The McGraw-Hill Companies

Technical Report for the California English Language Development Test (CELDT)

2004-2005 Form D

Submitted to the California Department of Education on October 27, 2005

Mc Graw Hill CTB

Developed and published under contract with the California Department of Education by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2005 by the California Department of Education. All rights reserved. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the California Department of Education. This work is based on the Bookmark Standard Setting Procedure, copyright © 2004 by CTB/McGraw-Hill LLC. Bookmark Standard Setting Procedure is a trademark of The McGraw-Hill Companies, Inc.

ACKNOWLEDGEMENTS

The following CTB/McGraw-Hill staff members are primarily responsible for the content and statistical quality of this report:

Donald Ross Green
Chief Research Psychologist

Keith Boughton
Research Scientist

Michelle Boyer
Erica Connelly
Marie Huchton
Launa Rodden
Research Associates

TABLE OF CONTENTS

TABLE OF TABLES & FIGURES

Overview

As stated in California Assembly Bill 748 (Statutes of 1997), the Superintendent of Public Instruction was required to select or develop a test that assesses the English language development of pupils whose primary language is a language other than English. Subsequently, California Senate Bill 638 (Statutes of 1999) required school districts to assess the English language development of all English Learners. The California English Language Development Test (CELDT) was the test designed to fulfill these requirements. As stated in the California Education Code, Section 60810(d), "The test shall be used for the following purposes: (1) To identify pupils who are limited English proficient. (2) To determine the level of English language proficiency of pupils who are limited English proficient. (3) To assess the progress of limited-English-proficient pupils in acquiring the skills of listening, reading, speaking, and writing in English."

Responding to these requirements, the California Department of Education, with the approval of the Superintendent of Public Instruction and the State Board of Education, developed the California English Language Development Test (CELDT). The test assesses English Learners in the skill areas of Listening/Speaking, Reading, and Writing. The test is administered to four separate grade span levels (K-2, 3-5, 6-8, and 9-12).

For the 2004-2005 administration, Form D was used for both initial identification and annual assessment and was designed for use with the four grade span levels listed above. The layouts of the test books varied by grade span, with each grade span containing either four or eight operational test booklets. Each booklet contained the operational test, with some booklets also containing field test items for the three skill areas.

For the 2004 operational test, students were scored in the skill areas of Listening/Speaking, Reading, and Writing. The resulting scores from these skill areas were then combined to create an overall score. The Listening/Speaking portion of the test was double-weighted, with the Listening portion of the test administered in groups, and the Speaking portion of the test administered individually. The Reading and Writing skill areas were single-weighted and given in group administrations.

This document provides technical details on the operational test for 2004-2005 only. As such, it is an extension of previous technical reports. For information regarding the CELDT standard setting, refer to the *California English Language Development Bookmark Standard Setting Technical Report*, published in 2001. For the 2000 field test or the 2001 operational test, refer to the *Technical Report for the California English Language Development Test (CELDT) 2000 – 2001,* published in 2002, for information regarding the 2002-2003 operational test, refer to the *CELDT 2002-2003 Form B Technical Report*, published in 2003, and for information regarding the 2003-2004 operational test, refer to the *CELDT 2003-2004 Form C Technical Report*, published in 2004. CTB endeavored to follow the testing guidelines published by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. Information regarding documentation and compliance can be found in Appendix Q.

California English Language Development Test, Form D

Test Development and Structure

Each booklet in the Form D series was divided into the three skill areas of Listening/Speaking, Reading, and Writing, following parallel specifications to Form C. All items included in the Form D operational test were administered in Form C as either operational or field test items. New items developed for Form D were included in each booklet as field test items. The layout of the booklets varied, with every booklet in the series containing the operational test for the given grade span, and most also containing embedded field test items for the three skill areas. For detail on the number of questions in each operational test and field test item section, please see Table 1, *CELDT Form D Test Structure*.

For Grade Span 1 (kindergarten-grade 2), there were a total of eight distinct booklets. There were four booklets for kindergarten and grade 1 (D1-D4), consisting only of the Listening/Speaking test. Kindergartners and first graders are not currently administered the Reading or Writing portions of the CELDT, and their overall scores are based solely on the results of their Listening/Speaking test. (Please see Table 2, *2004-2005 Operational Test Administration Structure*, for more detail.) Each of the four booklets contained the same operational items, as well as unique embedded field test items created for Form D.

There were eight booklets for grade 2 students; in addition to the same Listening/Speaking items administered to kindergarten and grade 1, the grade 2 booklets also contained Reading and Writing tests. Booklets for test forms D1-D4 contained Listening/Speaking sections identical to the kindergarten and grade 1 tests, as well as operational Reading and Writing items. Booklets D5-D8 contained only the operational Listening/Speaking items, as well as operational and field test items for both Reading and Writing items.

Grade Spans 2, 3, and 4 (for grades 3-5, 6-8, and 9-12, respectively) each had parallel booklet layouts. Each grade span had eight booklets, called D1-D8. Within each grade span, one set of operational items was used across all booklets. In addition to the operational items, booklets D1-D4 contained embedded field test items for Listening/Speaking and booklets D5-D8 contained field test items for Reading and Writing.

Regarding the items field tested in booklets D1-D8 each grade span, it should be noted that each booklet usually contained different embedded field test items, though there were some cases of overlap. Forms D1-D8 were randomly distributed across districts and specific precautions were also taken to ensure that no more than 30% of the sample for any field test item came from a single school district. For a detailed structure of the embedded items, please see Appendix A: Form D Item Map.

Each individual question in each skill area had a set number of obtainable score points. For most questions, either 0 or 1 score point could be obtained on the question. For some questions, the number of score points was higher; in such cases the scoring was based on a scoring rubric. This was the case for the *constructed response–Speech Functions* questions in Speaking with three

score points (0, 1, or 2); for the *constructed response–Choose and Give Reasons* questions in Speaking with three score points (0, 1, or 2); for the *constructed response–4-Picture Narrative* questions in Speaking with five score points (0, 1, 2, 3, or 4); for the *constructed response–Writing Sentence*s questions in Writing with four score points (0, 1, 2, or 3); and for the *constructed response–Short Composition* question in Writing with five score points (0, 1, 2, 3, or 4). For each section the points achieved on each question were then summed to provide a total raw score. The total raw score had a particular scale score associated with it, based on the raw score and the item parameters.

