring tables.

The equated results were used to create new raw-to-scale score tables for each of the six content areas (Reading, Writing, Listening, Speaking, Oral, and Comprehension) for Grade Spans K-2 and 3-5. 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 new scoring tables for Grade Spans K-2 and 3-5 are included in Appendix E. The tests for Grade Spans 6-8 and 9-12 were scored using the standard *LAS Links* scoring tables, which can be found in the *LAS Links Technical Manual.*

Results of the Calibration and Equating

Tables B1 to B20 and Figures B1 to B40 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 Listening, Speaking, Reading, and Writing. In these figures, the original parameters are the 2006 CELA item parameters, and the equated parameters are the new CELA 2007 parameters. The item-by-item differences between the CELA and *LAS Links* parameters are shown in Tables B1 to B25 in that appendix. The "a" differences are the differences between the log of "a" values. With one notable exception (Item 15 in the Grade 6-8 Writing), the 2007 CELA parameters are very similar to the original item parameters, suggesting that the tests are functioning consistently across different years and student populations, and across the reconfigured grade spans.

Figures C1 to C12 in Appendix C show the CELA test characteristic curves (TCCs) and the conditional standard errors of measurement (CSEMs) for each grade span and content domain. A comparison of these figures with the corresponding TCCs and CSEMs in the *LAS Links Technical Manual* is further evidence that the tests are generally behaving similarly. For a vertically scaled test such as the CELA/*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., with tests increasing in difficulty as grade level increases). With the exception of the Speaking and Oral scales (where the TCCs are very close together), the TCCs show this expected pattern.

The correlations between the equated and input anchor item parameters are shown in Table 14. For selected response scales, these represent the correlations of the a and b parameters. For constructed response items, these represent the alpha and gamma correlations, respectively.

Table 14. Stocking and Lord Parameter Correlations

	Grade Span K-2			Grade Span 3-5		
	P	Discrimination	Location	P	Discrimination	Location
Speaking	0.91	0.51	0.95	0.97	0.64	0.97
Listening	0.91	0.80	0.83	0.97	0.93	0.97
Reading	0.97	0.73	0.98	0.94	0.75	0.92
Writing	0.83	0.66	0.91	0.75	0.82	0.82
Oral	0.73	0.78	0.83	0.95	0.74	0.97
Comprehension	0.84	0.62	0.80	0.96	0.79	0.95

For Speaking, Listening, and Reading, the P-value correlations are all greater than .90 in both grade spans. This is also the case for Oral and Comprehension in Grade Span 3-5. Correlations are notably lower for the K-2 Writing, Oral, and Comprehension scales and for the 3-5 Writing scale.

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 for the equated total test. With the exception of the K-2 Reading, K-2 Writing, and K-2 Comprehension scales, every item in the test was also used as an anchor item, so the total test and the anchor are indistinguishable in most of these plots.

Item Analysis

Classical item analysis statistics were computed for the 2007 CELA 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, item *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 SR 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 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 SR 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 inflate the correlation artificially.

Item Difficulty Statistics (p-values)

The 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 selected-response items, the p-value is the proportion of students answering the item correctly. For constructed response 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 statistics for individual items at each grade span are provided in the item analysis tables in Appendix A. The p-values in Appendix A are above .20 except for one item in Kindergarten Reading and one item in K-1 Writing which are both .17, and most are in the desired difficulty range between .30 and .90.

The range of p-values varies by grade span and content domain. Across grade spans, the p-values range from .33 to .97 for Listening; .23 to .97 for Speaking; .17 to .98 for Reading; .17 to .93 for Writing; .22 to .98 for Comprehension; and .23 to .97 for Oral. Within grade spans, p-values range from .17 to .98 in Grade Span K-2; from .35 to .98 in Grade Span 3-5; from .25 to .93 in Grade Span 6-8; and from .36 to .92 in Grade Span 9-12.

Average item difficulty for each content area, grade and grade span is summarized in Table 15, below. In this table, item difficulty is expressed in terms of p-values. For selected-response items, the p-value is the proportion of students answering the item correctly. For constructed response 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).

Table 15. Mean P-Values by Grade Span and by Grade

	Speaking	Listening	Reading	Writing	Oral	Compre-hension
Grade Span	0.63	0.66	.	.	0.65	.
K	0.49	0.47	0.46	0.26	0.48	0.48
1	0.66	0.70	0.60	0.47	0.68	0.62
2	0.76	0.83	0.77	0.65	0.80	0.78
Grade Span	0.77	0.70	0.63	0.72	0.74	0.66
3	0.73	0.65	0.55	0.66	0.69	0.59
4	0.78	0.71	0.65	0.74	0.75	0.68
5	0.81	0.77	0.72	0.79	0.79	0.74
Grade Span	0.76	0.77	0.65	0.74	0.77	0.71
6	0.75	0.76	0.61	0.73	0.76	0.68
7	0.76	0.78	0.66	0.75	0.77	0.71
8	0.75	0.79	0.68	0.76	0.77	0.74
Grade Span	0.77	0.74	0.62	0.74	0.76	0.68
9	0.76	0.73	0.58	0.73	0.75	0.64
10	0.78	0.75	0.62	0.74	0.77	0.68
11	0.79	0.77	0.65	0.75	0.78	0.70
12	0.79	0.76	0.67	0.75	0.78	0.71

Item-Total Correlations

An important indicator of item quality is the correlation of scores on that item with scores on the total test. These item total correlations (point biserial correlation coefficients) are summarized below in Table 16. 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 Listening item was correlated with the total Listening score minus the score on the item in question, 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 Reading, Writing, Oral, and Comprehension scales.

Individual item-total correlations for each content area and grade span are provided in the item analysis tables in Appendix A. Across all grade spans, item-total correlations for the Listening items range from .23 to .63. Item-total correlations for Speaking range from .28 to .85. For Reading, the correlations range from .15 to .61, with two items below .20, and for Writing the correlations range from .08 to .79 with three items below .20. Comprehension item-total correlations range from .14 to .59, with three items below .20 and Oral item-total correlations range from .09 to .83 with 5 items below .20.

Across all grade spans and content domains, there are only two item-total correlation coefficients below .10: These are item #15 in Grade Span 6-8 Writing (r = .08) and item #1 in Grade Span 6-8 Oral (r = .09).

The average (mean) item-total correlation coefficients for each content area, grade span and grade are shown in Table 16. The average item-total correlation coefficients ranged from 0.57

to 0.69 for Speaking, from .36 to .52 for Listening, .38 to .45 for Reading, .38 to .50 for Writing, .41 to .54 for Oral, and .36 to .43 for Comprehension.

Table 16. Average Item-Total Correlations by Grade Span and Grade.

Grade	Speaking	Listening	Reading	Writing	Oral	Compre-hension
Grade Span 1	0.67	0.52	.	.	0.54	.
K	0.66	0.42	0.42	0.38	0.47	0.39
1	0.62	0.44	0.38	0.43	0.47	0.36
2	0.61	0.44	0.42	0.49	0.46	0.39
Grade Span 2	0.58	0.39	0.45	0.50	0.43	0.40
3	0.57	0.36	0.41	0.48	0.41	0.36
4	0.57	0.38	0.44	0.48	0.42	0.38
5	0.59	0.41	0.45	0.49	0.45	0.41
Grade Span 3	0.64	0.44	0.42	0.45	0.49	0.41
6	0.61	0.42	0.40	0.45	0.46	0.38
7	0.65	0.44	0.42	0.46	0.50	0.41
8	0.66	0.46	0.42	0.43	0.51	0.42
Grade Span 4	0.67	0.45	0.42	0.48	0.52	0.42
9	0.69	0.44	0.41	0.48	0.52	0.41
10	0.68	0.46	0.42	0.49	0.53	0.43
11	0.65	0.45	0.42	0.47	0.51	0.42
12	0.62	0.44	0.41	0.47	0.48	0.42

Item Omit Rates

The item analysis tables in Appendix A also show the rate at which students omit 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 Span 9-12 were below 5 percent for all of the content areas. For the 6-8 grade span, two Speaking items had omit rates between 5 and 6 percent, but omit rates were below 5 percent for all of the other items in all content areas. For Grade Span 3-5, two Reading items had omit rates of 13.29 percent and 20.31 percent, but omit rates were below 5 percent for all of the other items in all of the content areas.

