BEST Literacy™

Technical Report

JUNE 2008

BEST Literacy™

WWW.CAL.ORG/BESTLITERACY

CAL CENTER FOR APPLIED LINGUISTICS

BEST Literacy™

Technical Report

CAL CENTER FOR APPLIED LINGUISTICS

June 2008

BEST Literacy Technical Report

June 2008

Writing: Dorry M. Kenyon, Ph.D., and Jeong Ran (Willow) Ryu
Editing: Carolyn Fidelman, Ph.D., and Jeannie Rennie
Proofreading: Amy Fitch
Review: Akihito Kamata, Ph.D., and Stephen Sireci, Ph.D.
Coordination: Daniel Lieberson
Design: Based on an original design by Falls River Arts
Layout: Ellipse Design
Printing: Balmar, Inc.

Recommended reference citation: Center for Applied Linguistics. (2008). *BEST Literacy Technical Report*. Washington, DC: Author.

Printed in the United States of America

Contents

1. Introduction 1

1.1 Overview 1

1.2 Purpose of the *Technical Report* 1

2. Description of *BEST Literacy* 2

2.1 Purpose of the Test 2

2.2 Test Forms 2

2.3 Test Scoring 3

2.4 Interpreting *BEST Literacy* Scores 3

3. The *BEST*: Predecessor of *BEST Literacy* 4

3.1 Relationship of *BEST Literacy* to the *BEST* 4

3.2 Development of the *BEST Literacy* Skills Section (1981-1984) 4

3.2.1 Development of Test Specifications for the *BEST* 4

3.2.2 Test Preparation, Field Testing, and Development of Final *BEST* Forms (1983-1984) 6

3.3 Reliability and Validity Research on the *BEST Literacy* Skills Section 6

3.3.1 Estimates of Internal Consistency 6

3.3.2 Estimates of Interrater Reliability 7

3.3.3 Validity 7

3.4 Development of *BEST Literacy* 8

4. *BEST Literacy* Field Test 9

4.1 Purpose of the Field Test 9

4.2 Procedures Used in the Field Test 9

4.2.1 Programs and Students 9

4.2.2 Materials 10

4.2.3 Administration Procedures 11

4.2.4 Scoring Procedures 11

4.3 Student Results 12

4.3.1 Readings 13

4.3.2 Writing 18

5. *BEST Literacy* Equating Study 23

5.1 The Measurement Model 23

5.2 Procedures 24

5.2.1 Identification of Common Item Linkages 24

5.2.2 Concurrent Calibration 29

5.2.3 Final Equating 29

5.3 Results for Reading 29

5.3.1 Concurrent Calibration of Reading 29

5.3.2 Equating of Reading Forms 34

Contents continued...

5.4 Results for Writing 51

5.4.1 Concurrent Calibration of Writing. 51

5.4.2 Equating of Writing Forms. 53

5.5 Results for Total Scores 69

6. Development of Final Forms 70

6.1 Development of Final Updated Test Forms (2006) 70

6.1.1 Final Writing Form New B 71

6.1.2 Final Writing Form New C 73

6.1.3 Final Writing Form New D 75

6.2 Comparison of Final Writing Forms. 77

6.3 Item Statistics for Final *BEST Literacy* Test Forms. 78

7. *BEST Literacy* Standard Setting Study 79

7.1 Introduction 79

7.2 Purpose of the *BEST Literacy* Standard Setting Study 79

7.3 Procedure 80

7.3.1 Judges 80

7.3.2 General Procedures 82

7.3.3 The Rating Process 83

7.3.4 Materials 84

7.4 Analyses 86

7.5 Results 89

7.5.1 NRS Study 89

7.5.2 SPL Study 92

7.6 Cross-Validation of the Standard Setting Study 93

8. Reliability of *BEST Literacy* 97

8.1 Reliability 97

8.1.1 Estimates of Internal Consistency. 97

8.1.2 Estimates of Interrater Reliability 97

8.2 Measurement Precision 99

8.2.1 Reading 99

8.2.2 Writing. 102

8.3 Consistency and Accuracy of Classification 105

9. Validity of *BEST Literacy* 109
9.1 Establishing Validity 109
9.2 Evidence from Test Content 109
9.3 Evidence from Data Fit 109
9.4 Evidence From Teacher Judgments 109
9.5 Evidence from Intersubscale Correlations 110

References 111

APPENDIX A: Comparison of Old and New Forms B and C 112

APPENDIX B: Complete Item Data for *BEST Literacy* Forms B, C, and D 119

This page
intentionally
left blank

1. Introduction

1.1 Overview

BEST Literacy, developed by the Center for Applied Linguistics (CAL), is a competency-based assessment that measures functional literacy skills (reading and writing) of adult English language learners who are 16 years and older and are enrolled in an educational program within the United States. It has three parallel forms that measure survival-level skills and work-related language commonly used in the United States. For comprehensive information about administering and scoring the three forms (B, C, and D) of *BEST Literacy*, see the *BEST Literacy Test Manual* (Center for Applied Linguistics, 2008).

1.2 Purpose of the *Technical Report*

The *Technical Report for BEST Literacy* is intended for test administrators, test scorers, program administrators, and other interested individuals. It describes the development of *BEST Literacy*, including three research studies pertaining to its field testing, equating, and standard setting. It concludes with a discussion of the evidence for its reliability and validity. The report is divided into nine sections:

- Section 1: Introduction
- Section 2: Description of *BEST Literacy*
- Section 3: The *BEST*: Predecessor of *BEST Literacy*
- Section 4: *BEST Literacy* Field Test
- Section 5: *BEST Literacy* Equating Study
- Section 6: Development of Final Forms
- Section 7: *BEST Literacy* Standard Setting Study
- Section 8: Reliability of *BEST Literacy*
- Section 9: Validity of *BEST Literacy*

2. Description of *BEST Literacy*

2.1 Purpose of the Test

BEST Literacy is a competency-based assessment that uses a variety of functional literacy tasks to measure adult English language learners' ability to read and write in English in authentic situations in the United States. The results of *BEST Literacy* can be used by educational programs for placement decisions, assessment of student progress, diagnosis of student strengths and weaknesses, and program evaluation.

2.2 Test Forms

BEST Literacy has three parallel forms (Forms B, C, and D). The three forms are equated for difficulty and are parallel in structure and format. Some items are identical across the three forms.

Each form is comprised of 11 parts: 7 that assess reading (49 items) and 4 that assess writing (19 items). A variety of item types are found across the different parts of the test. Reading is measured by assessment tasks such as finding and circling an answer, finding and copying an answer, circling answers in multiple-choice rational-deletion cloze reading passages, and answering multiple-choice questions following reading passages. Writing is assessed through tasks such as completing a form with personal information, completing a rent check and an envelope, and writing short personal notes. Table 1 provides a detailed description of the *BEST Literacy* test format and task types.

Table 1
Format of *BEST Literacy*

Part	Domain	Number of Items	Possible Points*	Item Type	Topic
1	Writing	10	10	Complete a form	Personal information
2	Reading	4	4	Find and circle	Calendar
3	Reading	3	3	Find and copy	Food labels
4	Reading	4	4	Find and copy	Clothing labels
5	Writing	5	5	Complete a form	Rent check
6	Writing	2	4	Complete an envelope	Envelope
7	Reading	2	2	Find and circle	Telephone directory
8	Reading	3	3	Find and circle	Train schedule
9	Reading	15	15	3-option cloze	Reading passages
10	Reading	18	18	4-option multiple choice	Ads, signs, notices, etc.
11	Writing	2	10	Extended response	Personal notes

*Total possible points: reading items = 49; writing items = 29; maximum combined total = 78

Part 1 is a personal information form to be completed by the examinee. For each piece of information requested (e.g., name, age), there is a blank line for the examinee to write a response. Part 2 shows a year of monthly calendars and instructs the examinee to find and circle dates that are written out above the calendars. Parts 3 and 4 present product labels of food and clothing items and require the examinee to read a question, locate specific information on the labels (such as price or size), and write that information in the blank provided. Part 5 directs the examinee to fill out a personal check; Part 6 requires the examinee to write the addresses of both the sender and the intended recipient on a standard envelope. Part 7 presents a page from a phone book. The examinee is instructed to find and circle the phone numbers of specific individuals or businesses on the page. Part 8 presents a train schedule and requires the examinee to find and circle information in the train schedule in response to questions. Part 9 presents two short reading texts containing 15 three-option cloze items. The examinee is asked to circle the correct word in each item. Part 10 is comprised of 18 four-option multiple-choice items, with two or three items grouped together with a common stimulus such as a newspaper ad or a medicine label. The examinee reads each question and marks his or her answer in the booklet. Part 11 presents the examinee with two short personal note-writing tasks. Throughout the test, examinees mark all of their answers directly in their individual test booklet.

Each correct reading item is awarded one point; a total of 49 points can be awarded for reading. About 67% of the reading items (33 out of 49) are selected response items, found in Parts 9 and 10 of the test. For writing, each correct item on the personal information form and rent check is awarded one point, each part of the envelope written correctly is awarded two points, and each personal note can be awarded a maximum of five points, for a total of 29 possible points for writing. For further information about the format of *BEST Literacy*, see the *BEST Literacy Test Manual*.

2.3 Test Scoring

BEST Literacy is scored locally by the test administrators or by others qualified to score this test. Scoring instructions for each item can be found in the *BEST Literacy Test Manual*. In addition to the test booklet, scorers need a separate scoring sheet that contains answers to the questions in the test booklet. Raw scores for the reading and writing sections are totaled separately, then each total is converted to a scale score. The two scale scores are then summed to form the total scale score. The maximum possible scale score is 49 for reading, 29 for writing, and 78 total.

2.4 Interpreting *BEST Literacy* Scores

Performances on *BEST Literacy* are interpreted in terms of two sets of proficiency-level descriptors designed specifically for adult learners of English. Historically, the descriptors of the Student Performance Levels (SPLs) have been used to interpret performances on the literacy skills section of the *Basic English Skills Test (BEST)*, of which *BEST Literacy* represents an update. The SPLs may still be used to understand performances on *BEST Literacy*. Performances on *BEST Literacy* are also interpreted according to the descriptors contained in the educational functioning levels of the National Reporting System (NRS). Full information on interpreting *BEST Literacy* scores, including the two sets of descriptors, can be found in the *BEST Literacy Test Manual*. Section 7 of this report provides detailed information on how standards were set to relate performances on *BEST Literacy* to the SPL and NRS descriptors.

3. The *BEST*: Predecessor of *BEST Literacy*

3.1 Relationship of *BEST Literacy* to the *BEST*

BEST Literacy is built upon the *Basic English Skills Test*, or the *BEST*, which has been used in adult education programs across the United States as a reliable assessment of adult English language proficiency since the early 1980s. The *BEST* consisted of two sections: an oral interview section and a literacy skills section. *BEST Literacy* is an updated form of the literacy skills section of the *BEST*.

The *BEST* was developed as part of a cooperative venture among English as a second language (ESL) teachers, administrators, and test developers, with funding from the Office of Refugee Resettlement (ORR), U.S. Department of Health and Human Services. The *BEST* was developed as the assessment component of the new functionally based ESL curricula and materials created at that time by the Mainstream English Language Training (MELT) project. As such, the *BEST* was one of the first standardized criterion-referenced tests designed to measure the functional literacy skills of English language learners. The first form (Form A) appeared in 1982. In 1984, Form A was retired following the development of three new parallel forms, known as B, C, and D. Unlike Form A, the new forms included some work-related items.

In 2003, following a 3-year project, the oral interview section of the *BEST* was replaced by *BEST Plus*, although the *BEST* oral interview section continued to be available until October 2006. In 2006, CAL updated the literacy skills section of the *BEST*. The goal was to maintain the essential test and item properties of the *BEST* as much as possible, while bringing the look and feel of the test up to date with contemporary information, graphics, and photographs. This updated version is now available as *BEST Literacy*.

In the development of *BEST Literacy*, all items in the *BEST* literacy skills section were examined, and many were slightly modified. After successful nationwide field testing and analyses, *BEST Literacy* Forms B, C, and D were made publicly available in October 2006. Section 4 of this report includes the results of the field test, which demonstrate that the modifications did not greatly affect the difficulty level of the items or the measurement qualities of the test.

Through technically sophisticated standard setting procedures, performances on *BEST Literacy* have been related to two sets of proficiency-level descriptors: the Student Performance Level (SPL) descriptors and the ESL educational functioning level descriptors of the National Reporting System (NRS) for Adult Education. The details and results of the standard setting procedures are presented in section 7 of this report.

3.2 Development of the *BEST Literacy* Skills Section (1981-1984)

Because *BEST Literacy* is an update of the literacy skills section of the *BEST*, we summarize here the development and available technical information on the literacy skills section of the *BEST*.

3.2.1 Development of Test Specifications for the *BEST*

A conference was held early in the test-planning stage to identify the topical and linguistic elements to be tested. The following topic areas were identified as crucial to survival-level competency in English for both the oral interview and literacy skills sections of the *BEST*: personal identification, greetings, kinship terms, health terms, parts of the body, numbers, time, money, shopping for food and clothing, housing, emergencies, directions, use of the telephone, completion of simple forms, writing checks, addressing envelopes, and other similar writing activities. Grammatical structures identified as necessary for the accomplishment of these tasks included the simple present and present progressive tenses, yes/no and wh— questions, and negation. Language functions given top priority included imparting information, searching for information, and seeking clarification. It was also decided that test items would be set in a U.S. context and reflect U.S. culture, because prospective examinees would represent diverse linguistic and ethnic backgrounds and because the primary purpose of the test was to measure competency in functioning in a variety of U.S. settings.

The original 1982 version (Form A) of the *BEST* contained all the above topic areas except use of the telephone. In field testing, this topic area was found to present inordinate logistical problems. However, telephone-related items, such as locating phone numbers in the directory, were included.

In developing the three new forms of the *BEST* in 1984, test developers reexamined the topics and tasks included in Form A. They added a new section on work-related language and incorporated suggestions from users of Form A into the new forms. Table 2 shows how topics and language skills intersected in the literacy skills section of the *BEST*. It also shows the specific tasks that correspond to each intersection.

Table 2
Topic Areas and Language Skills of the Literacy Skills Section of the *BEST*

Topic Areas	Skills
Greetings Personal Information Interpersonal Communication	• Fill out forms • Write a personal note to a friend to say thank you for a gift, extend an invitation to dinner, or explain why you cannot go to the friend's house for dinner
Time/Numbers	• Locate dates on calendar • Find telephone numbers in a directory • Read train schedules • Read store hours • Read a bus notice with times and fares • Write date of birth on form • Write date, street number, and zip code on forms
Money/Shopping for Food and Clothing	• Read the price on clothing labels • Read the price per pound and other information on food labels • Read the price, size, etc., on clothing labels • Write how much items cost per package or per pound
Health	• Read a medical appointment card • Read a prescription medicine label • Read a notice from a clinic • Read a letter from a hospital
Emergencies/Safety	• Read excerpts from a driver's manual • Read instructions in a telephone directory on how to use 911
Housing	• Write a note to a landlord about fixing a problem in the apartment • Read a newspaper ad for an apartment • Fill out a rent check • Address an envelope to the landlord • Write a note to the landlord to explain why the rent payment is late
General Information	• Comprehend general reading materials (e.g., newspaper articles, school notices)
Employment/ Training	• Read a job want ad • Write a note to a teacher explaining an absence from class

BEST Literacy, as an update of the *BEST* literacy skills section, reflects the test specifications presented in Table 2.

3.2.2 Test Preparation, Field Testing, and Development of Final *BEST* Forms (1983-1984)

On the basis of the content specifications described above, field-test versions of *BEST* Forms B, C, and D were developed and field-tested over a 6-month period, from December 1983 to June 1984, at the ORR MELT centers. For the field test, all tests were administered and scored by ESL teachers and program supervisors on the basis of detailed written instructions and procedures presented by CAL at tester training sessions. A total of 632 students participated in the field testing of the *BEST* literacy skills section. Native languages represented in the sample included Vietnamese, Hmong, Lao, Cambodian/Khmer, Chinese, Spanish, and Polish, among others (see Clark & Grognet, 1985).

As in most field tests, the total number of items included in the field-test versions was intentionally greater than the number to be included in the operational test. This provided the opportunity to select items for the final version on the basis of statistical performance and other information gathered during the field test.

Selection of items for inclusion in the operational Forms B, C, and D of the *BEST* literacy skills section was based primarily on the statistical results of an item analysis. Level of difficulty and discrimination (r-biserial coefficients) were examined for each test item. The results of the item analysis showed that very few field-test items needed to be eliminated; however, a number of items were deleted to shorten the test. Comments from MELT field-test examiners were also taken into consideration in selecting items to be included in the operational forms of the *BEST*.

3.3 Reliability and Validity Research on the *BEST Literacy* Skills Section

3.3.1 Estimates of Internal Consistency

Based on the field-test data, reliability estimates for internal consistency were calculated for all the items included in the 1983–1984 field-test version of the *BEST* literacy skills section. These are presented in Table 3. High internal reliability estimates are desirable and show that the items for which the estimate is made are consistently testing the same skill. Reliability estimates are provided for the two subscales of each test form as well as for the total test. (Note that incomplete test responses were not included in this analysis.)

Table 3
Internal Consistency Estimates for 1983–1984 Field-Test Version of the *BEST Literacy* Skills Section

BEST Literacy **Skills Section**	**Form B** (n = 207)	**Form C** (n = 204)	**Form D** (n = 208)
Reading	0.957	0.968	0.956
Writing	0.899	0.909	0.903
Total	0.966	0.972	0.966

In this 1984 study, the internal consistency reliability estimates for the total score were very high: 0.966 for Form B, 0.972 for Form C, and 0.966 for Form D. In general, the demonstrated reliability estimates are quite high for a test like the *BEST*, which is composed mostly of free-response items.

3.3.2 Estimates of Interrater Reliability

To gather information on the interrater reliability of the test (i.e., how consistently different raters score the same examinee), 49 administrations of the *BEST* literacy skills section were scored by the same two raters across the forms in the 1984 study. Based on the scores of those two raters, interrater reliabilities for the final operational forms in terms of Pearson correlation are shown in Table 4.

Table 4
Interrater Reliabilities for the *BEST* (1984)

BEST Literacy **Skills Section**	**Form B** (n = 14)	**Form C** (n = 16)	**Form D** (n = 19)
Reading	0.999	0.999	0.999
Writing	0.999	0.982	0.984

The numbers in Table 4 show that the literacy skills section of the *BEST* could be scored quite consistently by different raters.

3.3.3 Validity

Validity refers to the degree to which evidence and theory support inferences made and actions taken concerning examinees on the basis of their test scores (*Standards for Educational and Psychological Testing, 1999*). The literacy skills section of the *BEST* purported to measure the survival-level competency in English of adult English language learners. A careful review of the content of the literacy skills section indicates that its content is quite similar to the real-life language-use tasks it aims to test (e.g., reading a label for information, writing a check).

The *BEST* was most often used to place students within programs and to measure their progress. In 1983-1984, in a study carried out with the assistance of MELT project training staff, *BEST* test scores were correlated with pre-assigned group ratings of the examinees' language proficiency that had been made by MELT project training staff. The examinees had already been placed in MELT language training program levels through the use of measures other than the *BEST*. Each of these program levels was assigned an SPL by relating the program levels to the SPL descriptions. Individual students were then assigned the SPL that corresponded to the program level in which they were enrolled; that is, students were assigned an SPL not according to anyone's assessment of their individual English skills or proficiency, but rather according to the instructional level of the class in which they were currently enrolled.

The obtained mean scores and standard deviations on the literacy skills section of the *BEST* for students in MELT SPLs 0 through 7 were calculated. (This procedure was also done, separately, for the oral interview section of the *BEST*.) For future placement and other planning purposes, *BEST* score ranges for each level were then derived from these data. This was done through a modified *centour* analysis, in which the cumulative frequency distributions of each performance level were compared. Each subscale score was assigned to an SPL according to the level for which that score was most typical. That is, the SPL for which the cumulative frequency was closest to 50% (the median) was selected as the most appropriate level to be predicted from that *BEST* score. The descriptive data and final SPL score ranges are shown in Table 5.

Table 5
***BEST* Scale Statistics (Reading and Writing) for Students at Student Performance Levels 0 to 7**

SPL	N	Mean	S.D.	*BEST* Score Ranges (Reading and Writing Total)
0	6	13.0	15.3	0–2
1	28	7.3	10.0	3–7
2	78	19.0	13.3	8–21
3	129	26.9	13.1	22–35
4	180	41.2	13.0	36–46
5	107	48.6	11.5	47–53
6	85	58.1	11.7	54–65
7	19	63.4	9.0	66 +

Source: *BEST Test Manual* (Center for Applied Linguistics, 1984)

With the exception of the 6 cases identified as SPL 0, Table 5 shows clearly that the mean total score of members of each SPL rises with each increasing SPL. The average score for the 28 students in SPL 1 was 7.3, while for SPL 2 it was 19.0, for level 3 it was 26.9, and so on. This analysis provides evidence that the *BEST* literacy skills section could be used to place students into hierarchical proficiency levels. (For additional technical information on the *BEST* literacy skills section Forms B, C, and D, see the *BEST Test Manual* [Center for Applied Linguistics, 1984].)

3.4 Development of *BEST Literacy*

In the fall of 2005, CAL began updating the *BEST* literacy skills section. This update is known as *BEST Literacy*. The goals of the project were to retain the psychometric characteristics (such as the difficulty of items and interrater reliability) of the *BEST* literacy skills section while bringing the format up to date with contemporary information, graphics, and photographs. Input and suggestions were solicited from long-time *BEST* users, and a team of experts in language testing and adult ESL education examined every item on the three forms of the *BEST* literacy skills section to determine if any aspects of the item needed to be updated or revised. Careful attention was given to improving the clarity of the layout and the quality of the graphics and photos used. At the same time, CAL staff with psychometric expertise advised the group as to the types of changes that might make the test perform differently and thus disrupt the comparability of the old and new forms. In this manner, Forms B, C, and D of the literacy skills section of the *BEST* became the three new Forms B, C, and D of *BEST Literacy*. The new forms were prepared for field testing in the spring of 2006.

The remainder of this report presents technical information on *BEST Literacy* based on data from the 2006 field test and the subsequent standard setting study. While comparisons with the technical information from the *BEST* literacy skills section (given above) are made in this report to demonstrate the relationship between the two tests, much more technical information is provided for *BEST Literacy* than for the original test.

4. *BEST Literacy* Field Test

4.1 Purpose of the Field Test

There were two main goals for the field test of *BEST Literacy*. The first was to examine the comparability between the new (updated) forms and the old forms. We wanted to verify that neither the difficulty of items nor the measurement precision of the old test forms was significantly changed by the updating process. Second, we wanted to collect data to do more thorough empirical analyses on the test than had been done in the past. Since the *BEST* appeared in the early 1980s, more technically sophisticated procedures for analyzing test data, equating test forms, and setting standards have been developed and widely adopted. We wanted to apply these to *BEST Literacy* to help users better understand the psychometric properties of the test and to provide stronger psychometric support for its use.

4.2 Procedures Used in the Field Test

4.2.1 Programs and Students

To recruit students for the study, CAL sent invitations to several adult education programs across the United States that were currently using the literacy skills section of the *BEST*. Each participating program was to test approximately 48 students, making certain that the sample included a balanced number of students from the program's lowest to highest instructional levels.

In the end, nine programs from seven states (Arizona, Colorado, Illinois, New Mexico, Ohio, Tennessee, and Texas) participated in the field test, which included 407 students. Table 6 presents complete information on the participating programs and students. Students and programs that participated in the study signed non-disclosure, consent, and voluntary participation forms.

Table 6
Programs and Students Participating in the *BEST Literacy* Field Test

State	Program	Number of participating Students
Arizona	Cochise College	48
	Mesa Public Schools	46
Colorado	Links to Literacy	16
	Spring Institute	49
Illinois	College of Du Page	49
New Mexico	Roosevelt Elementary Family Center	48
Ohio	Columbus Literacy Council	49
Tennessee	Center for Literacy	54
Texas	San Jacinto Adult Learning Center	48
Total		407

In keeping with current demographics in adult ESL programs in the United States, a wide variety of native languages were represented among the field-test students. Table 7 shows the number and percentage of students from each language background. As can be seen, the vast majority of students (76.4%) were native speakers of Spanish, as is typical in many areas of the United States.

Table 7
Student Participants by Native Language

	Number	Percent
Albanian	3	0.7
Amharic	8	2.0
Arabic	7	1.7
Bulgarian	2	0.5
Chinese	11	2.7
Creole	2	0.5
Farsi	3	0.7
French	3	0.7
Gujarati	1	0.2
Hausa	1	0.2
Hindi	3	0.7
Hungarian	1	0.2
Japanese	3	0.7
Korean	5	1.2
Latvian	1	0.2
Lithuanian	2	0.5
Oluf	1	0.2
Oromo	3	0.7
Panjabe	1	0.2
Patois	1	0.2
Persian	1	0.2
Portuguese	2	0.5
Russian	4	1.0
Slovak	1	0.2
Somali	17	4.2
Spanish	311	76.4
Teluge	1	0.2
Ukrainian	2	0.5
Urdu	1	0.2
Vietnamese	5	1.2
Total	407	100.0

4.2.2 Materials

Five test forms were administered during the field test: the three updated forms (i.e., *BEST Literacy* Forms B, C, and D) and the two older operational forms of the literacy skills section of the *BEST*, Forms B and C. (NOTE: Form D had never been operationally available from CAL.) In this section of the report, the updated test forms (i.e., *BEST Literacy* forms) will be designated as *New B*, *New C*, and *New D*. The older forms from the *BEST* literacy skills section will be referred to as *Old B* and *Old C*.

4.2.3 Administration Procedures

All tests were administered in locally scheduled test sessions in March and April of 2006 by local ESL teachers and administrators following the general directions for the literacy skills section of the *BEST*. In other words, no changes were made in test administration procedures from the then-current *BEST*, with which the programs were already familiar. These procedures include preparing the testing rooms, verifying the identity of each examinee, providing one hour to complete the test, and proctoring the test throughout the administration time.

However, because the goal of the field test was to compare performance on the two forms of the operational *BEST* literacy skills section (Old B and Old C) with performance on the three updated forms of *BEST Literacy* (New B, New C, and New D), these five forms were delivered to the programs in a spiraled order; that is, the booklets were in sets of Old B, Old C, New D, New B, New C, and New D. (Note: Because Form D had never been available from CAL, we wanted to ensure that Form D was administered to twice as many students in the field test as each of the other forms.) Administrators were asked to pass out the booklets in this order randomly among their students in each testing session. Thus, instead of all the students in one room working on the same version of the test booklet, five different versions of the test booklet were randomly distributed among students in the same room. This methodology ensured random assignment of test forms to programs and students.

4.2.4 Scoring Procedures

After the test was administered, the test booklets were shipped back to CAL. CAL then conducted a 2-day scoring session on May 11 and 12, 2006. The scoring session was led by Dr. Dorry Kenyon, director of the Language Testing Division (LTD) at CAL, assisted by Willow Ryu, an LTD research assistant. Three CAL staff members and two outside experienced adult ESL professionals served as scorers. Table 8 shows the names of the scorers.

Table 8
Scorers' Names and Affiliations

Name	Affiliation
Kathleen Jelinek	Georgetown University
Mary Lidinsky	Baltimore City Community College
Daniel Lieberson	CAL
Michelle Ueland	CAL
Bryan Woerner	CAL

The first day started with orientation to *BEST Literacy* and the field test. Dr. Kenyon provided training on scoring the reading section of the test using the scoring guide in the *BEST Test Manual*. Issues that arose needing further clarification during this part of the training were thoroughly documented in order to improve the new scoring guide for *BEST Literacy*.

The test booklets were organized by test form but then randomized and put in individual envelopes and distributed to the scorers. The scorers, working independently, began scoring the test booklets, marking their scores on specially prepared scannable scoring sheets. They were able at any time to ask questions or request clarification from the session leaders. Again, all questions and requests for clarification were thoroughly noted. Responses were shared by the session leaders with the entire group of scorers if they pertained to everyone. After each scorer finished scoring his or her test booklets, he or she then independently scored a certain number of test booklets that other scorers had already scored for the purpose of the interrater reliability study.

The second day started with training on how to score the writing section of the test, with a focus on how to score the note-writing tasks. For this purpose, CAL testing staff had revised the original writing rubric, having gone through all field-test responses, selecting anchor, training, and calibration papers to use for the scoring session and to include in the new *BEST Literacy Test Manual*. First, the revised writing rubric was reviewed and discussed. Then the anchor papers were distributed and reviewed together along with task-specific instructions for each of the six note-writing tasks. The scorers were then given the training set of papers to rate individually. The scores awarded were then discussed by the whole group. The comments and issues that were raised during this part of the training were noted to provide guidance in preparing the *BEST Literacy Test Manual*. After discussing the anchor papers and scoring and discussing the training papers, the scorers were asked to rate a calibration set of 10 papers. Scorers who met the standard by assigning the same rating to 8 or more papers as had been assigned earlier by CAL testing staff were then asked to start scoring the writing sections from the field test. For scorers who did not meet this standard, an individual training session followed, and the scorer was asked to rate another calibration set of 10 papers. Four of the five scorers passed on the first try. The remaining scorer passed after the second set of calibration papers. The same procedure as the first day was followed for recording scores. The questions needing clarification were noted, and double scoring of the writing section was conducted for the interrater reliability study.

4.3 Student Results

After all of the tests were scored, the scannable scoring sheets were electronically scanned at CAL to create the main database of test scores. The SPSS software program was used for statistical data analysis. In Tables 9 through 13 and Figures 1 through 5, we present the results on each reading form in terms of raw scores. In Tables 15 through 19 and Figures 6 through 10, we present the results on each writing form in terms of raw scores. For each form, a figure shows the raw score distribution and a table presents the number of students who took the form, minimum and maximum observed raw scores, the mean raw score, and the standard deviation of the raw scores. The total possible score on the reading section was 49 and on the writing section 29.

4.3.1 Readings

4.3.1.1 Reading Form Old B

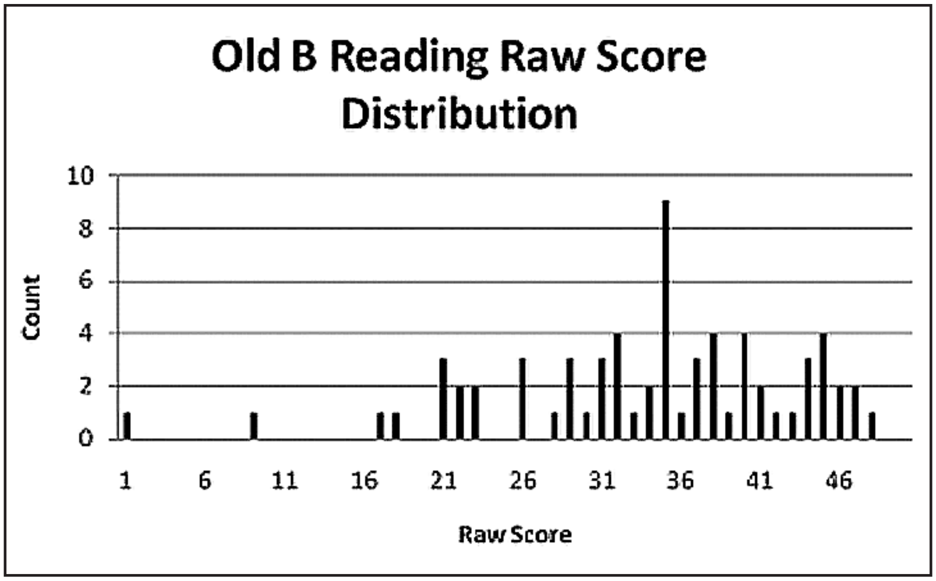

Figure 1. Old B Reading Raw Score Distribution

Table 9
Old B Reading Raw Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
67	0	47	32.87	9.36

4.3.1.2 Reading Form New B

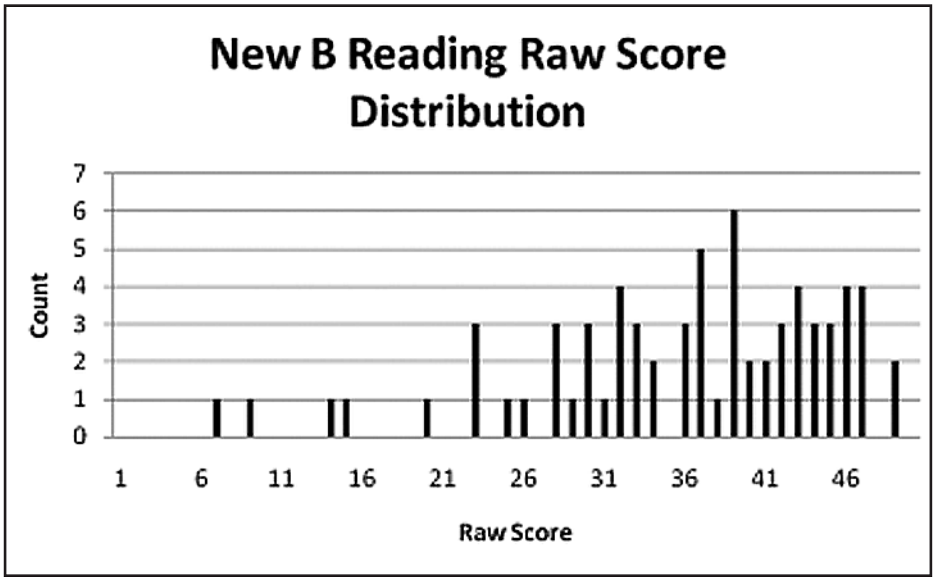

Figure 2. New B Reading Raw Score Distribution

Table 10
New B Reading Raw Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
69	6	48	34.88	9.44

4.3.1.3 Reading Form Old C

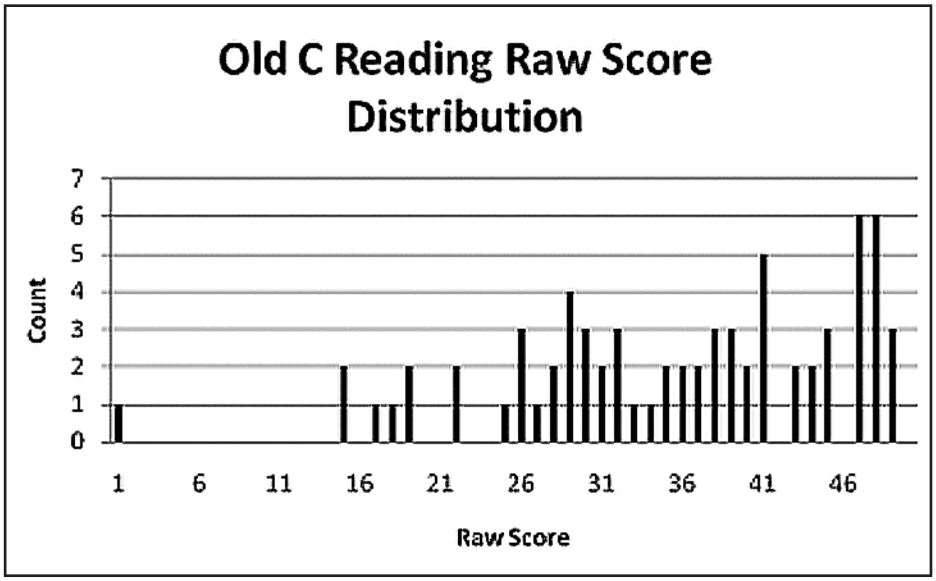

Figure 3. Old C Reading Raw Score Distribution

Table 11
Old C Reading Raw Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
71	0	48	34.54	10.39

4.3.1.4 Reading Form New C

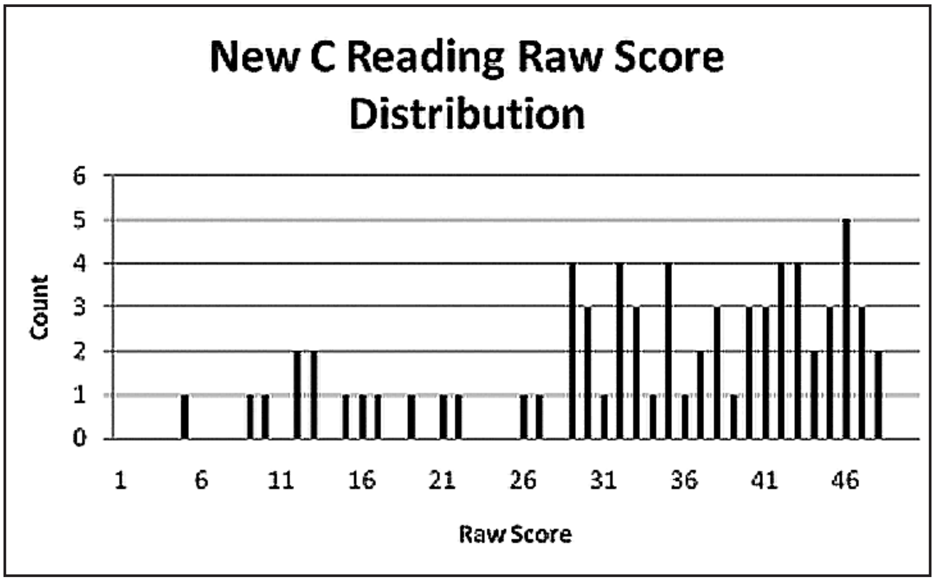

Figure 4. New C Reading Raw Score Distribution

Table 12
New C Reading Raw Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
71	4	47	32.96	11.22

4.3.1.5 Reading Form New D

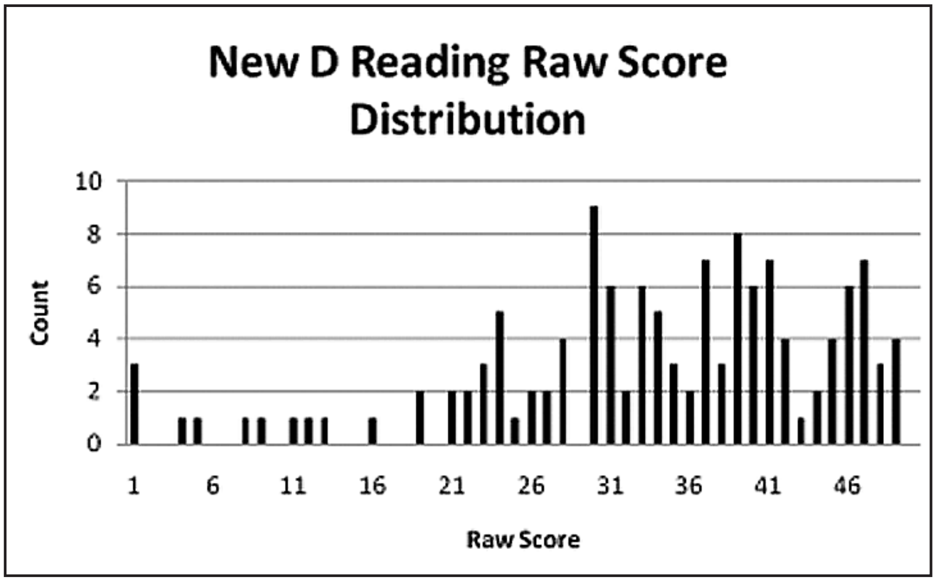

Figure 5. New D Reading Raw Score Distribution

Table 13
New D Reading Raw Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
129	0	48	32.63	11.22

4.3.1.6 Summary Across All Forms (Reading)

Across the five test forms, Table 14 compares the number of students who took each form, the average raw score, and the standard deviation of the raw scores. The largest difference in means, in terms of raw scores, was between New D (32.63), with the lowest average score, and New B (34.88), with the highest. To examine whether these differences were statistically significant, a one-way ANOVA was run. The results showed that the differences in mean raw scores for reading among the five groups was not statistically significant, $F(4,402) = .800$, $p = .526$. That means that the differences in the raw scores of the groups using each form were due to random error and not to genuine differences in abilities among the test takers in the five groups. This result suggests that the method of randomization of the test booklets used in the field test was successful. However, it must also be remembered that these performances in terms of raw scores are not equated to each other, so direct comparisons on the basis of raw scores are premature.