For Listening/Speaking on the 2004-2005 Operational CELDT, for grade span 1, there were 29 dichotomous items with two score points (0 or 1), one "Choose and Give Reasons" question with three score points (0, 1, or 2), and one "4-Picture Narrative" question with five score points (0, 1, 2, 3, or 4). In sum, the Listening/Speaking section of the test for grade span 1 had up to 35 (29x1 + 1x2 + 1x4) raw score points. For grade spans 2, 3, and 4, there were 29 dichotomous items with two score points (0 or 1), four "Speech Functions" questions with three score points (0, 1, or 2), one "Choose and Give Reasons" question with three score points (0, 1, or 2), and one "4-Picture Narrative" question with five score points (0, 1, 2, 3, or 4). In sum the Listening/Speaking section of the test for grade spans 2, 3, and 4 had up to 43 (29x1 + 4x2 + 1x2 + 1x4) raw score points.

For Reading on the 2004-2005 Operational CELDT, at each grade span, there were 35 dichotomous items with two score points (0 or 1). In sum the Reading section of the test had up to 35 (35x1) raw score points.

For Writing on the 2004-2005 Operational CELDT, at each grade span, there were 19 dichotomous items with two score points (0 or 1), four "Sentences" questions with four score points (0, 1, 2, or 3), and one "Short Composition" question with five score points (0, 1, 2, 3, or 4). In sum the Writing section of the test had up to 35 (19x1 + 4x3 + 1x4) raw score points.

Note that scale score and proficiency descriptors are appropriate only at the content area level. Interpretation of individual items or subsets of items within a content area is not recommended. For more detail on the structure of the Form D test, including the types of items and the distribution of field test items, please see Table 1.

Table 1 CELDT Form D Test Structure

Grade Span	Test Materials	Content of Materials (Items are not listed in order)			
		Skill Area	**No. of Operational Items**	**Item Type***	**Total No. of Field Test Items**
K-1 4 forms (D1-D4)	4 scannable test books	Listening/Speaking	9	MC	9
			20	DCR	16
			1	CGR-CR	2
			1	4PN-CR	2
Grade 2 8 forms (D1-D8)	8 scannable test books	Listening/Speaking	9	MC	9
			20	DCR	16
			1	CGR-CR	2
			1	4PN-CR	2
		Reading	35	MC	26
		Writing	19	MC	12
			4	S-CR	8
			1	SC-CR	3
3-5 8 forms (D1-D8)	8 nonscannable test books	Listening/Speaking	19	MC	15
			10	DCR	8
			4	SF-CR	4
			1	CGR-CR	2
			1	4PN-CR	2
		Reading	35	MC	24
		Writing	19	MC	12
			4	S-CR	8
			1	SC-CR	3
6-8 8 forms (D1-D8)	8 nonscannable test books	Listening/Speaking	19	MC	15
			10	DCR	8
			4	SF-CR	4
			1	CGR-CR	2
			1	4PN-CR	2
		Reading	35	MC	24
		Writing	19	MC	12
			4	S-CR	8
			1	SC-CR	3
9-12 8 forms (D1-D8)	8 nonscannable test books	Listening/Speaking	19	MC	15
			10	DCR	8
			4	SF-CR	4
			1	CGR-CR	2
			1	4PN-CR	2
		Reading	35	MC	24
		Writing	19	MC	12
			4	S-CR	8
			1	SC-CR	3

MC = Multiple Choice
DCR = Dichotomous Constructed Response
CR = Constructed Response
SF-CR = Speech Functions – Constructed Response
CGR-CR = Choose & Give Reasons – Constructed Response
4PN-CR = 4-Picture Narrative – Constructed Response
S-CR = Sentences – Constructed Response
SC-CR = Short Compositions – Constructed Response

Table 2 2004-2005 Operational Test Administration Structure

Subject	Grade Span				
	GS 1 : K and 1	GS 1 : 2	GS 2 : 3 – 5	GS 3 : 6 – 8	GS 4 : 9 – 12
Listening/Speaking	✓	✓	✓	✓	✓
Reading	Not Tested	✓	✓	✓	✓
Writing	Not Tested	✓	✓	✓	✓

✓= Subject Area Administered

Proficiency Levels

Cut-scores and Proficiency Level Descriptions remained the same as in previous operational test administrations. Cut-score information may be found in Table 3, and Proficiency Level Descriptions may be found in Table 4.

For more information on the development of the Cut-scores and Proficiency Level Descriptions, please refer to the *Technical Report for the California English Language Development Test (CELDT) 2000 – 2001*, published in 2002.

Table 3 CELDT Cut-scores

Listening & Speaking

Test Grade Span		Early Int. Cut	Int. Cut	Early Adv. Cut	Adv. Cut
	K	410	458	506	554
K-2	**1**	424	471	517	564
	2	454	495	536	577
3-5		438	482	526	569
6-8		438	482	526	569
9-12		438	482	526	569

Reading

Test Grade Span	Early Int. Cut	Int. Cut	Early Adv. Cut	Adv. Cut
2	438	475	511	548
3-5	466	499	533	566
6-8	466	499	533	566
9-12	466	499	533	566

Writing

Test Grade Span	Early Int. Cut	Int. Cut	Early Adv. Cut	Adv. Cut
2	424	469	514	559
3-5	445	488	530	573
6-8	445	488	530	573
9-12	445	488	530	573

Overall

Test Grade Span		Early Int. Cut	Int. Cut	Early Adv. Cut	Adv. Cut
	K	410	458	506	554
K-2	**1**	424	471	517	564
	2	443	483	524	565
3-5		447	488	529	569
6-8		447	488	529	569
9-12		447	488	529	569

Table 4 CELDT Proficiency Level Descriptions

Proficiency Level	Description
Advanced	Students performing at this level of English language proficiency communicate effectively with various audiences on a wide range of familiar and new topics to meet social and academic demands. In order to attain the English proficiency level of their native English-speaking peers, further linguistic enhancement and refinement are necessary.
Early Advanced	Students performing at this level of English language proficiency begin to combine the elements of the English language in complex, cognitively demanding situations and are able to use English as a means for learning in other academic areas.
Intermediate	Students performing at this level of English language proficiency begin to tailor the English language skills they have been taught to meet their immediate communication and learning needs.
Early Intermediate	Students performing at this level of English language proficiency start to respond with increasing ease to more varied communication tasks.
Beginning	Students performing at this level of English language proficiency may demonstrate little or no receptive or productive English skills. They may be able to respond to some communication tasks.

Administration of Form D

2004-2005 Operational Test Summary Statistics

Tables 5 and 6, on the following pages, show the 2004-2005 operational test scale score summary statistics. These statistics are based on the General Research Tape (GRT) data.[1] Data noted "annual" were collected from the 2004 CELDT annual administration, which occurred between July 1st and October 31st of 2004. Data noted "initial" were being collected from students whose home language is a language other than English, who have never taken the CELDT, and took the test for purposes of initial identification between July 1, 2004 and June 30, 2005. Students who took the CELDT for purposes of initial identification after July 1st, 2004 did not re-take the test during the 2004 annual administration. An overview of annual administration summary statistics from CELDT Form A, Form B, and Form C are available in Appendix B.