Omit rates were generally higher for Grade Span K-2. Omit rates were between 2.16 and 5.85 percent for all of the Listening items, with six items above 5 percent. For the Reading items, omit rates were above 5 percent for eleven items in grades 1-2 and for all but three of the items administered to Kindergarten students. Speaking K-2 had five items above 5 percent. Highest omit rates were for the K-1 Writing items, with omit rates ranging from 2.27 percent to 22.05 percent, and with all but one item above 5 percent. However, omit rates for the Writing items were all below 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 Winter 2007 CELA administration. For the CELA analyses, a separate IRT calibration and separate DIF analysis was conducted for each grade span and language domain (Listening, Speaking, Reading, Writing, Oral, and Comprehension). To calculate DIF for the CELA 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:

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

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 upon 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,

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

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 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 for a gender group and 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 & 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 gender 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 constructed-response items.

The results of the DIF analyses are shown in Table 17. Overall, very few items exhibited differential item functioning by gender or ethnicity. Across all grades and content areas, one item was flagged for DIF against males and no items were flagged for DIF against females.

Across all grades and content areas, one item was flagged for DIF against Hispanic students; a total of 8 items were flagged for DIF in favor of Black examinees, and 18 items were flagged for DIF against Black examinees.

Table 17. Number of Items Exhibiting Differential Item Functioning.

Subject	Grade Span	Male		Female		Hispanic		Black	
		For	Against	For	Against	For	Against	For	Against
Listening	K-2	0	0	0	0	0	0	1	0
	3-5	0	0	0	0	0	0	0	0
	6-8	0	0	0	0	0	0	0	0
	9-12	0	0	0	0	0	0	0	0
Speaking	K-2	0	0	0	0	0	0	0	0
	3-5	0	0	0	0	0	0	0	0
	6-8	0	0	0	0	0	0	1	4
	9-12	0	0	0	0	0	0	3	5
Reading	K	0	0	0	0	0	0	0	0
	1-2	0	1	0	0	0	1	0	1
	3-5	0	0	0	0	0	0	0	3
	6-8	0	0	0	0	0	0	1	1
	9-12	0	0	0	0	0	0	1	2
Writing	K-1	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0
	3-5	0	0	0	0	0	0	0	0
	6-8	0	0	0	0	0	0	1	0
	9-12	0	0	0	0	0	0	0	2

In the present case, it should be noted that the observed DIF for Black students may be a consequence of the fact that the representation of native language groups among Black students is very different from the distribution of languages in the total population. For example, more than 87% of the total CELA population but fewer than 9% of the Black examinees indicated that their first language is Spanish. Fewer than 2% of the total CELA population but more than 54% of the Black examinees indicated that their first language is Amharic, Somali, Arabic, French, or Tigrigna. Given the different linguistic properties of these languages, it is not surprising to find that the relationship between performance on individual items and overall test performance may vary between groups. That is, there is a near confounding of race and native language, and the DIF statistic is not designed to untangle it.

All items flagged for DIF will be 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.

Student Performance on the 2007 CELA

This section of the report will summarize the performance of students on the 2007 CELA. Results are presented for the total population and for various subgroups of interest. In addition, results will be compared with performance on the 2006 CELA. 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 2007 CELA are provided in Table 18.

Table 18. 2007 CELA Lowest and Highest Obtainable Scale Scores

						Comp	Oral	
		Speaking	Listening	Reading	Writing	(R+L)	(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. LOSSes and HOSSes that have changed this year because of the new grade span configurations are shown in ***Bold Italic.***

Table 19 shows the 2007 total scale score means and standard deviations by grade span, and Table 20 shows the results for each individual grade in 2006 and 2007.

Table 19. 2007 Total Scale Score Means and Standard Deviations by Grade Span.

	N	Mean	SD
Grade Span 1	31,368	429.73	57.75
Grade Span 2	23,486	512.32	45.31
Grade Span 3	14,241	536.32	45.28
Grade Span 4	11,625	537.98	47.53

Table 20. 2006 and 2007 Total Scale Score Means and Standard Deviations by Grade.

	2006			2007		
	N	Mean	SD	N	Mean	SD
KG	10,548	388.94	42.61	10,063	376.52	39.17
1	11,194	449.04	44.22	11,479	434.86	42.45
2	10,371	485.14	40.86	9,826	478.22	41.09
3	9,232	509.41	38.07	9,094	495.02	43.22
4	8,252	520.26	46.8	7,647	515.89	41.65
5	7,231	531.75	46.03	6,745	531.61	43.31
6	6,142	535.54	45.37	5,307	530.51	42.75
7	5,289	536.65	47.78	4,730	538.45	45.95
8	4,604	542.63	48.68	4,204	541.26	46.83
9	4,277	531.44	48.99	4,121	531.74	48.36
10	3,086	537.23	48.24	3,333	538.85	48.52
11	2,320	541.93	44.93	2,360	543.72	46.21
12	1,748	546.63	41.91	1,811	543.08	43.69

The 2007 total scale scores were lower than the 2006 scores in 9 of the 13 grades (K through 6, 8, and 12), and were higher than the 2006 scores in the remaining four grades (7, 9, 10, and 11). The greatest decline in scores occurred in Grades 1 and 3, and the greatest increase occurred in Grades 7 and 11.

The 2007 performance on the six component scales of Speaking, Listening, Reading, Writing, Comprehension, and Oral Proficiency is summarized by grade and by grade span in Table 21 and by grade and gender in Table 22.

Table 21. CELA 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	33,887	476.81	48.89	33,547	456.65	49.06	33,471	397.45	71.00	31,800	375.64	107.00	33,208	431.46	57.15	33,338	472.47	42.58
KG	12,156	453.25	50.74	11,845	420.88	39.64	11,789	339.69	54.69	10,306	277.00	79.30	11,597	386.76	47.29	11,752	446.43	38.84
1	11,779	482.01	42.33	11,758	462.13	38.70	11,741	404.72	48.38	11,594	388.74	82.26	11,698	436.13	40.70	11,694	476.81	34.56
2	9,952	499.41	40.66	9,944	492.79	40.13	9,941	457.35	55.27	9,900	462.98	65.95	9,913	478.26	42.62	9,892	498.26	37.60
Grade Span 2	23,744	525.11	47.18	23,716	507.30	47.61	23,654	499.86	64.40	23,646	516.70	61.22	23,611	504.95	46.90	23,641	518.95	41.25
3	9,232	514.08	44.52	9,207	490.86	42.92	9,171	476.71	65.77	9,165	497.87	61.63	9,147	486.99	43.27	9,182	506.24	36.57
4	7,717	527.12	45.73	7,711	509.43	44.87	7,697	505.00	59.29	7,692	521.74	57.25	7,685	507.96	43.40	7,686	520.74	39.09
5	6,795	537.79	48.79	6,798	527.15	48.61	6,786	525.30	56.76	6,789	536.41	57.61	6,779	525.78	46.07	6,773	534.16	44.02
Grade Span 3	14,385	537.58	57.41	14,387	535.48	52.03	14,396	534.76	53.03	14,362	538.02	55.06	14,357	533.87	50.69	14,290	536.33	56.88
6	5,348	535.02	53.71	5,351	528.93	48.20	5,350	524.85	51.86	5,346	534.02	53.80	5,341	525.51	48.05	5,323	531.64	51.72
7	4,789	539.29	58.62	4,783	538.16	52.51	4,795	536.83	52.84	4,781	539.70	56.52	4,771	535.97	50.65	4,747	538.75	57.99
8	4,248	538.87	60.36	4,253	540.71	55.19	4,251	544.89	52.53	4,235	541.19	54.67	4,245	542.02	52.36	4,220	539.54	61.28
Grade Span 4	11,928	533.06	60.44	11,883	537.97	59.89	11,961	544.99	54.29	11,911	536.25	50.65	11,837	545.18	54.65	11,691	529.31	58.34
9	4,226	528.82	62.19	4,223	531.25	58.99	4,250	535.05	56.11	4,226	531.88	51.82	4,207	536.20	55.30	4,148	523.64	58.94
10	3,407	534.34	62.27	3,396	538.75	60.58	3,421	544.91	54.57	3,404	537.30	51.18	3,389	545.43	55.27	3,349	530.40	59.89
11	2,426	537.94	58.35	2,412	545.01	59.74	2,418	552.95	51.40	2,414	539.89	50.12	2,399	553.20	52.33	2,371	535.49	57.64
12	1,869	533.99	54.86	1,852	542.69	59.30	1,872	557.40	48.74	1,867	539.49	46.83	1,842	554.78	51.63	1,823	532.16	53.68

Table 22. CELA Scale Score Means and Standard Deviations by Grade and Gender.