Table 14
Reading Raw Score Summary: Mean and Standard Deviation by Test Form

	Old B	New B	Old C	New C	New D
Number of students	67	69	71	71	129
Mean Score	32.87	34.88	34.54	32.96	32.63
Std. Deviation	9.358	9.438	10.391	11.218	11.218

4.3.2 Writing

4.3.2.1 Writing Form Old B

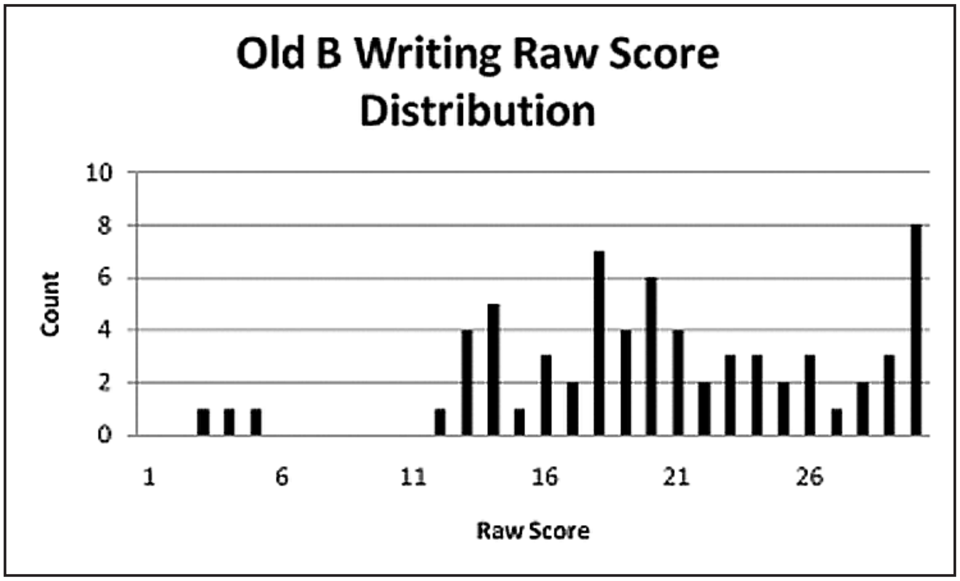

Figure 6. Old B Writing Raw Score Distribution

Table 15
Old B Writing Raw Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
67	2	29	19.48	6.49

4.3.2.2 Writing Form New B

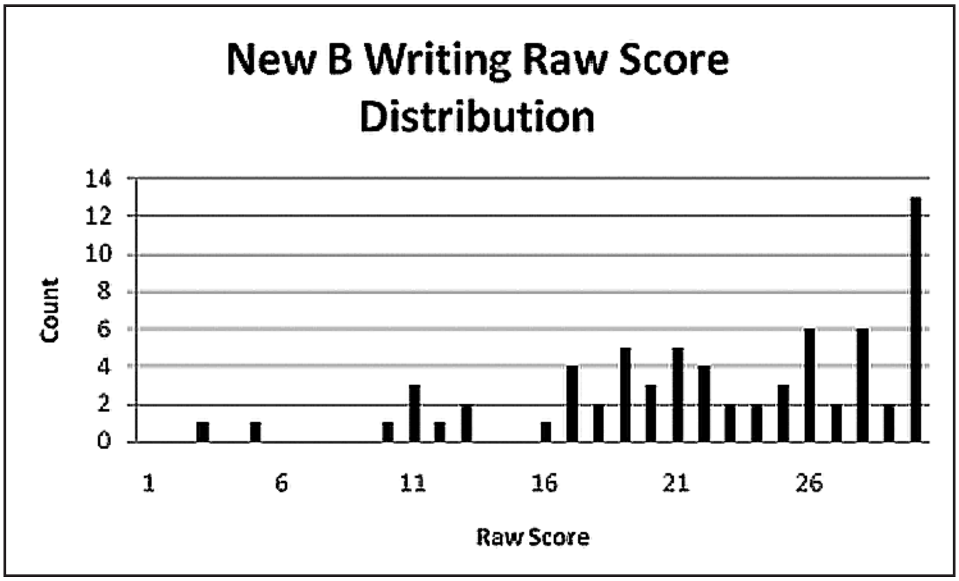

Figure 7. New B Writing Raw Score Distribution

Table 16
New B Writing Raw Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
69	2	29	21.49	6.60

4.3.2.3 Writing Form Old C

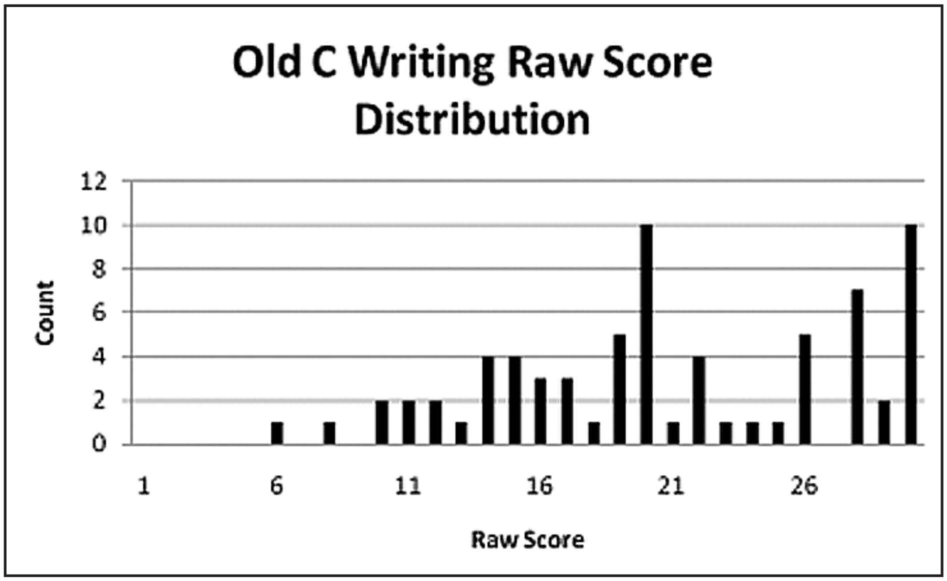

Figure 8. Old C Writing Raw Score Distribution

Table 17
Old C Writing Raw Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
71	5	29	19.93	6.55

4.3.2.4 Writing Form New C

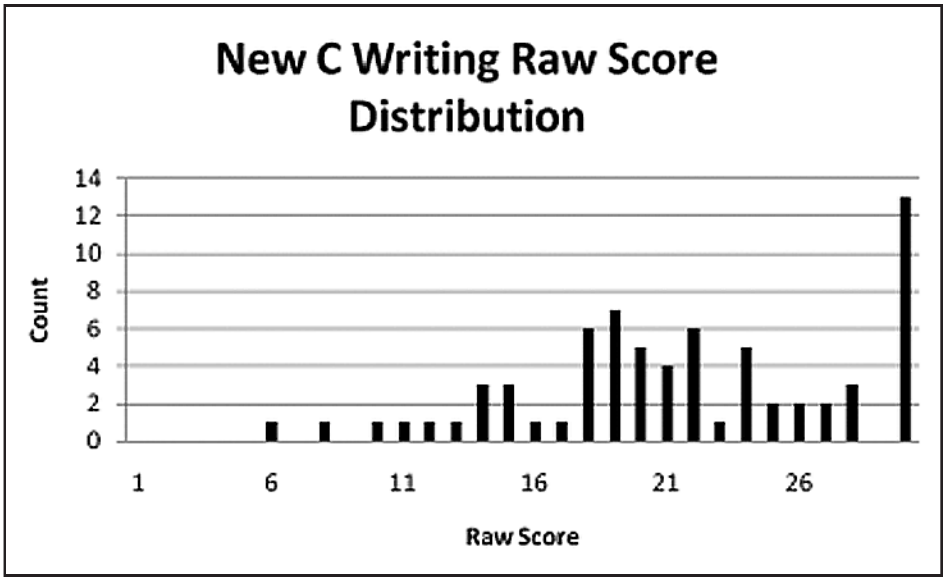

Figure 9. New C Writing Raw Score Distribution

Table 18
New C Writing Raw Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
70	5	29	20.57	6.05

4.3.2.5 Writing Form New D

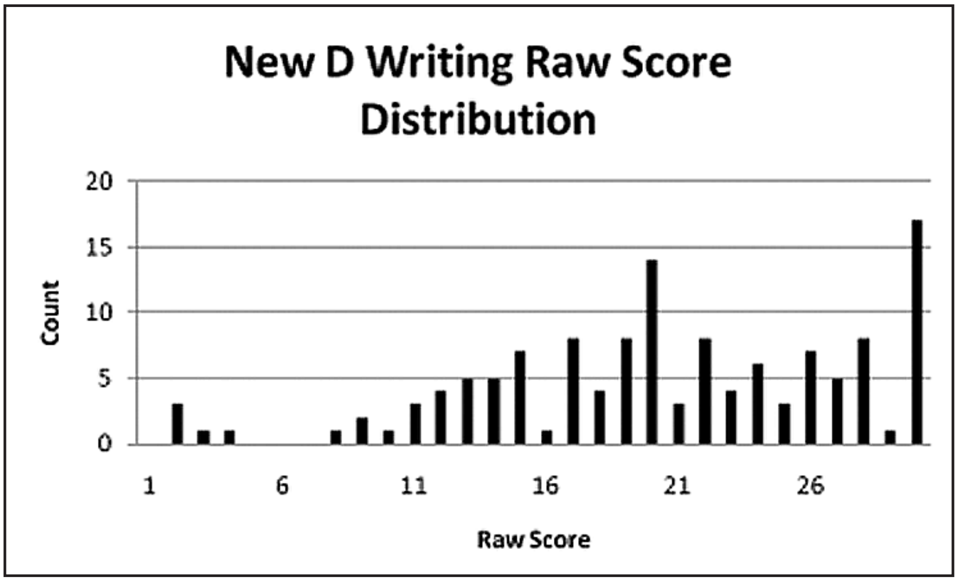

Figure 10. New D Writing Raw Score Distribution

Table 19
New D Writing Raw Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
130	1	29	19.43	6.94

4.3.2.6 Summary Across All Forms (Writing)

Across the five test forms, Table 20 compares the number of students who took each form, the average raw score, and the standard deviation of the raw scores. The largest difference in means, in terms of raw scores, was between New D (19.43), with the lowest average score, and New B (21.49), with the highest. These two forms had the lowest and highest average raw scores for reading as well. Again, to examine whether these differences were statistically significant, a one-way ANOVA was run. The results showed that the differences in mean raw scores for writing among the five groups was not statistically significant, $F(4,402) = 1.351$, $p = .250$. Again, this result suggests that the method of randomization of the test booklets used in the field test was successful, although again it must be remembered that these performances in terms of raw scores are not equated to each other, so direct comparisons on the basis of raw scores are premature.

Table 20
Writing Raw Score: Mean and Standard Deviation by Test Form

	Old B	New B	Old C	New C	New D
Number of people	67	69	71	70	130
Mean	19.48	21.49	19.93	20.57	19.43
Std. Deviation	6.493	6.599	6.549	6.054	6.945

5. *BEST Literacy* Equating Study

In order to examine the comparability of the old and new test forms and to enable users of the new forms to compare performances across Forms New B, New C, and New D, it was necessary to equate the five different test forms. Because the reading and writing scores are reported separately and then totaled, separate equating was conducted for reading and writing.

Using item response theory (IRT) for equating, the first step is to find an appropriate measurement model: that is, a mathematical model that can explain the observed data. In this study, the Rasch measurement model (1-parameter logistic model) was used. The second step is to use an approach to linking across separate test forms, generally through common people (i.e., the same people take two or more test forms) or through common items (i.e., the same items are common across two or more test forms). In this study, a common item approach was used.

5.1 The Measurement Model

The measurement model used for the equating of the *BEST Literacy* field test is the Rasch measurement model (Wright & Stone, 1979). Because all items in the reading section of *BEST Literacy* are scored as either correct or incorrect, the dichotomous Rasch model was used. Mathematically, this measurement model may be presented as

$$\log\left(\frac{P_{ni1}}{P_{ni0}}\right) = B_n - D_i$$

where

P_{ni1} = probability of a correct response by person *n* on item *i*
P_{ni0} = probability of an incorrect response by person *n* on item *i*
B_n = ability of person *n*
D_i = difficulty of item *i*

When the probability of a person getting a correct answer equals the probability of a person getting an incorrect answer (i.e., 50% probability of getting it right and 50% probability of getting it wrong), P_{ni1}/P_{ni0} is equal to 1. The log of 1 is 0. This is the point at which a person's ability equals the difficulty of an item. For example, a person whose ability is 1.56 on the Rasch logit scale encountering an item whose difficulty is 1.56 on the Rasch logit scale would have a 50% probability of answering that question correctly.

For the writing section, a mix of the dichotomous and polytomous Rasch models was used. Polytomous scoring applies to items that contain gradations of performance, such as a scale of 0 to 5. This approach combined the dichotomous Rasch model for the items scored correct or incorrect with the partial credit Rasch model for the note-writing tasks. Mathematically, the partial credit model can be represented as

$$\log\left(\frac{P_{nik}}{P_{nik-1}}\right) = B_n - D_i - F_{ik}$$

where

P_{nik} = probability of person *n* on task *i* receiving a rating at level *k* on the rating scale
P_{nik-1} = probability of person *n* on task *i* receiving a rating at level *k - 1* on the rating scale (i.e., the next lowest rating)
B_n = ability of person *n*
D_i = difficulty of task *i*
F_{ik} = calibration of step *k* on the rating scale used for scoring task *i*

All Rasch analyses were conducted using the Rasch measurement software program *Winsteps* (Linacre, 2006). When speaking of the measure of examinee ability, we use the term *ability measure* (rather than *theta*, used commonly when discussing models based on Item Response Theory). When speaking of the measure of item difficulty, we use the term *item difficulty measure* (rather than *b parameter*, used commonly when discussing models based on IRT). *Step measures* refer to the calibration of the steps in the Rasch partial credit model presented above. All three measures (ability, difficulty, and step) are expressed in terms of Rasch logits. For *BEST Literacy*, the logit ability measures are not converted into the reporting scale. Instead, the reporting scale is expressed in terms of the raw scores on the *BEST* Form B, which has been the standard for the scale score since the early 1980s. Nevertheless, these ability measures have been used for equating, as will be explained later in this section. These measures also appear in the discussion of the measurement precision in section 8.2.

5.2 Procedures

5.2.1 Identification of Common Item Linkages

Because all five test forms used in the field test had so many items in common, a common item approach to equating was used. Thus, the first step of the equating procedure was to identify common items across the test forms and to designate a unique item identifier for each of the common items. In considering items to be common, we were very conservative. That is, even when very small changes were made to older items (such as a slight change in price labels or a graphic change), the old and the new items were considered unique and were assigned different item identifications (IDs). Only if no change had been made to an item was it considered common across old and new forms.

The designation of common items was made by a group of three testing experts and two ESL experts at CAL. At the end of the process, there were three groups of common items:

- Items that were common across all five test forms (i.e., they were common items across the two forms of the *BEST* literacy skills section, were common items across the three forms of *BEST Literacy*, and were unchanged by the updating)
- Items that were common to all the new forms (i.e., they were identical on New B, New C, and New D) and common to the two old forms (i.e., they were identical on Old B and Old C) but were not common across the old and new forms (i.e., some changes were made to the items during updating)
- Items that were common to the old and new versions of a particular form (i.e., either Old B and New B or Old C and New C) but that were not common across different forms (e.g., Old B and Old C)

Unique items were those that were not common to Old B and Old C and that had undergone some type of updating between Old B and New B or Old C and New C (i.e., they were considered two unique items although the updating could have been very minor), and those items that appeared only on New D.

Tables 21a and 21c show the list of unique item IDs for each item that appeared on any of the five test forms. Table 21a is for reading items, and Table 21c is for writing items. The first column of each table shows the order of the items as they appear in the test booklet. Note that reading and writing are ordered independently of each other. The second column shows where in the booklet the item appears. For example, RP9I3 means that this is a reading item (R), that it is in part 9 (P9), and that it is the third item in that part (I3). Parts 2, 3, 4, 7, 8, 9, and 10 are reading; parts 1, 5, 6, and 11 are writing (see Table 1).

The next five columns in Tables 21a and 21c show the unique item IDs, which consist of three parts. The first part tells whether the item is a reading item or a writing item. The second part indicates the item location in the booklet. The third part indicates the booklets where the item appears, with Old B abbreviated as OB and New D abbreviated as ND, for example. Thus, WP1I4OBNBOCNCND means that this is a writing item (W), that it is in part 1 (P1), that it is the fourth item in that part (I4), and that it appears on all five forms (Old B, New B, Old C, New C, and New D). Thus, this item is common across all five forms. Again, it may be helpful to refer to Table 1 to understand Table 21 most clearly.

Tables 21b and 21d are graphic representations of Tables 21a and 21c. Within each row, cells that have the same shading indicate forms that share the same item, and those that have different shading have items that were, for equating purposes, considered to be different from one another.

Table 21a
Reading Item Identifiers (Text Version)

	Item	Old B	New B	Old C	New C	New D
1	RP2I1	RP2I1OBNBOCNCND	RP2I1OBNBOCNCND	RP2I1OBNBOCNCND	RP2I1OBNBOCNCND	RP2I1OBNBOCNCND
2	RP2I2	RP2I2OBNB	RP2I2OBNB	RP2I2OCNC	RP2I2OCNC	RP2I2ND
3	RP2I3	RP2I3OBOC	RP2I3NBNCND	RP2I3OBOC	RP2I3NBNCND	RP2I3NBNCND
4	RP2I4	RP2I4OB	RP2I4NB	RP2I4OC	RP2I4NC	RP2I4ND
5	RP3I1	RP3I1OB	RP3I1NB	RP3I1OC	RP3I1NC	RP3I1ND
6	RP3I2	RP3I2OB	RP3I2NB	RP3I2OC	RP3I2NC	RP3I2ND
7	RP3I3	RP3I3OBOC	RP3I3NBNCND	RP3I3OBOC	RP3I3NBNCND	RP3I3NBNCND
8	RP4I1	RP4I1OB	RP4I1NB	RP4I1OC	RP4I1NC	RP4I1ND
9	RP4I2	RP4I2OB	RP4I2NB	RP4I2OC	RP4I2NC	RP4I2ND
10	RP4I3	RP4I3OBOC	RP4I3NBNCND	RP4I3OBOC	RP4I3NBNCND	RP4I3NBNCND
11	RP4I4	RP4I4OBOC	RP4I4NBNCND	RP4I4OBOC	RP4I4NBNCND	RP4I4NBNCND
12	RP7I1	RP7I1OBOC	RP7I1NBNCND	RP7I1OBOC	RP7I1NBNCND	RP7I1NBNCND
13	RP7I2	RP7I2OBOC	RP7I2NBNCND	RP7I2OBOC	RP7I2NBNCND	RP7I2NBNCND
14	RP8I1	RP8I1OBOC	RP8I1NBNCND	RP8I1OBOC	RP8I1NBNCND	RP8I1NBNCND
15	RP8I2	RP8I2OB	RP8I2NB	RP8I2OC	RP8I2NC	RP8I2ND
16	RP8I3	RP8I3OB	RP8I3NB	RP8I3OC	RP8I3NC	RP8I3ND
17	RP9I1	RP9I1OBOC	RP9I1NBNCND	RP9I1OBOC	RP9I1NBNCND	RP9I1NBNCND
18	RP9I2	RP9I2OBOC	RP9I2NBNCND	RP9I2OBOC	RP9I2NBNCND	RP9I2NBNCND
19	RP9I3	RP9I3OBOC	RP9I3NBNCND	RP9I3OBOC	RP9I3NBNCND	RP9I3NBNCND
20	RP9I4	RP9I4OBOC	RP9I4NBNCND	RP9I4OBOC	RP9I4NBNCND	RP9I4NBNCND
21	RP9I5	RP9I5OB	RP9I5NB	RP9I5OC	RP9I5NC	RP9I5ND
22	RP9I6	RP9I6OB	RP9I6NB	RP9I6OC	RP9I6NC	RP9I6ND
23	RP9I7	RP9I7OB	RP9I7NB	RP9I7OC	RP9I7NC	RP9I7ND
24	RP9I8	RP9I8OB	RP9I8NB	RP9I8OC	RP9I8NC	RP9I8ND
25	RP9I9	RP9I9OB	RP9I9NB	RP9I9OC	RP9I9NC	RP9I9ND
26	RP9I10	RP9I10OB	RP9I10NB	RP9I10OC	RP9I10NC	RP9I10ND
27	RP9I11	RP9I11OB	RP9I11NB	RP9I11OC	RP9I11NC	RP9I11ND
28	RP9I12	RP9I12OB	RP9I12NB	RP9I12OC	RP9I12NC	RP9I12ND
29	RP9I13	RP9I13OB	RP9I13NB	RP9I13OC	RP9I13NC	RP9I13ND
30	RP9I14	RP9I14OB	RP9I14NB	RP9I14OC	RP9I14NC	RP9I14ND
31	RP9I15	RP9I15OB	RP9I15NB	RP9I15OC	RP9I15NC	RP9I15ND
32	RP10I1	RP10I1OB	RP10I1NB	RP10I1OC	RP10I1NC	RP10I1ND
33	RP10I2	RP10I2OB	RP10I2NB	RP10I2OC	RP10I2NC	RP10I2ND
34	RP10I3	RP10I3OB	RP10I3NB	RP10I3OC	RP10I3NC	RP10I3ND
35	RP10I4	RP10I4OBNB	RP10I4OBNB	RP10I4OCNC	RP10I4OCNC	RP10I4ND
36	RP10I5	RP10I5OBNB	RP10I5OBNB	RP10I5OCNC	RP10I5OCNC	RP10I5ND
37	RP10I6	RP10I6OBNB	RP10I6OBNB	RP10I6OCNC	RP10I6OCNC	RP10I6ND
38	RP10I7	RP10I7OB	RP10I7NB	RP10I7OC	RP10I7NC	RP10I7ND
39	RP10I8	RP10I8OB	RP10I8NB	RP10I8OC	RP10I8NC	RP10I8ND
40	RP10I9	RP10I9OB	RP10I9NB	RP10I9OC	RP10I9NC	RP10I9ND
41	RP10I10	RP10I10OBNBOCNCND	RP10I10OBNBOCNCND	RP10I10OBNBOCNCND	RP10I10OBNBOCNCND	RP10I10OBNBOCNCND
42	RP10I11	RP10I11OBNBOCNCND	RP10I11OBNBOCNCND	RP10I11OBNBOCNCND	RP10I11OBNBOCNCND	RP10I11OBNBOCNCND

Table 21a continued

	Item	Old B	New B	Old C	New C	New D
43	RP10I12	RP10I12OBNB	RP10I12OBNB	RP10I12OCNC	RP10I12OCNC	RP10I12ND
44	RP10I13	RP10I13OBNB	RP10I13OBNB	RP10I13OCNC	RP10I13OCNC	RP10I13ND
45	RP10I14	RP10I14OBOC	RP10I14NBNCND	RP10I14OBOC	RP10I14NBNCND	RP10I14NBNCND
46	RP10I15	RP10I15OBOC	RP10I15NBNCND	RP10I15OBOC	RP10I15NBNCND	RP10I15NBNCND
47	RP10I16	RP10I16OB	RP10I16NB	RP10I16OC	RP10I16NC	RP10I16ND
48	RP10I17	RP10I17OB	RP10I17NB	RP10I17OC	RP10I17NC	RP10I17ND
49	RP10I18	RP10I18OB	RP10I18NB	RP10I18OC	RP10I18NC	RP10I18ND

Table 21b
Reading Item Identifiers (Graphic Version)

	Item	Old B	New B	Old C	New C	New D
1	RP2I1					
2	RP2I2					
3	RP2I3					
4	RP2I4					
5	RP3I1					
6	RP3I2					
7	RP3I3					
8	RP4I1					
9	RP4I2					
10	RP4I3					
11	RP4I4					
12	RP7I1					
13	RP7I2					
14	RP8I1					
15	RP8I2					
16	RP8I3					
17	RP9I1					
18	RP9I2					
19	RP9I3					
20	RP9I4					
21	RP9I5					
22	RP9I6					
23	RP9I7					
24	RP9I8					
25	RP9I9					
26	RP9I10					
27	RP9I11					
28	RP9I12					
29	RP9I13					
30	RP9I14					

	Item	Old B	New B	Old C	New C	New D
31	RP9I15	□	░	▒	▓	█
32	RP10I1	□	░	▒	▓	█
33	RP10I2	□	░	▒	▓	█
34	RP10I3	□	░	▒	▓	█
35	RP10I4	□	□	░	░	▒
36	RP10I5	□	□	░	░	▒
37	RP10I6	□	□	░	░	▒
38	RP10I7	□	░	▒	▓	█
39	RP10I8	□	░	▒	▓	█
40	RP10I9	□	░	▒	▓	█
41	RP10I10	□	□	□	□	□
42	RP10I11	□	□	□	□	□
43	RP10I12	□	□	░	░	▒
44	RP10I13	□	□	░	░	▒
45	RP10I14	□	░	□	░	░
46	RP10I15	□	░	□	░	░
47	RP10I16	□	░	▒	▓	█
48	RP10I17	□	░	▒	▓	█
49	RP10I18	□	░	▒	▓	█

Table 21c
Writing Item Identifiers (Text Version)

	Item	Old B	New B	Old C	New C	New D
1	WP1I1	WP1I1OBNBOCNCND	WP1I1OBNBOCNCND	WP1I1OBNBOCNCND	WP1I1OBNBOCNCND	WP1I1OBNBOCNCND
2	WP1I2	WP1I2OBNBOCNCND	WP1I2OBNBOCNCND	WP1I2OBNBOCNCND	WP1I2OBNBOCNCND	WP1I2OBNBOCNCND
3	WP1I3	WP1I3OBNBOCNCND	WP1I3OBNBOCNCND	WP1I3OBNBOCNCND	WP1I3OBNBOCNCND	WP1I3OBNBOCNCND
4	WP1I4	WP1I4OBNBOCNCND	WP1I4OBNBOCNCND	WP1I4OBNBOCNCND	WP1I4OBNBOCNCND	WP1I4OBNBOCNCND
5	WP1I5	WP1I5OBNB	WP1I5OBNB	WP1I5OCNC	WP1I5OCNC	WP1I5ND
6	WP1I6	WP1I6OBNB	WP1I6OBNB	WP1I6OCNC	WP1I6OCNC	WP1I6ND
7	WP1I7	WP1I7OBNB	WP1I7OBNB	WP1I7OCNC	WP1I7OCNC	WP1I7ND
8	WP1I8	WP1I8OBNB	WP1I8OBNB	WP1I8OCNCND	WP1I8OCNCND	WP1I8OCNCND
9	WP1I9	WP1I9OBNBOCNCND	WP1I9OBNBOCNCND	WP1I9OBNBOCNCND	WP1I9OBNBOCNCND	WP1I9OBNBOCNCND
10	WP1I10	WP1I10OBNBOCNCND	WP1I10OBNBOCNCND	WP1I10OBNBOCNCND	WP1I10OBNBOCNCND	WP1I10OBNBOCNCND
11	WP5I1	WP5I1OB	WP5I1NB	WP5I1OC	WP5I1NC	WP5I1ND
12	WP5I2	WP5I2OB	WP5I2NB	WP5I2OC	WP5I2NC	WP5I2ND
13	WP5I3	WP5I3OB	WP5I3NB	WP5I3OC	WP5I3NC	WP5I3ND
14	WP5I4	WP5I4OB	WP5I4NB	WP5I4OC	WP5I4NC	WP5I4ND
15	WP5I5	WP5I5OB	WP5I5NB	WP5I5OC	WP5I5NC	WP5I5ND
16	WP6I1	WP6I1OB	WP6I1NB	WP6I1OC	WP6I1NC	WP6I1ND
17	WP6I2	WP6I2OB	WP6I2NB	WP6I2OC	WP6I2NC	WP6I2ND
18	WP11I1	WP11I1OBNB	WP11I1OBNB	WP11I1OCNC	WP11I1OCNC	WP11I1ND
19	WP11I2	WP11I2OBNB	WP11I2OBNB	WP11I2OCNC	WP11I2OCNC	WP11I2ND

Table 21d
Writing Item Identifiers (Graphic Version)

	Item	Old B	New B	Old C	New C	New D
1	WP1I1	░	░	░	░	░
2	WP1I2	░	░	░	░	░
3	WP1I3	░	░	░	░	░
4	WP1I4	░	░	░	░	░
5	WP1I5	░	░	▒	▒	▓
6	WP1I6	░	░	▒	▒	▓
7	WP1I7	░	░	▒	▒	▓
8	WP1I8	░	░	▒	▒	▒
9	WP1I9	░	░	░	░	░
10	WP1I10	░	░	░	░	░
11	WP5I1	░	▒	▓	█	■
12	WP5I2	░	▒	▓	█	■
13	WP5I3	░	▒	▓	█	■
14	WP5I4	░	▒	▓	█	■
15	WP5I5	░	▒	▓	█	■
16	WP6I1	░	▒	▓	█	■
17	WP6I2	░	▒	▓	█	■
18	WP11I1	░	░	▒	▒	▓
19	WP11I2	░	░	▒	▒	▓

As can be seen from Tables 21a-d, there was a lot of linkage through common items across the test forms. Of the 49 reading items, 3 items were common across all five forms, 13 were common across the three new forms, 13 were common across the two old forms, 6 were common to Old B and New B, and 6 were common to Old C and New C. Of the 19 writing items, 6 were common across all five forms, 6 were common to Old B and New B, and 6 were common to Old C and New C. In all, there were a 182 unique reading items and 58 unique writing items.

It should be noted that these common items were determined conservatively on the basis of a content review, as mentioned above, and if any aspect of an item was altered during updating, the item on the older form and the newer form were considered unique. In a very real sense, however, all the items on Old B and New B could have been considered common items, as could all the items on Old C and New C, because the updating of the item content was minimal. The tables in Appendix A show that, empirically speaking, fewer than half of the items on Old B and New B and fewer than half of the items on Old C and New C showed a statistical difference in difficulty given the calibration procedures followed here. For both test forms, about half the items became easier and half more difficult; the average item difficulty was not greatly affected. As results in the following sections show (i.e., Figure 13 comparing Old B and New B reading, Figure 19 comparing Old C and New C reading, Figure 24 comparing Old B and New B writing, and Figure 30 comparing Old C and New C writing), the method chosen here for identifying common items and equating reveals that there was very little difference, as could be expected, between estimates of examinee ability based on performances on Old B and New B, or on Old C and New C. These results give us grounds to believe that, had more items been linked for equating purposes on the basis of empirical (versus content) considerations, more item parameters would have been estimated with a larger number of students and thus with greater accuracy, but the overall results would not have differed greatly.

5.2.2 Concurrent Calibration

After all the items had been identified and linkages established, the entire data set was calibrated concurrently (though separately for reading and writing). This means that all item information was estimated at the same time from one calibration of the data, and the difficulty value of all items was placed onto a common measurement scale. From this calibration, item difficulty measures and fit statistics to the Rasch measurement model were obtained.

Fit statistics for the Rasch model are calculated by comparing the observed empirical data with the data that would be expected to be produced by the Rasch model. Of the several statistics available, the z-standardized fit statistics were used to flag items in the analysis of the *BEST Literacy* field test. Outfit z-standardized fit statistics are influenced by outliers. For example, a difficult item that for some reason some low ability examinees get correct will have a high outfit z-standardized fit statistic that indicates that the item may not be measuring the same thing as other items on the test. Infit mean square statistics are influenced by more aberrant response patterns and generally indicate a more serious measurement problem. The expectation for both these statistics is 0.00; values greater than -2.00 and less than 2.00 are acceptable.

5.2.3 Final Equating

After the concurrent calibration to obtain item difficulty values, a final analysis was conducted on each test form to determine the final equating values. First, in the independent analyses of each form, the difficulty value of all items was anchored to the value from the concurrent calibration. (This concurrent calibration put the test characteristic curves that appear in section 8.2 on the same scale as well.) Then, for each test form, the relationship of the raw score to the logit ability measure was established. Because the raw scores on Old B formed the basis of the scale score for the literacy skills section of the *BEST*, the same convention was used for the new forms. In other words, raw scores on the new forms were equated to raw scores on Old B if they had the same (or nearly the same) logit ability measure.

Following are the results of the equating process, first for reading and then for writing. The results for all items, calibrated together, are shown first. Then the equating results are shown form by form.

5.3 Results for Reading

5.3.1 Concurrent Calibration of Reading

Table 22 shows part of the output from the concurrent calibration of all reading items. The first column shows the entry number of the item in the data analyses. Note that the last column in the table gives the complete item name. The second column, Count, shows the number of students who took the item. For example, a total of 402 examinees had scores for the reading section, so items that were common across all five test forms should have a count of 402. If an item appeared on only one form, it would have a much lower count.

The next column, Score, shows the number of students who got that item correct, since each reading item was worth one point. So, for example, the first item was administered to 402 examinees and 384 examinees got it correct.

The fourth column, Measure, shows the item difficulty measure in terms of the Rasch model's logit values. Logit values center around 0.00; lower values indicate easier items, with negative values being easier than average; higher values indicate more difficult items, with positive values being more difficult than average. The easiest item was item 53 (RP3I2NB), with an item difficulty value of -3.80. It was administered to 69 examinees and 68 got it correct. The most difficult item was item 30 (RP9I14OB), with an item difficulty value of 3.67. It was administered to 66 people and only 9 people got it correct.

The next column, Error, shows the statistical standard error of the item difficulty measure. This value gives a sense of the precision with which the measure could be estimated. In general, the more examinees take an item, the smaller the error (see, e.g., items 41 and 42). The further the item is from the center and more toward the extreme of the measurement scale (i.e., many examinees got the item right or wrong), the larger the error (see, e.g., item 53).

The next two columns show the infit and outfit z-standardized fit statistics that were used to identify items that were not fitting the Rasch measurement model. While there is always variation, these values center around 0.00. Typically, when using these fit statistics, which are more appropriate for smaller samples such as this, most users of the Rasch model consider that values greater than 2.00 and less than -2.00 indicate that the item does not fit the Rasch measurement model (Linacre, 2005). It should be noted that large positive values (i.e., above 2.00) indicate that the item may be testing something different from the majority of items on the test, while large negative values (i.e., below -2.00) indicate that the test item is redundant (i.e., not providing more measurement information).

Outfit statistics are sensitive to extreme outliers: that is, when low-ability examinees (as based on their total test performance) unexpectedly get a difficult item correct, or when high-ability examinees unexpectedly get an easy item wrong. Infit statistics are statistically adjusted to be less sensitive to these extremes. They are thus, perhaps, the more important fit statistic.

For reading, the data were based on results from 402 people. Of the total of 182 items, only 7 items (3.8%) were misfitting using both infit and outfit criteria (i.e., both infit and outfit z-standardized fit statistics were above 2.0 or below -2.0). These results indicate that this set of reading items had an acceptable fit to the Rasch measurement model.

Table 22
Reading Item Properties

Entry	Count	Score	Measure	Error	In.zstd	Out.zstd	Item Name
1	402	384	-2.68	0.26	0.45	1.40	1,RP2I1OBNBOCNCND
2	135	131	-2.92	0.54	-0.09	0.12	2,RP2I2OBNB
3	136	122	-1.41	0.31	0.16	-0.23	3,RP2I3OBOC
4	66	62	-2.05	0.55	-0.11	0.19	4,RP2I4OB
5	66	64	-2.86	0.75	-0.14	-0.12	5,RP3I1OB
6	66	63	-2.40	0.62	-0.07	-0.03	6,RP3I2OB
7	136	102	-0.10	0.23	-0.39	0.55	7,RP3I3OBOC
8	66	61	-1.78	0.50	-0.31	1.35	8,RP4I1OB
9	66	62	-2.05	0.55	0.04	1.35	9,RP4I2OB
10	136	105	-0.25	0.23	1.86	0.84	10,RP4I3OBOC
11	136	127	-1.97	0.37	-0.48	-0.32	11,RP4I4OBOC
12	136	122	-1.41	0.31	0.17	0.69	12,RP7I1OBOC
13	136	49	2.18	0.21	-0.38	-0.58	13,RP7I2OBOC
14	136	122	-1.41	0.31	0.68	2.72	14,RP8I1OBOC
15	66	59	-1.35	0.43	0.59	2.55	15,RP8I2OB
16	66	55	-0.72	0.37	0.80	0.22	16,RP8I3OB
17	136	50	2.14	0.21	0.51	1.05	17,RP9I1OBOC
18	136	86	0.64	0.21	1.47	0.48	18,RP9I2OBOC
19	136	57	1.84	0.21	0.55	0.68	19,RP9I3OBOC
20	136	55	1.92	0.21	1.71	4.51	20,RP9I4OBOC
21	66	36	1.09	0.28	-1.39	-1.22	21,RP9I5OB
22	66	57	-1.01	0.39	-0.73	-1.03	22,RP9I6OB
23	66	39	0.85	0.29	2.33	1.76	23,RP9I7OB
24	66	43	0.52	0.29	0.30	0.47	24,RP9I8OB
25	66	26	1.89	0.29	-0.70	2.68	25,RP9I9OB

Entry	Count	Score	Measure	Error	In.zstd	Out.zstd	Item Name
26	66	25	1.97	0.29	0.92	1.47	26,RP9I10OB
27	66	39	0.85	0.29	-1.63	-1.55	27,RP9I11OB
28	66	38	0.93	0.28	0.35	1.25	28,RP9I12OB
29	66	36	1.09	0.28	2.71	1.94	29,RP9I13OB
30	66	9	3.67	0.39	0.15	0.18	30,RP9I14OB
31	66	45	0.34	0.30	0.07	0.86	31,RP9I15OB
32	66	57	-1.01	0.39	-0.45	-0.59	32,RP10I1OB
33	66	60	-1.55	0.46	0.39	0.26	33,RP10I2OB
34	66	44	0.43	0.3	0.02	0.57	34,RP10I3OB
35	135	102	-0.02	0.23	-0.79	-1.21	35,RP10I4OBNB
36	135	115	-0.82	0.27	-1.24	-0.46	36,RP10I5OBNB
37	135	68	1.45	0.20	-0.82	0.70	37,RP10I6OBNB
38	66	48	0.06	0.31	0.43	0.49	38,RP10I7OB
39	66	29	1.65	0.28	1.20	1.13	39,RP10I8OB
40	66	54	-0.59	0.35	-0.67	-0.90	40,RP10I9OB
41	402	272	0.34	0.12	-2.37	-2.43	41,RP10I10OBNBOCNCND
42	402	230	0.96	0.12	2.34	2.29	42,RP10I11OBNBOCNCND
43	135	101	0.03	0.23	-1.18	-0.68	43,RP10I12OBNB
44	135	45	2.40	0.21	-0.42	0.23	44,RP10I13OBNB
45	136	105	-0.25	0.23	-2.78	-2.02	45,RP10I14OBOC
46	136	77	1.01	0.20	0.05	-0.16	46,RP10I15OBOC
47	66	28	1.73	0.28	0.49	0.15	47,RP10I16OB
48	66	45	0.34	0.30	-0.65	-1.02	48,RP10I17OB
49	66	35	1.17	0.28	-1.93	-1.69	49,RP10I18OB
50	266	244	-1.97	0.26	0.33	3.34	50,RP2I3NBNCND
51	69	63	-1.54	0.49	-0.01	-0.13	51,RP2I4NB
52	69	67	-2.99	0.78	0.46	0.15	52,RP3I1NB
53	69	68	-3.80	1.06	0.36	0.47	53,RP3I2NB
54	266	198	-0.14	0.17	-0.03	-0.51	54,RP3I3NBNCND
55	69	62	-1.32	0.46	-0.38	-0.56	55,RP4I1NB
56	69	64	-1.80	0.53	-0.11	-0.54	56,RP4I2NB
57	266	210	-0.49	0.18	1.63	0.92	57,RP4I3NBNCND
58	266	242	-1.84	0.25	0.71	-0.42	58,RP4I4NBNCND
59	266	215	-0.65	0.18	1.59	2.01	59,RP7I1NBNCND
60	266	114	1.75	0.15	-1.17	-0.92	60,RP7I2NBNCND
61	266	225	-1.01	0.20	0.00	0.60	61,RP8I1NBNCND
62	69	60	-0.94	0.41	0.21	1.60	62,RP8I2NB
63	69	52	0.10	0.32	0.12	-0.64	63,RP8I3NB
64	266	102	2.01	0.15	2.26	3.20	64,RP9I1NBNCND
65	266	195	-0.06	0.16	1.09	-0.33	65,RP9I2NBNCND
66	266	95	2.16	0.15	-0.44	1.85	66,RP9I3NBNCND
67	266	112	1.79	0.15	2.48	1.07	67,RP9I4NBNCND
68	69	46	0.68	0.30	0.55	-0.39	68,RP9I5NB
69	69	61	-1.12	0.43	-0.95	-0.71	69,RP9I6NB
70	69	44	0.85	0.29	1.34	0.97	70,RP9I7NB
71	69	50	0.31	0.31	0.42	-0.24	71,RP9I8NB
72	69	38	1.34	0.28	0.43	0.65	72,RP9I9NB
73	69	27	2.22	0.29	0.75	0.70	73,RP9I10NB