Simple statistics for each skill area, as well as intercorrelations among skill area scores, are detailed in Appendix C.

Frequency distributions were run on the scale scores for annual and initial identification data for Listening/Speaking, Reading, and Writing for each of the four grade spans. These frequency distributions are located in Appendix F. Also available are frequency distributions based on student home language and primary ethnicity; these are located in Appendix G.

[1] The GRT data includes all Form D data received at CTB prior to July 15, 2005 (testing completed prior to June 30, 2005), without exclusions.

Table 5 2004-2005 Summary Statistics by Grade, Annual Data

		Listening/Speaking		Reading		Writing		Overall	
Grade	N	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
K	7,025	439.1	92.0	NA	NA	NA	NA	439.1	92.0
1	153,873	499.3	60.5	NA	NA	NA	NA	499.3	60.5
2	165,182	538.2	61.4	454.2	47.4	477.1	58.2	501.5	48.2
3	163,289	505.1	59.3	474.4	51.5	498.7	56.8	495.4	49.3
4	147,890	532.4	63.9	499.0	51.7	517.6	54.3	520.0	50.7
5	135,953	549.6	67.8	517.1	53.1	529.4	54.1	536.1	52.5
6	112,031	530.9	65.7	509.2	46.0	525.0	51.1	523.6	50.1
7	98,482	543.3	70.2	519.6	47.5	532.0	52.2	534.2	52.8
8	94,115	549.3	74.0	529.0	49.2	537.3	53.8	540.8	55.4
9	84,657	526.7	57.5	534.7	53.3	532.4	55.7	529.7	49.1
10	72,999	531.3	62.1	541.8	56.2	535.2	57.3	534.5	52.3
11	60,482	535.6	63.4	548.0	57.3	537.4	58.7	538.8	53.6
12	48,113	539.5	68.6	551.1	60.3	537.3	63.3	541.5	58.0

Table 6 2004-2005 Summary Statistics by Grade Span, Annual Data

		Listening/Speaking		Reading		Writing		Overall	
Grade Span	N	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
1: Grades K-2	326,080*	517.73	65.75	454.22	47.45	477.05	58.23	499.16	56.26
2: Grades 3-5	447,132	527.67	66.09	494.09	55.07	514.28	56.62	515.90	54.41
3: Grades 6-8	304,628	540.61	70.24	518.67	48.19	531.07	52.55	532.36	53.12
4: Grades 9 -12	266,251	532.26	62.42	542.66	56.67	535.21	58.28	535.22	52.88

* N-count for Grade Span 1 is 326,080 overall, but for Reading and Writing includes only Grade 2 data, for which the N-count is 165,182.

Table 7 2004-2005 Summary Statistics by Grade, Initial Data

		Listening/Speaking		Reading		Writing		Overall	
Grade	N	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
K	173,223	430.5	101.0	NA	NA	NA	NA	430.5	101.0
1	32,941	454.2	124.5	NA	NA	NA	NA	454.2	124.5
2	20,820	465.4	145.2	429.8	67.5	425.9	101.5	446.3	108.7
3	19,331	433.8	133.5	442.1	76.0	438.7	107.9	436.8	108.2
4	18,397	452.4	142.6	461.0	83.9	454.4	113.8	454.7	116.6
5	16,901	468.6	147.2	476.8	89.8	468.4	116.8	470.3	121.2
6	16,956	460.2	144.0	480.4	85.9	472.9	112.9	468.1	117.8
7	17,559	458.2	150.2	484.4	90.1	471.8	116.1	467.8	122.7
8	15,435	467.1	148.8	492.9	90.6	481.0	114.6	476.7	121.6
9	27,802	455.2	138.1	495.4	99.3	479.6	117.1	471.0	119.2
10	16,367	475.6	129.3	510.6	94.9	495.5	109.2	489.0	111.4
11	11,201	500.4	117.5	529.4	88.7	514.8	101.1	510.9	101.6
12	7,044	510.8	113.4	536.2	86.3	520.1	97.5	519.1	97.6

Table 8 2004-2005 Summary Statistics by Grade Span, Initial Data

		Listening/Speaking		Reading		Writing		Overall	
Grade Span	N	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
1: Grades K-2	226,984	437.1	110.1	429.8	67.5	425.9	101.5	435.4	105.8
2: Grades 3-5	54,629	450.8	141.6	459.2	84.3	453.2	113.4	453.2	116.0
3: Grades 6-8	49,950	510.0	147.7	503.0	89.0	508.0	114.6	511.0	120.8
4: Grades 9 -12	62,414	474.9	131.3	510.1	96.2	494.6	111.3	488.3	113.3

* N-count for Grade Span 1 is 226,984 overall, but for Reading and Writing includes only Grade 2 data, for which the N-count is 20,820.

Reliability and the Standard Error of Measurement

The reliability for a particular group of students' test scores is the extent to which the scores would remain consistent if those same students were retested with another parallel version of the same test, written to measure the same set of skills. If the test includes constructed-response questions, the reliability is the extent to which the students' scores would remain consistent if both the questions and the scorers were changed. Note that the constructed-response items are scored by raters who are locally trained on each item to increase the reliability of this scoring procedure. Additional data on rater consistency and reliability for handscored constructed response items are available in Appendix K.

The Reliability Coefficient

The reliability coefficient is the correlation between the students' scores and the scores that would result if the students were retested with a parallel form of the same test (and scored by different scorers, if the test includes constructed response questions). The reliability coefficient, in fact, cannot be computed directly unless the student actually takes two parallel forms of the same test. However, with some reasonable assumptions, it can be estimated from the students' responses to a single version of the test. Like other correlations, the reliability coefficient can vary substantially from one group of students to another. It tends to be larger in groups that are more diverse in the ability measured by the test and smaller in groups that are more homogeneous in the ability measured.

The reliability coefficients for the CELDT Form D are between 0.86 to 0.90 across all grades and subject areas, and these are typical coefficients for assessments of these lengths. Please see Table 7 for reliabilities for each skill area of the test by grade span.

The Standard Error of Measurement

The standard error of measurement is a measure of how much students' scores would vary from the scores they would earn on a perfectly reliable test. The "standard error of measurement" (SEM) is the difference between each student's score and the score that a student would earn on a perfectly reliable test. If it were possible to compute the error of measurement for each student's score, in a large group of students, these errors of measurement would have a mean of zero. The standard deviation of the errors of measurement would be an indication of how much the errors of measurement are affecting the students' scores. This statistic is the standard error of measurement. The standard error of measurement is expressed in the same units as the test scores, whether they are in raw-score or scale-score points. It is important to note that the SEM tends to be much more consistent across different groups of students than the reliability coefficient is. In a large group of students, about two-thirds of the students will earn scores within one SEM of the scores they would earn on a perfectly reliable test.