		Speaking			Listening			Reading			Writing			Comprehension			Oral			Total		
		N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
KG	F	5,934	453.73	52.43	5,783	422.79	38.92	5,768	342.20	54.87	5,097	279.70	80.50	5,676	388.95	46.86	5,738	447.38	39.73	4,982	378.34	39.82
	M	6,221	452.82	49.03	6,061	419.05	40.23	6,020	337.28	54.42	5,208	274.35	78.03	5,920	384.65	47.61	6,013	445.54	37.95	5,080	374.75	38.44
1	F	5,822	482.62	44.40	5,813	463.56	38.12	5,812	407.21	48.63	5,745	396.22	81.09	5,791	437.86	40.55	5,782	477.86	35.63	5,688	437.73	43.12
	M	5,954	481.45	40.14	5,942	460.74	39.21	5,926	402.31	47.99	5,846	381.44	82.72	5,904	434.45	40.78	5,909	475.80	33.45	5,788	432.07	41.56
2	F	4,722	500.40	42.10	4,720	492.08	39.09	4,722	460.09	53.97	4,706	470.71	61.35	4,707	479.45	41.56	4,695	498.98	38.34	4,666	480.83	40.12
	M	5,229	498.50	39.28	5,223	493.42	41.05	5,218	454.88	56.30	5,193	455.98	69.13	5,205	477.18	43.53	5,196	497.60	36.89	5,159	475.85	41.80
3	F	4,485	513.21	44.93	4,479	489.77	41.66	4,470	478.67	64.98	4,462	503.74	60.30	4,456	487.49	43.36	4,464	505.22	36.64	4,425	496.40	42.94
	M	4,742	514.89	44.14	4,723	491.87	44.04	4,696	474.81	66.49	4,698	492.27	62.40	4,686	486.48	43.19	4,713	507.18	36.49	4,664	493.67	43.45
4	F	3,748	526.59	46.40	3,741	508.62	44.05	3,736	508.14	57.65	3,735	528.17	56.50	3,728	509.44	42.50	3,732	520.04	39.14	3,715	518.00	41.24
	M	3,968	527.62	45.10	3,969	510.18	45.64	3,960	502.00	60.64	3,956	515.66	57.29	3,956	506.56	44.19	3,953	521.40	39.05	3,931	513.90	41.95
5	F	3,239	537.80	49.45	3,245	524.98	48.36	3,241	528.30	54.33	3,242	542.57	57.17	3,236	526.63	45.33	3,229	533.19	43.90	3,216	533.45	42.83
	M	3,554	537.77	48.19	3,551	529.08	48.72	3,543	522.50	58.71	3,545	530.72	57.38	3,541	524.93	46.66	3,542	535.00	44.09	3,527	529.88	43.63
6	F	2,419	532.85	54.80	2,418	531.14	48.24	2,416	528.69	50.43	2,416	539.55	53.37	2,412	528.18	47.36	2,406	530.55	53.04	2,399	532.71	43.09
	M	2,929	536.81	52.73	2,933	527.11	48.10	2,934	521.70	52.81	2,930	529.46	53.74	2,929	523.31	48.52	2,917	532.54	50.60	2,908	528.70	42.38
7	F	2,249	536.95	59.51	2,246	541.55	52.62	2,249	541.32	51.31	2,245	544.89	54.60	2,238	539.96	49.62	2,231	538.08	59.59	2,221	541.15	45.71
	M	2,540	541.36	57.76	2,537	535.16	52.24	2,546	532.87	53.85	2,536	535.11	57.78	2,533	532.45	51.30	2,516	539.34	56.55	2,509	536.06	46.04
8	F	1,926	536.78	59.89	1,925	544.00	55.78	1,924	548.12	52.63	1,917	545.04	54.41	1,923	544.98	52.28	1,911	539.54	61.05	1,904	543.35	47.05
	M	2,322	540.60	60.71	2,328	537.99	54.56	2,327	542.22	52.32	2,318	538.00	54.69	2,322	539.56	52.30	2,309	539.54	61.48	2,300	539.53	46.59
9	F	1,914	520.24	57.18	1,902	531.06	57.88	1,916	535.81	53.99	1,906	534.84	50.73	1,896	536.44	53.32	1,877	518.44	56.94	1,868	530.32	46.78
	M	2,311	535.91	65.22	2,320	531.40	59.91	2,333	534.43	57.81	2,319	529.44	52.61	2,310	536.01	56.90	2,270	527.92	60.23	2,252	532.90	49.61
10	F	1,630	527.68	60.63	1,616	540.43	59.87	1,629	545.85	54.51	1,627	539.94	52.27	1,614	546.58	55.32	1,600	526.99	59.66	1,596	538.30	49.17
	M	1,776	540.45	63.15	1,779	537.18	61.18	1,791	544.05	54.65	1,776	534.88	50.08	1,774	544.36	55.22	1,748	533.52	59.96	1,736	539.35	47.93
11	F	1,139	528.40	56.73	1,139	543.10	60.52	1,139	552.11	51.38	1,133	540.86	52.77	1,132	551.80	53.45	1,117	528.28	57.53	1,110	540.84	47.48
	M	1,287	546.37	58.48	1,273	546.72	59.00	1,279	553.70	51.43	1,281	539.04	47.66	1,267	554.45	51.29	1,254	541.91	56.99	1,250	546.29	44.92
12	F	872	529.44	53.53	862	542.58	56.83	877	558.97	48.06	877	541.63	47.60	857	555.95	49.25	848	528.68	52.60	843	542.49	43.48
	M	996	537.92	55.71	989	542.71	61.38	994	555.95	49.30	989	537.57	46.09	984	553.69	53.60	974	535.10	54.42	967	543.54	43.87

Overall, female students tended to score somewhat higher than male students from Kindergarten through Grade 8, with males scoring somewhat higher than females in Grades 9 through 12. The greatest gender differences were observed in Speaking and Writing. 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 elementary school years where the female score advantage ranged from 10 points to more than 14 points, with smaller differences observed at higher grade levels. Male students, on the other hand, tended to score substantially higher than females on the Speaking in Grades 9 through 12. The difference in mean Speaking scores was highest in Grade 11, where the mean score for male students was almost 18 points higher than the mean for female students. These results are displayed graphically in Figures 1 through 7.