Table 22 continued

Entry	Count	Score	Measure	Error	In.zstd	Out.zstd	Item Name
74	69	43	0.94	0.29	0.46	-0.09	74,RP9I11NB
75	69	43	0.94	0.29	2.45	1.17	75,RP9I12NB
76	69	36	1.50	0.28	0.76	0.32	76,RP9I13NB
77	69	17	3.11	0.32	2.04	2.28	77,RP9I14NB
78	69	45	0.77	0.29	0.85	0.52	78,RP9I15NB
79	69	59	-0.78	0.40	0.88	0.53	79,RP10I1NB
80	69	58	-0.63	0.38	-0.89	-0.75	80,RP10I2NB
81	69	55	-0.23	0.35	-0.67	-0.42	81,RP10I3NB
82	69	53	0.00	0.33	0.23	-0.39	82,RP10I7NB
83	69	30	1.98	0.28	-0.45	-0.54	83,RP10I8NB
84	69	55	-0.23	0.35	-1.72	-1.42	84,RP10I9NB
85	266	206	-0.37	0.17	-2.90	-1.92	85,RP10I14NBNCND
86	266	149	1.02	0.15	-0.80	-0.04	86,RP10I15NBNCND
87	69	38	1.34	0.28	-2.38	-1.65	87,RP10I16NB
88	69	50	0.31	0.31	-2.28	-1.76	88,RP10I17NB
89	69	38	1.34	0.28	-0.78	-0.84	89,RP10I18NB
90	140	128	-1.95	0.34	0.21	1.17	90,RP2I2OCNC
91	70	62	-1.30	0.41	0.51	0.21	91,RP2I4OC
92	70	68	-2.94	0.74	0.35	0.48	92,RP3I1OC
93	70	67	-2.49	0.61	0.34	0.81	93,RP3I2OC
94	70	59	-0.85	0.37	-1.04	-0.89	94,RP4I1OC
95	70	68	-2.94	0.74	0.08	0.04	95,RP4I2OC
96	70	64	-1.67	0.46	0.63	2.46	96,RP8I2OC
97	70	49	0.23	0.30	0.92	0.83	97,RP8I3OC
98	70	40	1.02	0.29	1.08	0.39	98,RP9I5OC
99	70	56	-0.48	0.34	0.55	1.12	99,RP9I6OC
100	70	35	1.44	0.29	-0.48	-1.03	100,RP9I7OC
101	70	30	1.86	0.29	1.00	1.05	101,RP9I8OC
102	70	59	-0.85	0.37	0.38	-0.16	102,RP9I9OC
103	70	34	1.52	0.29	0.09	0.43	103,RP9I10OC
104	70	34	1.52	0.29	0.14	0.52	104,RP9I11OC
105	70	48	0.32	0.30	-0.26	0.44	105,RP9I12OC
106	70	37	1.27	0.29	-1.13	-1.13	106,RP9I13OC
107	70	45	0.59	0.30	0.18	-0.49	107,RP9I14OC
108	70	58	-0.72	0.36	0.00	-0.18	108,RP9I15OC
109	70	59	-0.85	0.37	-0.20	-0.51	109,RP10I1OC
110	70	51	0.04	0.31	-1.92	-1.35	110,RP10I2OC
111	70	63	-1.47	0.43	0.03	-0.28	111,RP10I3OC
112	140	106	-0.32	0.23	-0.34	-0.96	112,RP10I4OCNC
113	140	124	-1.54	0.30	-0.93	-0.02	113,RP10I5OCNC
114	140	70	1.30	0.20	0.50	2.94	114,RP10I6OCNC
115	70	55	-0.37	0.33	-0.91	-0.65	115,RP10I7OC
116	70	35	1.44	0.29	-1.28	-1.34	116,RP10I8OC
117	70	47	0.41	0.30	-0.09	1.75	117,RP10I9OC
118	140	117	-0.99	0.26	-2.55	-1.84	118,RP10I12OCNC
119	140	98	0.09	0.22	-2.34	-1.12	119,RP10I13OCNC
120	70	44	0.68	0.29	-2.17	-1.37	120,RP10I16OC

Entry	Count	Score	Measure	Error	In.zstd	Out.zstd	Item Name
121	70	37	1.27	0.29	-0.06	0.05	121,RP10I17OC
122	70	51	0.04	0.31	-0.94	-0.69	122,RP10I18OC
123	70	61	-1.61	0.42	0.41	-0.39	123,RP2I4NC
124	70	62	-1.79	0.44	-0.44	-0.74	124,RP3I1NC
125	70	59	-1.28	0.39	-0.72	-0.83	125,RP3I2NC
126	70	55	-0.73	0.35	0.79	2.69	126,RP4I1NC
127	70	63	-1.99	0.46	0.36	2.63	127,RP4I2NC
128	70	56	-0.86	0.36	0.85	1.48	128,RP8I2NC
129	70	46	0.23	0.31	-0.62	-0.67	129,RP8I3NC
130	70	35	1.17	0.29	0.72	0.85	130,RP9I5NC
131	70	50	-0.17	0.32	0.00	0.40	131,RP9I6NC
132	70	32	1.42	0.29	-0.30	-0.34	132,RP9I7NC
133	70	24	2.08	0.29	1.76	1.81	133,RP9I8NC
134	70	53	-0.49	0.34	-0.44	-0.79	134,RP9I9NC
135	70	28	1.75	0.29	-0.84	-0.27	135,RP9I10NC
136	70	38	0.93	0.29	-0.14	-0.12	136,RP9I11NC
137	70	48	0.04	0.31	0.76	-0.18	137,RP9I12NC
138	70	30	1.58	0.29	0.34	0.00	138,RP9I13NC
139	70	48	0.04	0.31	-0.21	-0.71	139,RP9I14NC
140	70	57	-0.99	0.37	-0.93	-0.53	140,RP9I15NC
141	70	57	-0.99	0.37	0.07	0.72	141,RP10I1NC
142	70	50	-0.17	0.32	0.95	2.81	142,RP10I2NC
143	70	57	-0.99	0.37	0.33	-0.36	143,RP10I3NC
144	70	47	0.13	0.31	-0.36	-0.32	144,RP10I7NC
145	70	40	0.76	0.29	-1.01	-0.82	145,RP10I8NC
146	70	52	-0.38	0.33	0.35	-0.22	146,RP10I9NC
147	70	36	1.09	0.29	0.34	0.36	147,RP10I16NC
148	70	38	0.93	0.29	-0.88	-0.20	148,RP10I17NC
149	70	46	0.23	0.31	-0.07	-0.40	149,RP10I18NC
150	127	120	-2.54	0.44	0.39	0.70	150,RP2I2ND
151	127	117	-2.03	0.38	0.55	0.36	151,RP2I4ND
152	127	121	-2.75	0.47	-0.52	-0.46	152,RP3I1ND
153	127	119	-2.35	0.42	-0.26	-0.08	153,RP3I2ND
154	127	116	-1.89	0.36	-1.05	-1.05	154,RP4I1ND
155	127	120	-2.54	0.44	-0.62	-0.36	155,RP4I2ND
156	127	107	-0.98	0.28	-0.90	-1.05	156,RP8I2ND
157	127	107	-0.98	0.28	-1.02	-0.66	157,RP8I3ND
158	127	70	1.02	0.21	1.12	0.44	158,RP9I5ND
159	127	84	0.38	0.22	-0.42	-1.02	159,RP9I6ND
160	127	58	1.54	0.21	1.31	1.78	160,RP9I7ND
161	127	91	0.03	0.23	0.99	0.53	161,RP9I8ND
162	127	61	1.41	0.21	-0.36	-0.6	162,RP9I9ND
163	127	59	1.50	0.21	-0.31	-0.39	163,RP9I10ND
164	127	63	1.32	0.21	-0.70	0.68	164,RP9I11ND
165	127	80	0.57	0.21	-0.40	-0.40	165,RP9I12ND
166	127	97	-0.30	0.24	0.63	1.00	166,RP9I13ND
167	127	68	1.10	0.21	-2.45	-1.25	167,RP9I14ND
168	127	98	-0.36	0.25	0.15	-0.20	168,RP9I15ND

Table 22 continued

Entry	Count	Score	Measure	Error	In.zstd	Out.zstd	Item Name
169	127	61	1.41	0.21	5.64	4.64	169,RP10I1ND
170	127	108	-1.06	0.29	0.40	0.62	170,RP10I2ND
171	127	87	0.23	0.22	0.70	1.26	171,RP10I3ND
172	127	88	0.18	0.22	-0.49	-0.50	172,RP10I4ND
173	127	108	-1.06	0.29	-0.62	1.20	173,RP10I5ND
174	127	54	1.72	0.21	-2.45	-1.69	174,RP10I6ND
175	127	97	-0.30	0.24	0.03	-0.37	175,RP10I7ND
176	127	60	1.45	0.21	-0.31	-0.49	176,RP10I8ND
177	127	94	-0.13	0.24	-2.08	-1.71	177,RP10I9ND
178	127	71	0.97	0.21	-1.41	-0.83	178,RP10I12ND
179	127	67	1.15	0.21	-0.64	-0.86	179,RP10I13ND
180	127	92	-0.02	0.23	-2.38	-0.82	180,RP10I16ND
181	127	71	0.97	0.21	-0.31	-0.57	181,RP10I17ND
182	127	55	1.67	0.21	0.16	0.79	182,RP10I18ND

5.3.2 Equating of Reading Forms

We used the true score method of equating (e.g., Hambleton & Swaminathan, 1985), which is based on the test characteristic curves. In the following sections, we show the results for each of the reading forms. For each form, we present a graph that explains the equating. The left-hand side of each graph shows all possible raw scores (Score) for the reading section (0 to 49). At the bottom of each graph is the person ability measure (Person Measure) scale in terms of Rasch logits. This scale, like the item difficulty scale, is centered around 0.00. The curves on the graphs show the relationship between the raw scores and the examinee ability level. With the exception of the graph for Old Form B, to whose raw score scale performances on the new forms are equated, each graph has two curves. The two curves compare the relationship between raw scores and ability across two tests. The closer the curves are to one another, the more similar in difficulty the two test forms are.

Following each graph is a chart that shows the conversion from raw scores on a form other than Old B to the scale score, which is the same as the raw scores on Old B. In this methodology, two raw scores are equated when they represent the same person ability measure. For example, look at the graph comparing New C with Old B (Figure 17). Look at the person ability measure of 0.00. Follow the line up and find where it intersects with the graph for Old B. Then go over to the left-hand column and find the raw score that goes with that ability measure. In this case, the raw score corresponding to an ability measure of 0.00 is 24. Next, find the raw score that corresponds with the person ability measure using the curve for New C. In this case, it is 25. That is, an examinee taking Old B who scores 24 points is estimated to have the same underlying ability as a person taking New C who scores 25 points. Thus, in the New C Reading Conversion Chart, a raw score of 25 becomes a scale score of 24.

The next figure in each section shows the distribution of the scale scores of the examinees who took a particular test form, while the final table shows the descriptive statistics of the scale scores. The figure and the table replicate what was presented in Section 4.3.2.1 but with scale scores (i.e., raw scores from Old B).

5.3.2.1 Reading Form Old B

Because all other forms are equated to Old B, there is only one curve in Figure 11; in the chart in Table 23, the raw scores are the scale scores. Also, Figure 12 and Table 24 replicate the earlier figures for Old B (Figure 1 and Table 9).

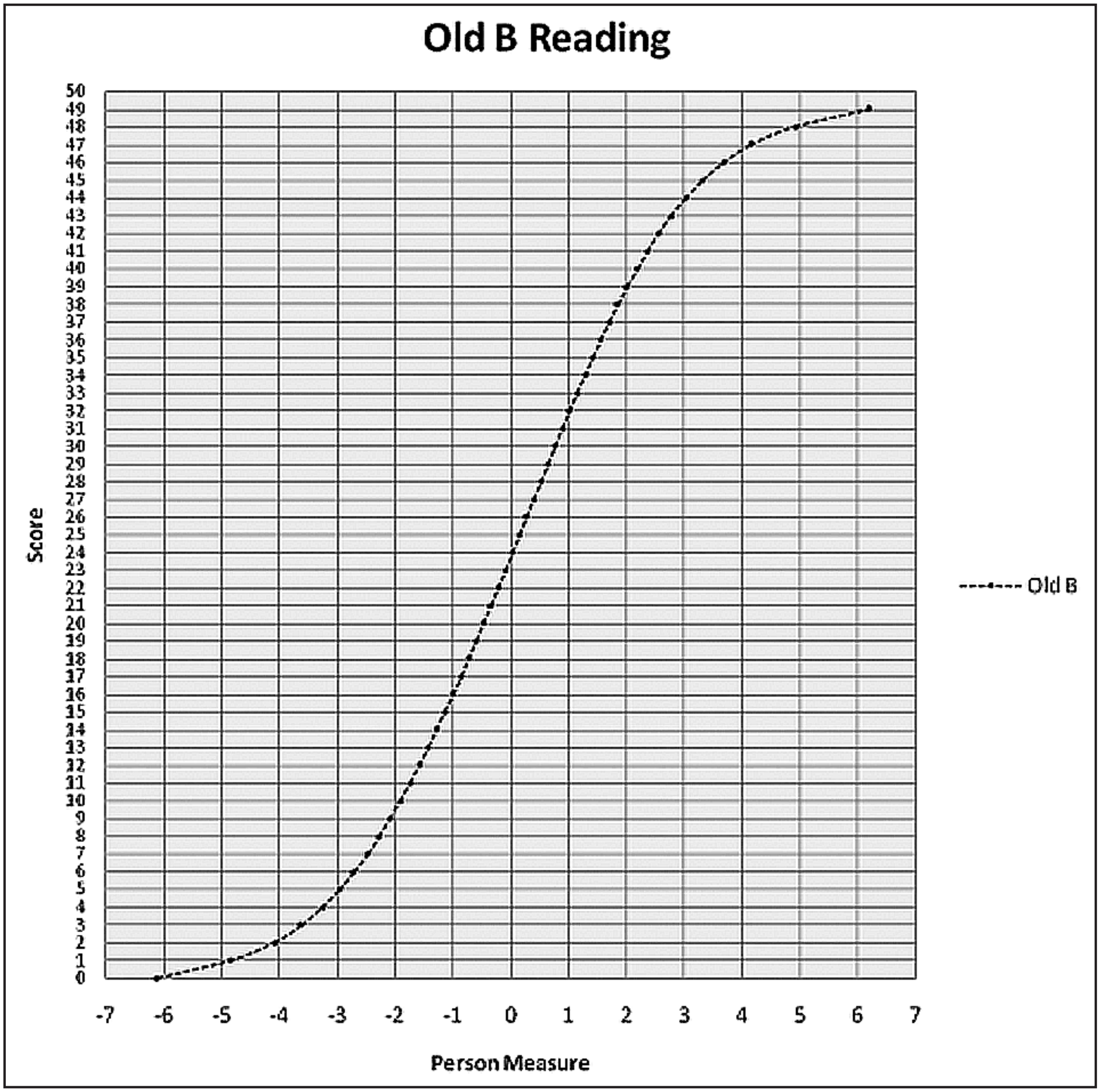

Figure 11. Old B Reading Graph

Table 23
Old B Reading Conversion Chart

Raw Score	Scale Score	Raw Score	Scale Score
0	0	25	25
1	1	26	26
2	2	27	27
3	3	28	28
4	4	29	29
5	5	30	30
6	6	31	31
7	7	32	32
8	8	33	33
9	9	34	34
10	10	35	35
11	11	36	36
12	12	37	37
13	13	38	38
14	14	39	39
15	15	40	40
16	16	41	41
17	17	42	42
18	18	43	43
19	19	44	44
20	20	45	45
21	21	46	46
22	22	47	47
23	23	48	48
24	24	49	49

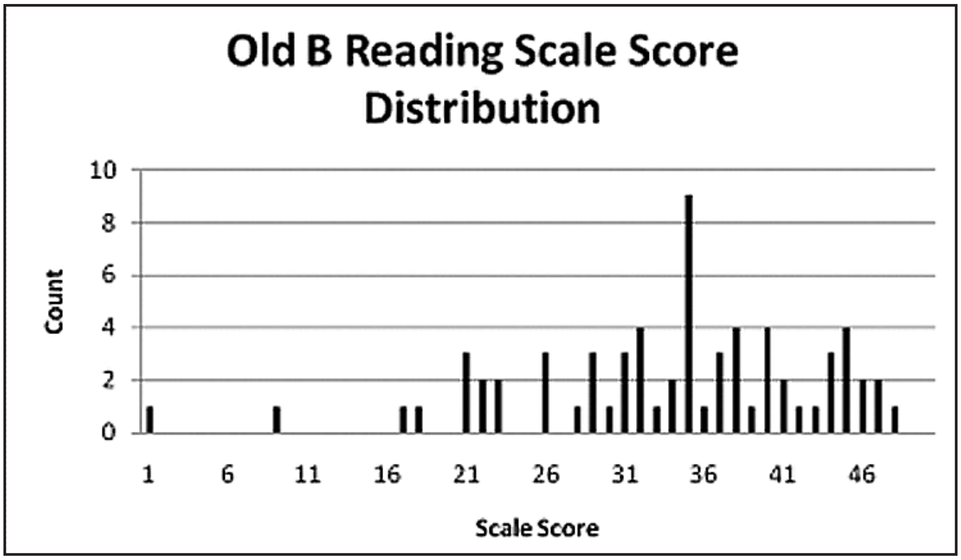

Figure 12. Old B Reading Scale Score Distribution

Table 24
Old B Reading Scale Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
67	0	47	32.87	9.36

5.3.2.2 Reading Form New B

As Figure 13 clearly shows, the equating results for New B indicate that this updated form is very similar to Old B. This shows that the updates to the *BEST* had a negligible effect on the overall difficulty of the test form, making it neither easier nor harder.

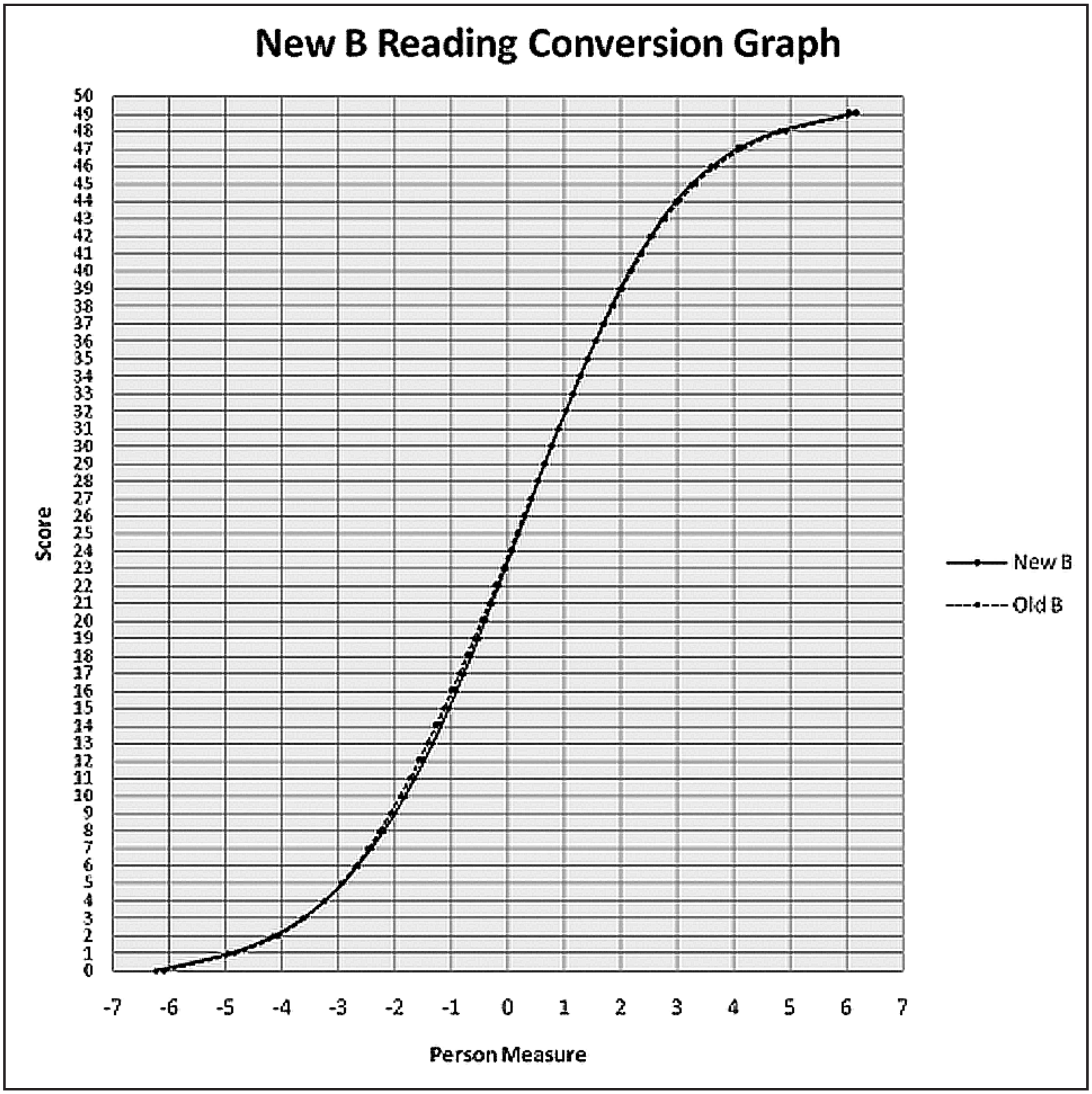

Figure 13. New B Reading Conversion Graph

Table 25
New B Reading Conversion Chart

Raw Score	Scale Score	Raw Score	Scale Score
0	0	25	25
1	1	26	26
2	2	27	27
3	3	28	28
4	4	29	29
5	5	30	30
6	6	31	31
7	7	32	32
8	8	33	33
9	9	34	34
10	10	35	35
11	11	36	36
12	13	37	37
13	14	38	38
14	15	39	39
15	16	40	40
16	17	41	41
17	18	42	42
18	19	43	43
19	20	44	44
20	20	45	45
21	21	46	46
22	22	47	47
23	23	48	48
24	24	49	49

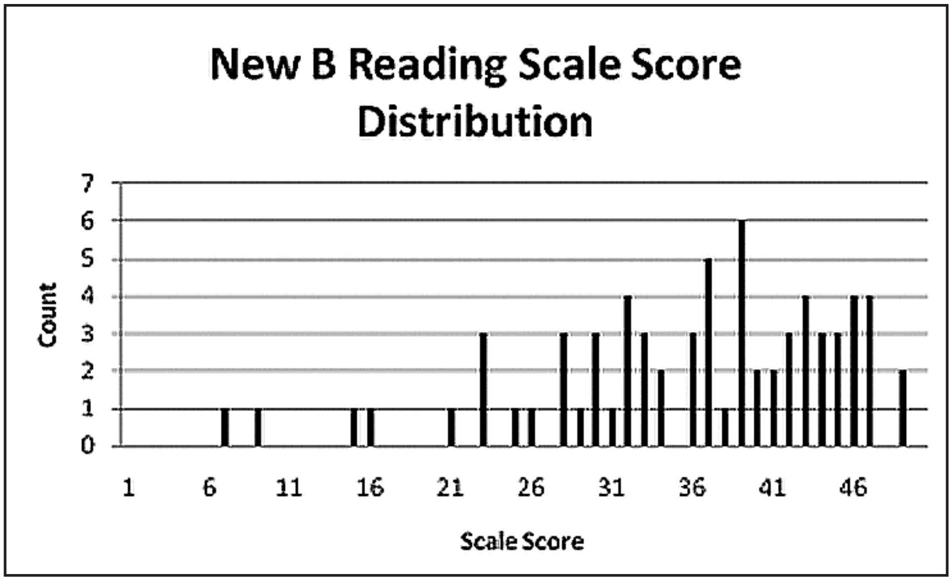

Figure 14. New B Reading Scale Score Distribution

Table 26
New B Reading Scale Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
69	6	48	34.93	9.35

5.3.2.3 Reading Form Old C

The *BEST Test Manual* (Center for Applied Linguistics, 1982, 1984, 1987, 1989) provided a conversion between Form C and Form B that was based on a linear approach to equating that was current at the time of the development of the *BEST*. In this technical report for *BEST Literacy*, we look only at the results of the equating procedures that were used with the current data, using more modern IRT methodology.

The two lines in Figure 15 diverge more than those in Figure 13, particularly at the upper score ranges. Old C is to the left of Old B, which means that Old C is a bit easier than Old B; that is, for example, a person of ability measure 2.00 would score only 39 points on Old B but 41 points on Old C. (This trend was also seen in the 1980s study.) Thus, a raw score of 41 points on Old C converts to a scale score of 39, as seen in Table 27.

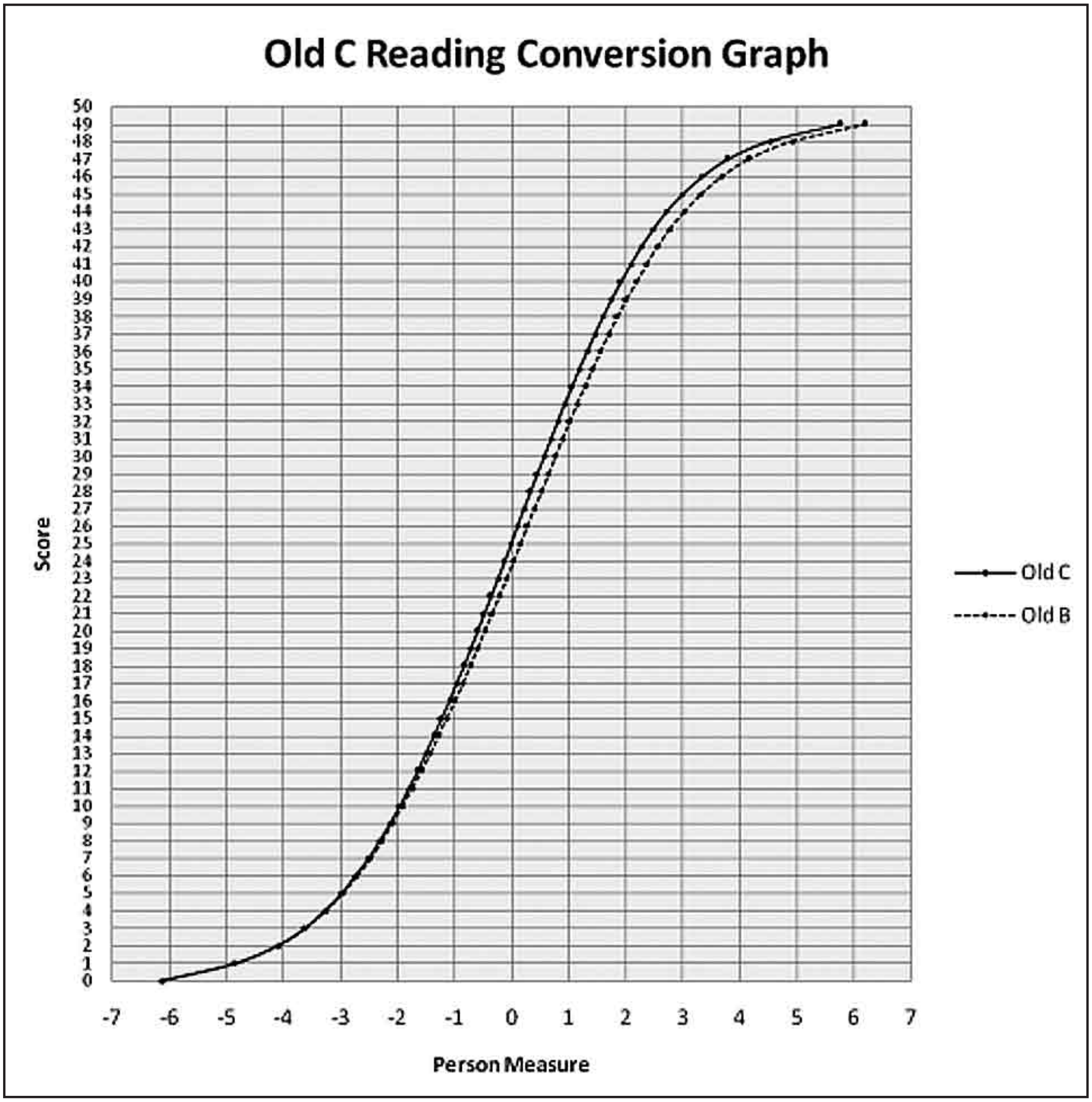

Figure 15. Old C Reading Conversion Graph

Table 27
Old C Reading Conversion Chart

Raw Score	Scale Score	Raw Score	Scale Score
0	0	25	24
1	1	26	25
2	2	27	26
3	3	28	26
4	4	29	27
5	5	30	28
6	6	31	29
7	7	32	30
8	8	33	31
9	9	34	32
10	10	35	33
11	11	36	34
12	12	37	35
13	13	38	36
14	14	39	37
15	14	40	38
16	15	41	39
17	16	42	40
18	17	43	42
19	18	44	43
20	19	45	44
21	20	46	45
22	21	47	46
23	22	48	47
24	23	49	49

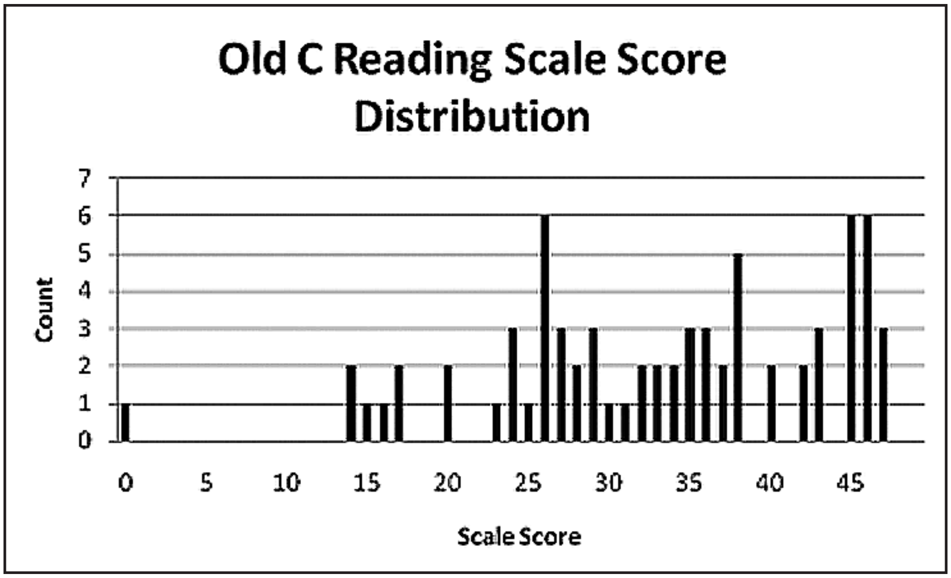

Figure 16. Old C Reading Scale Score Distribution

Table 28
Old C Reading Scale Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
71	0	47	33.08	10.30

5.3.2.4 Reading Form New C

As may be expected given the results of Old C, New C was also found to be easier than Old B at the upper end of the ability scale (see Figure 17). However, at the very lowest end (up to a raw score of about 10), New C appears a bit more difficult than Old B.

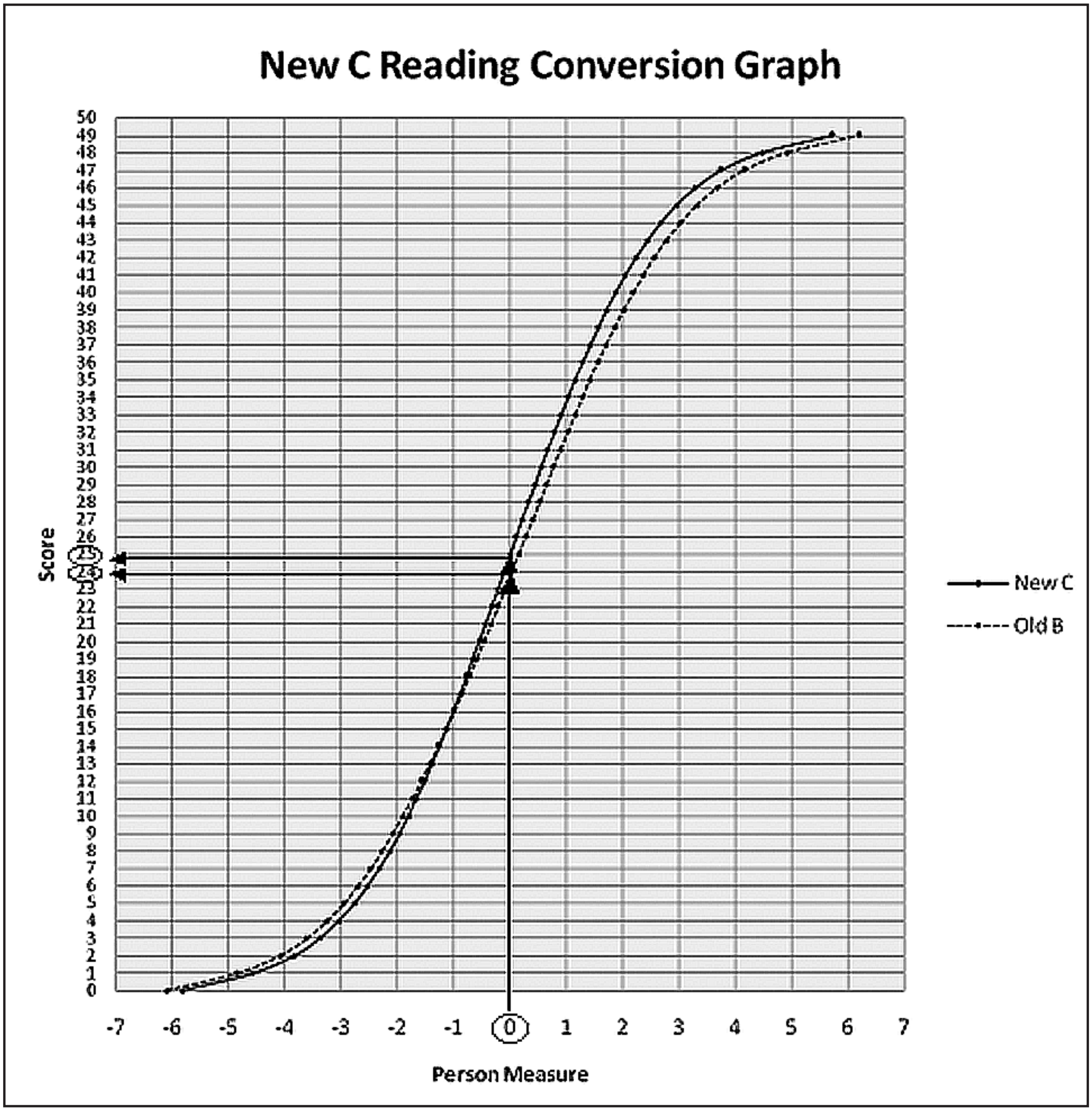

Figure 17. New C Reading Conversion Graph

Table 29
New C Reading Conversion Chart

Raw Score	Scale Score	Raw Score	Scale Score
0	0	25	24
1	1	26	25
2	3	27	26
3	4	28	27
4	5	29	28
5	6	30	28
6	7	31	29
7	8	32	30
8	9	33	31
9	10	34	32
10	11	35	33
11	11	36	34
12	12	37	35
13	13	38	36
14	14	39	37
15	15	40	38
16	16	41	39
17	17	42	40
18	18	43	41
19	19	44	43
20	19	45	44
21	20	46	45
22	21	47	46
23	22	48	47
24	23	49	49

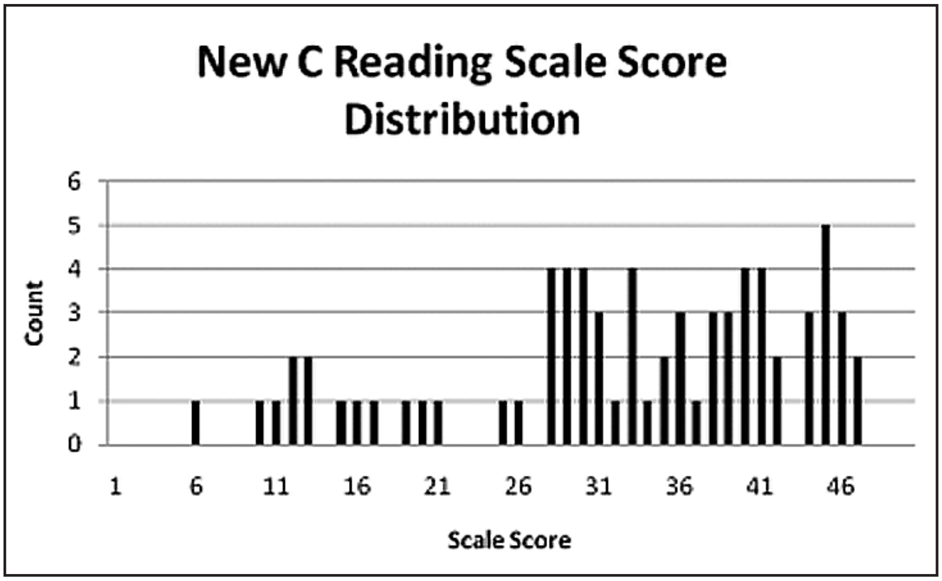

Figure 18. New C Reading Scale Score Distribution

Table 30
New C Reading Scale Score Descriptive Statistics

No. of Students	Min. Score	Max. Score	Mean Score	Std. Dev
71	5	46	31.65	10.66

We saw above that New B and Old B were essentially the same. In order to compare the effects of the updating process on Form C, the graph in Figure 19 compares Old C and New C. Except at the lowest end of the ability scale, they are almost identical, suggesting that the improvements and changes that were made to the test form did not affect the item properties, except perhaps making it slightly more difficult for the lowest ability examinees (i.e., those who score fewer than 16 points on Reading).

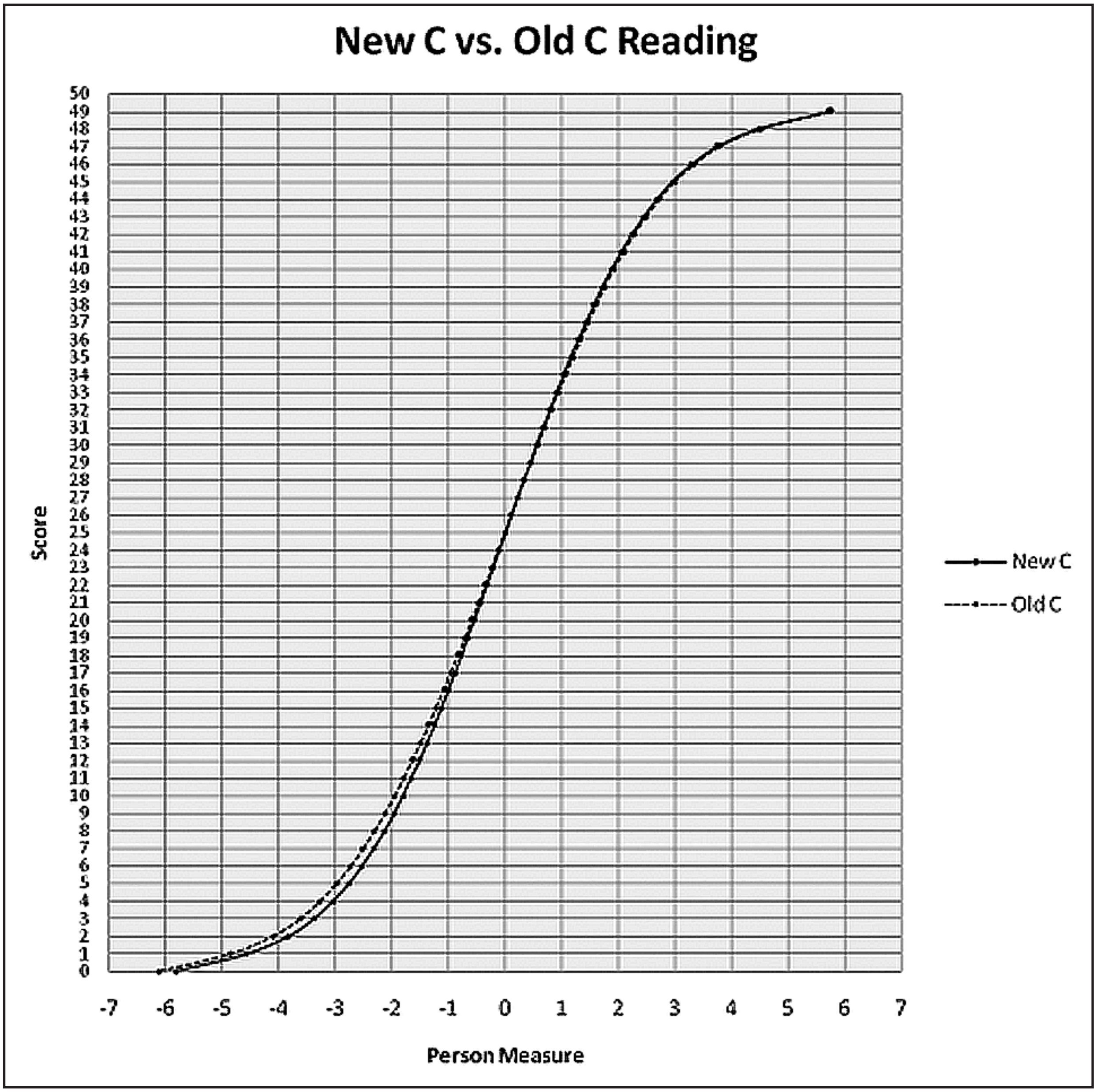

Figure 19. New C vs. Old C Reading Scale Score

5.3.2.5 Reading Form New D

Figure 20 indicates that New D reading, like New C reading, was also slightly easier than Old B reading at the upper ends of the ability scale.