The standard error of measurement is the margin of error associated with an examinees' score. The range of standard errors for the CELDT Form D is between 16 and 26 points across all grades and subject areas in scale score units. In general, this translates into an error band by about one to two raw score points, depending on the students' score. For example, if a student received a raw score of 25 with a standard error of 1 point, then on retesting, the student might have attained a score between 24 to 26, about two-thirds of the time. It is important to remember that assessments are not perfectly reliable and only offer an estimate of what the student is

capable of, in a specified domain of knowledge. CELDT standard errors of measurement for each skill area and overall are shown in Table 8, below. For scale score standard errors of measurement for each skill area, see Appendix E: Form D Raw Score to Scale Score Tables.

The reliability from year to year is maintained by equating each new test form to a previous form, thus producing a relationship in which one can compare students' proficiency levels across years.

Table 9 2004-2005 Operational Test Reliabilities

Subject	Number of Items	Grade Span K-2	3-5	6-8	9-12
	31 (K-2)				
Listening/Speaking	35 (3-12)	0.86	0.86	0.87	0.88
Reading	35	0.88	0.90	0.87	0.89
Writing	24	0.89	0.88	0.87	0.87

Based on Cronbach's Alpha

Table 10 2004-2005 Operational Test Standard Errors of Measurement by Skill Area

Grade Span	Standard Error of Measurement in Scale Score Units: Listening/ Speaking	Reading	Writing	Overall
Grades K-2	24.60	16.44	19.31	21.53
Grades 3-5	24.73	17.41	19.61	21.86
Grades 6-8	25.33	17.38	18.95	22.04
Grades 9-12	21.62	18.80	21.01	20.80

Standard Errors of Measurement for each skill area calculated according to the formula: $SEM = SD\sqrt{1-\alpha}$, where *SD* represents the standard deviation and α represents the test reliability. Overall Standard Error of Measurement calculated according to the formula:

$$SEM_{all} = \sqrt{\frac{2(SEM_{LS}{}^2) + SEM_{RD}{}^2 + SEM_{WT}{}^2}{4}}$$

Test Population

The 2004-2005 Annual Administration operational test was administered to all students in the state of California whose home language was a language other than English and who had previously taken the CELDT. During this administration 1,344,091 took the CELDT for Annual Assessment and 393,977 took the CELDT for Initial Identification. These statewide data serve as population norms for the CELDT test and can be considered precise due to the sample size and appropriate due to the population composition.

It should be noted that data for this technical report were collected prior to the Data Review Module window of October 2004 and does not reflect changes made during that time. As a result, the data may not necessarily reflect the final test purpose as entered by the district in the Data Review module window.

Item Analysis

An analysis of the statistics for each of the 136 operational Listening/Speaking, 140 operational Reading, and 96 operational Writing items was conducted (numbers given are for all items across all grade spans). In addition, the 122 field-tested Listening/Speaking items, 98 field-tested

Reading items and 92 field-tested Writing items were studied. The results of both the operational and field test item analyses are located in Appendix H.

In addition to the standard item analyses, operational test item p-values and correlations between multiple choice and constructed response items were also studied. The differences in p-values for the annual administration data minus the initial identification data are included in Appendix I. Correlations between multiple choice and constructed response items are available in Appendix J.

Item Response Theory Analyses

Calibration and scaling of the 2004-2005 Operational Test data was accomplished using the PARDUX and WINFLUX computer programs. This proprietary software, developed at CTB/McGraw-Hill, enabled scaling and linking of complex assessment data such as that produced for the CELDT.

Because the characteristics of selected response and constructed response items are different, two item response theory models were used in the analysis of the data. The three-parameter logistic model (Lord & Novick, 1968; Lord, 1980) was used in the analysis of selected response (multiple choice) items. In this model, the probability that a student with scale score θ responds correctly to item *i* is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}$$

where a_i is the item discrimination, b_i is the item difficulty, and c_i is the probability of a correct response by a very low-scoring student.

For analysis of the constructed response items in the CELDT, the two-parameter partial credit model (Muraki, 1992; Yen, 1993) was used. 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

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

where

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

For the special case of the 2PPC model used here, the following constraints were used:

$$A_{jk} = \alpha_j(k-1),$$

and

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

where α_j and γ_{ji} are parameters freely estimated from the data. The first constraint implies that higher item scores reflect higher ability levels and that items can vary in their discriminations. The 2PPC model estimates a total of m_j independent item parameters; for each item there are $m_j - 1$ independent γ_{ji} parameters and one α_j parameter.

Goodness-of-Fit

Goodness-of-fit statistics were computed for each item to examine how closely the item's data conform to the item response models. A procedure described by Yen (1981) was used to measure fit. In this procedure, students are rank ordered on the basis of their $\hat{\theta}$ values and sorted into ten cells with ten percent of the sample in each cell. Each item *j* in each decile *i* has a response from N_{ij} examinees. The fitted IRT models are used to calculate an expected proportion E_{ijk} of examinees who respond to item *j* in category *k*. The observed proportion O_{ijk} is also tabulated for each decile, and the approximate chi-square statistic

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

Q_{1j} should be approximately chi-square distributed with degrees of freedom (*DF*) equal to the number of "independent" cells, 10(m_j-1), minus the number of estimated parameters. The number of score levels for an item *j* are represented by m_j, so for the 3PL model m_j =2, and $DF = 10(2-1) - 3 = 7$. For the 2PPC model, $DF = 10(m_j - 1) - m_j = 9m_j - 10$. Since *DF* differs between multiple choice and performance assessment (PA) items and between PA items with different score levels m_j, Q_{1j} is transformed, yielding the test statistic

$$Z_j = \frac{Q_{1j} - DF}{\sqrt{2DF}}.$$

This statistic is useful for flagging items that fit relatively poorly. Z_j is sensitive to sample size, and cutoff values for flagging an item based on Z_j have been developed and were used to identify items for the item review. The cut-off value is (N/1500 x 4) for a given test, where N is the sample size.

Model fit information is obtained from the Z-statistic. The Z-statistic is a transformation of the chi-square (Q1) statistic that takes into account differing numbers of score levels as well as sample size:

$$Z_j = \frac{(Q_{1j} - DF)}{\sqrt{2DF}}, \text{ where } j = \text{item } j.$$

The Z statistic is an index of the degree to which obtained proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters. These values are computed for ten intervals corresponding to deciles of the theta distribution (Burket, 1991). The Z statistic is used to characterize item fit. The critical value of Z is different for each grade or grade span because it is dependent on sample size.