Figure 1. Mean Speaking Scale Scores by Grade and Gender

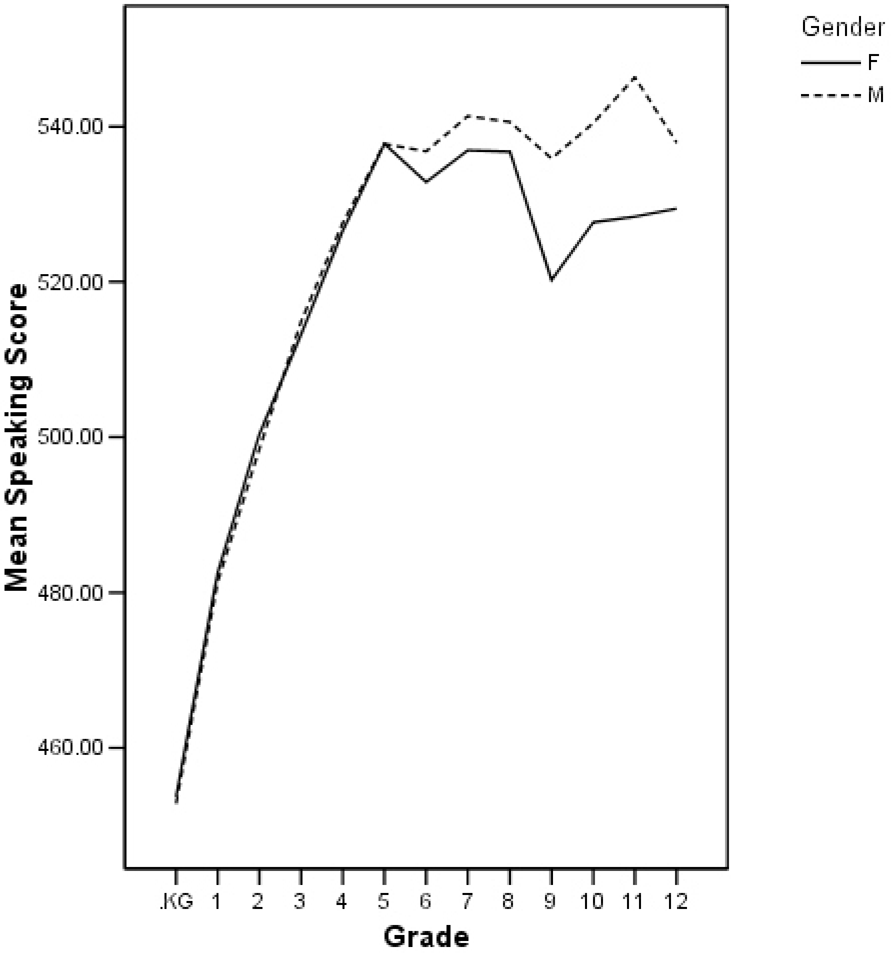


Figure 2. Mean Listening Scale Scores by Grade and Gender

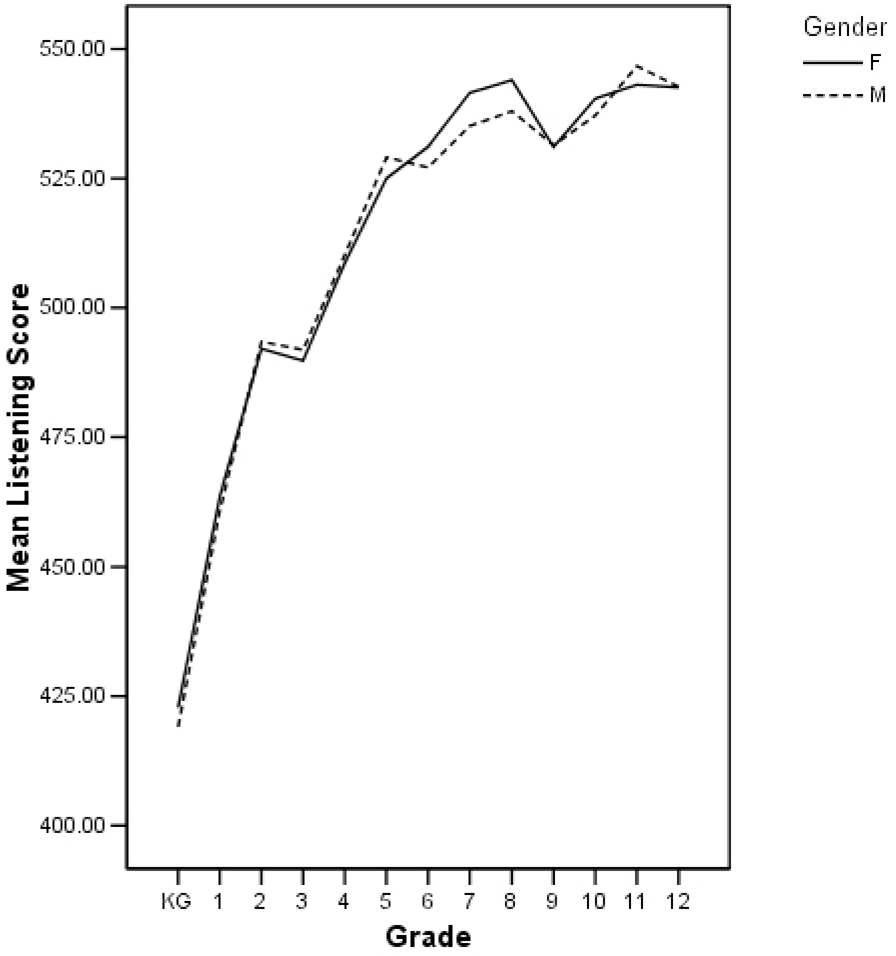


Figure 3. Mean Reading Scale Scores by Grade and Gender

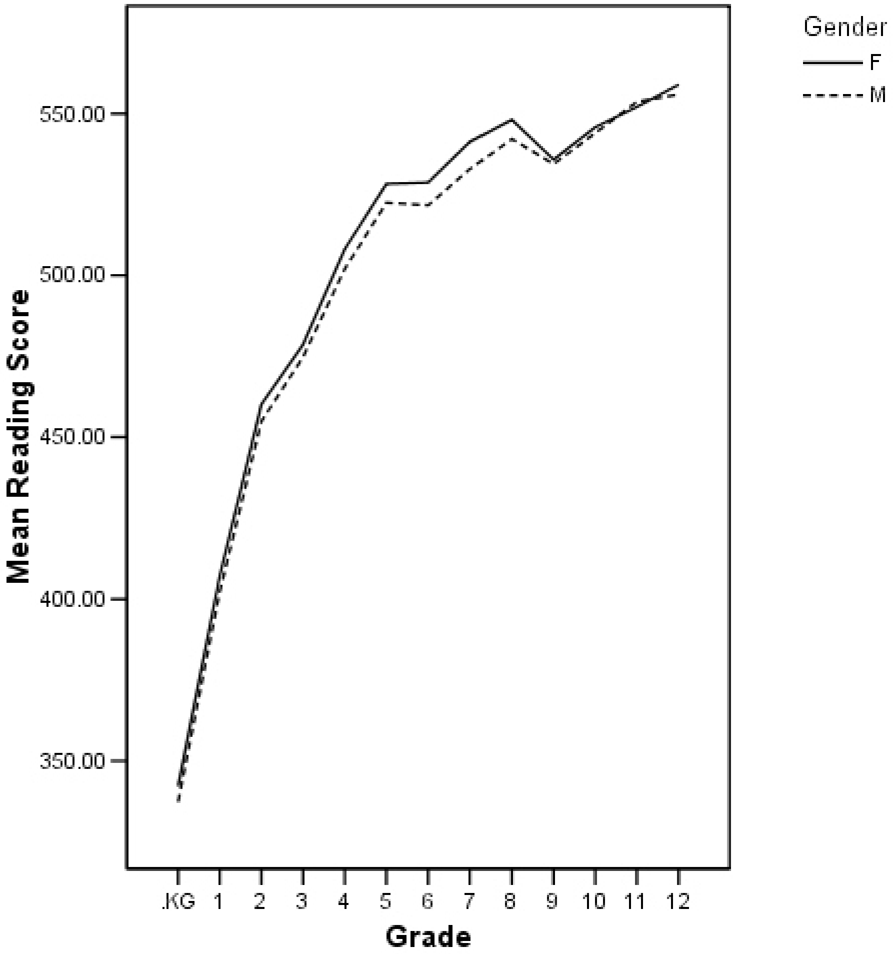


Figure 4. Mean Writing Scale Scores by Grade and Gender