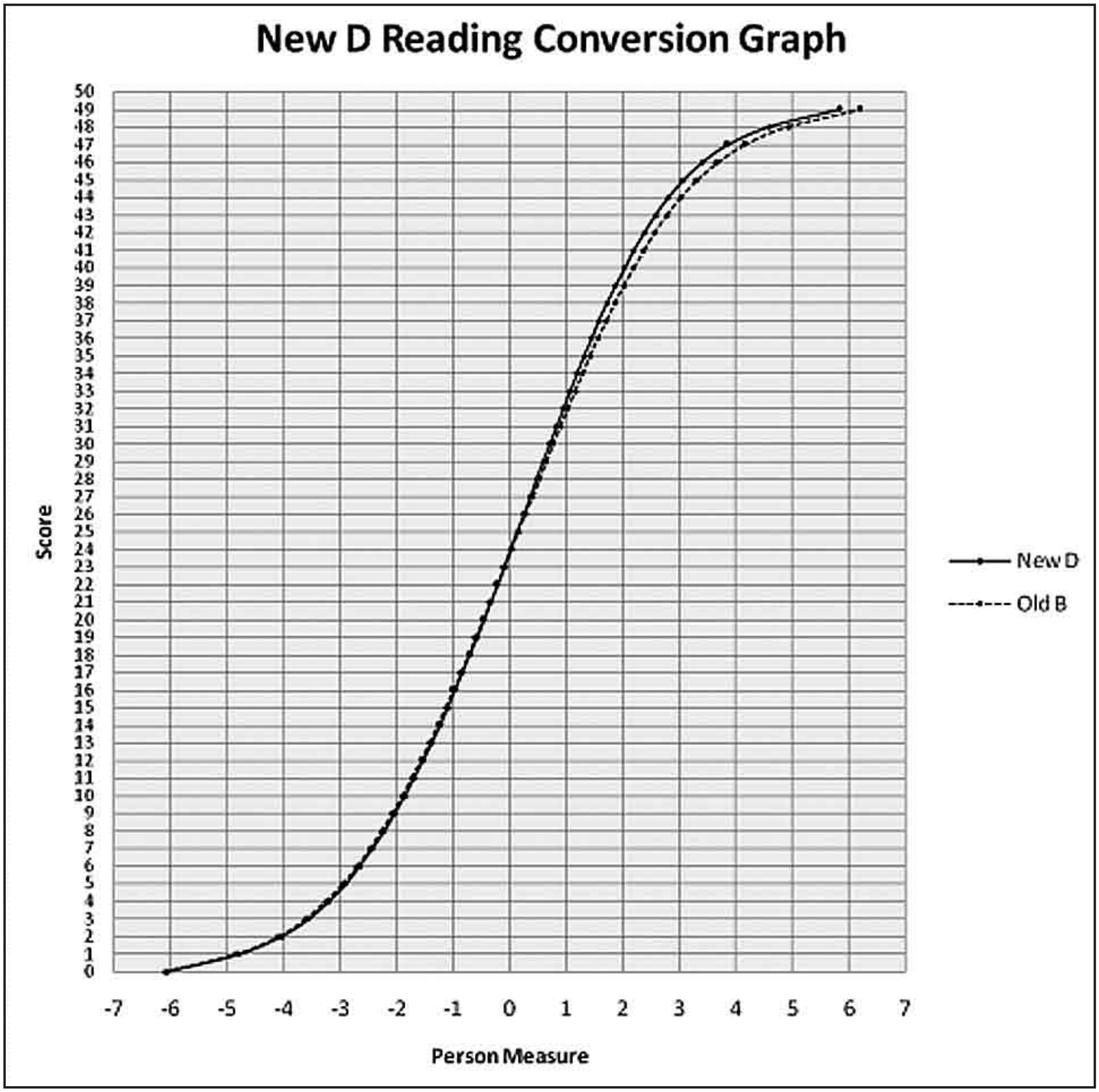

Figure 20. New D Reading Conversion Graph

Table 31
New D Reading Conversion Chart

Raw Score	Scale Score	Raw Score	Scale Score
0	0	25	25
1	1	26	26
2	2	27	27
3	3	28	28
4	4	29	29
5	5	30	30
6	6	31	30
7	7	32	31
8	8	33	32
9	9	34	33
10	10	35	34
11	11	36	35
12	12	37	36
13	13	38	37
14	14	39	38
15	15	40	39
16	16	41	40
17	17	42	41
18	18	43	42
19	19	44	43
20	20	45	44
21	21	46	45
22	22	47	46
23	23	48	48
24	24	49	49

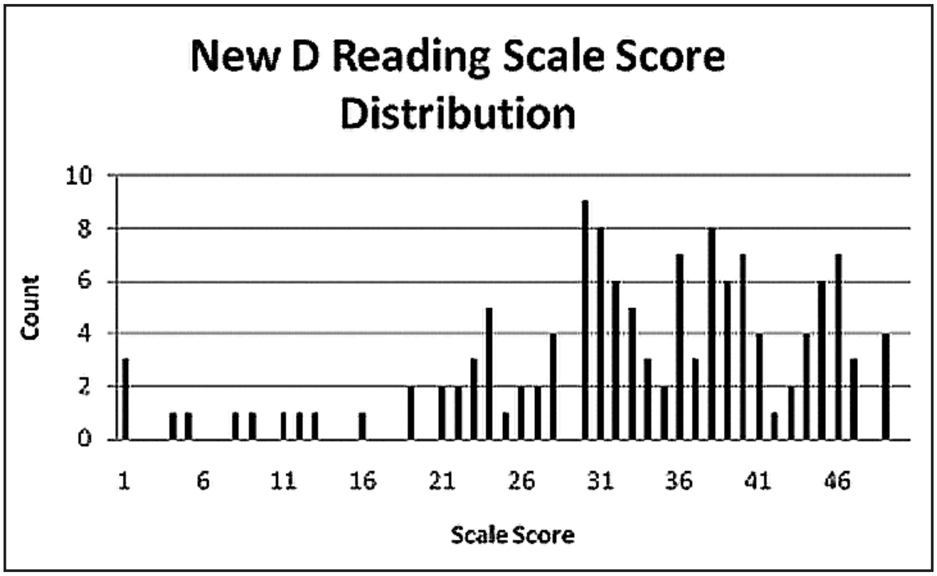

Figure 21. New D Reading Scale Score Distribution

Table 32
New D Reading Scale Score Descriptive Statistics

No. of Students	Min	Max	Mean	Std. Dev
129	0	48	32.04	10.88

5.3.2.6 Reading All Forms: Scale Scores

The maximum possible scale score for reading is 49. Table 33 shows the number of people who took each reading form and the mean and the standard deviation of the scale scores for each reading form. Because these results are now equated, a total across all five forms is also given. The results show that the mean scores for any of the five forms are within 1 point of the mean across all five forms (32.63), with the exception of New B. The mean reading score for New B (34.93) is 2.3 points higher than the mean across all forms. This suggests that, in general, the randomization procedure was effectual, but the group that took New B may have had a slightly higher ability in reading than those taking the other test forms. In order to verify this statistically, a one-way ANOVA was run. The results showed that the difference in the equated scale scores for reading among the five groups was not statistically significant: $F(4,402) = 1.159$, $p = .328$. This result suggests that the method of randomization of the test booklets used in the field test was successful and provides some evidence for the success of the equating procedure.

Table 33
Reading Scale Score Mean and Standard Deviation by Test Form

	Old B	New B	Old C	New C	New D	Total
Number of people	67	69	71	71	129	407
Mean	32.87	34.93	33.08	31.65	32.04	32.63
Std. Deviation	9.358	9.348	10.295	10.657	10.880	10.096

5.4 Results for Writing

5.4.1 Concurrent Calibration of Writing

Writing data from 407 people were analyzed, although a number of students with perfect scores were deleted from the complete analyses. Table 34 shows the properties of the writing items. As in Table 22, the first column shows the entry number of the item and the last column gives the complete item name. The Count column shows the number of students who were administered the item. For the writing items scored dichotomously, the Score column shows the number of students answering that item correctly. For polytomously scored items (e.g., the note-writing tasks), it shows the number of raw score points awarded across all examinees.

The next four columns (Measure, Error, In.ZStd, Out.ZStd) are the same as in Table 22. The following two columns provide additional information unique to the writing analyses. The Weight column shows the weight an item received in computing the total score. Tasks related to completing an envelope are scored dichotomously but are given a weight of 2 rather than 1. The next column, Grouping, indicates the items analyzed with the dichotomous Rasch model (D) or the items scored with the partial-credit Rasch model (0), which were the note-writing tasks.

Regarding fit to the Rasch model, of the total of 58 unique writing items, only 2 items (3.4%) had both infit and outfit z-standardized fit statistics above 2.0 or below -2.0. Again, this result shows appropriate fit of the data to the Rasch measurement model.

Table 34
Writing Item Properties

Entry	Count	Score	P-value / *Expected Scores	Measure	Error	In.zstd	Out.zstd	Weight	Grouping	Name
1	346	341	0.99	-4.71	0.54	1.31	0.26	1	D	1,WP1I1OBNBOCNCND
2	346	305	0.88	-1.14	0.21	3.1	4.52	1	D	2,WP1I2OBNBOCNCND
3	346	330	0.95	-2.82	0.33	-0.92	-0.3	1	D	3,WP1I3OBNBOCNCND
4	346	331	0.96	-2.93	0.34	-0.89	1.44	1	D	4,WP1I4OBNBOCNCND
5	115	109	0.95	-2.76	0.54	-0.32	0.67	1	D	5,WP1I5OBNB
6	115	107	0.93	-2.25	0.47	0.36	1.14	1	D	6,WP1I6OBNB
7	115	107	0.93	-2.25	0.47	0.16	1.28	1	D	7,WP1I7OBNB
8	115	86	0.75	0.08	0.26	0.65	0.79	1	D	8,WP1I8OBNB
9	231	208	0.90	-1.37	0.28	0.05	0.75	1	D	9,WP1I8OCNCND
10	346	311	0.90	-1.43	0.23	-1.04	-1.15	1	D	10,WP1I9OBNBOCNCND
11	346	312	0.90	-1.48	0.23	0.89	2.52	1	D	11,WP1I10OBNBOCNCND
12	59	54	0.92	-2.09	0.6	-0.5	-0.14	1	D	12,WP5I1OB
13	59	48	0.81	-0.68	0.4	0.34	1.01	1	D	13,WP5I2OB
14	59	52	0.88	-1.49	0.5	0.65	0.5	1	D	14,WP5I3OB
15	59	43	0.73	0.01	0.35	0.82	1.62	1	D	15,WP5I4OB
16	59	35	0.59	0.87	0.31	0.09	0.41	1	D	16,WP5I5OB
17	59	36	0.61	0.77	0.31	-1.59	-0.85	2	D	17,WP6I1OB
18	59	41	0.69	0.25	0.34	-0.4	-0.37	2	D	18,WP6I2OB
19	115	223	*1.94	2.11	0.08	-0.86	0.92	1	0	19,WP11I1OBNB
20	115	148	*1.29	2.58	0.08	-0.8	0.05	1	0	20,WP11I2OBNB
21	56	50	0.89	-1.28	0.54	1.43	0.38	1	D	21,WP5I1NB
22	56	48	0.86	-0.76	0.48	1.68	0.63	1	D	22,WP5I2NB
23	56	46	0.82	-0.34	0.43	-0.13	-0.18	1	D	23,WP5I3NB
24	56	45	0.80	-0.16	0.42	-0.22	-0.49	1	D	24,WP5I4NB
25	56	39	0.70	0.71	0.35	-1.07	-1.02	1	D	25,WP5I5NB

Table 34 continued

26	56	35	0.63	1.16	0.33	0.05	0.06	2	D	26,WP6I1NB
27	56	39	0.70	0.71	0.35	-0.76	-0.5	2	D	27,WP6I2NB
28	118	114	0.97	-2.72	0.57	-0.35	0.78	1	D	28,WP1I5OCNC
29	118	116	0.98	-3.56	0.76	-0.49	0.23	1	D	29,WP1I6OCNC
30	118	113	0.96	-2.43	0.52	0.52	1.18	1	D	30,WP1I7OCNC
31	118	178	*1.51	3.31	0.1	-0.08	0.27	1	0	31,WP11I1OCNC
32	118	127	*1.08	3.79	0.1	-1.31	5.29	1	0	32,WP11I2OCNC
33	61	55	0.90	-1.44	0.51	-0.76	0.8	1	D	33,WP5I1OC
34	61	46	0.75	0.22	0.38	-0.51	-0.23	1	D	34,WP5I2OC
35	61	52	0.85	-0.77	0.44	0.48	2.35	1	D	35,WP5I3OC
36	61	42	0.69	0.76	0.36	0.64	1.06	1	D	36,WP5I4OC
37	61	36	0.59	1.48	0.34	0.61	2.55	1	D	37,WP5I5OC
38	61	39	0.64	1.13	0.35	-4.08	-2.5	2	D	38,WP6I1OC
39	61	42	0.69	0.76	0.36	-0.55	0.39	2	D	39,WP6I2OC
40	57	48	0.84	-0.07	0.45	0.09	0.32	1	D	40,WP5I1NC
41	57	47	0.82	0.13	0.43	-0.36	-0.17	1	D	41,WP5I2NC
42	57	52	0.91	-1.08	0.57	-0.34	-0.13	1	D	42,WP5I3NC
43	57	33	0.58	1.97	0.32	-0.56	-0.3	1	D	43,WP5I4NC
44	57	37	0.65	1.54	0.34	0.24	-0.19	1	D	44,WP5I5NC
45	57	35	0.61	1.76	0.33	0.39	0.14	2	D	45,WP6I1NC
46	57	42	0.74	0.92	0.37	0.67	0.88	2	D	46,WP6I2NC
47	113	104	0.92	-2.08	0.49	-0.72	-0.41	1	D	47,WP1I5ND
48	113	104	0.92	-2.08	0.49	-0.35	-0.3	1	D	48,WP1I6ND
49	113	106	0.94	-2.65	0.57	-0.53	0.03	1	D	49,WP1I7ND
50	113	98	0.87	-0.99	0.38	0.28	0.78	1	D	50,WP5I1ND
51	113	87	0.77	0.21	0.3	0.24	0.4	1	D	51,WP5I2ND
52	113	101	0.89	-1.47	0.42	-0.42	0.17	1	D	52,WP5I3ND
53	113	80	0.71	0.79	0.28	1.21	0.34	1	D	53,WP5I4ND
54	113	69	0.61	1.54	0.25	1.97	1.28	1	D	54,WP5I5ND
55	113	60	0.53	2.08	0.24	-1.56	-0.24	2	D	55,WP6I1ND
56	113	70	0.62	1.48	0.25	-2.57	-0.98	2	D	56,WP6I2ND
57	113	143	*1.27	3.61	0.1	-1.79	0.11	1	0	57,WP11I1ND
58	113	156	*1.38	3.5	0.1	0.39	0.46	1	0	58,WP11I2ND

5.4.2 Equating of Writing Forms

The following sections replicate those of section 5.3.2, except these data are for writing rather than reading.

5.4.2.1 Writing Form Old B

Again, since all other forms are equated to Old B, there is only one curve in Figure 22, and in the conversion charted in Table 35, the raw scores are the scale scores. Figure 23 and Table 36 replicate the earlier figures for Old Form B (Figure 6 and Table 15).

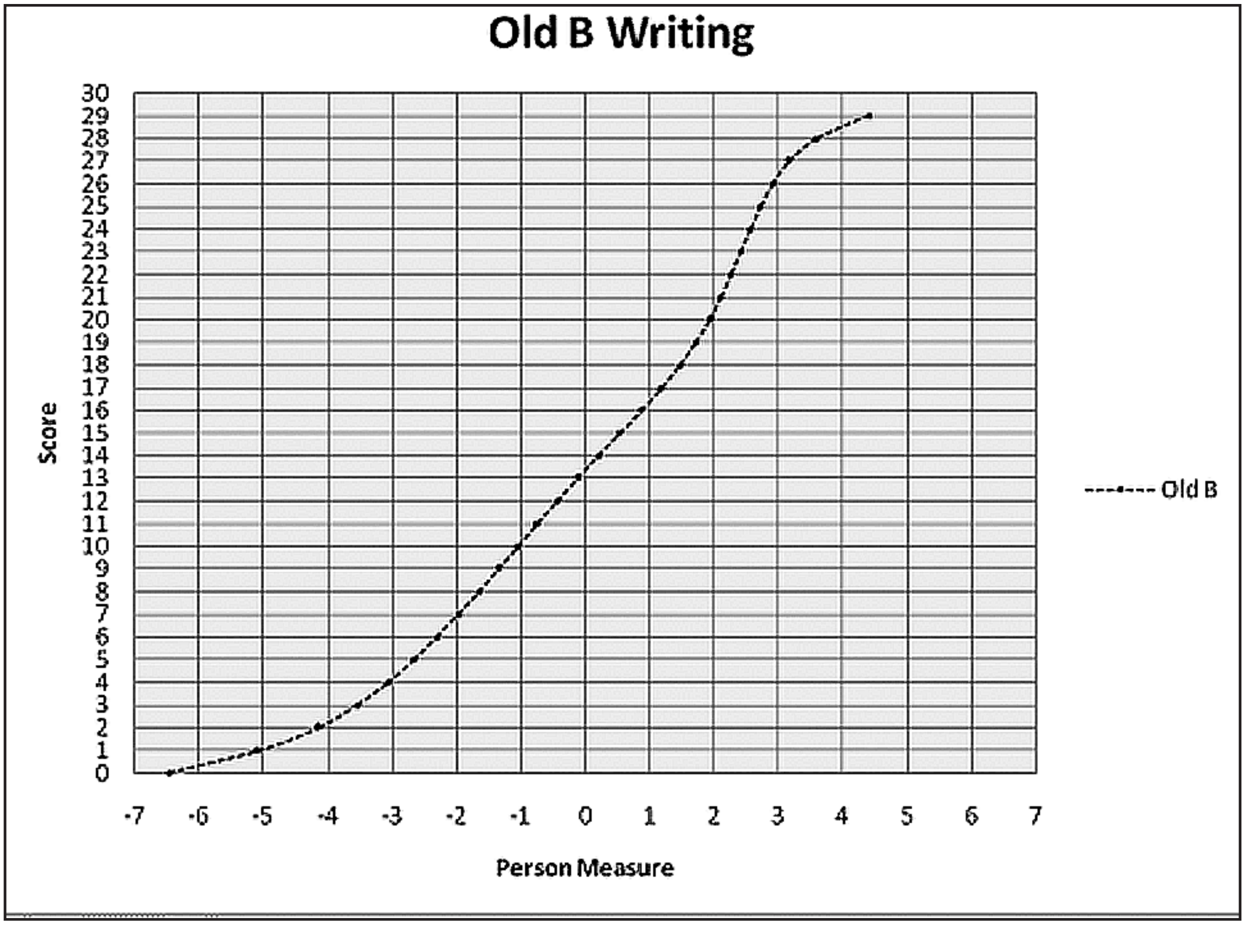

Figure 22. Old B Writing Graph

Table 35
Old B Writing Conversion Chart

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

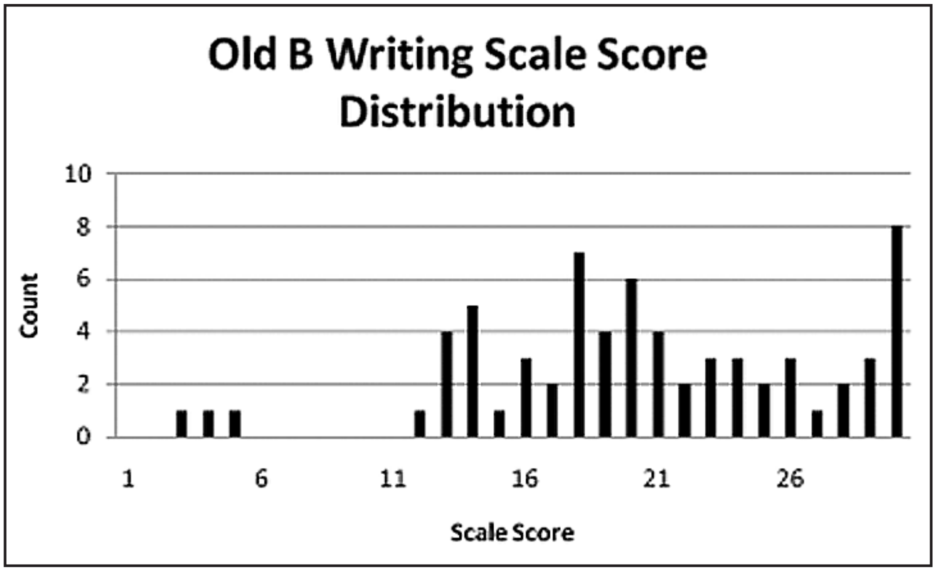

Figure 23. Old B Writing Scale Score Distribution

Table 36
Old B Writing Scale Score Descriptive Statistics

No. of Students	Min	Max	Mean	Std. Dev
67	2	29	19.48	6.49

5.4.2.2 Writing Form New B

As Figure 24 clearly shows, the equating results for New B indicate that the updated writing form is very similar to Old B. As with reading, these results indicate that the effects of the updates and modifications to the literacy skills section of the *BEST* were negligible on the overall difficulty of the test form, making it neither significantly easier nor harder. However, in the middle of the ability distribution (from about -2.5 logits to 1.5 logits), or between 7 and 17 raw score points, Form New B appears slightly more difficult than Form Old B. In other words, it took slightly more ability to achieve the same raw score. However, there was a 1-point difference for only a few scale score points; for the remaining scale points, the scale scores were identical to the raw scores.

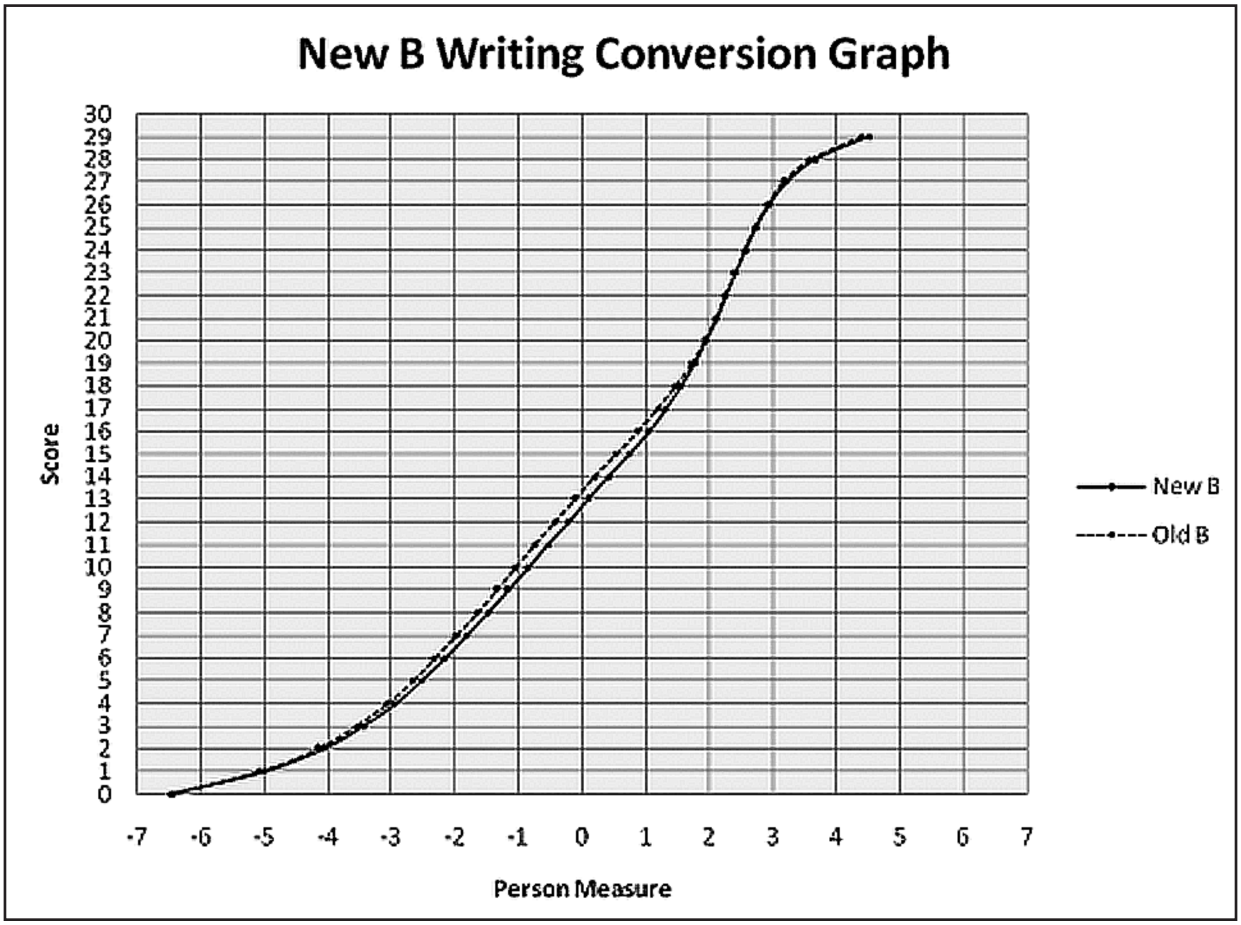

Figure 24. New B Writing Conversion Graph

Table 37
New B Writing Conversion Chart

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

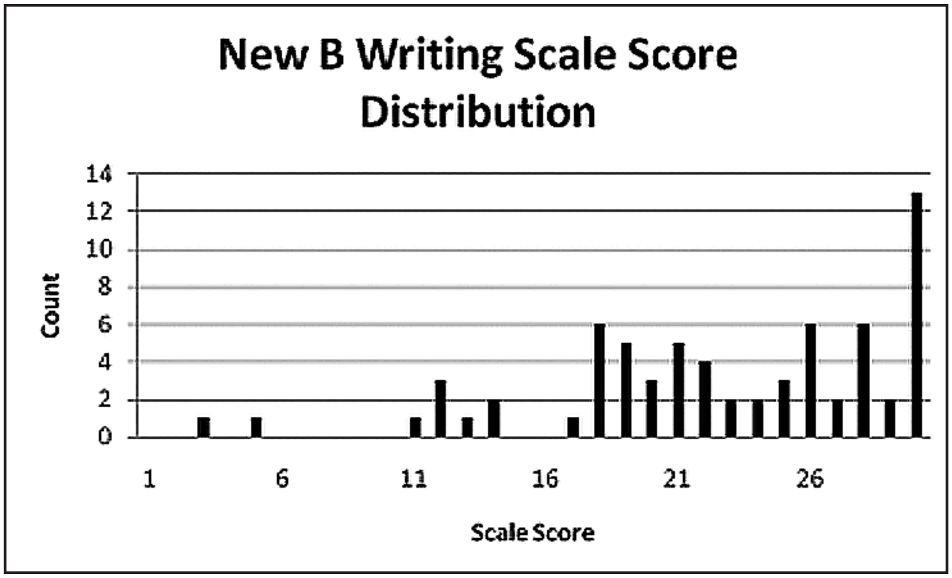

Figure 25. New B Writing Scale Score Distribution

Table 38
New B Writing Scale Score Descriptive Statistics

No. of Students	Min	Max	Mean	Std. Dev
69	2	29	21.67	6.37

5.4.2.3 Writing Form Old C

Figure 26 shows the relationship between writing on Form Old C and writing on Form Old B from this study. Although Form Old C will no longer be used, it may be noted that there is a wide difference in the degree of difficulty of the Old C and Old B that did not seem to be captured in the conversion provided by the 1984 *BEST Test Manual*. From the conversion table in that manual, based on an older linear conversion methodology, it would appear that the two forms were basically identical in difficulty. In this study, the divergence begins at around a raw score of 9 and continues to widen, particularly in the scoring area obtainable by points on the note-writing tasks (i.e., above 19).

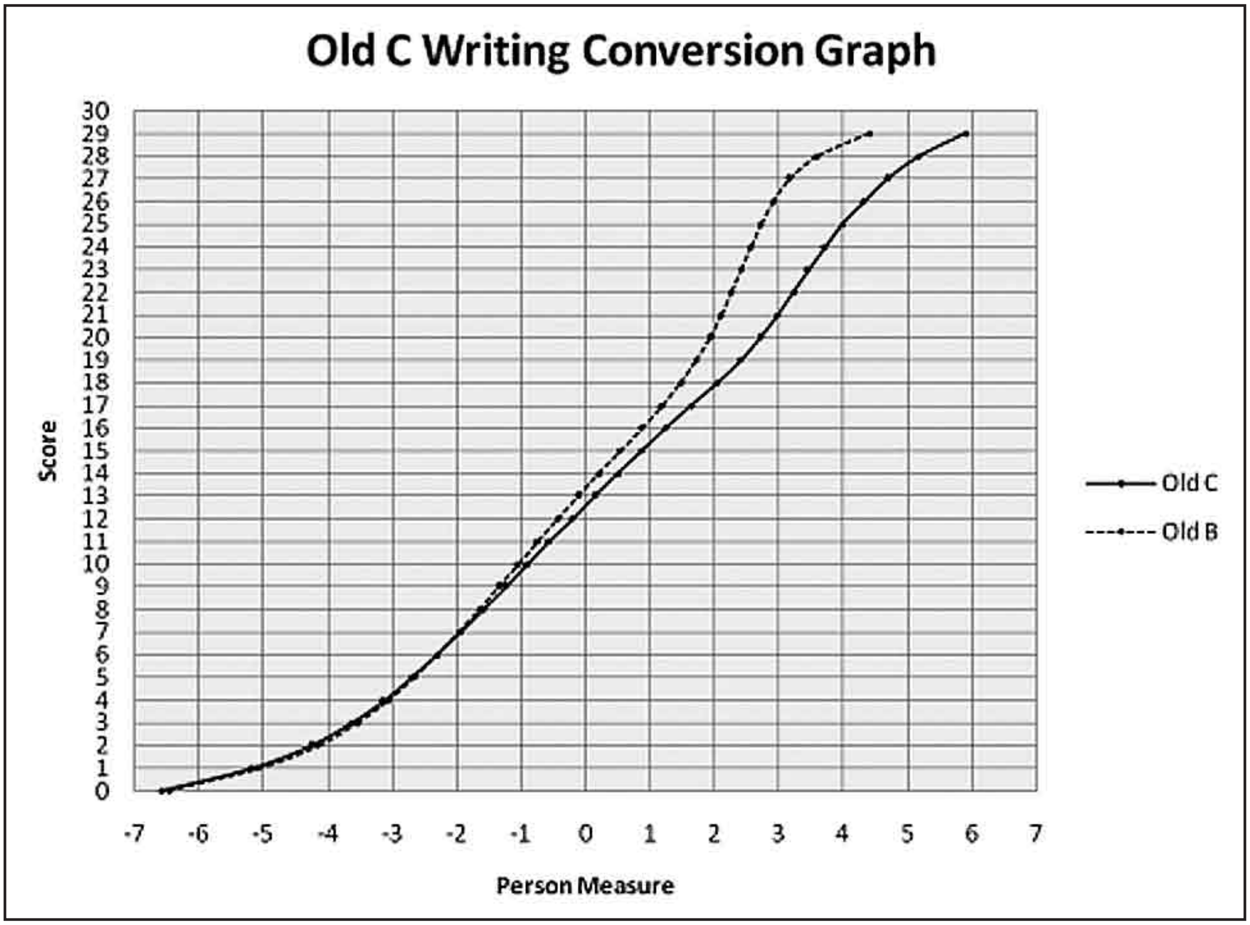

Figure 26. Old C Writing Conversion Graph

Table 39
Old C Writing Conversion Chart

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

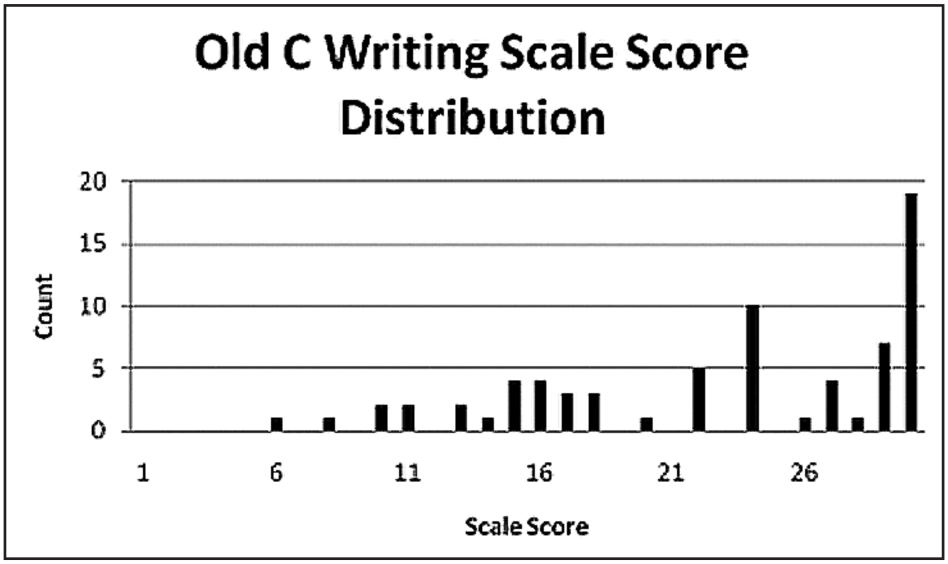

Figure 27. Old C Writing Scale Score Distribution

Table 40
Old C Writing Scale Score Descriptive Statistics

No. of Students	Min	Max	Mean	Std. Dev
71	5	29	21.96	6.93

5.4.2.4 Writing Form New C

As can be seen in Figure 28, the relationship between Form New C and Form Old B appears similar to that of Form Old C and Form Old B. New C appears to be much more difficult than Old B, particularly at the upper end of the score scale.

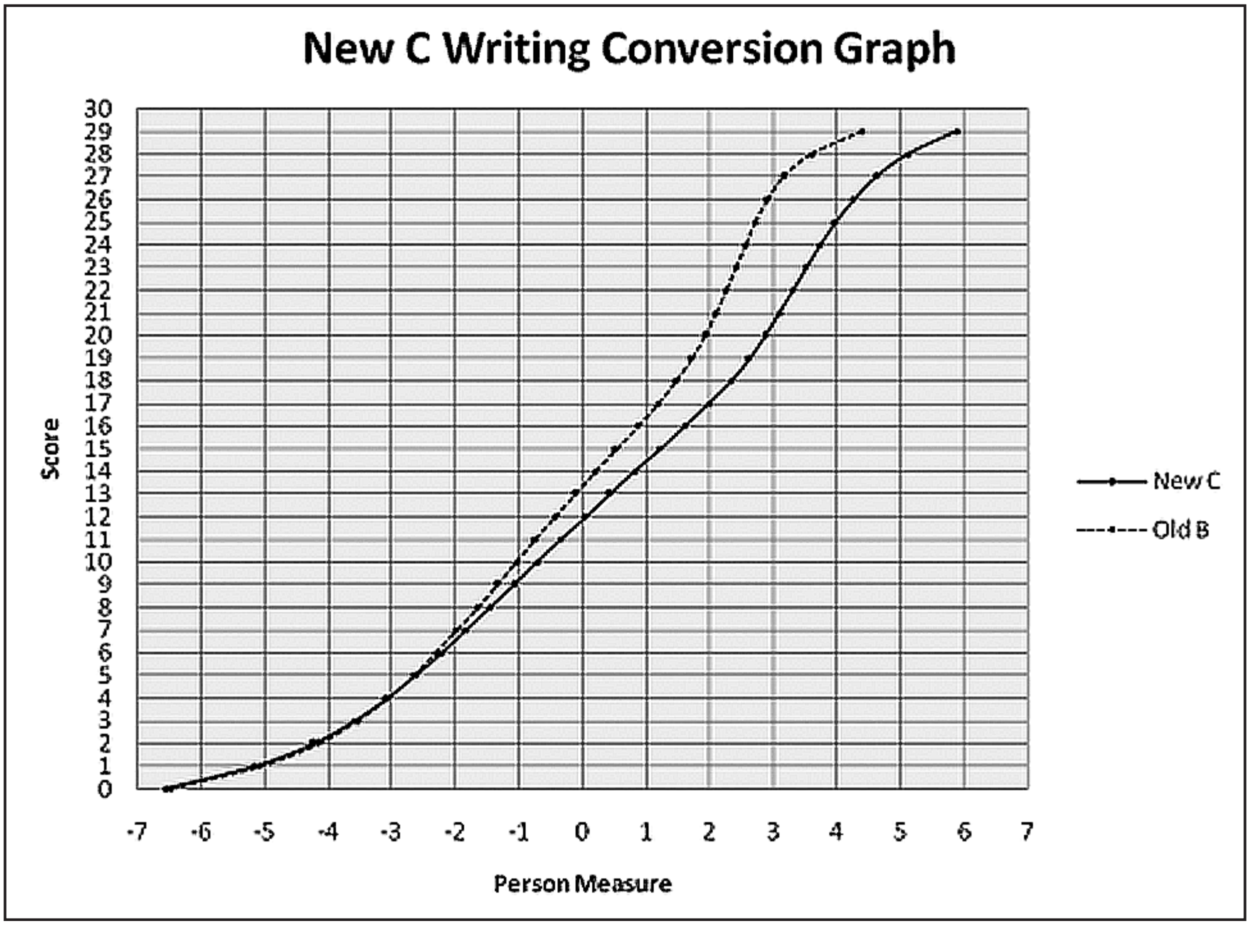

Figure 28. New C Writing Conversion Graph

Table 41
New C Writing Conversion Chart

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

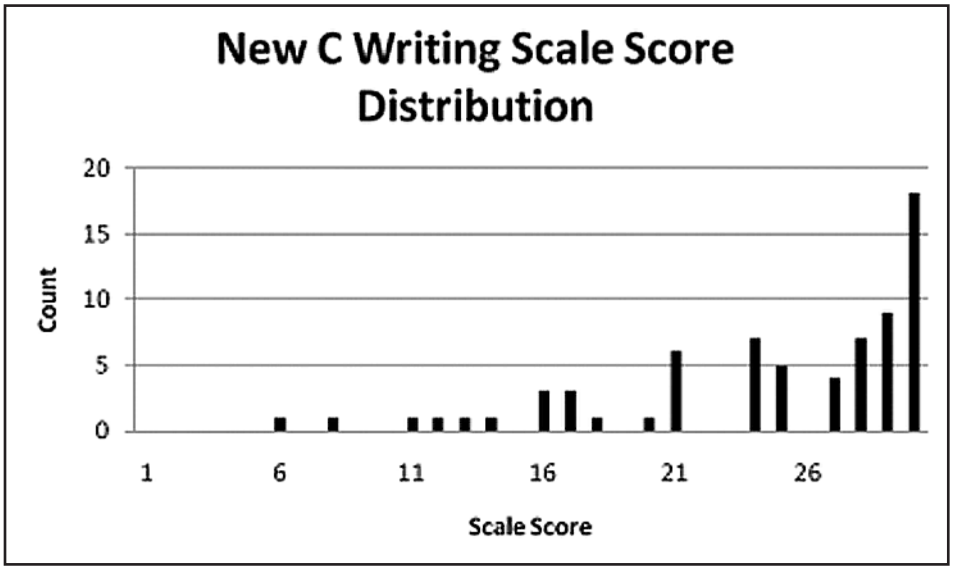

Figure 29. New C Writing Scale Score Distribution

Table 42
New C Writing Scale Score Descriptive Statistics

No. of Students	Min	Max	Mean	Std. Dev
70	5	29	23.64	6.09

In order to better compare the effect of the updates to the writing tasks on Form New C, Figure 30 compares equating results from Forms Old C and New C based on the current study. Like Old B and New B, they are almost identical, except in the middle (scores 10 to 20), suggesting that the improvements and changes that were made to the test form did not affect the item properties extensively.