Scaling and Equating

CTB uses an equating design based on common items to maintain the CELDT scales. Common items are used to equate field-test items onto the existing CELDT scales, and new operational test forms can then be selected from the field-test items and maintain the scale. In this way, the new form can be constructed on the CELDT scales of the previous form. The use of common items has become an industry-standard procedure for ensuring that a common scale can be established across the test forms. The linking and equating is conducted using the procedure by Stocking and Lord (Stocking and Lord, 1983). The Stocking and Lord procedure is based on determining the linear equating constants, M1 and M2, that minimize the difference between two test characteristic curves, such that, for a suitable group of examinees, the average squared difference between true-score estimates is as small as possible.

Probability of Classification

For CELDT, a scale score that was obtained from the "number-correct" scoring method of item calibration is assigned to any of the five scale score categories. Let c_i (i = 1, 2, 3, …, m) denote the cut-scores in increasing order that define the categories, let x_j denote an estimate of the scale score of examinee j, and let σ_j denote the standard error of the estimate. We assume that x_j (for all j) is normally distributed with mean x_j and variance σ_j^2.

A scale score may be located below the first cut-score (e.g., $x_j < c_1$), in between two cut-scores (e.g., $c_i < x_j < c_{i+1}$, for i=1,…,m-1), or above the last cut-score (e.g., $x_j > c_m$). Depending on which category a scale score is located, it is possible to obtain the probability of correct classification (PCC) and incorrect classification (PIC) from the standard normal distribution. For example, if $X_j < c_1$, the probability of correct classification is $\Pr(X_j < c_1)$ and the probability of misclassifications are: $\Pr(c_i \leq X_j < c_{i+1})$ for $i=1,\ldots,m-2$, and $\Pr(X_j \geq m)$. Once we have the PCC and PIC for each scale score, the category PCC and PIC is computed.

The computation of the category PCC and PIC involved two steps. In step one, the PCC and PIC for each scale score are computed. In the case of CELDT where m=4, one PCC and four PIC are computed for each scale. If there are N possible scales scores, e.g., N possible raw scores, at the end of step one, we will have a matrix of probabilities with dimensions N by m+1.

In step two, the probabilities within a category are weighted by the frequency of examinees that received a given scale score. These probabilities are then summed up row-wise to obtain a vector containing a PCC and PIC for a given category. At the end of step two, we will have an m+1 by m+1 table that summarized the PCC and PIC for each category. How to use the table? First locate the category where a scale score is located. For example, if a scale score is located in category 1, then the PCC is the first entry in row 1 and the PIC are the remaining entries. Similarly, if a scale score is located in category two, the PCC is the second entry in row 2 and the PIC are the remaining categories. Similar interpretation applies to scale scores from other categories.

The diagonal numbers in **bold** (Appendix D) should be interpreted as the probability of being correctly classified at each of the five cut-scores. The most important classification is whether someone is above or below the Early Advanced cut-score and should be close to 80% correct classification. However, this is only a general guideline and could be lower, depending on the where the distribution of scores lies, relative to the cut-scores.

The paper by Rogosa (1994) also discussed misclassification in student performance levels.

For probabilities of misclassification specific to CELDT, please see Appendix D: Probability of Classification.

Growth

The CELDT scale was established using data from the initial field test conducted in the fall of 2000, and modified using data from the 2001 operational administration (Form A). New items have been developed each year, field-tested with anchor items, and their item parameters placed on the scale developed from Form A in order to preserve the validity of the cut points that had been established by standard setting committees in the spring of 2001. These procedures allow reasonable comparisons of the results from each year.

The annual mean scores overall have shown an increase each year for all grades, except grade 12 between 2001 and 2002. These means are shown in Appendix B: CELDT Summary Statistics.

Note that the annual data does not include initial assessments and therefore do not include the lowest scoring students who often show substantial growth in their first months in their school. See tables 6 and 8 and note the much larger standard deviation for the initial group.

Correspondingly, the percentages of English Learners attaining proficiency have shown increases in each grade span each year, as shown in Table 12. Proficiency for CELDT is defined as an Overall score of Early Advanced or higher, and each skill area proficiency level (Listening/Speaking, Reading, Writing) as Intermediate or higher. For the tables and figures on the following pages, proficiency in a skill area is defined as a skill area score of Early Advanced or higher.

These percents also show an increase across the four grade spans, on the average, and grade-by-grade within each grade span. The grade-by-grade increases in these percents are shown in Tables 12, 13, 14, and 15 for each of the four years and are illustrated in Figures 1 through 4. It can be seen that the percent of students classified as Fluent English Proficient has increased each year for both Listening/Speaking and Overall. Writing has also shown such increases in grades 2 through 8, and Reading in grades 2 through 5. Writing grades 9 through 12 and Reading grades 6 through 12 do not show this year-to-year increase.

Table 11 Percent of English Learners Attaining English Language Proficiency on the CELDT, 2001, 2002, 2003, and 2004 Annual Assessments

Year	Grade Spans Tested				All Grades
	K–2	3–5	6–8	9–12	K–12
2004	28.7	37.0	54.0	62.5	43.9
2003	28.8	34.2	47.4	54.9	39.7
2002	21.7	25.1	39.5	46.7	31.5
2001	14.9	16.8	30.0	44.4	24.4

Table 12 Proficiency by Grade and Grade Span for Form A, Annual Data

Grade	N Tested	N Prof Listening/ Speaking	% Prof Listening/ Speaking	N Prof Reading	% Prof Reading	N Prof Writing	% Prof Writing	N Prof Overall	% Prof Overall
K*	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1	159986	28579	17.9	N/A	N/A	N/A	N/A	28579	17.9
2	166679	32758	19.7	12901	7.7	33274	20.0	20059	12.0
3	156520	21980	14.0	8924	5.7	25746	16.5	13078	8.4
4	135134	29629	21.9	18924	14.0	36860	27.3	23118	17.1
5	125877	36233	28.8	30117	23.9	47076	37.4	33796	26.9
6	108263	23356	21.6	31684	29.3	36376	33.6	25134	23.2
7	92351	23847	25.8	37685	40.8	36522	39.6	28813	31.2
8	85456	24932	29.2	42734	50.0	37334	43.7	31982	37.4
9	71239	26627	37.4	37974	53.3	28531	40.1	28277	39.7
10	67735	26593	39.3	39145	57.8	28557	42.2	29067	42.9
11	53768	22544	41.9	35081	65.3	24891	46.3	25533	47.5
12	39288	17528	44.6	28105	71.5	19859	50.6	20241	51.5
K-2	326665	61337	18.8	12901	7.7	33274	20.0	48638	14.9
3-5	417531	87842	21.0	57965	13.9	109682	26.3	69992	16.8
6-8	286070	72135	25.2	112103	39.2	110232	38.5	85929	30.0
9-12	232030	93292	40.2	140305	60.5	101838	43.9	103118	44.4
Overall	**1262296**	**314606**	**24.9**	**323274**	**29.3**	**355026**	**32.2**	**307677**	**24.4**

* Form A was the first year of operational testing; as such, all kindergartener data were treated as initial identification.