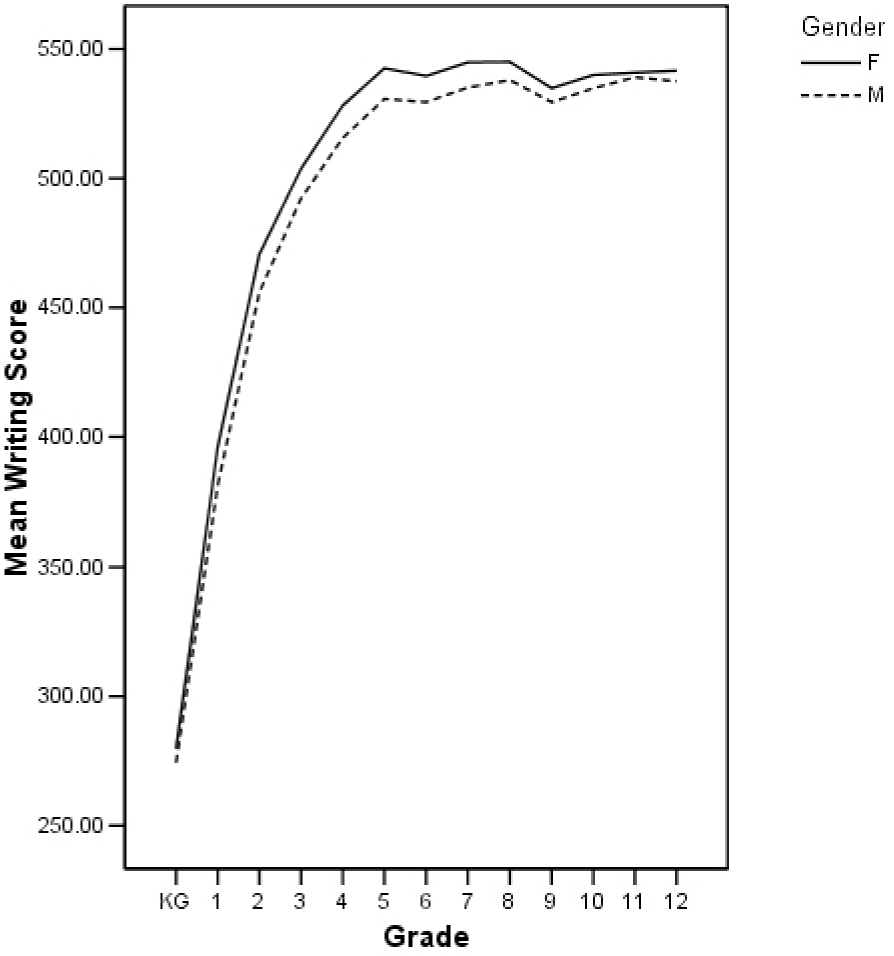
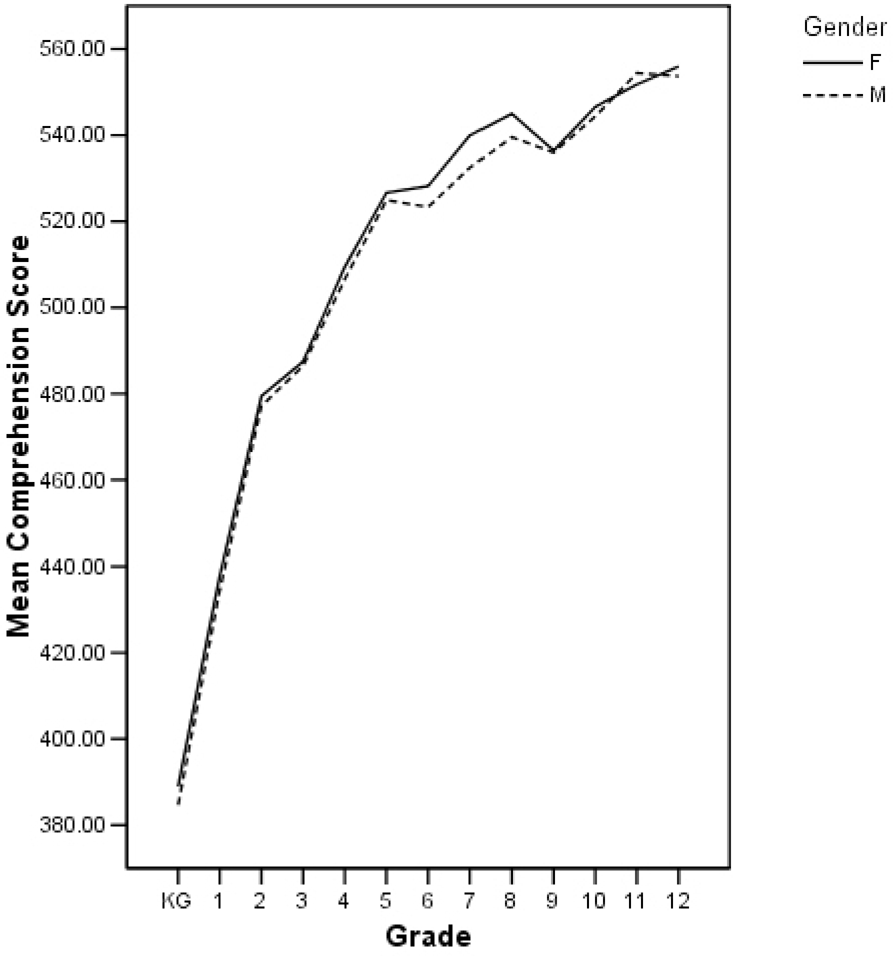


Figure 5. Mean Comprehension Scale Scores by Grade and Gender



Copyright © 2007 by the Colorado Department of Education.

Figure 6. Mean Oral Scale Scores by Grade and Gender

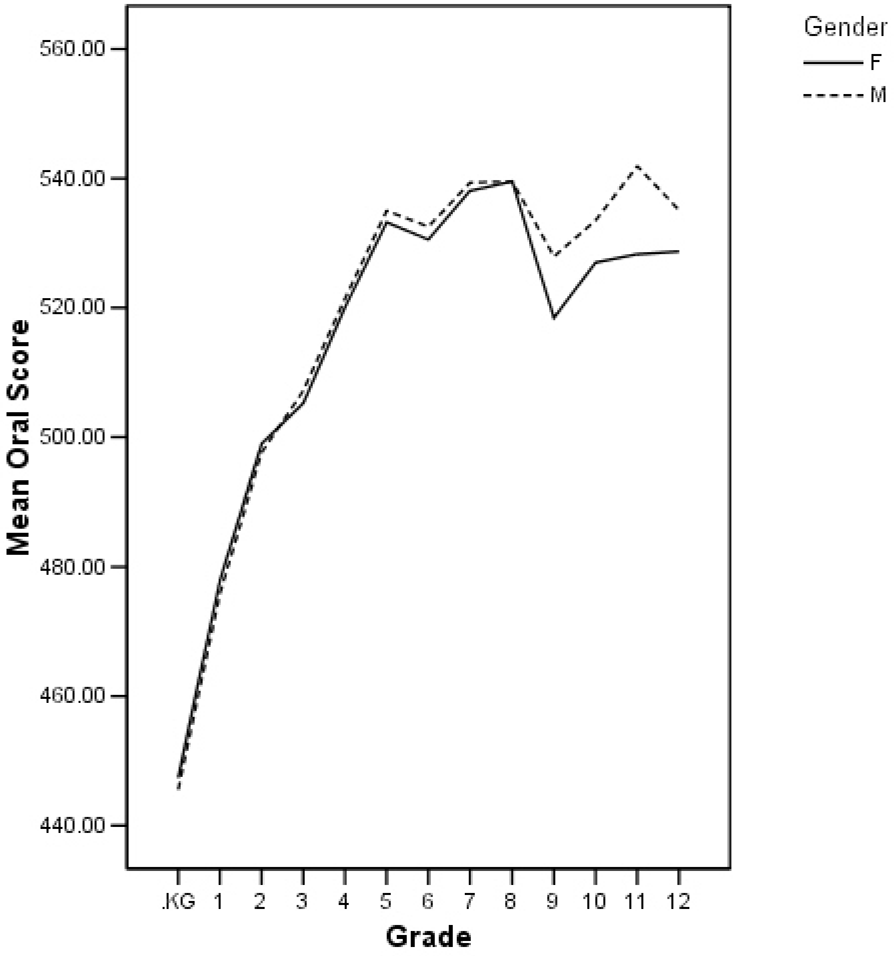
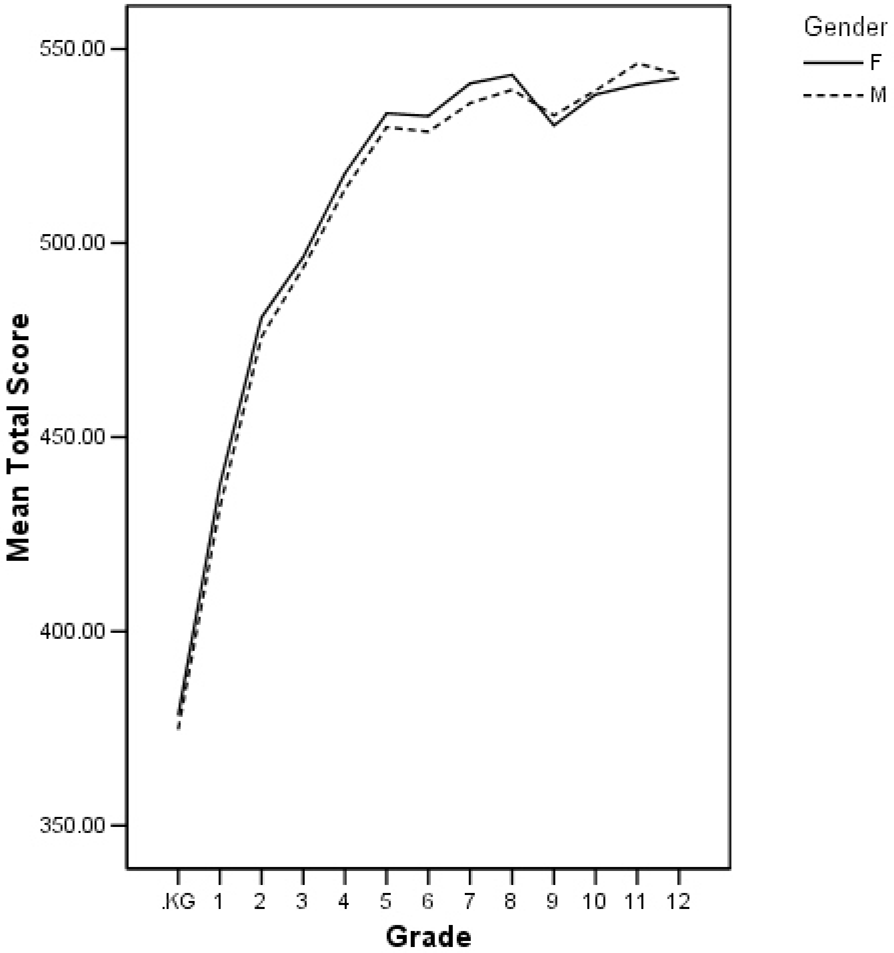