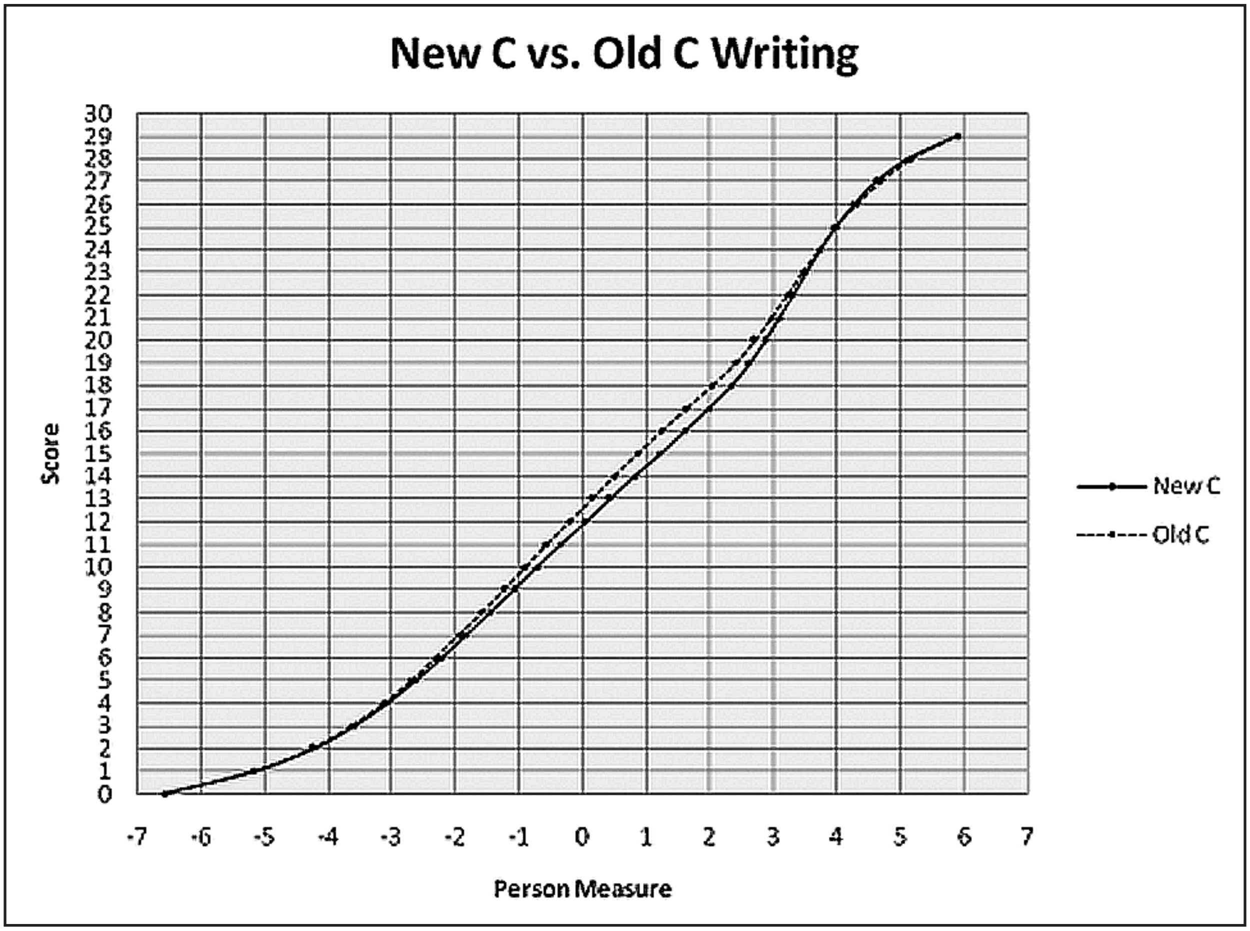

Figure 30. New C vs. Old C Writing Scale Score

5.4.2.5 Writing Form New D

Like Form New C, Form New D appeared much more difficult than Old B. This is clearly seen in Figure 31.

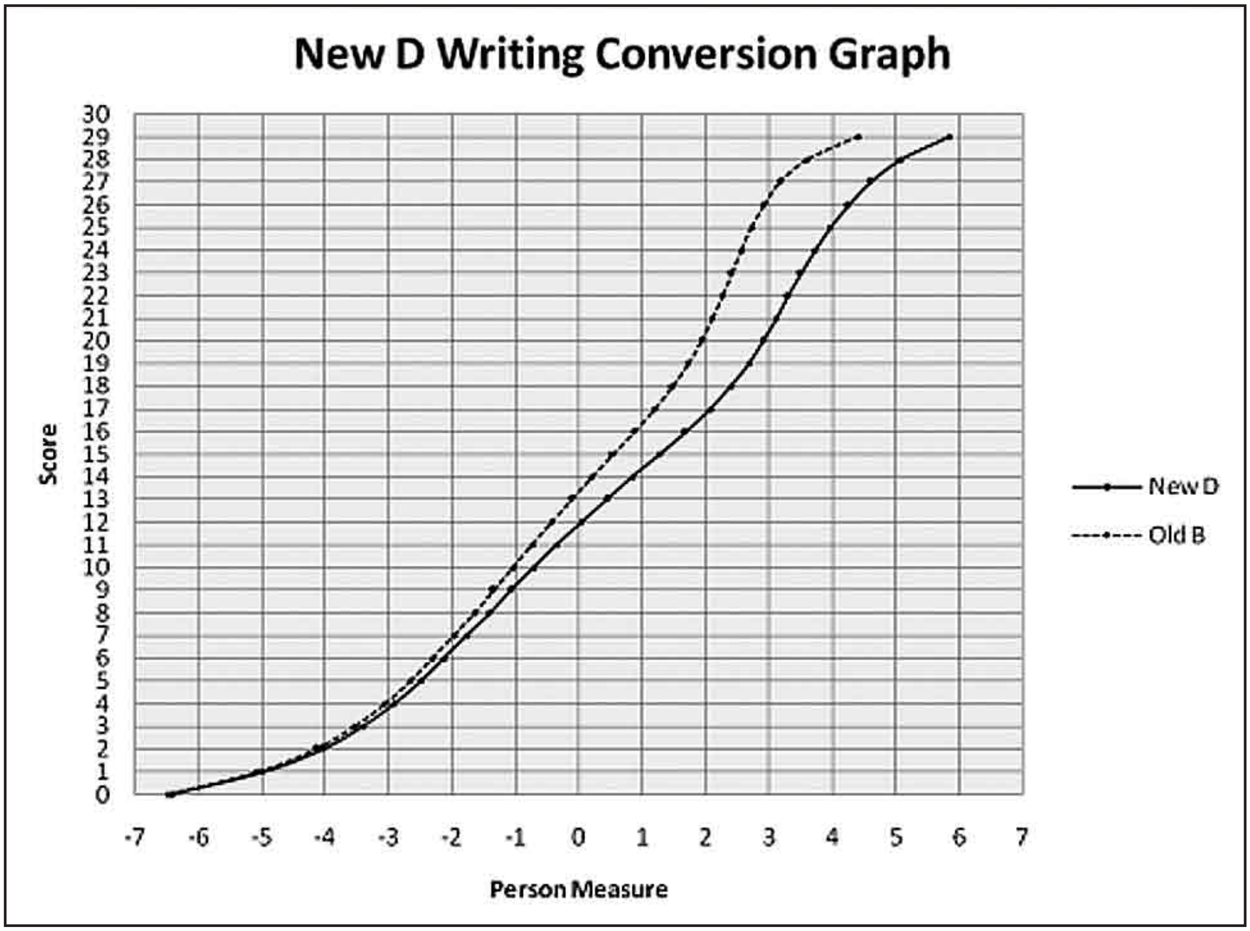

Figure 31. New D Writing Conversion Graph

Table 43
New D Writing Conversion Chart

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

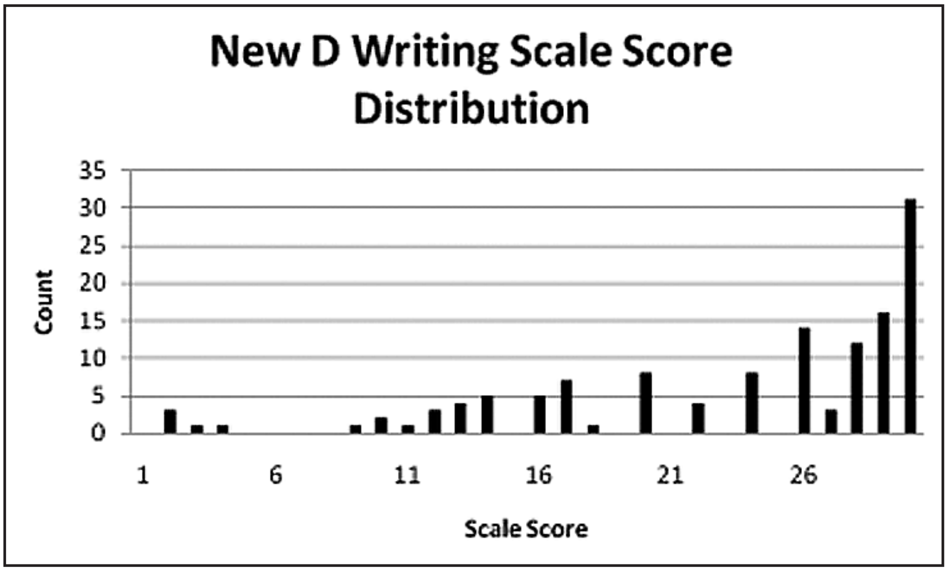

Figure 32. New D Writing Scale Score Distribution

Table 44
New D Writing Scale Score Descriptive Statistics

No. of Students	Min	Max	Mean	Std. Dev
130	1	29	22.41	7.41

5.4.2.6 Writing All Forms

The maximum possible scale score for writing is 29. Table 45 shows the number of people who took each form, together with the mean and the standard deviation of the scale scores for each writing form. Because these forms have been equated, the final column shows the mean across all test forms.

Table 45
Writing Scale Score Mean and Standard Deviation by Test Form

	Old B	New B	Old C	New C	New D	Total
Number of people	67	69	71	70	130	407
Mean	19.48	21.67	21.96	23.64	22.41	21.59
Std. Deviation	6.493	6.372	6.929	6.086	7.410	6.825

The results presented in Table 45 show that performances on New B, Old C, and New D are close to one another and within one point of the mean for all test takers. This provides some evidence that the randomization process used in the field test was successful. However, the results also suggest that the students who took Old B were somewhat weaker in their writing ability than the total group, as their mean (19.48) was more than two points lower than the mean for the total group (21.59). However, those who took New C appeared to be somewhat stronger in their writing ability than the total group, as their mean (23.64) was more than two points higher than the group total.

Again, to examine whether these differences were statistically significant, a one-way ANOVA was run. The results showed that the difference in the mean equated scale scores for writing among the five groups was statistically significant, $F(4,402) = 3.484$, $p = .008$. Post hoc analyses using the Scheffé post hoc criterion for significance indicated that only the mean scale score for writing on Old B (19.48) was statistically significantly different (lower) from the mean scale score for New C (23.64), $p = .013$. The effect size (Cohen's *d*) was .67, indicating a medium effect. This result suggests that although on the basis of writing raw scores the method of randomization of the test booklets used in the field test was successful, the group of students who were administered Old B may well have been somewhat weaker writers than those in the New C group when differences in difficulty of the test forms are taken into account via equating. The Scheffé post hoc criterion for significance indicated no statistical differences between any other pairs of groups.

5.4.2.7 Performance of the Writing Scales (Note-Writing Tasks)

As mentioned earlier, the partial credit model was used to analyze performances on the note-writing tasks. A rating scale model, in which a single empirical rating scale is modeled to score all tasks, was tried initially for analyses, but it produced poor model fit. The partial credit model treats the empirical scales used for each task independently of each other, as if a different rubric were used for each. Although only one general rubric is used to score the note-writing tasks on *BEST Literacy*, there are also specific scoring instructions for individual tasks. Note that data from only 346 people (rather than 407) are used in this analysis; the ability of participants with perfect scores in the writing section cannot be estimated, so their scores are not counted in the analysis.

Tables 46 to 51 provide information on the technical properties of the empirical rating scales used in the scoring of each note-writing task. The tables show how well the rating scales fit the data. In each table, the first column, Points, shows the number of points awarded for each category in the common *BEST Literacy* rubric for the note-writing tasks. In other words, these are the category names in terms of the *BEST Literacy* rubric for the note-writing tasks. The next column, Observed Count, shows the number of examinees in the analysis who received a rating of that scale step. The next column, Observed Percentage, shows the percentage of students scoring at that scale step. Note that the bottom row gives the total count and the total percentage (i.e., 100%).

The next column in each table shows the observed average measure of all examinees receiving that rating, while the following column, Expected Measure, shows what the partial credit measurement model would predict as the measure corresponding to that rating. These two measures, observed and expected, should be close to one another if the rating scale fits the data. The last two columns are infit and outfit mean-square statistics. These fit statistics, provided by the Winsteps computer program, are different from the z-standardized fit statistics in that they have an expectation of 1 rather than 0 when the model fits the data. While there is greater variability in these statistics when compared to items, and they can be significantly affected by the number of examinees falling into each level (i.e., they are generally less reliable the fewer students there are in the rating category), the statistics in these tables, particularly the infit statistics, show that the scales in general had a good fit to the Rasch measurement model.

Table 46
Score Scale Statistics for *WP11I1OBNB*

Points	Observed Count	Observed Percentage	Observed Average	Expected Measure	Infit Mnsq	Outfit Mnsq
0	44	38	.09	.16	.55	.64
1	18	16	1.59	1.45	.74	.32
3	30	26	2.30	2.25	.96	2.64
5	23	20	2.85	2.89	1.17	1.45
Total	115	100				

Table 47
Score Scale Statistics for *WP11I2OBNB*

Points	Observed Count	Observed Percentage	Observed Average	Expected Measure	Infit Mnsq	Outfit Mnsq
0	73	63	.75	.77	.79	.73
1	7	6	1.91	1.85	.29	.17
3	17	15	2.59	2.54	.80	1.14
5	18	16	3.04	3.02	1.06	.97
Total	115	100				

Table 48
Score Scale Statistics for *WP11I1OCNC*

Points	Observed Count	Observed Percentage	Observed Average	Expected Measure	Infit Mnsq	Outfit Mnsq
0	50	42	.68	.76	.56	.75
1	27	23	2.40	2.26	.96	.58
3	27	23	3.50	3.43	.82	1.33
5	14	12	4.24	4.36	1.68	1.93
Total	118	100				

Table 49
Score Scale Statistics for *WP11I2OCNC*

Points	Observed Count	Observed Percentage	Observed Average	Expected Measure	Infit Mnsq	Outfit Mnsq
0	71	60	1.27	1.23	.55	.71
1	16	14	2.42	2.62	.86	9.90
3	22	19	3.77	3.78	.83	4.02
5	9	8	4.55	4.49	.94	.88
Total	118	100				

Table 50
Score Scale Statistics for *WP11I1ND*

Points	Observed Count	Observed Percentage	Observed Average	Expected Measure	Infit Mnsq	Outfit Mnsq
0	62	55	.75	.80	.76	.80
1	16	14	2.67	2.63	.27	.14
3	24	21	3.72	3.63	.71	1.36
5	11	10	4.32	4.29	.94	.88
Total	113	100				

Table 51
Score Scale Statistics for *WP11I2ND*

Points	Observed Count	Observed Percentage	Observed Average	Expected Measure	Infit Mnsq	Outfit Mnsq
0	59	52	.73	.70	.81	.86
1	15	13	2.32	2.54	.75	1.16
3	27	24	3.67	3.55	.86	.99
5	12	11	4.07	4.25	1.76	1.51
Total	113	100				

5.5 Results for Total Scores

Decisions are made about examinees who take *BEST Literacy* on the basis of a total score that combines reading and writing. First, the separate scores are scaled using raw score to scale score conversion tables, then the two scores are added for the composite total scale score.

We want to examine potential differences in performance in terms of total scores among the groups of students taking each test form in the field-test study. Across the five test forms, Table 52 shows the number of students who took each form, the average total scale score, and the standard deviation of the total scale scores. The largest difference in means, in terms of total scale scores, was between Old B (52.34), with the lowest average score, and New B (56.59), with the highest. To examine whether these differences were statistically significant, a one-way ANOVA was run. The results showed that the differences in the mean total scale scores among the five groups were not statistically significant, $F(4,402) = .647$, $p = .629$. That means that the differences in the ability of the groups, in terms of total scale scores, were due to random error and not to genuine differences in abilities in the groups. This result provides further evidence that the method of randomization of the test booklets used in the field test was successful.

Table 52
Total Scale Score Average and Standard Deviation by Test Form

	Old B	New B	Old C	New C	New D
Number of students	67	69	71	70	130
Mean	52.34	56.59	55.04	55.33	54.42
Std. Deviation	14.93	14.42	16.36	15.94	17.22

6. Development of Final Forms

6.1 Development of Final Updated Test Forms (2006)

The analyses reported in section 4.3 and section 5 revealed that performances on the three forms of *BEST Literacy* (New B, C, and D) were generally equivalent to those on the forms of the *BEST* literacy skills section (Old B and C). In other words, reading performances on Old B (the *BEST*) were equivalent to reading performances on New B (*BEST Literacy*), and reading performances on Old C (the *BEST*) were equivalent to reading performances on New C (*BEST Literacy*). Performances on the reading and writing sections of New D (*BEST Literacy*) were very similar to those on New C. The difficulty of items from the old *BEST* forms was replicated on the new versions of the forms used in *BEST Literacy*.

However, this analysis indicated that the six note-writing tasks that appear on the three test forms were clearly not of equal difficulty. The two note-writing tasks on Form B appeared significantly easier than the four (two each) on Forms C and D. For this reason, in developing the final versions of *BEST Literacy*, the note-writing tasks were redistributed from their original locations on Forms B, C, and D to new locations over the three test forms to ensure that the updated Forms B, C, and D were of more equivalent difficulty. (Note that the redistribution, however, does not affect the use of Old B or Old C as a pretest with New B or New C as a posttest, as may occur as *BEST Literacy* gradually replaces the literacy skills section of the *BEST*. In other words, a student taking Old B as a pretest and New C as a posttest will not encounter the same note-writing tasks.) Any remaining differences in degree of difficulty among the three forms of *BEST Literacy* are resolved by using the *BEST Literacy* scale scores as the basis for comparison of student performances on the different test forms. The *BEST Literacy* scale scores are based on the raw scores used on Old B.

Figures 33 to 35 and Tables 53 to 55 show the effect of the redistribution of the note-writing tasks on the writing section of the final three forms of *BEST Literacy*. Each figure shows Old B and the comparison form before redistribution of the note-writing tasks. The darkest curve in each figure shows the final version of the comparison form. After each figure, the table shows the final raw to scale score conversion for each form.

6.1.1 Final Writing Form New B

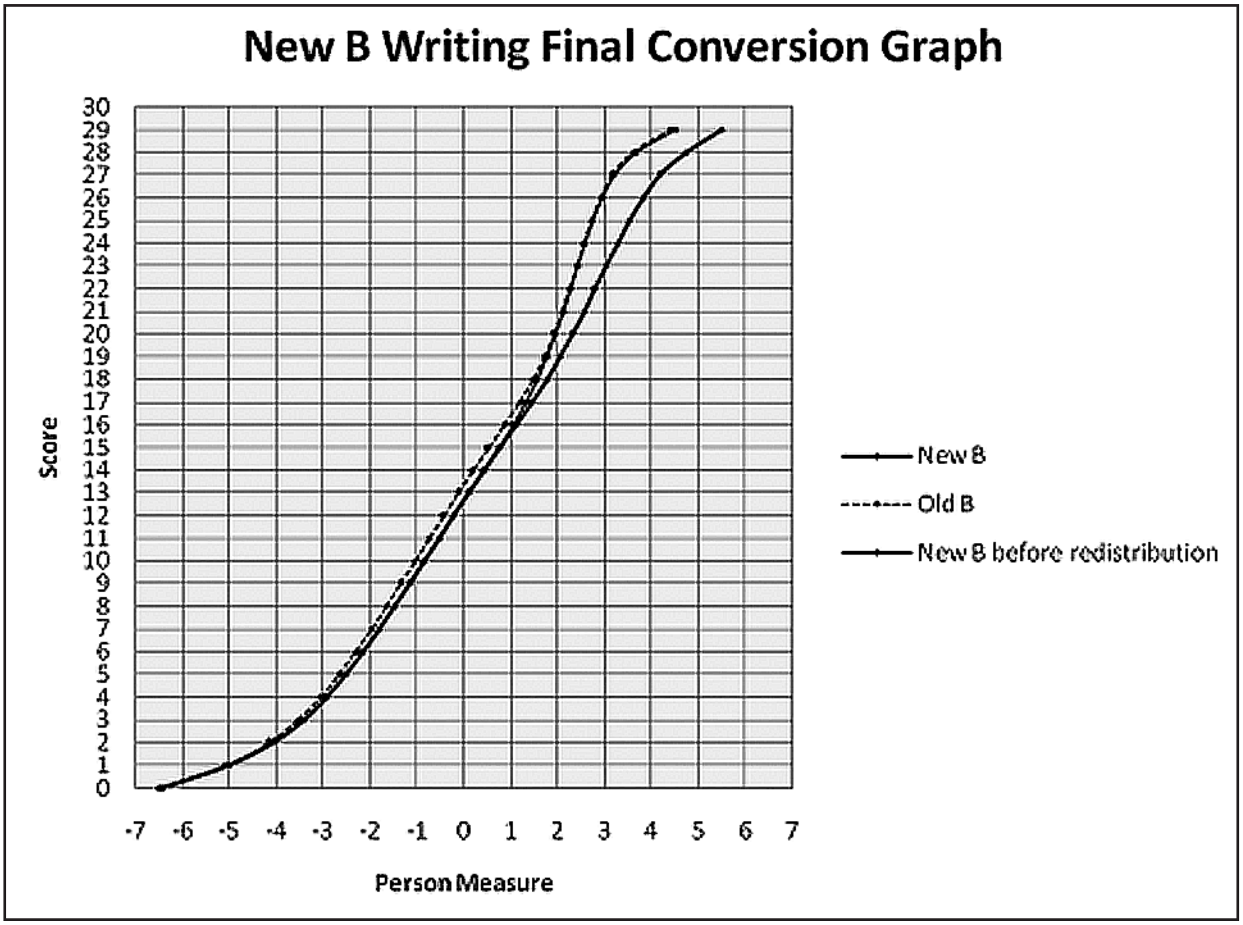

Figure 33. New B Writing Final Conversion Graph

Table 53
New B Writing Final Conversion Chart

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

6.1.2 Final Writing Form New C

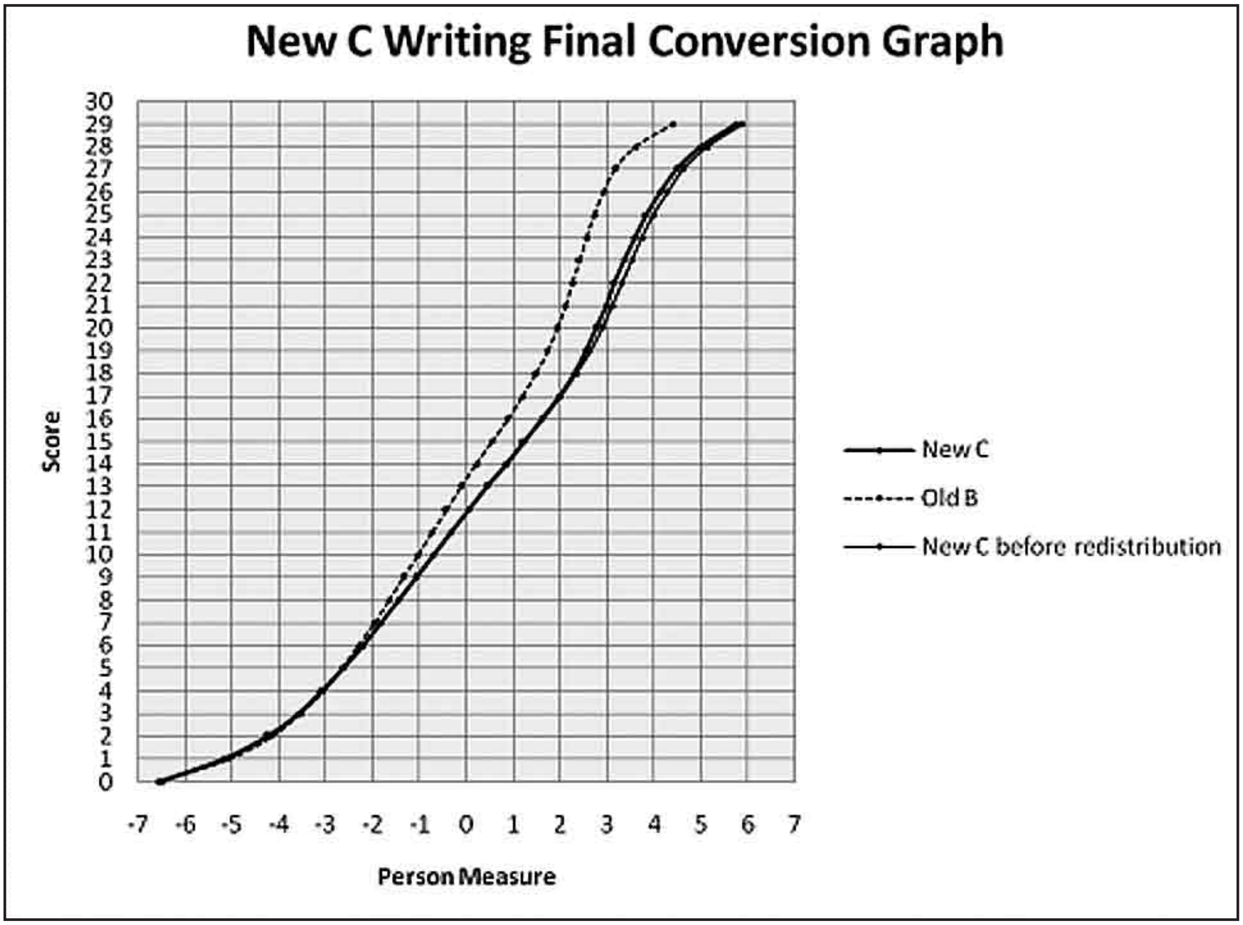

Figure 34. New C Writing Final Conversion Graph

Table 54
New C Writing Final Conversion Chart

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

6.1.3 Final Writing Form New D

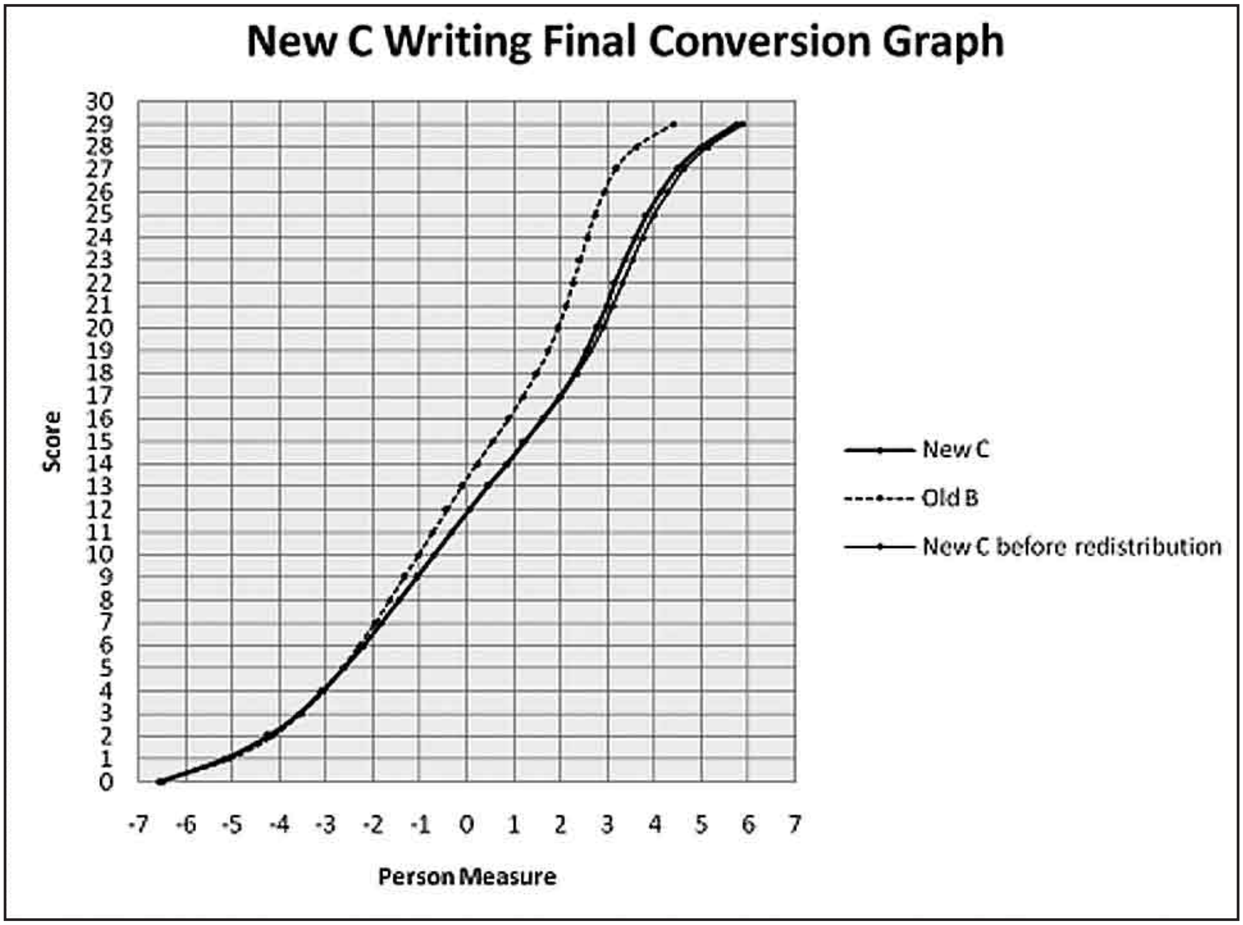

Figure 35. New D Writing Final Conversion Graph

Table 55
New D Writing Final Conversion Chart

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

6.2 Comparison of Final Writing Forms

Figures 36 and 37 compare the difficulty of the three writing forms of *BEST Literacy* before and after the redistribution of the note-writing tasks. Figure 36 shows how, before the redistribution, New B was very close to Old B, meaning that the scale scores were not too distant from the raw scores. However, the other two forms (New C and New D) were far away from Old B, meaning that the gap between the raw scores and scale scores was great.

Figure 37, however, shows that after the redistribution of the writing tasks, all three forms (New B, New C, and New D) are very close to each other, meaning that they are very similar in difficulty. They are each also somewhat closer to Form Old B, making the gap between the raw score and scale scores smaller.

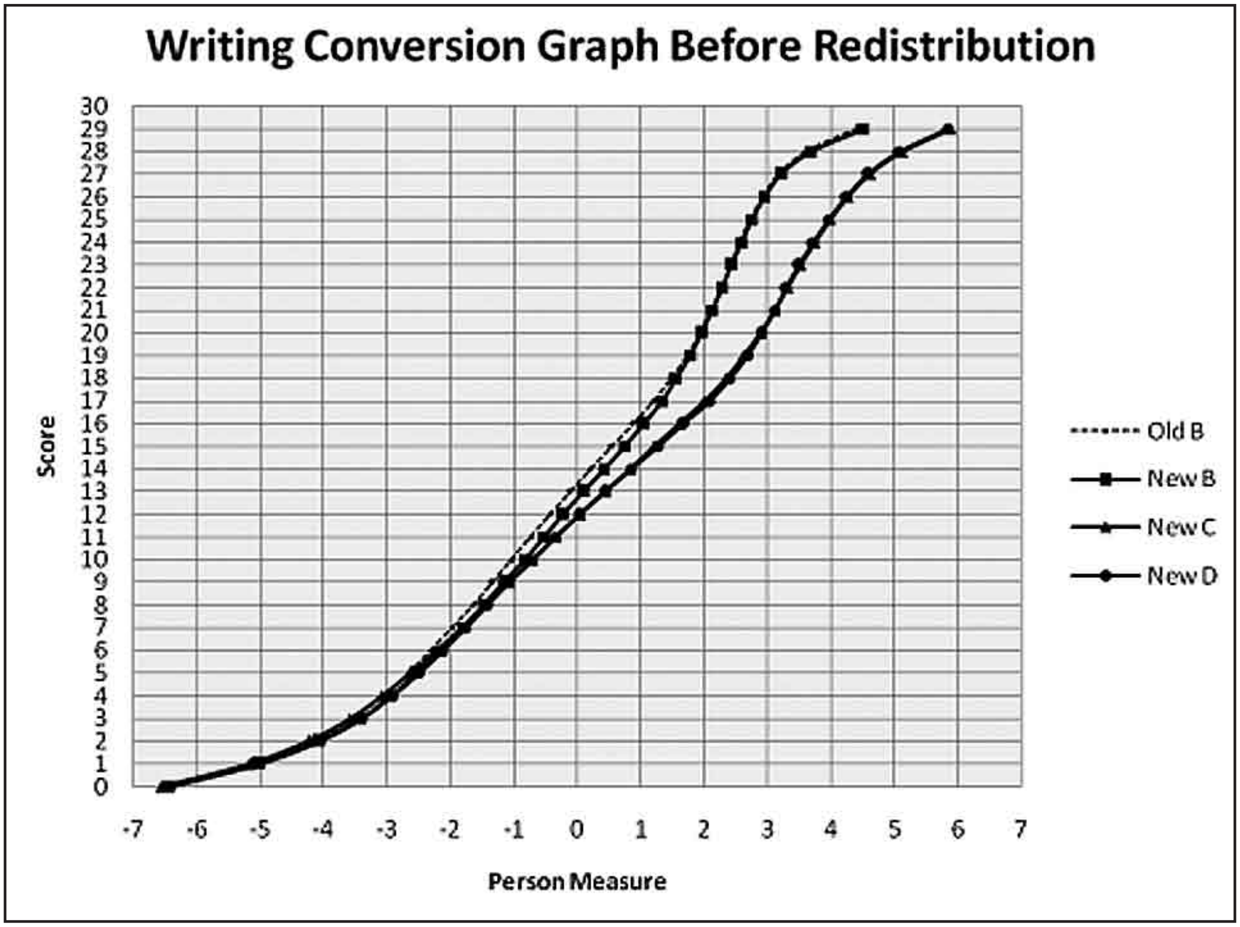

Figure 36. Writing Conversion Graph Before Redistribution

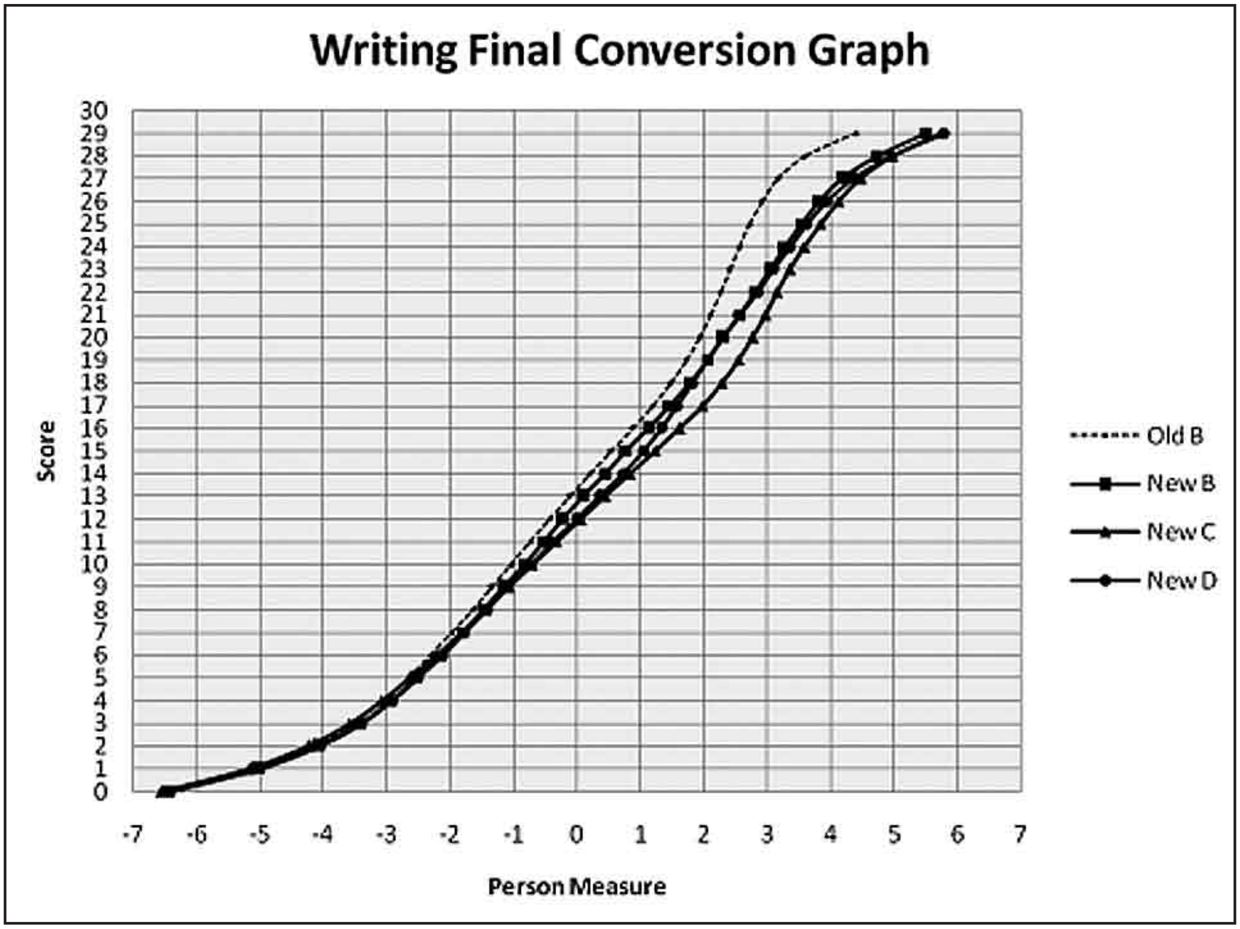

Figure 37. Writing Final Conversion Graph

6.3 Item Statistics for Final *BEST Literacy* Test Forms

Appendix B contains the item statistics for the final three forms of *BEST Literacy*. These forms are Form B, Form C, and Form D. The statistics come from the field-test data. Users of the test can use these tables, along with all the other information in this report, to get an indication of the overall quality of each test form.

7. *BEST Literacy* Standard Setting Study

7.1 Introduction

A standard setting study is a replicable, fully documented process used to establish one or more cut scores on a test in a defensible manner. Several new approaches to conducting standard setting studies have been developed over the past decade to complement more traditional approaches that were developed for multiple-choice testing. To evaluate any study, it is necessary to provide evidence of the reasonableness and replicability of the standard setting process used (Cizek & Bunch, 2007).

Reasonableness refers to "the degree to which cut scores derived from the standard setting study classify examinees into groups in a manner consistent with other information about the examinees" (Cizek & Bunch, 2007, p. 61). For example, if cuts derived from a standard setting study on one test classify 70% of students as proficient when cuts for another test classify 40% of students from the same population as proficient, the reasonableness of one or both of the standard setting processes may be questionable.

Standard setting is a costly and labor-intensive process. Resources generally do not allow studies to be replicated. However, to ensure valid outcomes, the procedures and processes used in every study must be replicable. Thus, a report on a standard setting needs to contain complete details in order to judge the study's replicability.

This section of the report provides a detailed description of a standard setting study conducted for *BEST Literacy*. Materials used, procedures followed, and outcomes are fully described.