Table 13 Proficiency by Grade and Grade Span for Form B, Annual Data

Grade	N Tested	N Prof Listening/ Speaking	% Prof Listening/ Speaking	N Prof Reading	% Prof Reading	N Prof Writing	% Prof Writing	N Prof Overall	% Prof Overall
K	8135	1317	16.2	N/A	N/A	N/A	N/A	1317	16.2
1	160579	45080	28.1	N/A	N/A	N/A	N/A	45080	28.1
2	160257	58154	36.3	12705	7.9	35939	22.4	25118	15.7
3	160107	44422	27.8	10433	6.5	27743	17.3	20059	12.5
4	147640	60245	40.8	26751	18.1	46514	31.5	40416	27.4
5	125227	60696	48.5	35116	28.0	51730	41.3	48251	38.5
6	112594	43796	38.9	34548	30.7	40729	36.2	36805	32.7
7	98844	44757	45.3	42333	42.8	42247	42.7	40689	41.2
8	84780	41041	48.4	43196	51.0	39614	46.7	39471	46.6
9	76959	31249	40.6	43264	56.2	32643	42.4	32536	42.3
10	67284	30987	46.1	41674	61.9	30591	45.5	31219	46.4
11	54396	25737	47.3	36708	67.5	26506	48.7	26790	49.3
12	40633	20347	50.1	29537	72.7	21132	52.0	21310	52.5
K-2	328971	104551	31.8	12705	7.9	35939	22.4	71515	21.7
3-5	432974	165363	38.2	72300	16.7	125987	29.1	108726	25.1
6-8	296218	129594	43.7	120077	40.5	122590	41.4	116965	39.5
9-12	239272	108320	45.3	151183	63.2	110872	46.3	111855	46.7
Overall	**1297435**	**507828**	**39.1**	**356265**	**31.6**	**395388**	**35.0**	**409061**	**31.5**

Table 14 Proficiency by Grade and Grade Span for Form C, Annual Data

Grade	N Tested	N Prof Listening/ Speaking	% Prof Listening/ Speaking	N Prof Reading	% Prof Reading	N Prof Writing	% Prof Writing	N Prof Overall	% Prof Overall
K	6664	1550	23.3	N/A	N/A	N/A	N/A	1550	23.3
1	166704	59042	35.4	N/A	N/A	N/A	N/A	59042	35.4
2	170782	89450	52.4	16172	9.5	45257	26.5	38360	22.5
3	159439	56642	35.5	12600	7.9	37308	23.4	26870	16.9
4	153602	83827	54.6	32643	21.3	60895	39.6	56189	36.6
5	137167	90615	66.1	47561	34.7	69751	50.9	70965	51.7
6	112653	57564	51.1	34369	30.5	46422	41.2	44397	39.4
7	104276	59639	57.2	42111	40.4	50195	48.1	50448	48.4
8	94262	58279	61.8	46760	49.6	50706	53.8	52589	55.8
9	77889	37718	48.4	41011	52.7	32830	42.2	37783	48.5
10	74559	39112	52.5	45022	60.4	33619	45.1	40302	54.1
11	59229	33517	56.6	39469	66.6	28432	48.0	34822	58.8
12	45211	27172	60.1	32061	70.9	22558	49.9	28175	62.3
K-2	344150	150042	43.6	16172	9.5	45257	26.5	98952	28.8
3-5	450208	231084	51.3	92804	20.6	167954	37.3	154024	34.2
6-8	311191	175482	56.4	123240	39.6	147323	47.3	147434	47.4
9-12	256888	137519	53.5	157563	61.3	117439	45.7	141082	54.9
Overall	**1362437**	**694127**	**50.9%**	**389779**	**32.8%**	**477973**	**40.2%**	**541492**	**39.7%**

Table 15 Proficiency by Grade and Grade Span for Form D, Annual Data

Grade	N Tested	N Prof Listening/ Speaking	% Prof Listening/ Speaking	N Prof Reading	% Prof Reading	N Prof Writing	% Prof Writing	N Prof Overall	% Prof Overall
K	7025	1382	19.7	N/A	N/A	N/A	N/A	1382	19.7
1	153873	51870	33.7	N/A	N/A	N/A	N/A	51870	33.7
2	165182	91264	55.3	13039	7.9	44694	27.1	40294	24.4
3	163289	50992	31.2	13105	8.0	53958	33.0	30332	18.6
4	147890	78634	53.2	29906	20.2	72721	49.2	58475	39.5
5	135953	90721	66.7	47062	34.6	82757	60.9	76617	56.4
6	112031	61587	55.0	35374	31.6	59016	52.7	49482	44.2
7	98482	64336	65.3	42302	43.0	59326	60.2	55555	56.4
8	94115	64992	69.1	49922	53.0	61276	65.1	59585	63.3
9	84657	48766	57.6	49311	58.2	49705	58.7	49095	58.0
10	72999	45144	61.8	46251	63.4	44140	60.5	45416	62.2
11	60482	39050	64.6	40921	67.7	37286	61.6	39339	65.0
12	48113	32853	68.3	33960	70.6	30122	62.6	32637	67.8
K-2	326080	144516	44.3	13039	4.0	44694	13.7	93546	28.7
3-5	447132	220347	49.3	90073	20.1	209436	46.8	165424	37.0
6-8	304628	190915	62.7	127598	41.9	179618	59.0	164622	54.0
9-12	266251	165813	62.3	170443	64.0	161253	60.6	166487	62.5
Overall	**1344091**	**721591**	**53.7%**	**401153**	**33.9%**	**595001**	**50.3%**	**590079**	**43.9%**