Figure 7. Mean Total Scale Scores by Grade and Gender



The performance of students tested with and without accommodations is provided in Tables 23 and 24. 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. A comparison of the means indicates that students receiving accommodations tended to obtain somewhat lower scores than students who did not require accommodations.

Table 23. Total Scale Score Means by Grade and Accommodations

Grade	Total Scale Score					
	Without Accommodations			With Accommodations		
	N	Mean	SD	N	Mean	SD
KG	8,811	376.62	39.25	1,252	375.81	38.65
1	10,334	434.97	42.95	1,145	433.89	37.58
2	9,510	478.71	40.91	316	463.35	43.66
3	8,827	496.06	42.86	267	460.64	40.88
4	7,429	516.83	41.33	218	483.81	39.80
5	6,571	532.31	43.04	174	505.26	45.03
6	5,186	530.82	42.72	121	517.22	41.73
7	4,674	538.69	45.99	56	518.16	37.77
8	4,149	541.41	46.74	55	530.04	52.57
9	4,103	531.78	48.38	18	522.44	41.90
10	3,317	538.93	48.52	16	523.25	45.85
11	2,357	543.75	46.22	3	524.33	46.82
12	1,808	543.17	43.61	3	488.67	66.49

Table 24. Component Scale Score Means by Grade and Accommodations

Grade	Speaking Scale Scores						Listening Scale Scores					
	No Speaking Accommodations			With Speaking Accommodations			No Listening Accommodations			With Listening Accommodations		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
KG	10,833	453.69	51.90	1,323	449.64	39.72	10,532	421.44	40.14	1,313	416.36	35.08
1	10,626	482.65	43.09	1,153	476.10	33.99	10,608	462.80	38.93	1,150	455.95	35.90
2	9,640	499.69	40.74	312	490.86	37.04	9,634	493.11	39.90	310	482.82	45.79
3	8,980	514.64	44.62	252	494.34	35.55	8,955	491.55	42.80	252	466.62	40.09
4	7,507	527.81	45.65	210	502.40	42.00	7,497	510.08	44.63	214	486.50	47.30
5	6,627	538.34	48.74	168	516.30	46.06	6,631	527.63	48.49	167	508.19	49.84
6	5,233	535.38	53.76	115	518.63	48.69	5,233	529.18	48.13	118	517.58	50.03
7	4,737	539.54	58.74	52	516.67	41.31	4,730	538.43	52.50	53	513.92	48.66
8	4,190	538.90	60.21	58	536.84	71.13	4,196	540.94	55.17	57	523.70	54.36
9	4,208	528.81	62.26	18	531.67	45.08	4,207	531.30	59.02	16	517.69	47.44
10	3,390	534.28	62.29	17	545.71	57.95	3,380	538.75	60.64	16	538.25	46.53
11	2,423	537.93	58.38	3	541.33	14.57	2,409	544.98	59.75	3	567.33	53.20
12	1,865	534.04	54.88	4	510.25	43.57	1,849	542.79	59.22	3	481.67	97.00

Grade	Reading Scale Scores						Writing Scale Scores					
	No Reading Accommodations			With Reading Accommodations			No Writing Accommodations			With Writing Accommodations		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
KG	10,488	340.00	55.26	1,301	337.20	49.78	9,034	274.03	78.57	1,272	298.12	81.26
1	10,591	405.15	48.83	1,150	400.70	43.91	10,445	387.02	83.36	1,149	404.33	69.59
2	9,623	458.15	55.03	318	433.11	57.07	9,583	463.53	65.74	317	446.26	70.25
3	8,913	478.34	64.94	258	420.30	69.61	8,894	499.03	61.24	271	460.05	62.59
4	7,486	506.50	58.23	211	451.68	71.03	7,471	522.79	57.03	221	486.10	53.09
5	6,615	526.23	56.16	171	489.58	67.56	6,612	537.37	57.14	177	500.55	63.73
6	5,229	525.22	51.72	121	508.88	55.69	5,223	534.26	53.83	123	523.86	51.83
7	4,742	537.11	52.90	53	512.30	40.78	4,722	539.99	56.47	59	516.76	55.83
8	4,194	545.10	52.32	57	529.44	65.32	4,179	541.38	54.62	56	526.57	56.43
9	4,234	535.12	56.09	16	517.75	60.66	4,209	531.91	51.86	17	522.94	41.73
10	3,406	545.11	54.47	15	501.27	63.17	3,389	537.38	51.22	15	520.93	40.46
11	2,415	553.04	51.30	3	484.00	97.14	2,411	539.93	50.13	3	506.67	40.55
12	1,869	557.54	48.50	3	466.33	114.51	1,864	539.56	46.81	3	495.67	46.14

As discussed in detail in the 2006 *CELA Performance Summary and Technical Addendum*, preliminary cut scores for the Colorado FEP category were set using the Spring 2006 CELA data[4]. These preliminary cut scores are only one of the many pieces of information used to classify Colorado students as proficient. These preliminary cut scores are for each grade level in Table 25, along with the corresponding LAS Links cut scores.

Table 25. LAS Links and CELA Cut Scores on Total Test

	LAS Links Cut Scores				Preliminary CELA FEP Cut Score
Grade	Early Intermediate Level 2	Intermediate Level 3	Proficient Level 4	Above Proficient Level 5	
KG	389	425	468	515	503
1	394	433	471	521	508
2	436	470	501	546	534
3	438	475	511	553	539
4	452	490	525	578	564
5	453	492	528	579	566
6	465	498	537	586	573
7	465	499	538	587	574
8	467	501	539	587	575
9	469	508	547	602	588
10	471	508	549	603	589
11	472	510	551	604	590
12	473	511	553	606	592

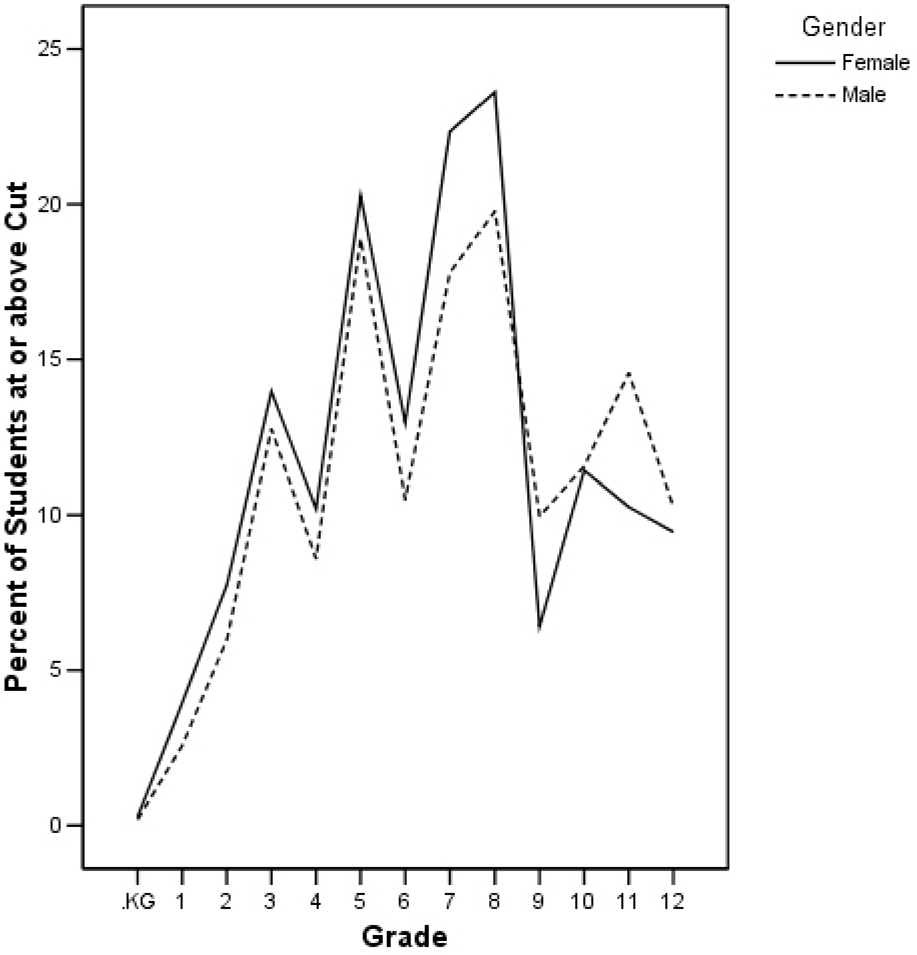
The percentages of male and female students at each grade level who scored at or above the preliminary FEP cut score are shown below in Table 26 and Figure 8. From Kindergarten through Grade 8, the percentage of females scoring at or above the cut was greater than the percentage of males. However, this pattern is reversed in Grades 9 through 12, with male students outperforming females.