7.2 Purpose of the *BEST Literacy* Standard Setting Study

The standard setting study for *BEST Literacy* was conducted at CAL on April 3 and 4, 2007. The goal of the study was to relate performances on *BEST Literacy* (i.e., total scale scores) to the performance descriptors of the recently revised ESL Educational Functioning Levels of the National Reporting System (NRS) and to the revised performance descriptors of the reading and writing Student Performance Levels (SPLs). (For complete descriptors and further information about them, see the *BEST Literacy Test Manual* [Center for Applied Linguistics, 2008].) The original SPLs had been used since the 1980s to interpret performances on the literacy skills section of the *BEST*, the predecessor to *BEST Literacy*. In recent years, however, the NRS has become increasingly important to programs for federal reporting purposes. Thus this study included both SPLs and NRS levels.

Prior to this standard setting study, the *BEST* cut scores corresponding to the NRS levels were determined using the original *BEST* cut scores for the SPLs. That is, each of the NRS levels was matched with appropriate SPLs (NRS Beginning ESL Literacy with SPLs 0-1, Beginning ESL with SPLs 2-3, Low Intermediate ESL with SPL 4, High Intermediate ESL with SPL 5, Low Advanced ESL with SPL 6, and High Advanced ESL with SPL 7). The NRS cut scores were the same as those corresponding to the SPLs, as reported in the *BEST Test Manual* of 1984.

Since that time, however, both the NRS levels and the SPLs have been revised (independently of each other). In addition, technical advances have been made in setting cut scores through standard setting studies. And of course, the literacy skills section of the *BEST* was revised to create *BEST Literacy*, so a new standard setting study was called for.

7.3 Procedure

7.3.1 Judges

Ten adult ESL experts from around the country were invited to serve as the judges for the standard setting study. Judges were invited on the basis of their long experience in the field of adult ESL education, their familiarity with the *BEST* tests, and their familiarity with the SPLs and NRS levels. Table 56 shows their names and affiliations.

Table 56
Judges' Names and Affiliations

Name	Affiliations
Allene Grognet	Center for Applied Linguistics (retired), Florida
Barbara Sample	Spring Institute, Colorado
Cindy Shermeyer	Christina Adult Programs, Delaware
David Red	Fairfax County Public Schools Adult ESOL, Virginia
Donna Moss	Arlington Education and Employment Program (REEP), Virginia
Jane C. Miller	Department of Education, Colorado
Jane Schwerdtfeger	Department of Education, Adult & Community Learning Services, Massachusetts
Kate Diggins	Guadalupe School's VIP Program, Utah
Phillip Anderson	Department of Education, Florida
Sarah Young	Center for Applied Linguistics, DC, and REEP, Virginia

The following three tables provide information on the demographics of the judges. Eight of the ten judges were female; two were male (Table 57). Table 58 shows the number of years judges had spent in an adult ESL career; half of them had more than 20 years of experience. Table 59 shows the highest level of education achieved by the judges: eight had a master's degree and two had a doctoral degree. There is no table showing ethnicity because all of the judges were white.

Table 57
Gender of the Judges

	Frequency	Percent
Male	2	20.0
Female	8	80.0
Total	10	100.0

Table 58
Number of Years Experience Judges Had in Adult ESL Careers

	Frequency	Percent
1-5 years	1	10.0
6-10 years	1	10.0
11-15 years	2	20.0
16-20 years	1	10.0
21 or more years	5	50.0
Total	10	100.0

Table 59
Highest Education Level Achieved by Judges

	Frequency	Percent
Master's Degree	8	80.0
Doctoral Degree	2	20.0
Total	10	100.0

Table 60 shows how familiar the judges reported they were with the NRS level descriptors, with the SPL descriptors, and with *BEST Literacy*. They rated their familiarity on a 3-point scale: (3) *Very Familiar*, (2) *Somewhat Familiar*, or (1) *Not Familiar*. Table 60 presents means and standard deviations for each category. As the results suggest, the judges felt quite familiar with both the SPLs and the NRS levels, although slightly more so with the SPLs. They felt somewhat less familiar with *BEST Literacy*.

Table 60
Judges' Familiarity with the NRS, SPL, and *BEST Literacy*

	N	Minimum	Maximum	Mean	Std. Dev
Familiar with NRS levels	10	2	3	2.70	.483
Familiar with SPL levels	10	2	3	2.80	.422
Familiar with *BEST Literacy*	10	1	3	2.20	.789

7.3.2 General Procedures

The standard setting study was facilitated by Dr. Dorry Kenyon, director of the Language Testing Division at CAL. Table 61 shows the agenda for the standard setting study.

Table 61
Standard Setting Study Agenda

Tuesday, April 3, 2007	
8:30 – 9:00	Continental breakfast
9:00 – 9:15	Welcome and introductions (complete background data form)
9:15 – 9:30	Background to the study
9:30 – 9:50	Take the test
9:50 – 10:00	BREAK
10:00 – 10:45	Scoring the test
10:45 – 11:15	Review of NRS descriptors
11:15 – 11:30	Explanation of the process
11:30 – 12:15	Sample and Students 1-3
12:15 – 1:00	LUNCH
1:00 – 2:45	Students 4-13
2:45 – 3:00	BREAK
3:00 – 4:45	Students 14-23
4:45 – 5:00	Evaluation and wrap-up
5:00	END
Wednesday, April 4, 2007	
8:00 – 8:30	Continental breakfast
8:30 – 9:15	Review of NRS results
9:15 – 9:30	BREAK
9:30 – 9:45	Review of SPL descriptors
9:45 – 11:45	SPL study (with odd-numbered students)
11:45 – 12:00	Evaluation and wrap-up
12:00	END (optional lunch)

The first day started with an overview of the study and what was expected of the judges. One of the first activities was for the judges to take New Form D of *BEST Literacy* themselves to become familiar with the test. After the judges took the test, they were walked through the scoring procedure as they scored their own test, to ensure that they understood the criteria for correct answers and the rubric for the note-writing tasks. The judges were then trained on their tasks (see section 7.3.3 for details) and spent the rest of the first day assigning ratings to student tests using the NRS level descriptors. On the second day, the results of the NRS cuts based on the first day's work were reported back to the judges, who then assigned ratings using the SPL descriptors.

7.3.3 The Rating Process

The procedure used was a modified Body of Work method (Kingston, Kahl, & Sweeney, 2001). Test booklets were carefully selected in advance (see Table 63) to serve as portfolios of student work for the judges. Body of Work method was chosen because all the writing items and one third of the reading items were constructed-response items. For those reading items that were selected response (i.e., 15 three-option cloze items and 18 four-option multiple-choice items), examinees still made their marks directly in their test booklets.

The traditional Body of Work method generally uses two separate steps: a range-finding round and a pinpointing round. As Sweeney and Ferdous (2007) point out, doing two separate rounds requires a much larger number of portfolios than the number actually used in the standard setting study. Considering the limitations of the traditional Body of Work method and the number of cuts to be made in our study, and in accord with the approach recommended by Sweeney and Ferdous, we modified the method by including more papers for the range-finding round (i.e., papers that represented the entire range of performances) and eliminated the pinpointing round.

The choice of a modified Body of Work method, in which judges examined actual student work, further proved appropriate when we observed the behaviors of judges in the study. Judges used not only the information about whether a student got an item right or wrong, but also paid attention to and defended their ratings according to other nuances of the student's performance. These nuances included the student's handwriting and, for the selected-response tasks, eraser marks, patterns of right and wrong answers on the page, skip patterns, and different ways of marking the correct answer (e.g., circling the letter, circling the answer, crossing out the correct answer). These additional sources of evidence of student proficiency would have been unavailable to the judges had we not used the Body of Work method.

For the purposes of this report, each NRS level may be referred to interchangeably by either its full name or its level number. For example, the lowest NRS level, Beginning ESL Literacy, may be referred to as level 1; the Low Beginning ESL level may be referred to as level 2, and so forth. The names and corresponding numbers of the NRS levels are shown below:

- Level 1 Beginning ESL Literacy
- Level 2 Low Beginning ESL
- Level 3 High Beginning ESL
- Level 4 Low Intermediate ESL
- Level 5 High Intermediate ESL
- Level 6 Advanced ESL

For each student portfolio, the judges were to apply three steps as they gave a rating. These are the three steps as presented in the instructions to the judges:

1. Decide at which NRS level (1 – Exit) *(or SPL 0-8 for the SPL study)* you feel the student whose work is represented in the test booklet is currently functioning.
2. Think about how confident you are that this student is currently functioning at that NRS level *(or SPL)* (100% - 50%).
3. If not 100% confident of your selection in #2, decide at which adjacent NRS level *(or SPL)* (higher or lower) this student might also be functioning.

Thus, after reading through a student's test booklet, a judge in the NRS study may award, for example, 50% to level 2 and 50% to level 3, or 80% to level 2 and 20% to level 1, or 100% to level 3.

After completing a practice round with a sample portfolio from the middle of the performance continuum, the judges began their work. First, all of the judges read through the same portfolio and marked their decisions on a scoring sheet. When they all had finished, they shared their decisions, which were input into a table and shown immediately to the entire group on a screen, which displayed the average rating in each category. The judges then considered the outcome and discussed it as necessary. In general, judges at the extremes commented on why they awarded the scores they did. Finally, after this discussion, the judges made their final individual ratings. These were shared with the group, entered into a table, and shown on the screen, but no discussion followed. Only the results from the second round were analyzed for the final cuts.

Table 62 shows how the judges rated the general aspects of the standard setting study. The rating options were *Excellent* (4), *Good* (3), *Satisfactory* (2), and *Poor* (1). Means and standard deviations were calculated. As the results suggest, the judges were unanimous in their excellent rating for the introduction to the study and for the workshop leaders. The judges were also highly satisfied with the facilities and food. It appears that CAL staff were successful in helping them feel positively inclined toward their participation in the study.

Table 62
Judges' Evaluation of General Aspects of the Study

	Overall Introduction to the Study	Workshop Leaders	Facilities	Food
Mean	4.00	4.00	3.90	3.90
St. Dev.	.000	.000	.316	.316

7.3.4 Materials

As mentioned earlier, a modified Body of Work method was used. This entailed carefully selecting student portfolios (i.e., student test booklets) for the judges to rate.

Because the final NRS cuts and SPL cuts were going to be expressed in terms of scale scores, we used scale scores in selecting representative samples. The results from the field test were examined to select appropriate student portfolios. Because the NRS and SPL descriptors are applied to adult literacy in general and not separately to reading and writing, the total score (i.e., the reading and writing scale scores combined) was used. The goal was to find test booklets from one test form spanning the range from 0 to 78 (maximum total score) at 3 scale point intervals. However, the reading score and writing score of each test booklet were examined to avoid selecting test booklets that were unbalanced (e.g., had a very high reading score but a very low writing score). In the study, the judges were presented with the student portfolios in order from lowest to highest scale score.

Table 63 provides a complete description of the portfolios used in this study. All test booklets were from Form New D. The first column shows the portfolio number. (The sample portfolio used with the judges following training is also included.) The second column shows the student ID, the third the total scale score, the fourth the reading scale score, the fifth the writing scale score, the sixth the score awarded for the first note-writing task, and the seventh the score awarded on the second note-writing task. It will be noted that, because we did not find any appropriate test booklets to illustrate scores between 6 and 19, we developed two portfolios based on composites of actual student test booklets. It will also be clear that there is a threshold effect in the test in that students generally either do not attempt the note-writing tasks or are unable to score anything on them until they reach a certain number of points on the other reading and writing items.

Table 63
Distribution of Student Portfolios

Portfolio	ID	Total SS	Reading SS	Writing SS	Note 1 Score	Note 2 Score
Sample	*B9*	*55*	*30*	*25*	*1*	*1*
1	A12	1	1	0	0	0
2	B24	6	3	3	0	0
3	composite	11	6	5	0	0
4	composite	16	9	7	0	0
5	C48	19	7	12	0	0
6	G25	23	20	3	0	0
7	B14	27	18	9	0	0
8	I24	30	11	19	0	0
9	G29	35	23	12	0	0
10	B18	39	26	13	0	0
11	G28	45	29	16	0	0
12	E23	47	31	16	0	0
13	I11	51	26	25	0	0
14	G27	54	33	21	0	0
15	D33	55	29	26	3	1
16	D30	60	31	29	5	3
17	D8	63	35	28	3	5
18	F34	65	39	26	3	3
19	E34	67	39	28	3	3
20	G14	68	39	29	5	5
21	I30	71	42	29	5	5
22	E43	75	46	29	5	5
23	E5	78	49	29	5	5

Test booklets for inclusion in the study were selected by CAL staff in an iterative process. Once the final test booklets were chosen, the scores on the variety of items included in the test were clearly marked in the margins with 1s and 0s; a 1 was used if the item was correct, a 0 was used if incorrect. This marking system helped judges understand the student's performance at a glance without interfering with the student's work. For the writing tasks, the rating (0, 1, 3, or 5) was noted at the top of the page containing the student's response. Test booklets were then photocopied and bound into two volumes for the judges' use.

As mentioned earlier, during the judgment process, it became clear that the judges were using ancillary information available in the test booklets to shed further light on student literacy. Handwriting was noted at times, as well as patterns of erasing answers and making corrections. Having just taken and scored the test themselves, the judges compared strategies they used in completing some of the tasks—in particular in composing the notes—with those used by the students.

7.4 Analyses

Following the modified Body of Work method, logistic regression was used to determine, from the data collected from the judges, the point along the underlying proficiency continuum at which at least 50% of the judges would be expected to agree that the portfolio represents the work of the next higher proficiency level rather than the current proficiency level. Logistic regression is used when the outcome variable (dependent variable) is dichotomous, and in this study, the dichotomy is either being rated at the next lower proficiency level (or lower), or being rated at the next higher proficiency level (or higher). In other words, when conducting the analysis between two levels (e.g., 2 and 3), the input data were treated dichotomously as the percent at the lower level (2) or below and the percent at the higher level (3) or above. The total scale score of the student who produced each portfolio was used as the indicator of student proficiency.

To further illustrate how logistic regression was used, Table 64 shows an example of the data that were input to determine the cut score between proficiency levels 3 and 4 for the NRS. The first column shows the portfolio number, the second column shows the corresponding total scale score, the third column shows the observed percent of weighting from the judges for whom that portfolio did not yet demonstrate ability at NRS level 4, and the fourth column shows the observed percent of weighting from the judges for whom that portfolio represented ability at least at level 4. These numbers represent an average across all judges.

In this example, we see that until a score of 60 was reached, the judges were unanimous that the performance did not yet meet the criteria for NRS level 4. We also see that once a score of 67 was reached, judges were unanimous that the performance was at level 4 or above. Somewhere between those two scores is a point at which at least 50% of the weighting would be at least at level 4. From the observed data, that point most likely lies between portfolio 17 (total scale score of 63 with 28% agreement) and portfolio 18 (total scale score of 65 with 71% agreement).

Table 64
Example Data for Determination of NRS Level 4

Portfolio Number	Total Scale Score	Observed	
		% Agreeing Not Yet 4	% Agreeing 4 or Above
1	1	100	0
2	6	100	0
3	11	100	0
4	16	100	0
5	19	100	0
6	23	100	0
7	27	100	0
8	30	100	0
9	35	100	0
10	39	100	0
11	45	100	0
12	47	100	0
13	51	100	0
14	54	100	0
15	55	100	0
16	60	90	10
17	63	72	28
18	65	29	71
19	67	1	99
20	68	0	100
21	71	0	100
22	75	0	100
23	78	0	100

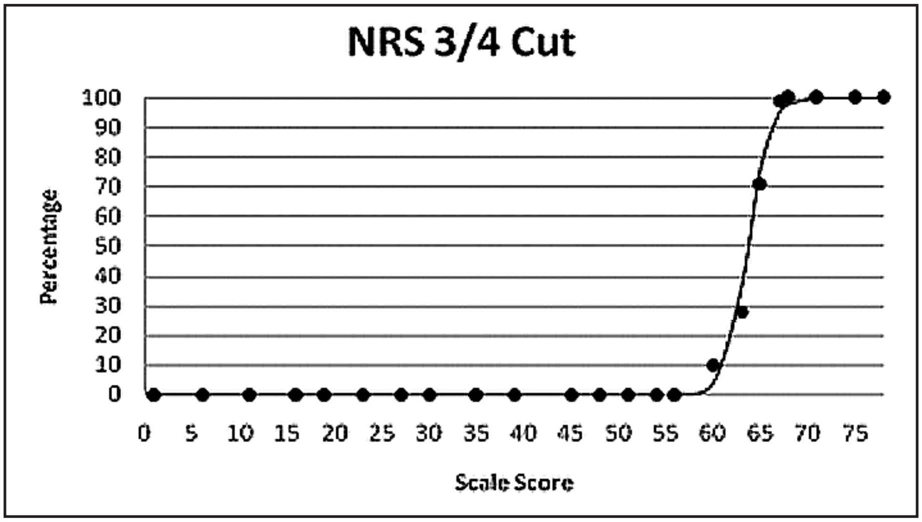

Figure 38. Example of NRS 3/4 Cut

Figure 38 is the graphic representation of Table 64, including the predicted logistic regression line. The vertical axis represents percentages. The horizontal axis represents the total scale score. The 23 dots in Figure 38 represent the observed percentage of agreement among the judges that each of the 23 portfolios (each at its own scale score) represents a performance at level 4 or higher on the NRS. The curve in Figure 38 represents the predicted percentages fitting the logistic regression line to the data. To find the point at which at least 50% of the weighting would be at least at level 4 using this figure, find 50 on the vertical axis, follow the horizontal line across to the point where it meets the curve, and go down to find the corresponding scale score on the horizontal axis. This scale score represents the cut between level 3 and level 4, because a group of judges (as represented by the judges in this study) would be more likely to rate a performance at that scale score at level 4 or above rather than at level 3 or below. In actuality, the exact point is found by solving a mathematical equation as follows: From the parameter estimates output (Table 65), we take the coefficients estimated for the model, presented in the second column (B, Lower Bound).

Table 65
Parameter Estimates for NRS 3/4 Cut

	B	**Std. Error**	**Wald**	**df**	**Sig.**	**Exp(B)**	**95% Confidence Interval for Exp(B)**	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
Intercept	56.139	4.648	145.888	1	.000			
Scale Score	-.882	.072	147.871	1	.000	.414	.359	.477

Now, from the logistic regression equation,

$$\text{Ln(Odds)} = 56.139 + (-0.882) \text{ x (scale score)}$$

we are looking for the scale score at which the odds for a portfolio to be put in level 3 (or lower) or in level 4 (or higher) are equal. The odds at that scale score are 1 (1/1=1). Therefore, the left side of the equation is 0 as the natural logarithm of 1is 0.

$$0 = 56.139 - 0.882 \text{ x (scale score)}$$

$$56.139 = 0.882 \text{ x (scale score)}$$

$$63.681 = \text{scale score}$$

Because the results are typically not whole numbers, results were rounded to the nearest whole number scale score as the closest approximation of the cut score set by the judges. This decision was made for two reasons. First, always rounding up to the next higher whole number, while ensuring that the student would be above the judges' cut, would have set the cut further away from the judges' decision point. Second, this method allowed for more whole score points to be used for making cuts (i.e., the point below the cut is included), which was important to ensure that the range between cut scores toward the end of the score scale distribution was not too narrow.

7.5 Results

7.5.1 NRS Study

7.5.1.1 Quantitative

Using the above procedure to analyze the ratings of the judges, cut scores were obtained for both the NRS levels and the SPLs. Table 66 presents the NRS cuts.

Table 66
NRS Cuts

NRS Levels	Beginning ESL Literacy/ Low Beginning ESL (1/2)	Low Beginning ESL/High Beginning ESL (2/3)	High Beginning ESL/Low Intermediate ESL (3/4)	Low Intermediate/ High Intermediate ESL (4/5)	High Intermediate/ Advanced ESL (5/6)	Advanced ESL/Exit NRS (6/Exit NRS)
Cut	21	53	64	68	76	(79)

As explained earlier, the maximum total scale score on *BEST Literacy* is 78. Therefore, the predicted NRS exit score of 79 cannot be observed. The gap between the cuts gets smaller as the levels go up. These results suggest that tasks on *BEST Literacy* may more accurately allow students who are in the lower NRS levels to demonstrate what they can do and thus perhaps be measured more appropriately than students in the higher NRS levels.

CAL staff reviewed these results with the judges on the second day before starting the SPL study. The judges found it interesting that the cut scores from their work in the study (presented in Table 67) were generally higher than the cuts that were in use at that time (presented in Table 68). Figure 39, which was shown to the judges, displays these differences visually. The top horizontal line in Figure 39 represents the cuts then being used, and the bottom horizontal line represents the new cuts derived from the study. The five arrows show where the new cuts are in relation to the cuts being used at the time of the study.

These results were not surprising to the judges, who expressed greater confidence in the new cuts than in the old ones. They felt the old cuts were too low vis-à-vis the NRS descriptors. They noted that it takes a long time to go from NRS level 2 (Low Beginning ESL) to the next level, as represented in Figure 39. They also noted that only the note-writing tasks on *BEST Literacy* allow students to display the ability to "write some simple sentences," a requirement of NRS level 3 (High Beginning ESL), and that it is not until examinees earn 55 points or more that they are successful at the note-writing tasks. Overall, the judges were highly confident in the results of the study, as will be seen later in section 7.5.1.2.

Table 67
NRS Cuts from the Study (2007)

BEST Literacy Scale Score	NRS Level
0 - 20	Beginning ESL Literacy (1)
21 - 52	Low Beginning ESL (2)
53 - 63	High Beginning ESL (3)
64 - 67	Low Intermediate ESL (4)
68 - 75	High Intermediate ESL (5)
76 and above	Advanced ESL (6)

Table 68
Earlier NRS Cuts (Based on SPL Linking)

BEST Literacy Scale Score	NRS Level
0 - 7	Beginning ESL Literacy (1)
8 - 35	Low Beginning ESL (2)
36 - 46	High Beginning ESL (3)
47 - 53	Low Intermediate ESL (4)
54 - 65	High Intermediate ESL (5)
66 and above	Advanced ESL (6)

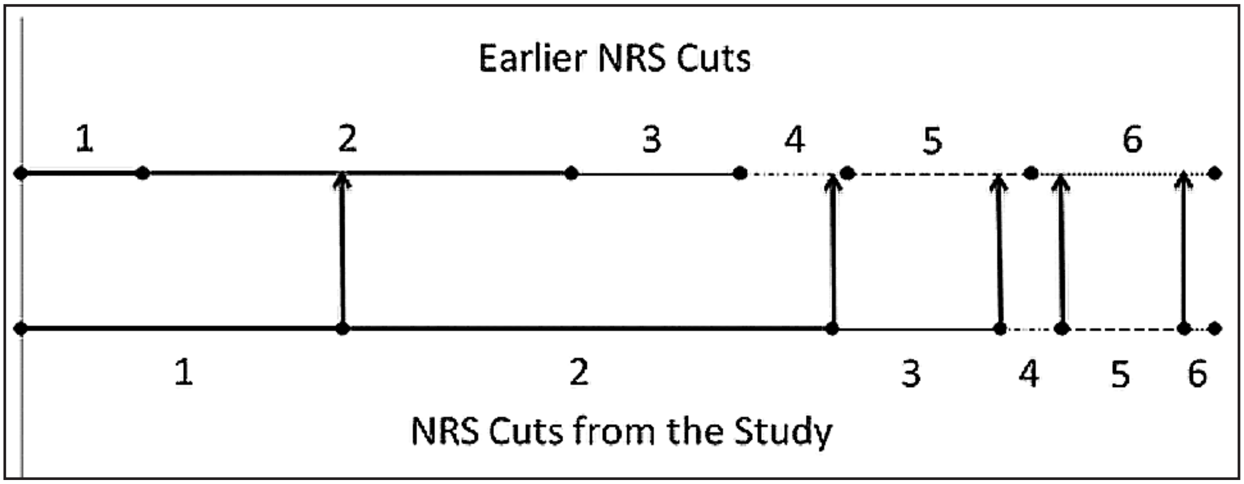

Figure 39. Comparison of Earlier Cuts (top line) and New NRS Cuts (bottom line)

About a month after the standard setting study, on May 9, 2007, CAL's Adult ESL Assessments Advisory Committee held a telephone conference during which committee members reviewed and confirmed the cuts derived from the standard setting study. The committee was composed of 10 adult ESL professionals, 5 of whom also served as judges in the *BEST Literacy* standard setting study.

As CAL staff were preparing the NRS cuts to present to the committee, they became concerned that, given the conversion charts between raw scores on the new forms and raw scores on Old Form B (i.e., the scale scores), it might be possible for examinees to be rated inappropriately at NRS level 6, for which the cut score was 76. That is, it was clear from the judges' discussions that only examinees who scored 5 on both note-writing tasks should be placed in level 6. In theory, however, if an examinee had all other reading and writing items correct, it would be possible to score a 3 on one of the note-writing tasks and still get a scale score of 76 and therefore be placed into NRS level 6. It was also theoretically possible that an examinee who missed a few items in reading but received a raw score of 26 or 27 in writing could get a scale score of 29 in writing, which would enable them to be classified at NRS level 6.

To investigate the extent of this problem with actual data, we went back to the data from the field test. We examined the performances of all students who scored a 3 and a 5 on the note-writing tasks and also achieved a perfect score on the rest of the writing section. That is, we selected those students who might have gotten a total scale score placing them in NRS level 6 but who did not score a 5 on both note-writing tasks. When we examined how these students scored in reading, we found that none of them achieved a scale score of 47 or higher. This finding means that no student without a score of 5 on both note-writing tasks had a total scale score of 76 or above, no matter which test form they took. From this examination of the field-test participants, we concluded that it is very unlikely for students who do not score 5 on both note-writing tasks to score high enough in reading to be placed into NRS level 6. Therefore, we were confident in presenting the study results as they were to the Adult ESL Assessments Advisory Committee.

7.5.1.2 Qualitative

The judges were asked to rate specific aspects of the NRS study, namely clarity of the NRS level descriptors, training, materials, process, and time allotted for setting the cut scores. They used a rating scale of *Excellent* (4), *Good* (3), *Satisfactory* (2), and *Poor* (1). Table 69 summarizes the judges' opinions. They were unanimous in rating the materials and process used to set the cut scores as excellent. Training and time allotted also received high ratings (3.6 and 3.4 respectively), although it appears that the judges might have wished for a bit more time. They gave very low ratings for the clarity of the NRS level descriptors—a mean rating of 1.8, which is below the *satisfactory* level. This dissatisfaction was expressed in discussions during the study as well. The judges felt that the NRS level descriptors were not clear in separating abilities in reading and writing and did not provide enough specification. Following the study with the SPL descriptors, the judges reported that the NRS descriptors were more abstract than the SPL descriptors and not as user friendly.

Table 69
Judges' Evaluation of NRS Portion of the Standard Setting Study

NRS Levels	**Clarity of the NRS level *descriptors***	***Training* on setting the cut scores for NRS levels**	***Materials* used in setting the cut scores for NRS levels**	***Process* used in setting the cut scores for NRS levels**	***Time* allotted for setting the cut scores for NRS levels**
Mean	1.80	3.60	4.00	4.00	3.40
St. Dev	.789	.516	.000	.000	.843

Table 70 summarizes the confidence level of the judges on each NRS cut. It shows that they felt more confident about the cut scores at the lower levels than at the higher levels. This was most likely due to the fact that *BEST Literacy* provides examinees with ample opportunity to demonstrate skills described by the lower NRS level descriptors but relatively few opportunities to demonstrate skills at the upper level. These ratings also show that the judges did not have confidence in the ability of performances on *BEST Literacy* to provide evidence of examinees reaching a level beyond NRS 6 (i.e., to be exited).

Table 70
Judges' Confidence in NRS Cuts

NRS Levels	Beginning ESL Literacy/ Low Beginning ESL (1/2)	Low Beginning ESL/High Beginning ESL (2/3)	High Beginning ESL/Low Intermediate ESL (3/4)	Low Intermediate/ High Intermediate ESL (4/5)	High Intermediate/ Advanced ESL (5/6)	Advanced ESL/Exit NRS (6/Exit NRS)
Mean	3.50	3.30	3.30	3.22	2.80	2.10
St. Dev	.707	.675	.675	.833	1.135	.994

7.5.2 SPL Study

The SPL study on the second day replicated the NRS study, except that, for the most part, only every second student portfolio was evaluated. Also, the cut scores were not able to be calculated and presented to the judges before the study ended.

7.5.2.1 Quantitative

Table 71 presents the cut scores obtained for the 8 SPL levels. That is, with a perfect score on *BEST Literacy*, an examinee might be appropriately placed into SPL 8, but the test could not exit them from that level.

Table 71
SPL Cuts

SPLs	0/1	1/2	2/3	3/4	4/5	5/6	6/7	7/8
Cut	1	13	31	54	67	72	75	78

As with the NRS results, the gap between the cuts gets quite small toward the higher levels. Most of the *BEST Literacy* data used to make decisions among the higher level SPLs come from the hardest reading items and from the two note-writing tasks.

7.5.2.2 Qualitative

The judges were asked to rate specific aspects of the SPL study in the same manner as for the NRS study. Table 72 summarizes their opinions. Again, the judges gave high ratings to the training, materials, process, and time allotted, although only about half as much time was spent on the SPL study as on the NRS study. The judges' familiarity with the student portfolios may be the reason they felt they needed less time on the second day.

An interesting finding was that the judges found the SPL descriptors to be much clearer than the NRS level descriptors. The SPL descriptors were rated 3.50 for clarity, whereas the NRS descriptors were rated only 1.80 (see Table 69). This happened despite the fact that the judges reported about the same degree of familiarity with both sets of descriptors (see Table 60).

Table 72
Judges' Evaluation of SPL Portion of the Standard Setting Study

SPLs	Clarity of the SPL *descriptors*	*Training* on setting the cut scores for SPLs	*Materials* used in setting the cut scores for SPLs	*Process* used in setting the cut scores for SPLs	*Time* allotted for setting the cut scores for SPLs
Mean	3.50	3.90	4.00	4.00	3.70
St. Dev	.527	.316	.000	.000	.675

Table 73 summarizes the confidence level of the judges on each SPL cut score. It shows that their confidence level is quite high on all but the final cut, which turned out to be a perfect score. It does appear that the judges feel confident that *BEST Literacy* can be used to assess up to SPL 7. In general, confidence in the SPL cuts is higher than confidence in the NRS cuts, although the only noticeable difference between the two, as far as the judges' ratings go, is in the clarity of the descriptors. Thus, the lack of clarity in the NRS descriptors may be a source of the judges' lower confidence in the NRS cuts.

Table 73
Judges' Confidence in SPL Cuts

SPLs	0/1	1/2	2/3	3/4	4/5	5/6	6/7	7/8
Mean	3.80	3.70	3.70	3.30	3.70	3.40	3.60	2.70
St. Dev	.422	.483	.483	.949	.483	.966	.516	1.160

7.6 Cross-Validation of the Standard Setting Study

As part of the field test, teachers in the programs from which the students were drawn were sent a copy of the NRS level descriptors and asked to place each of their students in a level. Based on the teachers' classification of their students, CAL staff calculated the average total scale score for students in each category. Originally, we were quite concerned that these averages appeared high compared to the cut scores in use at the time. As it turned out, however, the results from the teachers helped to cross-validate the results from the standard setting study.

Table 74 presents the data from the teachers (field-test data) and data from the standard setting study. The first four columns show the results from the field test. The first column shows the NRS level according to placement by the teachers, the second shows the number of students at each level, the third shows the mean total scale scores, and the fourth shows the standard deviation of the mean scores. (Note that NRS levels 5 and 6 are not distinguished in these results.) The final two columns show the score range based on the cut scores derived from the standard setting study and the NRS level designations. Although teacher judgment can often be inaccurate, the results of these two studies indicate that the cut scores set in the standard setting study were much more in agreement with the perception of the teachers than were the cut scores in use at that time. At all levels except Beginning ESL Literacy, the means of the total scale scores in the third column fall in the range derived from the standard setting study. For example, the mean of 54.04 for High Beginning ESL falls between 53 and 63, which is the range derived from the standard setting study. These results provide further evidence of the validity of the cut scores that were derived from the standard setting study.

Table 74
Comparison of Field-Test Results and Standard Setting Study Results

Field-Test Results				Standard Setting Study Results	
NRS Level		Total Scale Score		From the Standard Setting Study 2007	
	N	*Mean*	*S.D.*	BEST Literacy *Scale Score*	*NRS Level*
1	62	36.52	16.58	0 - 20	Beginning ESL Literacy (1)
2	84	46.07	13.92	21 - 52	Low Beginning ESL (2)
3	114	54.04	10.74	53 - 63	High Beginning ESL (3)
4	79	62.13	10.16	64 - 67	Low Intermediate ESL (4)
5/6	67	69.66	5.76	68 - 75	High Intermediate ESL (5)
				76 and above	Advanced ESL (6)

Table 75a presents data comparing the teachers' ratings of their students' NRS levels with the NRS levels into which the students were placed based on their performance on *BEST Literacy* using the cut scores derived from the standard setting study. Each of the rows labeled 1-6 shows the NRS level the student achieved based on his or her performance on *BEST Literacy*. At the end of the row labeled 1, for example, we see that scores on *BEST Literacy* placed 12 students in NRS level 1. Each of the columns labeled 1-6 shows how the teachers rated the students. The bottom of the column labeled 1, for example, shows that teachers placed 62 students in NRS level 1.

The important numbers are where the rows and columns intersect. The number 10 in the box where row 1 and column 1 intersect tells us that 10 students were placed in NRS level 1 by their *BEST Literacy* scores and by their teachers. In this case there was agreement between the test scores and the teachers' judgments. The number 2 in the adjacent box to the right shows that two students were placed in NRS level 1 by their *BEST Literacy* score and in NRS level 2 by their teachers. This result shows there was disagreement between the two placements.

Table 75a
Cross-Tabulation of NRS Levels Based on *BEST Literacy* and Field-Test Teacher Judgment

		NRS Level Based on Field-Test Teacher Judgment						Total
		1	2	3	4	5	6	
NRS Level Based on Performance on *BEST Literacy* (new cut scores)	1	10	2	0	0	0	0	12
	2	43	52	50	15	0	2	162
	3	6	21	38	23	3	2	93
	4	2	5	12	11	7	3	40
	5	1	4	14	30	26	19	94
	6	0	0	0	0	1	4	5
Total		62	84	114	79	37	30	406

From Table 75a, we can calculate several indices of agreement between the teachers' ratings and the *BEST Literacy* outcomes. These indices are presented in Table 75b. The "Exact Agreement" column in Table 75b shows the level of exact agreement between the teachers' ratings and the *BEST Literacy* outcomes (which is also reflected in the shaded cells of Table 75a). The "Exact and Adjacent" column shows the indices where teachers' ratings were either exact or one level adjacent to the *BEST Literacy* outcomes. Some disagreement between the teachers' ratings and the *BEST Literacy* outcomes was to be expected; the teachers' ratings were necessarily subjective (e.g., there was no training on how to accomplish the rating task), and the teachers were new to the revised NRS level descriptors.

The first part of Table 75b shows overall level of agreement. Of the 406 students, 34.7% were placed at the same NRS level by both *BEST Literacy* and the teacher rating; 86% were placed at either the same or adjacent NRS levels.

The second part of the table shows the level of agreement by NRS level based on the *BEST Literacy* results. For example, of the 12 children placed at NRS level 1 by *BEST Literacy*, 10 were also placed at that level by the teachers, for an exact agreement rate of 83.3%. However, when the teachers' adjacent ratings were also considered to be accurate, the agreement rate was 100%; that is, all of the students placed at NRS level 1 by *BEST Literacy* were placed at NRS level 1 or 2 by the teachers.

The third part of the table shows the agreement indices at each cut point. That is, for any cut point, it shows the percentage of the 406 students who were placed both by their performance on *BEST Literacy* and by the teacher ratings at either *below* or *above* that cut point.

Table 75b
Comparison of Classifications of NRS Level: Teachers' Ratings and *BEST Literacy* (Using New Cut Scores)

		Exact Agreement	Exact and Adjacent
Overall Indices		0.347	0.860
Conditional on Level	**Level**		
	1	0.833	1.000
	2	0.321	0.895
	3	0.409	0.882
	4	0.275	0.750
	5	0.277	0.798
	6	0.800	1.000
Indices at Cut Points	**Cut**		
	1/2	0.867	
	2/3	0.739	
	3/4	0.796	
	4/5	0.837	
	5/6	0.933	

The results in Table 75b are satisfactory considering the number of teachers involved and the fact that they had no training in their rating task. The results show a large degree of agreement between the rating of the students by the teachers and by their performance on *BEST Literacy*. They also provide support for the reasonableness of the cut scores set in the standard setting study.

Further evidence for the reasonableness of the outcomes of the standard setting study was found in an analysis of the same data, but using the earlier cut scores to place students into NRS levels based on their *BEST Literacy* performance. Tables 75c and 75d present the same results as Tables 75a and 75b, except that they use the former NRS cut scores rather than those derived from this study. Results show that the overall exact agreement (.222), reflected in the shaded diagonal, and overall exact and adjacent agreement (.559) are much lower with the former NRS cut scores than with the new cut scores (.347 and .860, respectively). In particular, agreement rates for the three highest NRS levels (4, 5, and 6), which are .600, .343, and .455 for adjacent and exact with the former cuts, are much higher with the new cuts at .750, .798, and 1.00 respectively. These results again provide strong evidence for the reasonableness of the outcome of the *BEST Literacy* standard setting study.

Table 75c
Cross-Tabulation of Field-Test Teacher Judgment and Earlier NRS Cuts

		NRS Level Based on Field-Test Teacher Judgment						**Total**
		1	2	3	4	5	6	
NRS Level Based on Performance on *BEST Literacy* (old cut scores)	1	5	0	0	0	0	0	5
	2	21	23	5	2	0	0	51
	3	19	18	24	5	0	1	67
	4	10	13	26	10	0	1	60
	5	4	23	40	25	5	5	102
	6	3	7	19	37	32	23	121
Total		62	84	114	79	37	30	406

Table 75d
Comparison of Classifications of NRS Level: Teachers' Ratings and *BEST Literacy* Using Old Cut Scores

		Exact Agreement	**Exact and Adjacent**
Overall Indices		0.222	0.559
Conditional on Level	**Level**		
	1	1.000	1.000
	2	0.451	0.961
	3	0.358	0.701
	4	0.167	0.600
	5	0.049	0.343
	6	0.190	0.455
Indices at Cut Points	**Cut**		
	1/2	0.860	
	2/3	0.744	
	3/4	0.623	
	4/5	0.606	
	5/6	0.741	

8. Reliability of *BEST Literacy*

8.1 Reliability

8.1.1 Estimates of Internal Consistency

Internal consistency reliability estimates (coefficient alpha) for the total score based on the 49 reading items and the 19 writing items on the three forms of *BEST Literacy* were all high: .917 for new Form B (n = 69), .937 for new Form C (n = 70), and .943 for new Form D (n = 130). The estimates of internal consistency on these new forms were very similar to those obtained on the old forms. In the field-test study on the old Form B (n = 67), coefficient alpha was .921; .917 was the result on the new Form B (n = 69). Similarly, on the old Form C (n = 71), the result was .949; on the new Form C (n = 70), it was .937. The similarity of the results between the old and new forms indicates that the internal consistency reliability was not affected by the updating of the test forms.

8.1.2 Estimates of Interrater Reliability

In the field test, the interrater reliability of the total scores for the reading and writing sections was examined on both the old and new forms. Interrater reliability was also analyzed separately on the note-writing tasks that appear at the end of the test. As explained earlier, examinee test booklets from the sites across the United States were scored in a 2-day session at CAL. Approximately 30% of the total forms administered in the study were double-scored (i.e., scored by two scorers).

Table 76 presents the Pearson correlation between scores awarded by the pair of scorers. While the pair is different for each test form, within a form, the same two scorers rated the same booklets. In addition to the interrater reliability observed for the total scores for reading and writing, Table 76 also presents the interrater reliability obtained for each of the note-writing tasks that appear on each form. Table 77 shows interrater reliability on the two note-writing tasks in terms of the percentage of agreement between the raters in each pair: either the exact match or the exact and adjacent match of scores that two raters gave. From the data, we can see that almost all of the double-scored note-writing tasks received the same or adjacent scores from both raters. Finally, Table 78 compares the raters in terms of mean of all the scores given by each rater.