Figure 1 Listening/Speaking Percent Proficient, Based on Annual Assessment Data

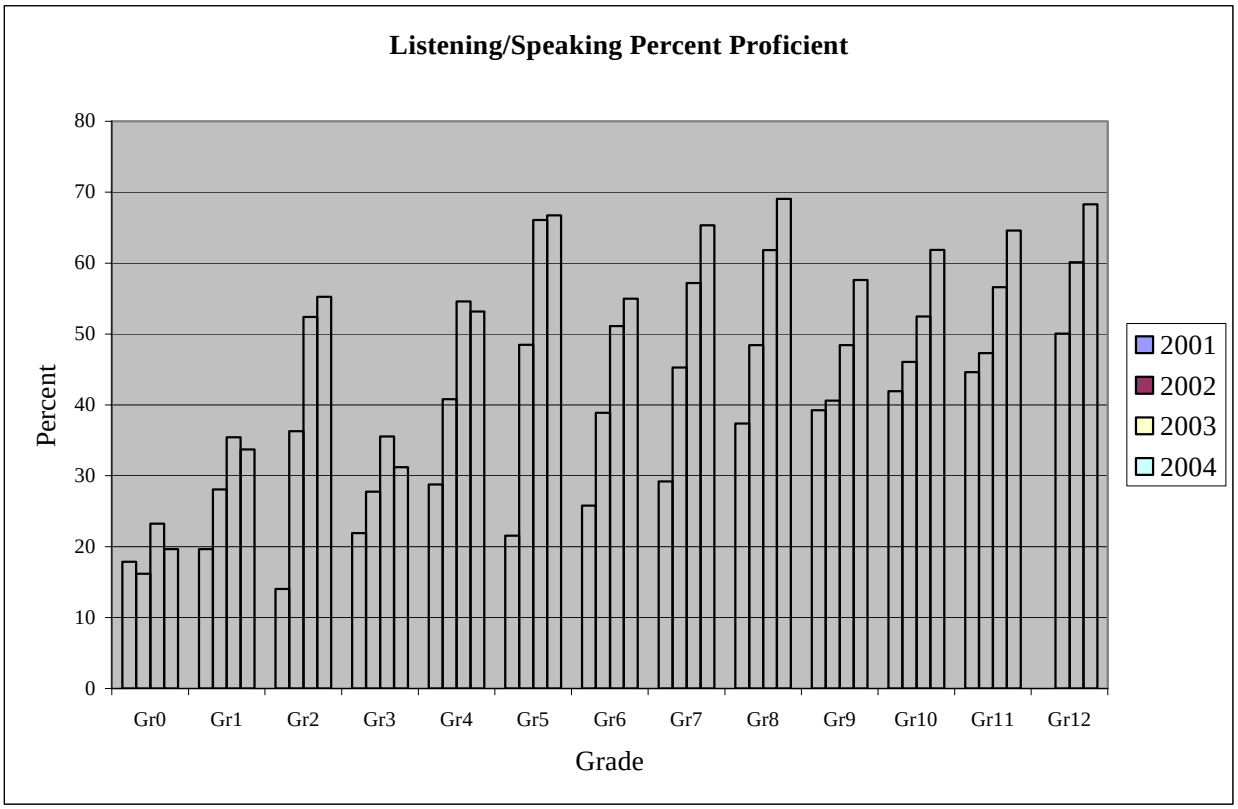


Figure 2 Reading Percent Proficient, Based on Annual Assessment Data

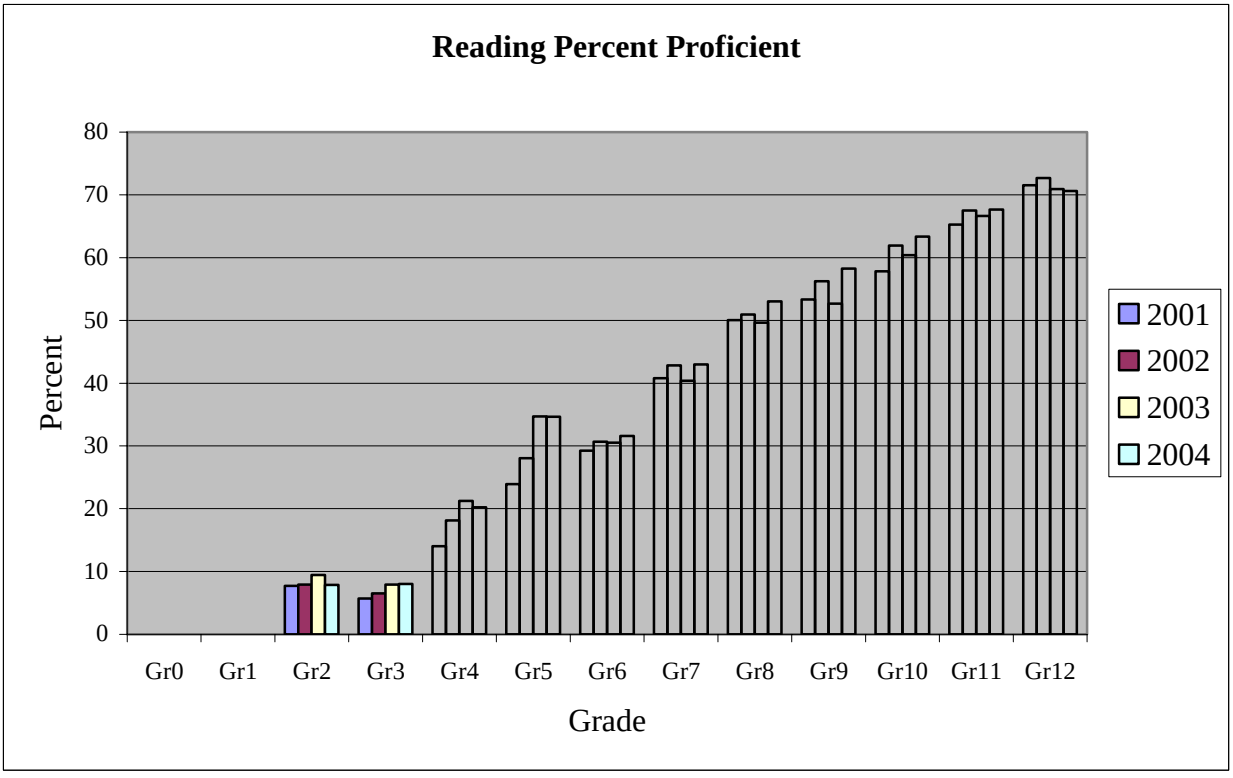


Figure 3 Writing Percent Proficient, Based on Annual Assessment Data

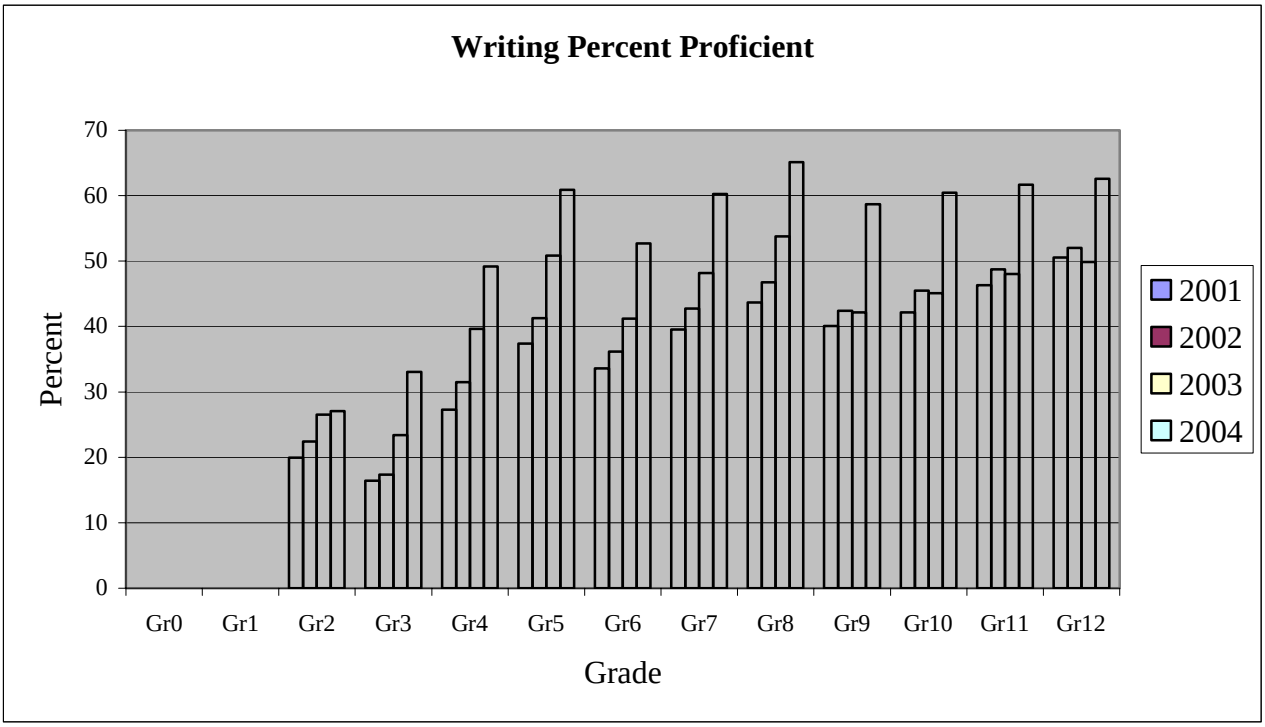
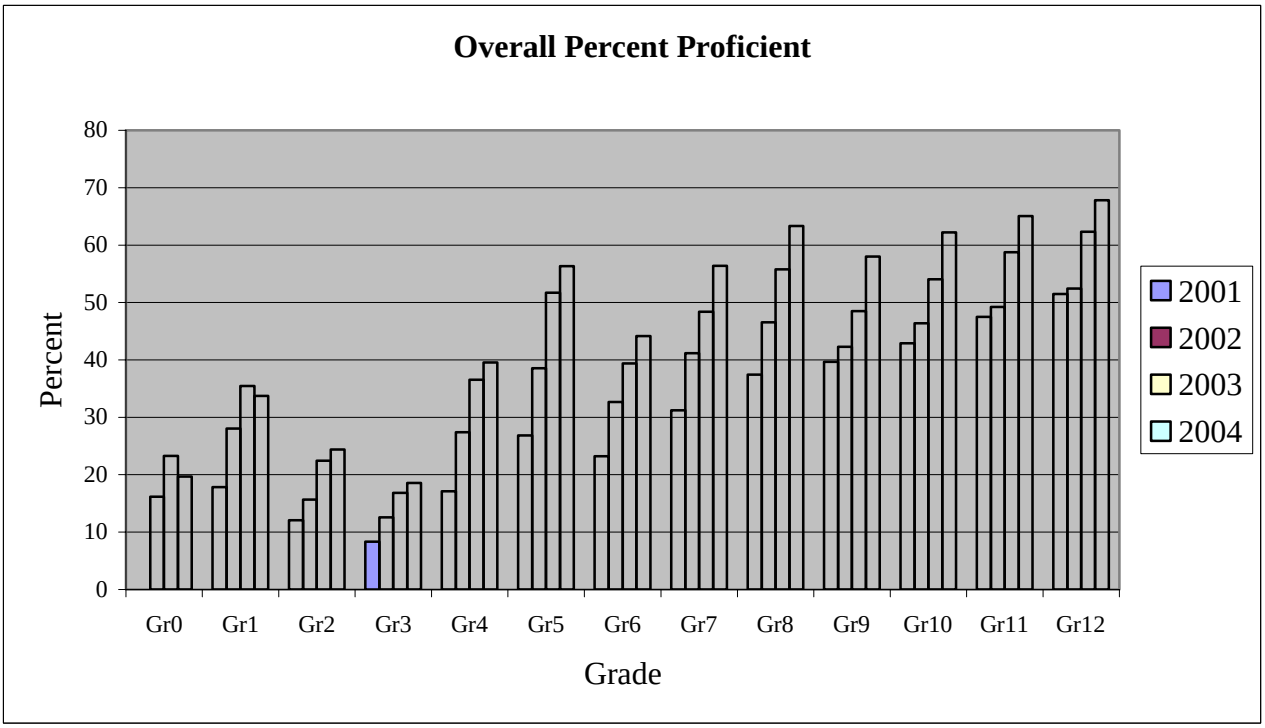


Figure 4 Overall Percent Proficient, Based on Annual Assessment Data



References

Burket, G. R. (1998). *PARDUX*. Monterey, CA: CTB/McGraw-Hill.

Burket, G. R. (1999). *WINFLUX*. Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill. (2001). *California English Language Development Bookmark Standard Setting Technical Report.* Monterey, CA: Author.

CTB/McGraw-Hill. (2002). *Technical Report for the California English Language Development Test (CELDT) 2000-2001.* Monterey, CA: Author.

CTB/McGraw-Hill. (2003). *California English Language Development Test 2002-2003 Form B Technical Report.* Monterey, CA: Author.

CTB/McGraw-Hill. (2004). *California English Language Development Test 2003-2004 Form C Technical Report.* Monterey, CA: Author.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., Patz, R. J. (1998). *The Bookmark Standard Setting Procedure: Methodology and Recent Implementations.* Paper presented at the 1998 annual meeting of the National Council on Measurement in Education.

Lewis, D.M., Mitzel, H.C., Green, D.R. (1996). *Standard Setting: A Bookmark Approach*. In D.R. Green (Chair), IRT-Based Standard-Setting Procedures Utilizing Behavioral Anchoring. Symposium presented at the 1996 Council of Chief State School Officers National Conference on Large Scale Assessment.

Rogosa, D. R. Misclassification in student performance levels. In *Technical Report: California Learning Assessment System 1993.* CTB/McGraw-Hill, May 1994.