[4] These preliminary cut scores will be replaced with new cut scores which will be established in a formal Bookmark standard-setting workshop in 2008.

Table 26. Percent of Students Scoring at or above Preliminary FEP Cut Score by Grade and Gender

Grade	Gender	FEP	
		N	Percent at or above 2006 Preliminary Cut
KG	F	4,982	0.34
	M	5,080	0.24
1	F	5,688	4.10
	M	5,788	2.70
2	F	4,666	7.93
	M	5,159	6.16
3	F	4,425	14.33
	M	4,664	13.14
4	F	3,715	10.42
	M	3,931	8.80
5	F	3,216	20.65
	M	3,527	19.19
6	F	2,399	13.34
	M	2,908	10.87
7	F	2,221	23.19
	M	2,509	18.45
8	F	1,904	24.68
	M	2,300	20.65
9	F	1868	6.96
	M	2252	10.92
10	F	1,596	12.22
	M	1,736	12.62
11	F	1,110	11.44
	M	1,250	16.08
12	F	843	10.91
	M	967	11.58

Figure 8. Percent of Students Scoring at or above the Preliminary Cut Score, by Grade and Gender



Although the profiles are quite jagged, there is a general upward trend in overall proficiency from kindergarten through grade 8 for both sexes. However, the proportion of students classified as proficient drops markedly in Grade 9. It should be noted that a drop in proficiency in Grade 9 was also apparent in the 2006 CELA proficiency data, in the Colorado historical proficiency data from 2003 through 2005, as well as in the original *LAS Links* standardization data.

For comparative purposes, the percentage of students meeting the preliminary CELA FEP cut score in 2006 and 2007 are shown in Table 27.

Table 27. Colorado FEP Proficiency Classification, 2006 vs.2007

Grade	2006		2007		Difference (2007 minus 2006)
	Total N	% at or above Cut	Total N	% at or above Cut	
KG	10,548	0.81	10,063	0.29	-0.52
1	11,195	8.73	11,479	3.39	-5.34
2	10,371	9.53	9,826	7.00	-2.52
3	9,232	21.14	9,094	13.72	-7.42
4	8,253	16.77	7,647	9.59	-7.18
5	7,231	22.61	6,745	19.91	-2.70
6	6,143	18.31	5,307	11.98	-6.33
7	5,289	19.89	4,730	20.68	0.79
8	4,603	24.88	4,204	22.48	-2.40
9	4,277	9.91	4,121	9.12	-0.79
10	3,086	12.80	3,333	12.42	-0.38
11	2,320	12.72	2,360	13.90	1.18
12	1,748	12.81	1,811	11.32	-1.49

Part 7: Reliability and Validity Evidence

Test reliability and validity statistics for the *LAS Links* standardization sample are provided in the *LAS Links Technical Manual*. Validity and reliability statistics were also computed for each *LAS Links* grade span using the data from the Spring 2007 CELA administration. Overall, these CELA analyses yielded results that were consistent with the *LAS Links* standardization data.

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 can not 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 error 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 coefficient alpha (1951) 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 CELA Speaking, Listening, Reading, Writing, Oral and Comprehension 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 CELA total scale score is a composite (the unweighted mean of the four component scale scores on Reading, Writing, Listening, and Speaking), the internal consistency reliability of the total score was computed using the following 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 CELA, *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 2007 CELA tests are shown in Table 28. Achievement tests are typically considered to be of sound reliability when their reliability coefficients are in the range of .80 and above. All of the reliability coefficients for Speaking, Reading, Writing, Oral, and Comprehension meet or exceed this criterion, with the exception of the Kindergarten Writing score. However, the reliability coefficients for the Listening scale are below .80 for every grade and grade span, with the sole exception of Grade Span 1. 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 .90 for every grade and grade span.

Table 28. Internal Consistency Reliability Coefficients by Grade Span and Grade.

	Speaking	Listening	Reading	Writing	Oral	Compre-hension	Total Score
Grade Span 1	0.94	0.85	.*	.*	0.94	.*	.*
K	0.94	0.76	0.81	0.79	0.93	0.83	0.90
1	0.92	0.77	0.82	0.85	0.92	0.84	0.93
2	0.91	0.76	0.86	0.88	0.91	0.86	0.95
Grade Span 2	0.90	0.68	0.88	0.87	0.89	0.87	0.94
3	0.89	0.62	0.86	0.87	0.88	0.84	0.93
4	0.89	0.64	0.87	0.86	0.89	0.86	0.93
5	0.90	0.70	0.88	0.86	0.90	0.88	0.94
Grade Span 3	0.93	0.76	0.85	0.84	0.92	0.88	0.95
6	0.92	0.72	0.84	0.84	0.91	0.87	0.94
7	0.93	0.76	0.86	0.85	0.93	0.89	0.95
8	0.93	0.78	0.86	0.84	0.93	0.89	0.95
Grade Span 4	0.94	0.78	0.86	0.86	0.94	0.89	0.95
9	0.94	0.77	0.85	0.86	0.94	0.89	0.95
10	0.94	0.79	0.86	0.87	0.94	0.90	0.95
11	0.93	0.77	0.86	0.86	0.93	0.90	0.95
12	0.92	0.77	0.85	0.85	0.92	0.89	0.94

* For Reading, Writing, Comprehension, and Total, different items were administered at different grades within Grade Span 1. Therefore, Grade Span 1 reliability coefficients are not provided for these content areas.

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 smaller the SEM, the smaller the variability and the higher the reliability.

The SEMs for the Spring 2007 CELA assessments are shown in Table 29.

Table 29. Standard Errors of Measurement by Grade Span and Grade.

	Speaking	Listening	Reading	Writing	Oral	Compre-hension	Total
Grade Span 1	11.98	19.00	.*	.*	10.43	.*	.*
K	12.43	19.42	23.84	36.34	10.28	19.50	12.30
1	11.97	18.56	20.53	31.86	9.78	16.28	10.97
2	12.20	19.66	20.68	22.85	11.28	15.95	9.63
Grade Span 2	14.92	26.93	22.31	22.07	13.68	16.91	10.99
3	14.77	26.46	24.61	22.22	12.67	17.31	11.23
4	15.17	26.92	21.38	21.42	12.97	16.24	10.81
5	15.43	26.63	19.66	21.56	13.92	15.96	10.60
Grade Span 3	15.19	25.49	20.54	22.02	16.09	17.56	10.57
6	15.19	25.51	20.74	21.52	15.52	17.33	10.53
7	15.51	25.73	19.77	21.89	15.34	16.80	10.52
8	15.97	25.89	19.66	21.87	16.21	17.36	10.58
Grade Span 4	14.80	28.09	20.31	18.95	14.29	18.12	10.55
9	15.23	28.29	21.73	19.39	14.44	18.34	10.84
10	15.25	27.76	20.42	18.45	14.67	17.48	10.49
11	15.44	28.65	19.23	18.75	15.25	16.55	10.55
12	15.52	28.44	18.88	18.14	15.18	17.89	10.41

* For Reading, Writing, Comprehension, and Total, different items were administered at different grades within Grade Span 1. Therefore, Grade Span 1 SEMs are not provided for these content areas.