Table 76
Interrater Reliability for *BEST Literacy*

	Old Form B (1984 test)	New Form B (2006 test)	Old Form C (1984 test)	New Form C (2006 test)	Form D (2006 test)
Section	(n = 27)	(n = 28)	(n = 29)	(n = 28)	(n = 38)
Reading (Total)	0.996	0.983	0.995	0.998	0.997
Writing (Total)	0.974	0.972	0.979	0.969	0.989
Task 1	0.884	0.967	0.963	0.952	0.975
Task 2	0.999	0.890	0.968	0.953	0.934

Table 77
Rater Agreement on Note-Writing Tasks of *BEST Literacy*

	Old Form B (1984 test)		**New Form B (2006 test)**		**Old Form C (1984 test)**		**New Form C (2006 test)**		**Form D (2006 test)**	
	(n = 27)		(n = 28)		(n = 29)		(n = 28)		(n = 38)	
	Exact	Exact and adjacent	Exact	Exact and adjacent	Exact	Exact and adjacent	Exact	Exact and adjacent	Exact	Exact and adjacent
Writing Task 1	85.2%	96.3%	92.9%	100.0%	89.7%	100.0%	89.3%	100.0%	92.0%	100.0%
Writing Task 2	100.0%	100.0%	78.6%	96.4%	93.1%	96.6%	85.7%	100.0%	84.0%	100.0%

Table 78
Comparison of Mean Scores Awarded by Raters for *BEST Literacy*

	Old Form B (1984 test)		**New Form B (2006 test)**		**Old Form C (1984 test)**		**New Form C (2006 test)**		**Form D (2006 test)**	
Section	(n = 27)		(n = 28)		(n = 29)		(n = 28)		(n = 38)	
	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
Reading Total	31.78	31.78	36.25	36.11	35.54	35.36	32.25	32.39	33.18	32.75
Writing Total	19.22	18.96	22.96	23.36	20.66	20.72	20.64	20.00	19.92	19.72
Task 1	1.85	1.81	3.32	3.46	2.24	2.31	2.29	2.07	1.48	1.44
Task 2	1.11	1.11	3.18	2.93	2.28	2.21	1.89	1.86	1.68	1.52

In the field test, the scorers were trained to score the note-writing tasks using the revised rubric and the scoring guide. The results presented here suggest that scorers who carefully follow the instructions given for scoring in the *BEST Literacy Test Manual* are able to score with a high degree of consistency across raters (or scorers). The results show that the combination of open-ended and selected-response items that make up the reading section may be scored with near-perfect interrater reliability. Likewise, the scores based on the short and extended written responses show that they are also able to be scored with a high degree of consistency: For the updated forms, the results were 0.972, 0.969, and 0.989 for Forms B, C, and D, respectively. The scoring of the note-writing tasks (included in the total) is also impressive, ranging in this study from 0.884 to 0.999.

The consistency of results presented in Tables 76, 77, and 78 for the reading and writing totals across the old and new forms also suggests that the updating of the materials in *BEST Literacy* had no negative effect on the reliability or difficulty of the test.

8.2 Measurement Precision

With the Rasch measurement model, as with any measurement model following Item Response Theory (IRT), the relationship between the ability measure (in logits) and the accuracy of test scores can be modeled. It is recognized that tests measure most accurately when the ability of the examinee and the difficulty of the items are appropriate for each other. If a test is too difficult for an examinee (i.e., the examinee gets almost no items correct), or if the test is too easy for an examinee (i.e., the examinee gets almost all items correct), the measurement does not give us much information about the examinee.

One way to look at measurement precision visually is through graphs of the test information function (TIF). These are presented in the figures that follow. The test information function shows graphically how well the test is measuring across the ability measure spectrum. High values indicate more accuracy in measurement. Thus, for each test form, the figures show the relationship between the ability measure (in logits) on the horizontal axis and measurement accuracy, represented as the Fisher information value (which is the inverse squared of the standard error), on the vertical axis. The test information function, then, reflects the conditional standard error of measurement.

Figures 40 to 44 show TIF graphs for each reading form, and Figures 45 to 49 show TIF graphs for each writing form. Figures 50 to 52 show TIF graphs for writing forms New B, New C, and New D after the note-writing tasks were redistributed.

8.2.1 Reading

The horizontal axis is difficult to interpret in the TIF graphs for the reading forms because *BEST Literacy* is not scored using logit measures. To aid interpretation, we have inserted six vertical lines in Figures 40 through 44. The lines represent, respectively, the ability measure points of the raw scores of 0 (lowest possible score), 10, 20, 30, 40, and 49 (maximum possible score).

As expected, measurement is less accurate at the low and high extreme scores. When comparing the old forms with the new forms on these graphs, a slight improvement in measurement accuracy can be seen in the new forms versus the old forms; the peak of the new form curves is slightly higher than the peak of the old form curves.

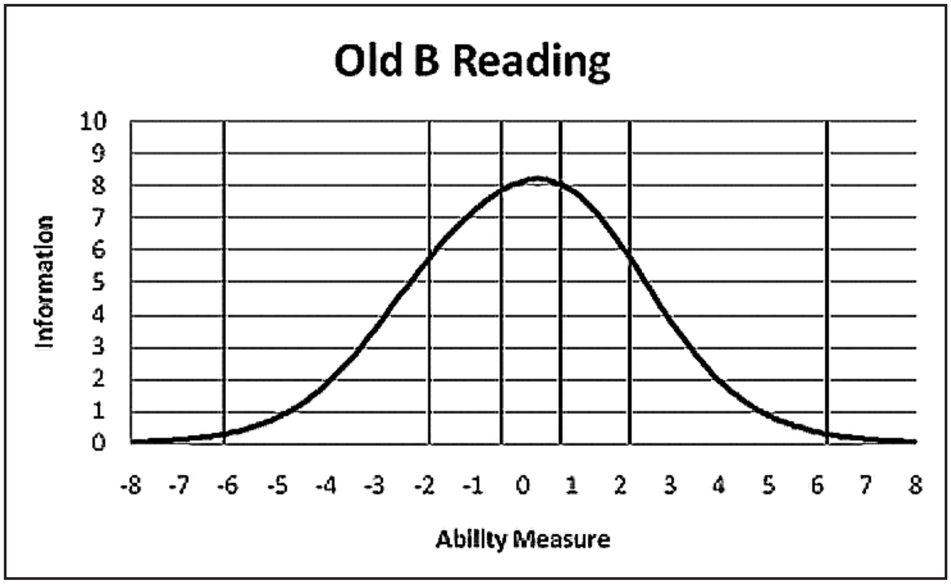

Figure 40. Old B Reading TIF

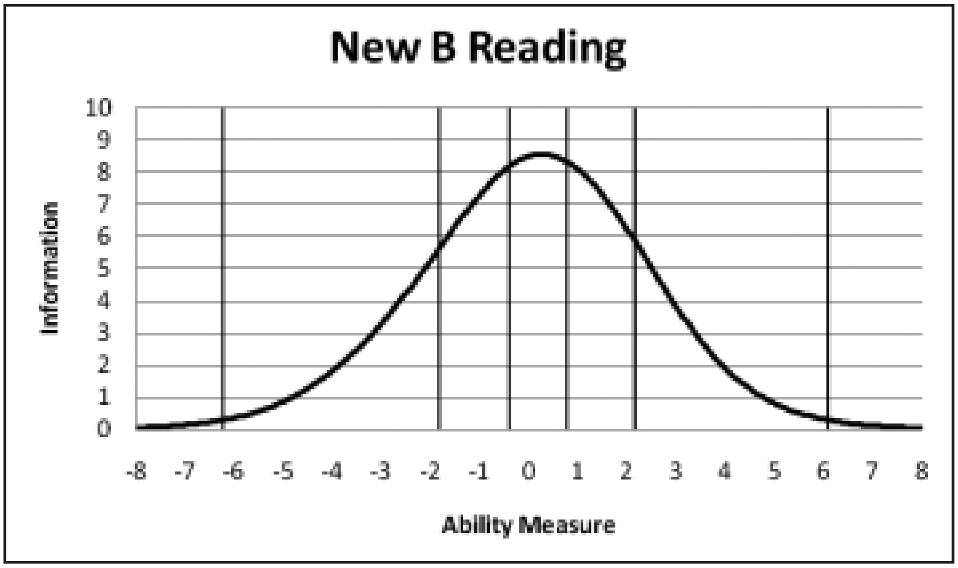

Figure 41. New B Reading TIF

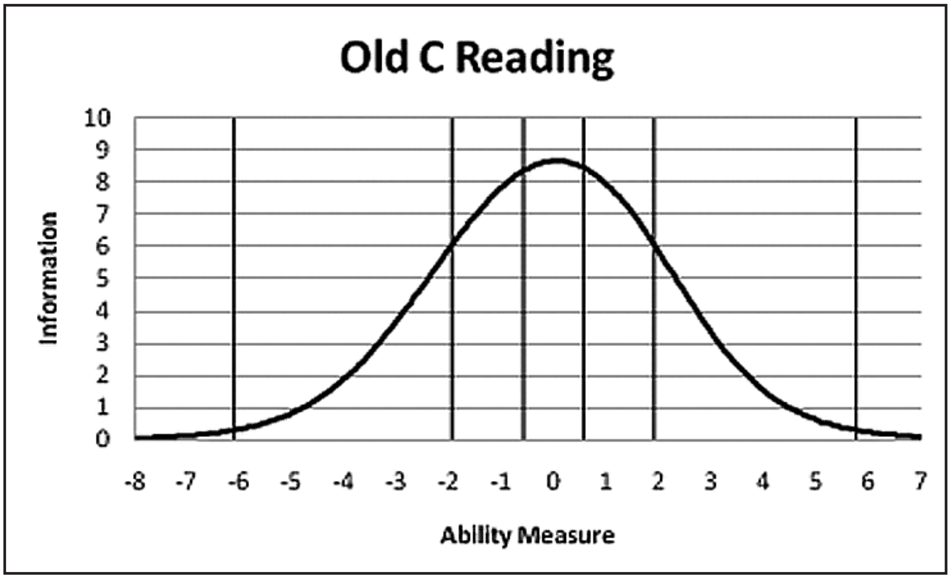

Figure 42. Old C Reading TIF

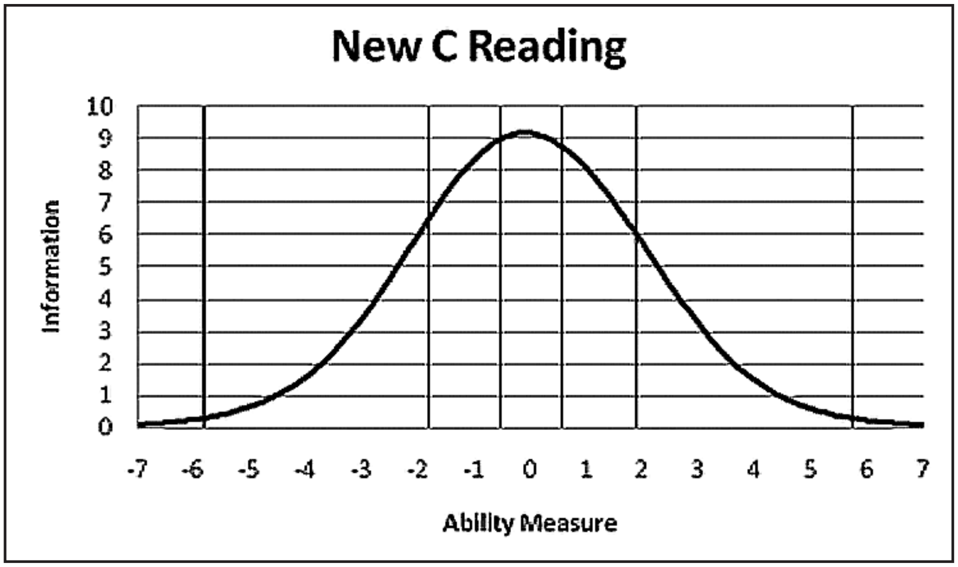

Figure 43. New C Reading TIF

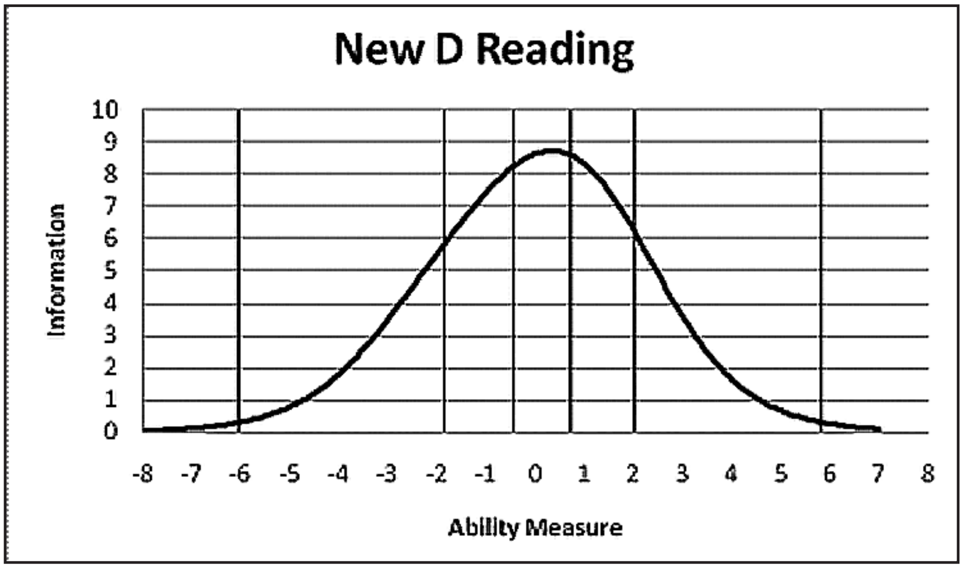

Figure 44. New D Reading TIF

8.2.2 Writing

To aid in the interpretation of the writing scores, we have inserted four vertical lines in Figures 45 to 52. These represent the ability measure points of raw scores of 0 (lowest possible score), 10, 19, and 29 (highest possible score). The raw score of 10 in writing is significant because many students are successful only in the first part of writing—filling out a personal information form—and therefore score 10. Another significant raw score point in writing is 19—the score received by students who are successful in all the writing parts other than the note-writing tasks. In other words, scores above 19 can be awarded only to examinees who have earned some points on the note-writing tasks.

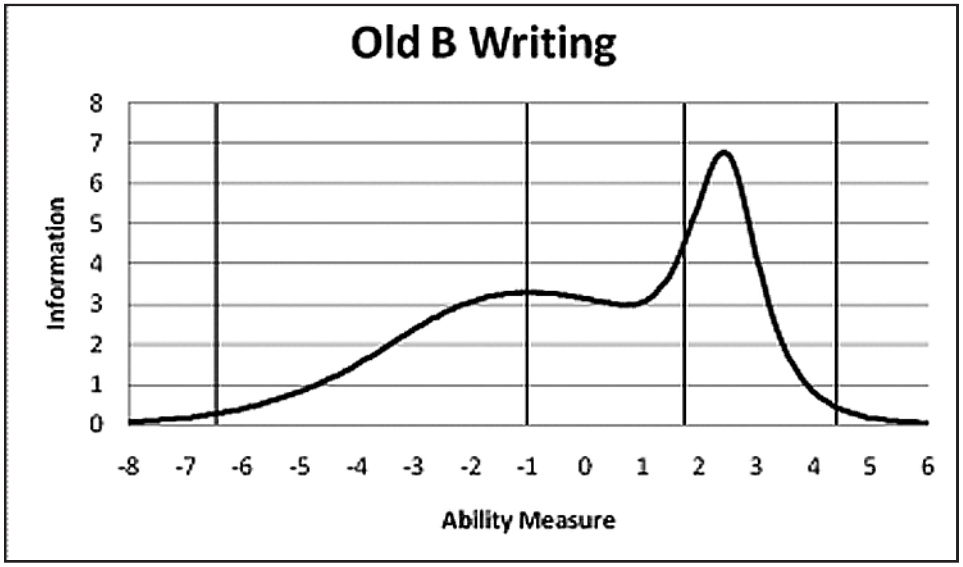

Figure 45. Old B Writing TIF

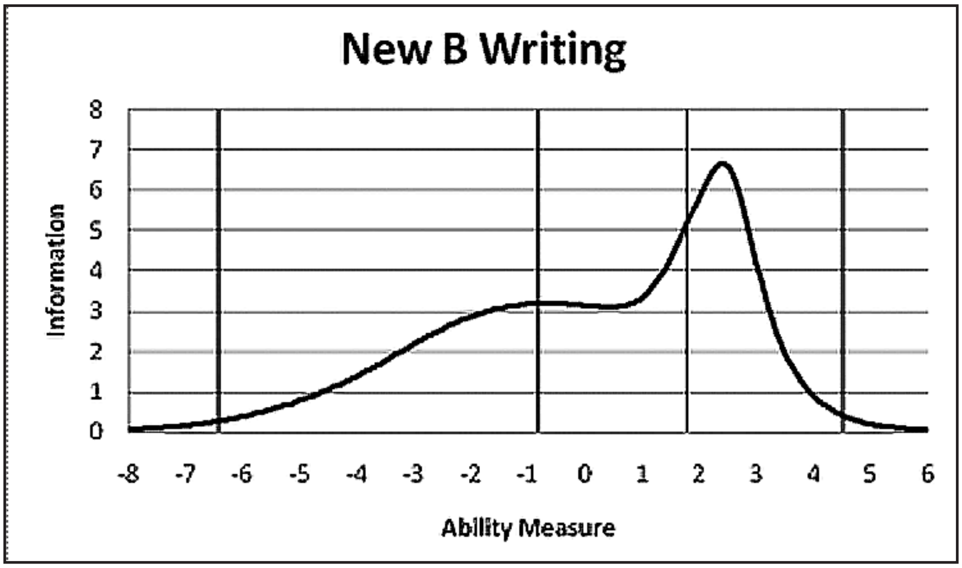

Figure 46. New B Writing TIF

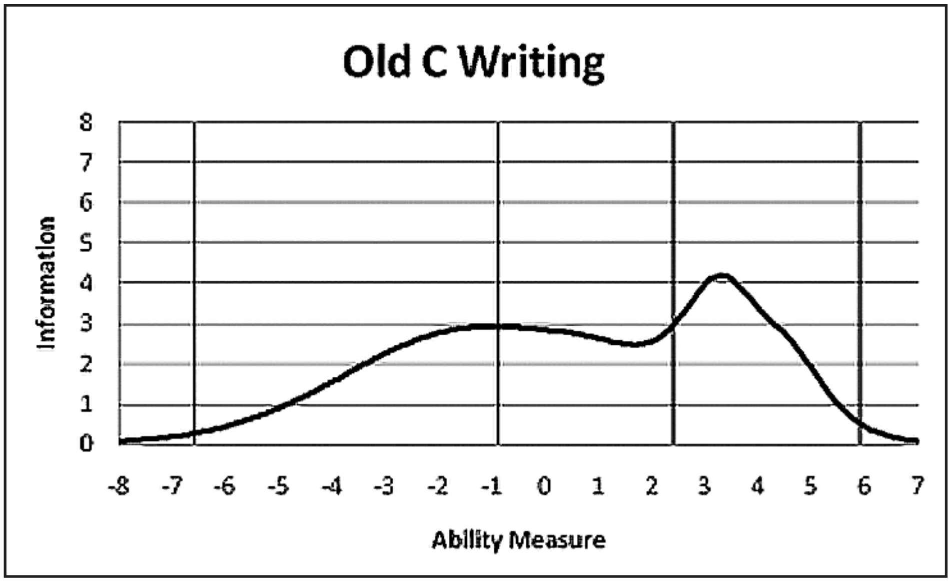

Figure 47. Old C Writing TIF

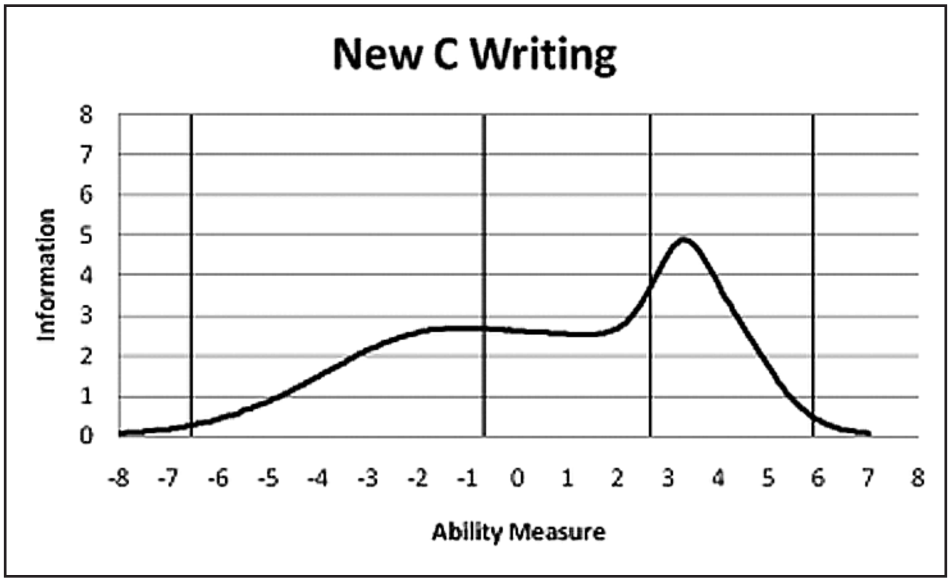

Figure 48. New C Writing TIF

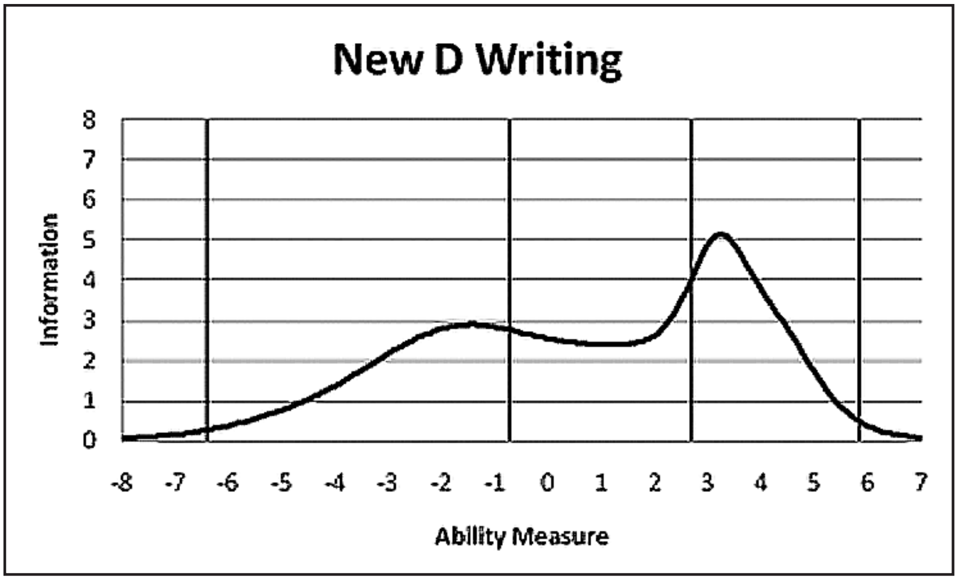

Figure 49. New D Writing TIF

Figures 50 through 52 show the TIF graphs after the redistribution of the note-writing tasks.

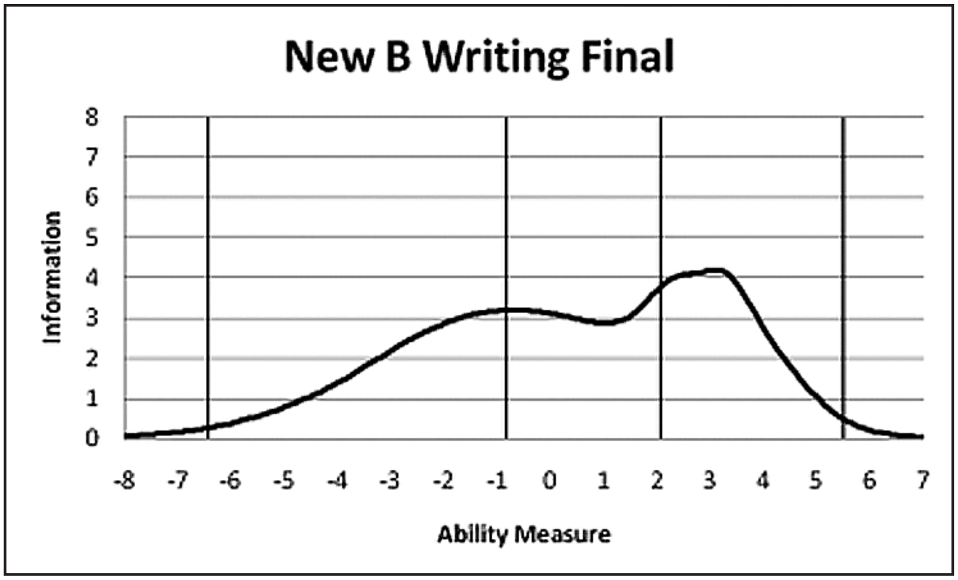

Figure 50. New B Writing Final TIF

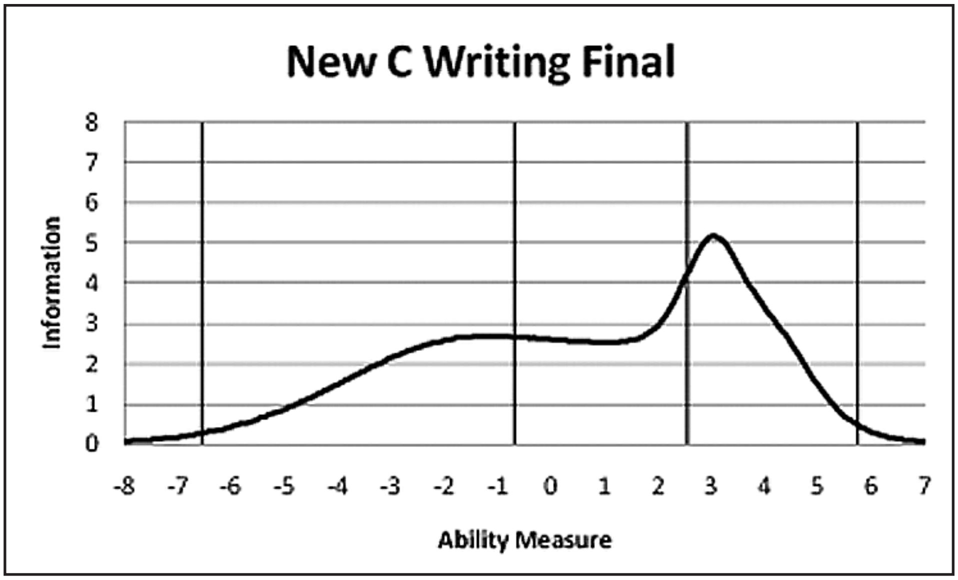

Figure 51. New C Writing Final TIF

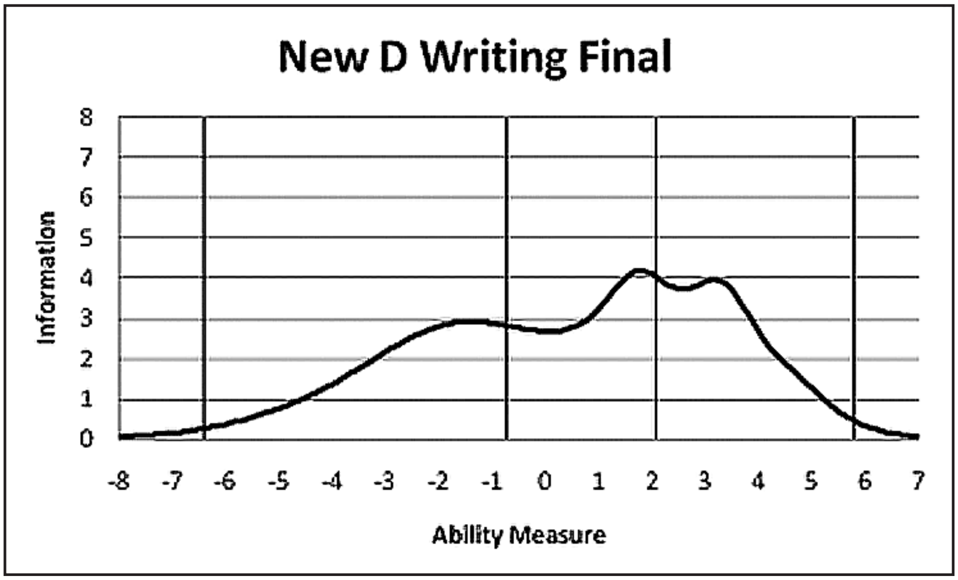

Figure 52. New D Writing Final TIF

8.3 Consistency and Accuracy of Classification

It is important to know the reliability of any student's test score (e.g., internal consistency and interrater reliability; see section 8.1) and the degree of precision with which it has been measured (i.e., the estimate of the conditional standard error of the Rasch measurement model; see section 8.2). However, because *BEST Literacy* is used for placement into SPLs and NRS levels, it is important to know how well these classifications are made. The analyses that we used make use of the methods outlined in Livingston and Lewis (1995) and Young and Yoon (1998) and implemented in the software program *BB-CLASS* (Brennan, 2004; see also Lee, Hanson, & Brennan, 2002).

In the approach of Livingston and Lewis (1995), the *accuracy* of a decision is the extent to which decisions made on the basis of the administered test (i.e., the observed scores) would agree with the decisions that would be made if each student could somehow be tested with all possible parallel forms of the assessment—that is, decisions based on the examinees' "true score." On the other hand, the *consistency* of a decision is the extent to which decisions made on the basis of the administered test would agree with the decisions that would be made if the students had taken a different but parallel form of the test. Thus, in every analysis of classification, two parallel analyses are made: accuracy (that is, vis-à-vis true scores) and consistency (that is, vis-à-vis a second form).

In terms of classifications around a single cut point, students can be misclassified in one of two ways. Students who were below the proficiency cut score (based on their true score) but were classified on the basis of the assessment as being above the cut score are considered to be *false positives*. Students who were above the proficiency cut score (based on their true score) but were classified as being below the cut score are considered to be *false negatives*. All other students are considered to be accurately placed either above or below the cut score.

True scores are, of course, unknown. The approach taken by Livingston and Lewis (1995) and implemented here uses information about the reliability of the test, the cut scores, and the observed distribution of scores and—using a two parameter beta distribution—models the distribution of the true scores and of scores on a parallel form. Overall accuracy and consistency indices are produced by comparing the percentage of students classified the same way across all categories by both the observed distribution and the modeled distribution. These indices indicate the percentage of all students who would be classified into the same proficiency level by both the administered test and either the true score distribution (accuracy) or a parallel test (consistency). Our tables also provide an estimate of Cohen's kappa statistic, which is a very conservative estimate of the overall classification because it corrects for chance.

We also looked at accuracy and consistency conditional on the proficiency level. These indices examine the percentage of students placed by the actual test (observed) into one level divided by the total number of students placed into that level either according to the true score distribution (accuracy) or based on a parallel test (consistency).

Finally, we look at what may be the most important set of indices, which are the indices at the cut points. While accuracy and consistency *conditional on level* provide information about the percentage of students who are classified into *one* level, indices *at the cut points* divide the data into four groups, with the considered cut as the middle point, and look at the percentage of students who are consistently placed either *above* or *below* the cut score by both the observed distribution and either the true score distribution or the parallel test distribution. That is, at every cut point, using the true score distribution (e.g., accuracy), we provide the percentage of students who are consistently placed above or below the cut score, as well as those who are false positives and false negatives. For consistency, only the percentage of students classified consistently above or below the cut score is calculated. For example, to evaluate the degree of confidence in a decision made regarding placement into an NRS level 6 based on scores on *BEST Literacy*, one can look at the accuracy index provided in the table for the cut score 5/6.

In Tables 79 and 80 there are three sections providing information related to the accuracy and consistency of placement into proficiency categories based on the NRS descriptors (Table 79) and the SPL descriptors (Table 80). The first section provides overall indices related to the accuracy and consistency of classification, as well as Cohen's kappa. The second section shows accuracy and consistency information conditional on level. The third section provides indices of classification accuracy and consistency at the cut points. These indices are perhaps the most important of all when using any of these as an absolute cut point, that is, asking the question, Which students have reached level 5 and which have not?

Table 79
Accuracy and Consistency of Classification Indices: NRS

<table>
<tr><th rowspan="2">Overall Indices</th><th>Accuracy</th><th colspan="2">Consistency</th><th colspan="2">Kappa (k)</th></tr>
<tr><td>0.751</td><td colspan="2">0.664</td><td colspan="2">0.548</td></tr>
<tr><th rowspan="7">Conditional on Level</th><th>Level</th><th colspan="2">Accuracy</th><th colspan="2">Consistency</th></tr>
<tr><td>1</td><td colspan="2">0.798</td><td colspan="2">0.647</td></tr>
<tr><td>2</td><td colspan="2">0.886</td><td colspan="2">0.841</td></tr>
<tr><td>3</td><td colspan="2">0.652</td><td colspan="2">0.547</td></tr>
<tr><td>4</td><td colspan="2">0.438</td><td colspan="2">0.337</td></tr>
<tr><td>5</td><td colspan="2">0.807</td><td colspan="2">0.716</td></tr>
<tr><td>6</td><td colspan="2">0.450</td><td colspan="2">0.274</td></tr>
<tr><th rowspan="7">Indices at Cut Points</th><th rowspan="2">Cut Point</th><th colspan="3">Accuracy</th><th rowspan="2">Consistency</th></tr>
<tr><td>Accuracy</td><td>False Positives</td><td>False Negatives</td></tr>
<tr><td>1/2</td><td>0.986</td><td>0.005</td><td>0.009</td><td>0.980</td></tr>
<tr><td>2/3</td><td>0.929</td><td>0.034</td><td>0.038</td><td>0.899</td></tr>
<tr><td>3/4</td><td>0.919</td><td>0.049</td><td>0.032</td><td>0.887</td></tr>
<tr><td>4/5</td><td>0.921</td><td>0.046</td><td>0.033</td><td>0.891</td></tr>
<tr><td>5/6</td><td>0.984</td><td>0.010</td><td>0.006</td><td>0.969</td></tr>
</table>

Table 80
Accuracy and Consistency of Classification Indices: SPL

Overall Indices	Accuracy	Consistency		Kappa (k)	
	0.728	0.634		0.519	
Conditional on Level	**Level**	**Accuracy**		**Consistency**	
	1	0.917		0.807	
	2	0.672		0.557	
	3	0.846		0.787	
	4	0.749		0.662	
	5	0.517		0.410	
	6	0.570		0.439	
	7	0.706		0.485	
Indices at Cut Points		**Accuracy**			
	Cut Point	Accuracy	False Positives	False Negatives	**Consistency**
	1/2	0.991	0.001	0.008	0.989
	2/3	0.973	0.013	0.013	0.960
	3/4	0.924	0.038	0.039	0.892
	4/5	0.927	0.042	0.031	0.896
	5/6	0.937	0.043	0.020	0.915
	6/7	0.969	0.020	0.011	0.952

In general, the more categories there are for placement, the harder it is to have high statistics for the overall indices and those conditional on level. Generally, Tables 79 and 80 show that the accuracy and consistency of *BEST Literacy* for placing students into the levels of the NRS and SPLs are quite good. Surprisingly, they are slightly higher for the SPLs (with seven levels) than for the six NRS levels. (Note that because classification into Level 8 of the SPLs is based on a perfect score, it was not included in this analysis.)

9. Validity of *BEST Literacy*

9.1 Establishing Validity

Validity research investigates the question of whether there is evidence that supports the appropriateness and adequacy of the interpretations and decisions made about examinees on the basis of their performance on a test. Examining the validity of an assessment is a task of collecting evidence in diverse ways and from a multitude of angles—evidence that supports the use of a test for its intended purpose. In this report, we present the validity evidence of *BEST Literacy* collected to date in diverse ways: evidence from test content, from internal structure, from teacher judgments, and from intersubscale correlations. Because establishing validity is an ongoing task, the results summarized here are only the preliminary findings to date.

9.2 Evidence from Test Content

BEST Literacy, like the literacy skills section of the *BEST*, claims to measure adult English language learners' ability to read and write through a variety of functional literacy tasks. Table 1 in section 2.2 presented the item type and topic of each item in *BEST Literacy*. Table 2 in section 3.2.1 presented the topic areas and language skills in the literacy skills section of the *BEST*. The topics included greetings, personal information, interpersonal communication, time, numbers, money, shopping for food and clothing, health, emergencies/ safety, housing, general information, and employment/training. A comparison of the two tables reveals that *BEST Literacy* covers the same topics and skills as the literacy skills section of the *BEST*.

9.3 Evidence from Data Fit

The Rasch model used to analyze the *BEST Literacy* field-test data is based on the understanding that for valid measurement to occur, item-level data from the measurement instrument must conform to some reasonable hierarchy on a single continuum of interest. Whether this is the case is investigated by examining the fit of the data to the measurement model. This investigation provides empirical evidence that all the items work together to measure a single variable (Bond & Fox, 2001). Good fit to the Rasch model, then, is another source of evidence that a test is measuring the single variable that it is claimed to measure. If the test appears to measure more than one variable or construct, its validity may be cast in doubt.

For *BEST Literacy*, there are two variables of interest, the reading and the writing ability of adult English language learners. Each is measured and scored separately, and each was analyzed separately for fit to the Rasch model. Approximately 5% of items on a measure may be misfitting by statistical probability (chance) on a measurement instrument fitting the Rasch model. The data in Tables 22 and 34 show that only 7 out of 182 reading items (3.8%) and only 2 out of 58 writing items (3.4%) were mifitting using both infit and outfit criteria (i.e., both infit and outfit z-standardized fit statistics were above 2.0 or below -2.0). These empirical results indicate that the reading and writing items on *BEST Literacy* show good fit to the Rasch measurement model, which shows that each part of the test is successfully measuring a single construct.

9.4 Evidence From Teacher Judgments

As mentioned in section 7.6, teachers whose students participated in the 2006 field test, which compared the older *BEST* literacy skills section, Forms B and C, to the updated *BEST Literacy*, Forms B, C, and D, were asked to provide their estimate of their students' proficiency levels according to the verbal proficiency descriptors of the six NRS educational functioning levels as revised in 2006 (U.S. Department of Education, 2007, pp. 19-20). Teacher ratings were available for 406 of the 407 students participating. Based on the *BEST Literacy* scale score for all students (no matter which form of the test they took), the correlations between the reading, writing, and total scale scores and the teachers' judgment of their student's placements were 0.635, 0.619, and 0.670, respectively. These results show that there was a substantial correlation between the teachers' ratings of student proficiency and student performances on the test. This finding provides support for the ability of *BEST Literacy* to place students accurately into hierarchical proficiency levels.

Table 81 presents the means for the reading and writing subscales as well as the total scale scores for students participating in the field test; the table groups the students by the NRS levels in which their teachers placed them. Although teachers may have been unfamiliar with the exact criteria that match observed student performances in classrooms to the NRS descriptors, it may be assumed that teachers did at least consistently rank their students in a relative ordering in this study.

Table 81
Means for Reading, Writing, and Total Scale Scores

NRS Level		Reading Scale Score		Writing Scale Score		Total Scale Score	
	N	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	62	22.73	10.94	13.79	6.86	36.52	16.58
2	84	27.33	9.09	18.74	6.25	46.07	13.92
3	114	32.49	7.01	21.54	5.16	54.04	10.74
4	79	37.68	6.71	24.44	4.94	62.13	10.16
5/6	67	42.27	4.69	27.39	2.51	69.66	5.76

Table 81 shows that mean performances on both sections of *BEST Literacy* and the total score clearly increase as teacher placement of the students increases on the NRS scale. Again, this study provides support for the use of *BEST Literacy* as a tool for student placement and for measuring student progress along a hierarchical continuum that is external to the test.

Finally, the agreement indices between placement in NRS levels based on performances on *BEST Literacy* and by teachers presented in section 7.6 also support the validity of the test as a measure of the reading and writing skills described by the performance level descriptors. Although the teachers' ratings were not collected under any type of controlled conditions, the amount of agreement is particularly impressive, particularly when teachers' exact and adjacent agreements are considered.

9.5 Evidence from Intersubscale Correlations

Examining correlations between the reading and writing sections of *BEST Literacy* can also provide evidence that the two sections are measuring different aspects of language. If correlations are too high, then it may be the case that the two subsections are not assessing distinct skills (here, reading and writing in English). On the other hand, as measures of English proficiency, the two subsections will be expected to be correlated.

The correlations between scores on the reading section and the writing section obtained from the field-test data are presented in Table 82. These results provide evidence that while correlated (as expected), the two sections are distinct enough to support the use of separate scores for each skill. (It is interesting to note how, with the exception of the new Form B, the correlations are very consistent across all the forms.)

Table 82
Correlation Between Reading and Writing Sections of Each Test Form

	Old Form B (1984 test)	New Form B (2006 test)	Old Form C (1984 test)	New Form C (2006 test)	New Form D (2006 test)
	(n = 67)	(n = 69)	(n = 71)	(n = 71)	(n = 129)
Correlation	0.771	0.662	0.804	0.787	0.799

References

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Erlbaum.