Classification Consistency

As further evidence about the reliability and validity of the proficiency levels, we reviewed the classification consistency of the placement of students into the FEP proficiency level, using the estimation methods described by Subkoviak (1988). Subkoviak (1988) provides tables from which approximate values of the Agreement Coefficient and Kappa can be obtained based on an estimation from a single administration of the test. In order to use Subkoviak's tables, the cut score is expressed as a standard score (z) and the reliability of the test (alpha) is taken as the internal consistency estimate provided in this report.

Classification consistency was estimated for only one cut score at each grade level – the FEP cut point on the total score scale – because this was the only cut score that was used for decision making.

Table 30 shows the kappa coefficients and the agreement coefficients at each grade level for the Spring 2007 CELA examinees. The agreement coefficients in this table indicate the consistency with which students would be classified above or below the preliminary FEP cut score that was established using the procedures discussed in a previous section of this report. Overall, these coefficients indicate consistent classification for 95 to 98 percent of students in Grades K-2 and for 90 to 92 percent of students in Grades 3-12.

The kappa coefficients are more sensitive than the agreement coefficients to the contribution of test score reliability to classification consistency (Subkoviak, 1988). The kappa coefficients for the 2007 CELA administration are consistently high, ranging from .58 for Kindergarten students to .70 for students in Grade 8.

Table 30. Subkoviak Agreement Coefficient and Kappa for the Overall Test by Grade

Grade	FEP Cut Score	Scale Score Mean	Scale Score SD	Z-Score	Agreement Coefficient	Kappa Coefficient
KG	503	376.52	39.17	3.22	0.98	0.58
1	508	434.86	42.45	1.71	0.97	0.61
2	534	478.22	41.09	1.35	0.95	0.64
3	539	495.02	43.22	1.01	0.91	0.68
4	564	515.89	41.65	1.14	0.92	0.67
5	566	531.61	43.31	0.78	0.90	0.69
6	573	530.51	42.75	0.98	0.91	0.68
7	574	538.45	45.95	0.76	0.90	0.69
8	575	541.26	46.83	0.71	0.90	0.70
9	588	531.74	48.36	1.15	0.92	0.66
10	589	538.85	48.52	1.02	0.91	0.68
11	590	543.72	46.21	0.99	0.91	0.68
12	592	543.08	43.69	1.11	0.91	0.67

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 CELA 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 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 previously in Table 2, a study of the alignment of the CELA 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 CELA 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 CELA scale scores, the pattern of correlations within CELA provides preliminary validity evidence. The intercorrelations among the CELA scales for each grade and grade span are shown in Tables 31 through 34.

Table 31. CELA Scale Score Correlations, Grade Span K-2.

		Listening	Reading	Writing	Comprehension	Oral	Total
KG	Speaking	0.49	0.49	0.19	0.54	0.91	0.66
	Listening	--	0.58	0.25	0.89	0.73	0.69
	Reading		--	0.39	0.81	0.57	0.80
	Writing			--	0.31	0.23	0.74
	Comprehension				--	0.73	0.79
	Oral					--	0.75
1	Speaking	0.51	0.49	0.45	0.55	0.93	0.71
	Listening	--	0.59	0.48	0.82	0.75	0.74
	Reading		--	0.67	0.89	0.59	0.85
	Writing			--	0.63	0.52	0.89
	Comprehension				--	0.71	0.86
	Oral					--	0.80
2	Speaking	0.48	0.51	0.52	0.55	0.93	0.73
	Listening	--	0.55	0.51	0.77	0.70	0.74
	Reading		--	0.73	0.91	0.58	0.88
	Writing			--	0.72	0.56	0.89
	Comprehension				--	0.69	0.91
	Oral					--	0.81

Table 32. CELA Scale Score Correlations, Grade Span 3-5.

		Listening	Reading	Writing	Comprehension	Oral	Total
3	Speaking	0.48	0.47	0.52	0.53	0.94	0.72
	Listening	--	0.52	0.53	0.78	0.73	0.74
	Reading		--	0.69	0.86	0.55	0.87
	Writing			--	0.72	0.59	0.87
	Comprehension				--	0.70	0.91
	Oral					--	0.83
4	Speaking	0.48	0.48	0.50	0.54	0.93	0.73
	Listening	--	0.55	0.52	0.80	0.74	0.77
	Reading		--	0.71	0.89	0.57	0.87
	Writing			--	0.72	0.56	0.86
	Comprehension				--	0.71	0.92
	Oral					--	0.84
5	Speaking	0.49	0.53	0.54	0.57	0.93	0.76
	Listening	--	0.60	0.55	0.82	0.75	0.79
	Reading		--	0.72	0.90	0.63	0.88
	Writing			--	0.73	0.60	0.86
	Comprehension				--	0.74	0.92
	Oral					--	0.86

Table 33. CELA Scale Score Correlations, Grade Span 6-8.

		Listening	Reading	Writing	Compre-hension	Oral	Total
6	Speaking	0.54	0.52	0.56	0.60	0.94	0.80
	Listening	--	0.58	0.58	0.79	0.76	0.80
	Reading		--	0.70	0.91	0.60	0.85
	Writing			--	0.73	0.64	0.86
	Comprehension				--	0.73	0.91
	Oral					--	0.88
7	Speaking	0.57	0.56	0.60	0.62	0.94	0.82
	Listening	--	0.63	0.60	0.80	0.78	0.82
	Reading		--	0.70	0.93	0.62	0.85
	Writing			--	0.73	0.65	0.86
	Comprehension				--	0.73	0.91
	Oral					--	0.89
8	Speaking	0.58	0.58	0.61	0.64	0.94	0.83
	Listening	--	0.65	0.60	0.82	0.78	0.83
	Reading		--	0.69	0.93	0.65	0.86
	Writing			--	0.72	0.66	0.86
	Comprehension				--	0.75	0.92
	Oral					--	0.90

Table 34. CELA Scale Score Correlations, Grade Span 9-12.

		Listening	Reading	Writing	Compre-hension	Oral	Total
9	Speaking	0.62	0.57	0.63	0.65	0.93	0.84
	Listening	--	0.65	0.63	0.84	0.83	0.85
	Reading		--	0.70	0.92	0.66	0.85
	Writing			--	0.74	0.69	0.86
	Comprehension				--	0.79	0.92
	Oral					--	0.92
10	Speaking	0.61	0.58	0.64	0.64	0.92	0.84
	Listening	--	0.67	0.65	0.84	0.82	0.86
	Reading		--	0.71	0.93	0.67	0.86
	Writing			--	0.75	0.71	0.87
	Comprehension				--	0.78	0.92
	Oral					--	0.92
11	Speaking	0.57	0.54	0.60	0.61	0.90	0.81
	Listening	--	0.66	0.62	0.83	0.81	0.86
	Reading		--	0.70	0.93	0.64	0.85
	Writing			--	0.73	0.68	0.86
	Comprehension				--	0.76	0.92
	Oral					--	0.91
12	Speaking	0.56	0.53	0.58	0.60	0.91	0.80
	Listening	--	0.64	0.63	0.83	0.80	0.86
	Reading		--	0.69	0.92	0.63	0.84
	Writing			--	0.73	0.66	0.85
	Comprehension				--	0.75	0.91
	Oral	--				--	0.90

Overall, the pattern of correlations among the four content domains of Listening, Speaking, Reading, and Writing is similar to the pattern observed in the 2006 data, and is consistent with theoretical expectations for the CELA 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 all grade levels, the highest correlation with the Listening scale is the Reading scale. And at all levels above Grade 2, the highest correlation with the Speaking scale is the Writing scale. The failure to find a similar pattern in Kindergarten and first grade is consistent with the less developed writing abilities at these lower grades.

It is also noteworthy that the highest single correlation coefficient among the four domains at each grade level except Kindergarten is the correlation between the two orthographic domains of Reading and Writing.

Part 8. Special Studies

No special studies were conducted during the 2006-2007 testing year.

References

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

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.

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

CTB/McGraw-Hill. (2006). *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 71*, 179–181. Hillsdale, NJ: Lawrence Erlbaum.

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

MacMillan/McGraw-Hill (1993a). *Guidelines for Bias-Free Publishing.* New York: Author.

MacMillan/McGraw-Hill (1993b). Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals. New York: Author.

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

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

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.