Brennan, R. L. (2004). BB-CLASS: *A computer program that uses the beta-binomial model for classification consistency and accuracy* [Computer software]. Iowa City, IA: CASMA.

Center for Applied Linguistics. (1982, 1984, 1987, 1989). *BEST test manual*. Washington, DC: Author.

Center for Applied Linguistics. (2008). *BEST Literacy test manual.* Washington, DC: Author.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on test.* Thousand Oaks, CA: Sage.

Clark, J. L. D., & Grognet, A. G. (1985). Development and validation of a performance-based test of ESL "survival skills." In P. C. Hauptmann, R. Leblanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 89–110). Ottawa, Ontario, Canada: University of Ottawa.

Hambleton, R. H., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Kingston, N. M., Kahl, S. R., & Sweeney, K. P. (2001). Setting performance standards using the Body of Work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Erlbaum.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26,* 412-432.

Linacre, J. M. (2005). Winsteps Rasch measurement [Computer software]. Chicago: Winsteps.

Linacre, J. M. (2006). Winsteps Rasch analysis version 3.60.1 [Computer program]. Available from http://www.winsteps.com

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179-197.

Standards for educational and psychological testing. (1999). Washington, DC: American Educational Research Association.

Sweeney, K. P., & Ferdous, A. (2007, April). *Variations of the "Body of Work" standard setting method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

U.S. Department of Education, Office of Vocational and Adult Education, Division of Adult Education and Literacy. (2007, June). *Implementation guidelines: Measures and methods for the National Reporting System for adult education.* Washington, DC: Author. Retrieved November 29, 2007, from http://www.nrsweb.org/docs/ImplementationGuidelines.pdf

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

Young, M. J., & Yoon, B. (1998, April). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment* (CSE Tech. Rep. No. 475). Los Angeles: University of California, Los Angeles; National Center for Research on Evaluation, Standards, and Student Testing.

APPENDIX A: Comparison of Old and New Forms B and C

The following tables provide information comparing the empirical difficulty of the old and the new Forms B and C to better understand the effect of the revisions to the test forms. The analyses presented in the report are based on an anchoring design that selected as items to be anchored only those that were totally unchanged between the old and new forms. Any minor revision meant that an item was no longer considered as an anchoring item for the concurrent calibration.

However, empirically, there were many more items that showed no statistical difference in their item difficulty on the old and new versions of the forms. To identify these items, we used two statistical tests. The first is a test commonly used in Rasch measurement to detect statistical differences between item difficulties calibrated, for example, by two different sets of one sample (e.g., boys versus girls). This methodology can be used to detect item differential functioning (DIF). It is fully explained in the online user's guide to the Rasch measurement computer program *Winsteps* (**www.winsteps.com/winman/table30dif.htm**). The author of the program, John Michael Linacre, proposes a t test that is the difference between the two item difficulties divided by their pooled standard error. (The pooled standard error is the square root of the sum of the two errors squared.) In this report, this is called the Linacre t test, and the probability value (with degrees of freedom equal to the sum of the number of cases used to estimate item difficulty for each item minus two) is called the Linacre *p*, or LinP.

The second statistic used to identify items with a statistically significant difference in difficulty was based on the effect size as calculated by Cohen's *d*. This formula is commonly represented like this:

$$\boxed{d = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\dfrac{(n_t-1)s_t^2 + (n_c-1)s_c^2}{n_t + n_c}}}}$$

where
x_t= the mean of the treatment group
n_t= the number of subjects in the treatment group
s_t= the standard deviation of the scores in the treatment group
x_c= the mean of the control group
n_c= the number of subjects in the control group
s_c= the standard deviation of the scores in the control group

For purposes of this report, the formula was adapted such that the two item difficulties replaced the means, the two numbers of cases used to estimate the item difficulties replaced the number of subjects, and the two standard errors replaced the two standard deviations.

As mentioned in the report, items that had undergone no changes were anchored to each other in a concurrent calibration across all five forms. This placed all item difficulties on the same scale. All unanchored items had been changed, albeit minimally, in the updating process. The statistical tests were used to test the hypothesis that the revision to the item did not affect item difficulty. If either or both tests are negative, then we conclude that the difficulty of the item had not been changed due to the revision process. In other words, for statistical purposes the items are the same. Such items could also have been anchored if an empirical approach to anchoring had been followed. If both tests are positive, however, we conclude that the revisions to the item did make a significant change in item difficulty and the items are different.

The criteria used for the two statistical tests were as follows. For the LinP, the test was positive if the value was below .50. For the effect size, the test was positive if the value was greater than 1.00. Thus, for an item to be marked as different, the LinP value needed to be below .50 and the effect size above 1.00.

Tables A1 through A4 compare the old and new Forms B and C separately for reading and writing. Each table gives the item name, the item difficulty measure (Measure) and its standard error (S.E.) for the item on the old form and the new form, and then the results of the two statistical tests. The next column gives the item status, whether it was anchored (and thus forced to have the same difficulty on each form), whether it was identified has having two statistically significant item difficulty values (different), or whether one or both of the statistical tests were negative and the two item difficulty values can be considered the same (same). The final column shows, for items marked as different, whether the revision process made the item easier or harder.

Table A1
Reading Form Old B vs. Form New B

	Item	Old B		New B		LinP	Effect Size	Item Status	Item Difficulty
		Measure	S.E.	Measure	S.E.				
1	P2I1	-2.68	0.26	-2.68	0.26	1.000	0.000	Anchored	
2	P2I2	-2.92	0.54	-2.92	0.54	1.000	0.000	Anchored	
3	P2I3	-1.41	0.31	-1.97	0.26	0.167	2.020	Different	Easier
4	P2I4	-2.05	0.55	-1.54	0.49	0.490	-0.988	Same	
5	P3I1	-2.86	0.75	-2.99	0.78	0.905	0.171	Same	
6	P3I2	-2.4	0.62	-3.8	1.06	0.256	1.615	Different	Easier
7	P3I3	-0.1	0.23	-0.14	0.17	0.889	0.208	Same	
8	P4I1	-1.78	0.5	-1.32	0.46	0.500	-0.966	Same	
9	P4I2	-2.05	0.55	-1.8	0.53	0.744	-0.467	Same	
10	P4I3	-0.25	0.23	-0.49	0.18	0.412	1.213	Different	Easier
11	P4I4	-1.97	0.37	-1.84	0.25	0.771	-0.440	Same	
12	P7I1	-1.41	0.31	-0.65	0.18	0.035	-3.282	Different	Harder
13	P7I2	2.18	0.21	1.75	0.15	0.096	2.498	Different	Easier
14	P8I1	-1.41	0.31	-1.01	0.2	0.279	-1.652	Different	Harder
15	P8I2	-1.35	0.43	-0.94	0.41	0.491	-0.984	Same	
16	P8I3	-0.72	0.37	0.1	0.32	0.096	-2.392	Different	Harder
17	P9I1	2.14	0.21	2.01	0.15	0.615	0.755	Same	
18	P9I2	0.64	0.21	-0.06	0.16	0.008	3.933	Different	Easier
19	P9I3	1.84	0.21	2.16	0.15	0.216	-1.859	Different	Harder
20	P9I4	1.92	0.21	1.79	0.15	0.615	0.755	Same	
21	P9I5	1.09	0.28	0.68	0.3	0.320	1.422	Different	Easier
22	P9I6	-1.01	0.39	-1.12	0.43	0.850	0.270	Same	
23	P9I7	0.85	0.29	0.85	0.29	1.000	0.000	Same	
24	P9I8	0.52	0.29	0.31	0.31	0.622	0.704	Same	
25	P9I9	1.89	0.29	1.34	0.28	0.175	1.945	Different	Easier
26	P9I10	1.97	0.29	2.22	0.29	0.543	-0.869	Same	
27	P9I11	0.85	0.29	0.94	0.29	0.827	-0.313	Same	
28	P9I12	0.93	0.28	0.94	0.29	0.980	-0.035	Same	
29	P9I13	1.09	0.28	1.5	0.28	0.302	-1.475	Different	Harder
30	P9I14	3.67	0.39	3.11	0.32	0.269	1.585	Different	Easier
31	P9I15	0.34	0.3	0.77	0.29	0.305	-1.469	Different	Harder
32	P10I1	-1.01	0.39	-0.78	0.4	0.681	-0.586	Same	
33	P10I2	-1.55	0.46	-0.63	0.38	0.125	-2.202	Different	Harder
34	P10I3	0.43	0.3	-0.23	0.35	0.155	2.036	Different	Easier
35	P10I4	-0.02	0.23	-0.02	0.23	1.000	0.000	Anchored	
36	P10I5	-0.82	0.27	-0.82	0.27	1.000	0.000	Anchored	
37	P10I6	1.45	0.2	1.45	0.2	1.000	0.000	Anchored	
38	P10I7	0.06	0.31	0	0.33	0.895	0.189	Same	
39	P10I8	1.65	0.28	1.98	0.28	0.406	-1.187	Different	Harder
40	P10I9	-0.59	0.35	-0.23	0.35	0.468	-1.036	Different	Harder
41	P10I10	0.34	0.12	0.34	0.12	1.000	0.000	Anchored	
42	P10I11	0.96	0.12	0.96	0.12	1.000	0.000	Anchored	
43	P10I12	0.03	0.23	0.03	0.23	1.000	0.000	Anchored	
44	P10I13	2.4	0.21	2.4	0.21	1.000	0.000	Anchored	
45	P10I14	-0.25	0.23	-0.37	0.17	0.675	0.625	Same	
46	P10I15	1.01	0.2	1.02	0.15	0.968	-0.059	Same	

Table A1 continued

	Item	Old B		New B		LinP	Effect Size	Item Status	Item Difficulty
		Measure	S.E.	Measure	S.E.				
47	P10I16	1.73	0.28	1.34	0.28	0.326	1.403	Different	Easier
48	P10I17	0.34	0.3	0.31	0.31	0.945	0.099	Same	
49	P10I18	1.17	0.28	1.34	0.28	0.668	-0.612	Same	

Table A2
Writing Form Old B vs. Form New B

	Item	Old B		New B		LinP	Effect Size	Item Status	Item Difficulty
		Measure	S.E.	Measure	S.E.				
1	P1I1	-4.71	0.54	-4.71	0.54	1.000	0.000	Anchored	
2	P1I2	-1.14	0.21	-1.14	0.21	1.000	0.000	Anchored	
3	P1I3	-2.82	0.33	-2.82	0.33	1.000	0.000	Anchored	
4	P1I4	-2.93	0.34	-2.93	0.34	1.000	0.000	Anchored	
5	P1I5	-2.76	0.54	-2.76	0.54	1.000	0.000	Anchored	
6	P1I6	-2.25	0.47	-2.25	0.47	1.000	0.000	Anchored	
7	P1I7	-2.25	0.47	-2.25	0.47	1.000	0.000	Anchored	
8	P1I8	0.08	0.26	0.08	0.26	1.000	0.000	Anchored	
9	P1I9	-1.43	0.23	-1.43	0.23	1.000	0.000	Anchored	
10	P1I10	-1.48	0.23	-1.48	0.23	1.000	0.000	Anchored	
11	P5I1	-2.09	0.6	-1.28	0.54	0.318	-1.430	Different	Harder
12	P5I2	-0.68	0.4	-0.76	0.48	0.898	0.183	Same	
13	P5I3	-1.49	0.5	-0.34	0.43	0.084	-2.483	Different	Harder
14	P5I4	0.01	0.35	-0.16	0.42	0.756	0.445	Same	
15	P5I5	0.87	0.31	0.71	0.35	0.733	0.489	Same	
16	P6I1	0.77	0.31	1.16	0.33	0.391	-1.230	Different	Harder
17	P6I2	0.25	0.34	0.71	0.35	0.348	-1.345	Different	Harder
18	P11I1	2.11	0.08	2.11	0.08	1.000	0.000	Anchored	
19	P11I2	2.58	0.08	2.58	0.08	1.000	0.000	Anchored	

Table A3
Reading Form Old C vs. Form New C

	Item	Old C		New C		LinP	Effect Size	Item Status	Item Difficulty
		Measure	S.E.	Measure	S.E.				
1	P2I1	-2.68	0.26	-2.68	0.26	1.000	0.000	Anchored	
2	P2I2	-1.95	0.34	-1.95	0.34	1.000	0.000	Anchored	
3	P2I3	-1.41	0.31	-1.97	0.26	0.167	2.020	Different	Easier
4	P2I4	-1.3	0.41	-1.61	0.42	0.598	0.752	Same	
5	P3I1	-2.94	0.74	-1.79	0.44	0.184	-1.903	Different	Harder
6	P3I2	-2.49	0.61	-1.28	0.39	0.097	-2.381	Different	Harder
7	P3I3	-0.1	0.23	-0.14	0.17	0.889	0.208	Same	
8	P4I1	-0.85	0.37	-0.73	0.35	0.814	-0.336	Same	
9	P4I2	-2.94	0.74	-1.99	0.46	0.277	-1.553	Different	Harder
10	P4I3	-0.25	0.23	-0.49	0.18	0.412	1.213	Different	Easier
11	P4I4	-1.97	0.37	-1.84	0.25	0.771	-0.440	Same	
12	P7I1	-1.41	0.31	-0.65	0.18	0.035	-3.282	Different	Harder
13	P7I2	2.18	0.21	1.75	0.15	0.096	2.498	Different	Easier
14	P8I1	-1.41	0.31	-1.01	0.2	0.279	-1.652	Different	Harder
15	P8I2	-1.67	0.46	-0.86	0.36	0.168	-1.975	Different	Harder

Table A3 continued

	Item	Old C		New C		LinP	Effect Size	Item Status	Item Difficulty
		Measure	S.E.	Measure	S.E.				
16	P8I3	0.23	0.3	0.23	0.31	1.000	0.000	Same	
17	P9I1	2.14	0.21	2.01	0.15	0.615	0.755	Same	
18	P9I2	0.64	0.21	-0.06	0.16	0.008	3.933	Different	Easier
19	P9I3	1.84	0.21	2.16	0.15	0.216	-1.859	Different	Harder
20	P9I4	1.92	0.21	1.79	0.15	0.615	0.755	Same	
21	P9I5	1.02	0.29	1.17	0.29	0.715	-0.521	Same	
22	P9I6	-0.48	0.34	-0.17	0.32	0.508	-0.946	Same	
23	P9I7	1.44	0.29	1.42	0.29	0.961	0.069	Same	
24	P9I8	1.86	0.29	2.08	0.29	0.593	-0.764	Same	
25	P9I9	-0.85	0.37	-0.49	0.34	0.475	-1.020	Different	Harder
26	P9I10	1.52	0.29	1.75	0.29	0.576	-0.799	Same	
27	P9I11	1.52	0.29	0.93	0.29	0.153	2.049	Different	Easier
28	P9I12	0.32	0.3	0.04	0.31	0.517	0.925	Same	
29	P9I13	1.27	0.29	1.58	0.29	0.451	-1.077	Different	Harder
30	P9I14	0.59	0.3	0.04	0.31	0.204	1.816	Different	Easier
31	P9I15	-0.72	0.36	-0.99	0.37	0.602	0.745	Same	
32	P10I1	-0.85	0.37	-0.99	0.37	0.789	0.381	Same	
33	P10I2	0.04	0.31	-0.17	0.32	0.638	0.671	Same	
34	P10I3	-1.47	0.43	-0.99	0.37	0.399	-1.205	Different	Harder
35	P10I4	-0.32	0.23	-0.32	0.23	1.000	0.000	Anchored	
36	P10I5	-1.54	0.3	-1.54	0.3	1.000	0.000	Anchored	
37	P10I6	1.3	0.2	1.3	0.2	1.000	0.000	Anchored	
38	P10I7	-0.37	0.33	0.13	0.31	0.271	-1.573	Different	Harder
39	P10I8	1.44	0.29	0.76	0.29	0.100	2.362	Different	Easier
40	P10I9	0.41	0.3	-0.38	0.33	0.079	2.523	Different	Easier
41	P10I10	0.34	0.12	0.34	0.12	1.000	0.000	Anchored	
42	P10I11	0.96	0.12	0.96	0.12	1.000	0.000	Anchored	
43	P10I12	-0.99	0.26	-0.99	0.26	1.000	0.000	Anchored	
44	P10I13	0.09	0.22	0.09	0.22	1.000	0.000	Anchored	
45	P10I14	-0.25	0.23	-0.37	0.17	0.675	0.625	Same	
46	P10I15	1.01	0.2	1.02	0.15	0.968	-0.059	Same	
47	P10I16	0.68	0.29	1.09	0.29	0.319	-1.424	Different	Harder
48	P10I17	1.27	0.29	0.93	0.29	0.409	1.181	Different	Easier
49	P10I18	0.04	0.31	0.23	0.31	0.665	-0.617	Same	

Table A4
Writing Form Old C vs. Form New C

	Item	Old C		New C		LinP	Effect Size	Item Status	Item Difficulty
		Measure	S.E.	Measure	S.E.				
1	P1I1	-4.71	0.54	-4.71	0.54	1.000	0.000	Anchored	
2	P1I2	-1.14	0.21	-1.14	0.21	1.000	0.000	Anchored	
3	P1I3	-2.82	0.33	-2.82	0.33	1.000	0.000	Anchored	
4	P1I4	-2.93	0.34	-2.93	0.34	1.000	0.000	Anchored	
5	P1I5	-2.72	0.57	-2.72	0.57	1.000	0.000	Anchored	
6	P1I6	-3.56	0.76	-3.56	0.76	1.000	0.000	Anchored	
7	P1I7	-2.43	0.52	-2.43	0.52	1.000	0.000	Anchored	
8	P1I8	-1.37	0.28	-1.37	0.28	1.000	0.000	Anchored	
9	P1I9	-1.43	0.23	-1.43	0.23	1.000	0.000	Anchored	
10	P1I10	-1.48	0.23	-1.48	0.23	1.000	0.000	Anchored	
11	P5I1	-1.44	0.51	-0.07	0.45	0.046	-2.804	Different	Harder
12	P5I2	0.22	0.38	0.13	0.43	0.876	0.228	Same	
13	P5I3	-0.77	0.44	-1.08	0.57	0.668	0.617	Same	
14	P5I4	0.76	0.36	1.97	0.32	0.013	-3.576	Different	Harder
15	P5I5	1.48	0.34	1.54	0.34	0.901	-0.178	Same	
16	P6I1	1.13	0.35	1.76	0.33	0.193	-1.866	Different	Harder
17	P6I2	0.76	0.36	0.92	0.37	0.757	-0.442	Same	
18	P11I1	3.31	0.1	3.31	0.1	1.000	0.000	Anchored	
19	P11I2	3.79	0.1	3.79	0.1	1.000	0.000	Anchored	

Finally, Table A5 summarizes in one chart the results presented in Tables A1 through A4. For reading, we see negligible differences between the average item difficulty in Form B and slightly more in Form C. The percentage of items on each form that were anchored or the same was well over half: 61% for Form B and 57% for Form C. Of those items whose difficulty changed, about half became easier and half became harder.

For writing, the revisions appeared to make the items somewhat more difficult. Although the vast majority of items were anchored or the same (79% for Form B and 86% for Form C), in every case in each form, where there was a statistical difference, the revisions served to make the item harder.

Overall, however, these results suggest that while many items on Forms B and C were unanchored in the methodological approach chosen in this report, and thus calibrated with smaller numbers of cases than the items in Form D, because of the high percentage of items that were statistically the same on both parts of the test, the results would not have been dramatically different had all these items been anchored and thus calibrated with a larger sample size.

Table A5.
Summary of Item Differences Between Old and New Forms B and C

READING	No. Items = 49	Form B	Form C
Average Item Difficulty (SD)	Old	.059 (1.574)	-.105 (1.409)
	New	.067 (1.542)	-.054 (1.251)
Item Status (% total number of items)	Anchored	9 (18%)	9 (18%)
	Same	21 (43%)	19 (39%)
	Anchored or Same	30 (61%)	28 (57%)
	Different	19 (39%)	21 (43%)
Change in Item Difficulty (if different)	Easier	10	9
	Harder	9	12
WRITING	**No. Items = 19**	**Form B**	**Form C**
Average Item Difficulty (SD)	Old	-1.019 (1.854)	-.808 (2.265)
	New	-.893 (1.866)	-.648 (2.367)
Item Status (% total number of items)	Anchored	12 (63%)	12 (63%)
	Same	3 (13%)	4 (21%)
	Anchored or Same	15 (79%)	16 (84%)
	Different	4 (21%)	3 (13%)
Change in Item Difficulty (if different)	Easier	0	0
	Harder	4	3

APPENDIX B: Complete Item Data for *BEST Literacy* Forms B, C, and D

Table B1
Reading New B Item Properties

Item Number	Item Name	Count	Score	P-value	Measure	Error	In.ZSTD	Out.ZSTD
1	RP2I1OBNBOCNCND	402	384	0.96	-2.68	0.26	0.45	1.40
2	RP2I2OBNB	135	131	0.97	-2.92	0.54	-0.09	0.12
3	RP2I3NBNCND	266	244	0.92	-1.97	0.26	0.33	3.34
4	RP2I4NB	69	63	0.91	-1.54	0.49	-0.01	-0.13
5	RP3I1NB	69	67	0.97	-2.99	0.78	0.46	0.15
6	RP3I2NB	69	68	0.99	-3.80	1.06	0.36	0.47
7	RP3I3NBNCND	266	198	0.74	-0.14	0.17	-0.03	-0.51
8	RP4I1NB	69	62	0.90	-1.32	0.46	-0.38	-0.56
9	RP4I2NB	69	64	0.93	-1.80	0.53	-0.11	-0.54
10	RP4I3NBNCND	266	210	0.79	-0.49	0.18	1.63	0.92
11	RP4I4NBNCND	266	242	0.91	-1.84	0.25	0.71	-0.42
12	RP7I1NBNCND	266	215	0.81	-0.65	0.18	1.59	2.01
13	RP7I2NBNCND	266	114	0.43	1.75	0.15	-1.17	-0.92
14	RP8I1NBNCND	266	225	0.85	-1.01	0.20	0.00	0.60
15	RP8I2NB	69	60	0.87	-0.94	0.41	0.21	1.60
16	RP8I3NB	69	52	0.75	0.10	0.32	0.12	-0.64
17	RP9I1NBNCND	266	102	0.38	2.01	0.15	2.26	3.20
18	RP9I2NBNCND	266	195	0.73	-0.06	0.16	1.09	-0.33
19	RP9I3NBNCND	266	95	0.36	2.16	0.15	-0.44	1.85
20	RP9I4NBNCND	266	112	0.42	1.79	0.15	2.48	1.07
21	RP9I5NB	69	46	0.67	0.68	0.30	0.55	-0.39
22	RP9I6NB	69	61	0.88	-1.12	0.43	-0.95	-0.71
23	RP9I7NB	69	44	0.64	0.85	0.29	1.34	0.97
24	RP9I8NB	69	50	0.72	0.31	0.31	0.42	-0.24
25	RP9I9NB	69	38	0.55	1.34	0.28	0.43	0.65
26	RP9I10NB	69	27	0.39	2.22	0.29	0.75	0.70
27	RP9I11NB	69	43	0.62	0.94	0.29	0.46	-0.09
28	RP9I12NB	69	43	0.62	0.94	0.29	2.45	1.17
29	RP9I13NB	69	36	0.52	1.50	0.28	0.76	0.32
30	RP9I14NB	69	17	0.25	3.11	0.32	2.04	2.28
31	RP9I15NB	69	45	0.65	0.77	0.29	0.85	0.52
32	RP10I1NB	69	59	0.86	-0.78	0.40	0.88	0.53
33	RP10I2NB	69	58	0.84	-0.63	0.38	-0.89	-0.75
34	RP10I3NB	69	55	0.80	-0.23	0.35	-0.67	-0.42
35	RP10I4OBNB	135	102	0.76	-0.02	0.23	-0.79	-1.21
36	RP10I5OBNB	135	115	0.85	-0.82	0.27	-1.24	-0.46
37	RP10I6OBNB	135	68	0.50	1.45	0.20	-0.82	0.70
38	RP10I7NB	69	53	0.77	0.00	0.33	0.23	-0.39
39	RP10I8NB	69	30	0.43	1.98	0.28	-0.45	-0.54
40	RP10I9NB	69	55	0.80	-0.23	0.35	-1.72	-1.42

Table B1 continued

Item Number	Item Name	Count	Score	P-value	Measure	Error	In-ZSTD	Out-ZSTD
41	RP10I10OBNBOCNCND	402	272	0.68	0.34	0.12	-2.37	-2.43
42	RP10I11OBNBOCNCND	402	230	0.57	0.96	0.12	2.34	2.29
43	RP10I12OBNB	135	101	0.75	0.03	0.23	-1.18	-0.68
44	RP10I13OBNB	135	45	0.33	2.40	0.21	-0.42	0.23
45	RP10I14NBNCND	266	206	0.77	-0.37	0.17	-2.90	-1.92
46	RP10I15NBNCND	266	149	0.56	1.02	0.15	-0.80	-0.04
47	RP10I16NB	69	38	0.55	1.34	0.28	-2.38	-1.65
48	RP10I17NB	69	50	0.72	0.31	0.31	-2.28	-1.76
49	RP10I18NB	69	38	0.55	1.34	0.28	-0.78	-0.84

Table B2
Writing New B Item Properties

Item Number	Item Name	Count	Score	P-value / *Expected scores	Measure	Error	In.ZSTD	Out.ZSTD
1	WP1I1OBNBOCNCND	346	341	0.99	-4.71	0.54	1.31	0.26
2	WP1I2OBNBOCNCND	346	305	0.88	-1.14	0.21	3.1	4.52
3	WP1I3OBNBOCNCND	346	330	0.95	-2.82	0.33	-0.92	-0.3
4	WP1I4OBNBOCNCND	346	331	0.96	-2.93	0.34	-0.89	1.44
5	WP1I5OBNB	115	109	0.95	-2.76	0.54	-0.32	0.67
6	WP1I6OBNB	115	107	0.93	-2.25	0.47	0.36	1.14
7	WP1I7OBNB	115	107	0.93	-2.25	0.47	0.16	1.28
8	WP1I8OBNB	115	86	0.75	0.08	0.26	0.65	0.79
9	WP1I9OBNBOCNCND	346	311	0.90	-1.43	0.23	-1.04	-1.15
10	WP1I10OBNBOCNCND	346	312	0.90	-1.48	0.23	0.89	2.52
11	WP5I1NB	56	50	0.89	-1.28	0.54	1.43	0.38
12	WP5I2NB	56	48	0.86	-0.76	0.48	1.68	0.63
13	WP5I3NB	56	46	0.82	-0.34	0.43	-0.13	-0.18
14	WP5I4NB	56	45	0.80	-0.16	0.42	-0.22	-0.49
15	WP5I5NB	56	39	0.70	0.71	0.35	-1.07	-1.02
16	WP6I1NB	56	35	0.63	1.16	0.33	0.05	0.06
17	WP6I2NB	56	39	0.70	0.71	0.35	-0.76	-0.5
18	WP11I1OBNB	115	223	*1.94	2.11	0.08	-0.86	0.92
19	WP11I2OBNB	115	148	*1.29	2.58	0.08	-0.8	0.05

Table B3
Reading New C Item Properties

Item Number	Item Name	Count	Score	P-value	Measure	Error	In.ZSTD	Out.ZSTD
1	RP2I1OBNBOCNCND	402	384	0.96	-2.68	0.26	0.45	1.40
2	RP2I2OCNC	140	128	0.91	-1.95	0.34	0.21	1.17
3	RP2I3NBNCND	266	244	0.92	-1.97	0.26	0.33	3.34
4	RP2I4NC	70	61	0.87	-1.61	0.42	0.41	-0.39
5	RP3I1NC	70	62	0.89	-1.79	0.44	-0.44	-0.74
6	RP3I2NC	70	59	0.84	-1.28	0.39	-0.72	-0.83
7	RP3I3NBNCND	266	198	0.74	-0.14	0.17	-0.03	-0.51
8	RP4I1NC	70	55	0.79	-0.73	0.35	0.79	2.69
9	RP4I2NC	70	63	0.90	-1.99	0.46	0.36	2.63
10	RP4I3NBNCND	266	210	0.79	-0.49	0.18	1.63	0.92
11	RP4I4NBNCND	266	242	0.91	-1.84	0.25	0.71	-0.42
12	RP7I1NBNCND	266	215	0.81	-0.65	0.18	1.59	2.01
13	RP7I2NBNCND	266	114	0.43	1.75	0.15	-1.17	-0.92
14	RP8I1NBNCND	266	225	0.85	-1.01	0.20	0.00	0.60
15	RP8I2NC	70	56	0.80	-0.86	0.36	0.85	1.48
16	RP8I3NC	70	46	0.66	0.23	0.31	-0.62	-0.67
17	RP9I1NBNCND	266	102	0.38	2.01	0.15	2.26	3.20
18	RP9I2NBNCND	266	195	0.73	-0.06	0.16	1.09	-0.33
19	RP9I3NBNCND	266	95	0.36	2.16	0.15	-0.44	1.85
20	RP9I4NBNCND	266	112	0.42	1.79	0.15	2.48	1.07
21	RP9I5NC	70	35	0.50	1.17	0.29	0.72	0.85
22	RP9I6NC	70	50	0.71	-0.17	0.32	0.00	0.40
23	RP9I7NC	70	32	0.46	1.42	0.29	-0.30	-0.34
24	RP9I8NC	70	24	0.34	2.08	0.29	1.76	1.81
25	RP9I9NC	70	53	0.76	-0.49	0.34	-0.44	-0.79
26	RP9I10NC	70	28	0.40	1.75	0.29	-0.84	-0.27
27	RP9I11NC	70	38	0.54	0.93	0.29	-0.14	-0.12
28	RP9I12NC	70	48	0.69	0.04	0.31	0.76	-0.18
29	RP9I13NC	70	30	0.43	1.58	0.29	0.34	0.00
30	RP9I14NC	70	48	0.69	0.04	0.31	-0.21	-0.71
31	RP9I15NC	70	57	0.81	-0.99	0.37	-0.93	-0.53
32	RP10I1NC	70	57	0.81	-0.99	0.37	0.07	0.72
33	RP10I2NC	70	50	0.71	-0.17	0.32	0.95	2.81
34	RP10I3NC	70	57	0.81	-0.99	0.37	0.33	-0.36
35	RP10I4OCNC	140	106	0.76	-0.32	0.23	-0.34	-0.96
36	RP10I5OCNC	140	124	0.89	-1.54	0.30	-0.93	-0.02
37	RP10I6OCNC	140	70	0.50	1.30	0.20	0.50	2.94
38	RP10I7NC	70	47	0.67	0.13	0.31	-0.36	-0.32
39	RP10I8NC	70	40	0.57	0.76	0.29	-1.01	-0.82
40	RP10I9NC	70	52	0.74	-0.38	0.33	0.35	-0.22

Table B3 continued

Item Number	Item Name	Count	Score	P-value	Measure	Error	In.ZSTD	Out.ZSTD
41	RP10I10OBNBOCNCND	402	272	0.68	0.34	0.12	-2.37	-2.43
42	RP10I11OBNBOCNCND	402	230	0.57	0.96	0.12	2.34	2.29
43	RP10I12OCNC	140	117	0.84	-0.99	0.26	-2.55	-1.84
44	RP10I13OCNC	140	98	0.70	0.09	0.22	-2.34	-1.12
45	RP10I14NBNCND	266	206	0.77	-0.37	0.17	-2.90	-1.92
46	RP10I15NBNCND	266	149	0.56	1.02	0.15	-0.80	-0.04
47	RP10I16NC	70	36	0.51	1.09	0.29	0.34	0.36
48	RP10I17NC	70	38	0.54	0.93	0.29	-0.88	-0.20
49	RP10I18NC	70	46	0.66	0.23	0.31	-0.07	-0.40

Table B4
Writing New C Item Properties

Item Number	Item Name	Count	Score	P-value / *Expected scores	Measure	Error	In.ZSTD	Out.ZSTD
1	WP1I1OBNBOCNCND	346	341	0.99	-4.71	0.54	1.31	0.26
2	WP1I2OBNBOCNCND	346	305	0.88	-1.14	0.21	3.1	4.52
3	WP1I3OBNBOCNCND	346	330	0.95	-2.82	0.33	-0.92	-0.3
4	WP1I4OBNBOCNCND	346	331	0.96	-2.93	0.34	-0.89	1.44
5	WP1I5OCNC	118	114	0.97	-2.72	0.57	-0.35	0.78
6	WP1I6OCNC	118	116	0.98	-3.56	0.76	-0.49	0.23
7	WP1I7OCNC	118	113	0.96	-2.43	0.52	0.52	1.18
8	WP1I8OCNCND	231	208	0.90	-1.37	0.28	0.05	0.75
9	WP1I9OBNBOCNCND	346	311	0.90	-1.43	0.23	-1.04	-1.15
10	WP1I10OBNBOCNCND	346	312	0.90	-1.48	0.23	0.89	2.52
11	WP5I1NC	57	48	0.84	-0.07	0.45	0.09	0.32
12	WP5I2NC	57	47	0.82	0.13	0.43	-0.36	-0.17
13	WP5I3NC	57	52	0.91	-1.08	0.57	-0.34	-0.13
14	WP5I4NC	57	33	0.58	1.97	0.32	-0.56	-0.3
15	WP5I5NC	57	37	0.65	1.54	0.34	0.24	-0.19
16	WP6I1NC	57	35	0.61	1.76	0.33	0.39	0.14
17	WP6I2NC	57	42	0.74	0.92	0.37	0.67	0.88
18	WP11I1OCNC	118	178	*1.51	3.31	0.1	-0.08	0.27
19	WP11I2OCNC	118	127	*1.08	3.79	0.1	-1.31	5.29

Table B5
Reading New D Item Properties

Item Number	Item Name	Count	Score	P-value	Measure	Error	In.ZSTD	Out.ZSTD
1	RP2I1OBNBOCNCND	402	384	0.96	-2.68	0.26	0.45	1.40
2	RP2I2ND	127	120	0.94	-2.54	0.44	0.39	0.70
3	RP2I3NBNCND	266	244	0.92	-1.97	0.26	0.33	3.34
4	RP2I4ND	127	117	0.92	-2.03	0.38	0.55	0.36
5	RP3I1ND	127	121	0.95	-2.75	0.47	-0.52	-0.46
6	RP3I2ND	127	119	0.94	-2.35	0.42	-0.26	-0.08
7	RP3I3NBNCND	266	198	0.74	-0.14	0.17	-0.03	-0.51
8	RP4I1ND	127	116	0.91	-1.89	0.36	-1.05	-1.05
9	RP4I2ND	127	120	0.94	-2.54	0.44	-0.62	-0.36
10	RP4I3NBNCND	266	210	0.79	-0.49	0.18	1.63	0.92
11	RP4I4NBNCND	266	242	0.91	-1.84	0.25	0.71	-0.42
12	RP7I1NBNCND	266	215	0.81	-0.65	0.18	1.59	2.01
13	RP7I2NBNCND	266	114	0.43	1.75	0.15	-1.17	-0.92
14	RP8I1NBNCND	266	225	0.85	-1.01	0.20	0.00	0.60
15	RP8I2ND	127	107	0.84	-0.98	0.28	-0.90	-1.05
16	RP8I3ND	127	107	0.84	-0.98	0.28	-1.02	-0.66
17	RP9I1NBNCND	266	102	0.38	2.01	0.15	2.26	3.20
18	RP9I2NBNCND	266	195	0.73	-0.06	0.16	1.09	-0.33
19	RP9I3NBNCND	266	95	0.36	2.16	0.15	-0.44	1.85
20	RP9I4NBNCND	266	112	0.42	1.79	0.15	2.48	1.07
21	RP9I5ND	127	70	0.55	1.02	0.21	1.12	0.44
22	RP9I6ND	127	84	0.66	0.38	0.22	-0.42	-1.02
23	RP9I7ND	127	58	0.46	1.54	0.21	1.31	1.78
24	RP9I8ND	127	91	0.72	0.03	0.23	0.99	0.53
25	RP9I9ND	127	61	0.48	1.41	0.21	-0.36	-0.6
26	RP9I10ND	127	59	0.46	1.50	0.21	-0.31	-0.39
27	RP9I11ND	127	63	0.50	1.32	0.21	-0.70	0.68
28	RP9I12ND	127	80	0.63	0.57	0.21	-0.40	-0.40
29	RP9I13ND	127	97	0.76	-0.30	0.24	0.63	1.00
30	RP9I14ND	127	68	0.54	1.10	0.21	-2.45	-1.25
31	RP9I15ND	127	98	0.77	-0.36	0.25	0.15	-0.20
32	RP10I1ND	127	61	0.48	1.41	0.21	5.64	4.64
33	RP10I2ND	127	108	0.85	-1.06	0.29	0.40	0.62
34	RP10I3ND	127	87	0.69	0.23	0.22	0.70	1.26
35	RP10I4ND	127	88	0.69	0.18	0.22	-0.49	-0.50
36	RP10I5ND	127	108	0.85	-1.06	0.29	-0.62	1.20
37	RP10I6ND	127	54	0.43	1.72	0.21	-2.45	-1.69
38	RP10I7ND	127	97	0.76	-0.30	0.24	0.03	-0.37
39	RP10I8ND	127	60	0.47	1.45	0.21	-0.31	-0.49
40	RP10I9ND	127	94	0.74	-0.13	0.24	-2.08	-1.71

Table B5 continued

Item Number	Item Name	Count	Score	P-value	Measure	Error	In.ZSTD	Out.ZSTD
41	RP10I10OBNBOCNCND	402	272	0.68	0.34	0.12	-2.37	-2.43
42	RP10I11OBNBOCNCND	402	230	0.57	0.96	0.12	2.34	2.29
43	RP10I12ND	127	71	0.56	0.97	0.21	-1.41	-0.83
44	RP10I13ND	127	67	0.53	1.15	0.21	-0.64	-0.86
45	RP10I14NBNCND	266	206	0.77	-0.37	0.17	-2.90	-1.92
46	RP10I15NBNCND	266	149	0.56	1.02	0.15	-0.80	-0.04
47	RP10I16ND	127	92	0.72	-0.02	0.23	-2.38	-0.82
48	RP10I17ND	127	71	0.56	0.97	0.21	-0.31	-0.57
49	RP10I18ND	127	55	0.43	1.67	0.21	0.16	0.79

Table B6
Writing New D Item Properties

Item Number	Item Name	Count	Score	P-value / *Expected scores	Measure	Error	In.ZSTD	Out.ZSTD
1	WP1I1OBNBOCNCND	346	341	0.99	-4.71	0.54	1.31	0.26
2	WP1I2OBNBOCNCND	346	305	0.88	-1.14	0.21	3.1	4.52
3	WP1I3OBNBOCNCND	346	330	0.95	-2.82	0.33	-0.92	-0.3
4	WP1I4OBNBOCNCND	346	331	0.96	-2.93	0.34	-0.89	1.44
5	WP1I5ND	113	104	0.92	-2.08	0.49	-0.72	-0.41
6	WP1I6ND	113	104	0.92	-2.08	0.49	-0.35	-0.3
7	WP1I7ND	113	106	0.94	-2.65	0.57	-0.53	0.03
8	WP1I8OCNCND	231	208	0.90	-1.37	0.28	0.05	0.75
9	WP1I9OBNBOCNCND	346	311	0.90	-1.43	0.23	-1.04	-1.15
10	WP1I10OBNBOCNCND	346	312	0.90	-1.48	0.23	0.89	2.52
11	WP5I1ND	113	98	0.87	-0.99	0.38	0.28	0.78
12	WP5I2ND	113	87	0.77	0.21	0.3	0.24	0.4
13	WP5I3ND	113	101	0.89	-1.47	0.42	-0.42	0.17
14	WP5I4ND	113	80	0.71	0.79	0.28	1.21	0.34
15	WP5I5ND	113	69	0.61	1.54	0.25	1.97	1.28
16	WP6I1ND	113	60	0.53	2.08	0.24	-1.56	-0.24
17	WP6I2ND	113	70	0.62	1.48	0.25	-2.57	-0.98
18	WP11I1ND	113	143	*1.27	3.61	0.1	-1.79	0.11
19	WP11I2ND	113	156	*1.38	3.5	0.1	0.39	0.46

BEST Literacy™

Phone:	1-866-845-BEST (2378)
Fax:	1-888-700-3629
Email:	bestliteracy@cal.org
Web site:	www.cal.org/bestliteracy
Write:	Center for Applied Linguistics Attn: BEST Literacy 4646 40th Street, NW Washington, DC 20016-1859

CAL

www.cal.org

June 2008