Chapter 14

ASSESSMENT FRAMEWORKS AND INSTRUMENTS FOR THE 1998 NATIONAL AND STATE READING ASSESSMENTS[1]

Patricia L. Donahue and Terry L. Schoeps
Educational Testing Service

14.1 INTRODUCTION

The reading framework was originally developed through a broad-based consensus process conducted by the Council of Chief State School Officers (CCSSO) working under contract to the National Assessment Governing Board (NAGB). The development process involved a steering committee, a planning committee, and CCSSO project staff. Educators, scholars, and citizens, representative of many diverse constituencies and points of view, participated in the national consensus process to design objectives for the reading assessment. The framework that was used for the 1998 NAEP reading assessment was also used for the 1992 and 1994 assessments.

The instrument used in the 1998 reading assessment was composed of a combination of reading passages and questions from the 1992 and 1994 assessments and a set of passages and questions newly developed for 1998. A total of twenty-three unique blocks (a block is a reading passage with a set of questions) were administrated in 1998. Three of these blocks were developed for 1998 and the remaining twenty were carried over from the 1992 and 1994 assessments. Administering the same blocks across assessment years allows for the reporting of trends in reading performance. At the same time, developing new sets of passages and questions made it possible to release three blocks for public use. The framework for the reading assessment is available on the National Assessment Governing Board (NAGB) web site at *http://www.nagb.org*.

Sections 14.3 through 14.5 include a detailed description of the framework and the development of reading questions, or *items*, for the 1998 NAEP reading assessment. Section 14.8 also describes the student background questionnaires and the reading teacher questionnaire. Additional information on the structure and content of assessment booklets can be found in Section 14.9. The list of committee members who participated in the 1998 development process is provided in Appendix K.

Samples of assessment instruments and student responses are published in the *NAEP 1998 Reading Report Card for the Nation and the States: Findings from the National Assessment of Educational Progress* (Donahue, Voelkl, Campbell, & Mazzeo, 1999).

14.2 DEVELOPING THE READING ASSESSMENT FRAMEWORK

NAGB is responsible for setting policy for NAEP; this policymaking role includes the development of assessment frameworks and test specifications. Appointed by the Secretary of Education from lists of nominees proposed by the Board itself in various statutory categories, the 24-member board is composed of state, local, and federal officials, as well as educators and members of the public.

[1] Patricia L. Donahue manages the item development process for NAEP reading assessments. Terry L. Schoeps coordinates the production of NAEP technical reports.

NAGB began the development process for the 1992 reading objectives (which also served as the objectives for the 1994 and 1998 assessments) by conducting a widespread mail review of the objectives for the 1990 reading assessment and by holding a series of public hearings throughout the country. The contract for managing the remainder of the consensus process was awarded to the CCSSO. The development process included the following activities:

- A Steering Committee consisting of members recommended by each of 16 national organizations was established to provide guidance for the consensus process. The committee monitored the progress of the project and offered advice. Drafts of each version of the document were sent to members of the committee for review and reaction.
- A Planning Committee was established to identify the objectives to be assessed in reading and prepare the framework document. The members of this committee consisted of experts in reading, including college professors, an academic dean, a classroom teacher, a school administrator, state level assessment and reading specialists, and a representative of the business community. This committee met with the Steering Committee and as a separate group. A subgroup also met to develop item specifications. Between meetings, members of the committee provided information and reactions to drafts of the framework.
- The project staff at CCSSO met regularly with staff from NAGB and NCES to discuss progress made by the Steering and Planning committees.

During this development process, input and reactions were continually sought from a wide range of members of the reading field, experts in assessment, school administrators, and state staff in reading assessment. In particular, innovative state assessment efforts and work being done by the Center for the Learning and Teaching of Literature (Langer, 1989, 1990).

For more detail on the development and specifications of the reading framework, refer to the *Reading Framework and Specifications for the 1998 National Assessment of Educational Progress, 1992–1998* (NAGB, 1990).

14.3 READING FRAMEWORK AND ASSESSMENT DESIGN PRINCIPLES

The reading objectives framework was designed to focus on reading processes and outcomes, rather than reflect a particular instructional or theoretical approach. It was stated that the framework should focus not on the specific reading skills that lead to outcomes, but rather on the quality of the outcomes themselves. The framework was intended to embody a broad view of reading by addressing the increasing level of literacy needed for employability, personal development, and citizenship. The framework also specified a reliance on contemporary reading research and the use of nontraditional assessment formats that more closely resemble desired classroom activities.

The objectives development was guided by the consideration that the assessment should reflect many of the curricular emphases and objectives in various states, localities, and school districts in addition to what various scholars, practitioners, and interested citizens believed should be included in the curriculum. Accordingly, the committee gave attention to several frames of reference:

- The purpose of the NAEP reading assessment is to provide information about the progress and achievement of students in general rather than to test individual

students' ability. NAEP is designed to inform policymakers and the public about reading ability in the United States.

- The term "reading literacy" should be used in the broad sense of knowing when to read, how to read, and how to reflect on what has been read. It represents a complex, interactive process that goes beyond basic or functional literacy.
- The reading assessment should use valid and authentic tasks that are both broad and complete in their coverage of important reading behaviors so that the test will be useful and valid, and will demonstrate a close link to desired classroom instruction.
- Every effort should be made to make the best use of available methodology and resources in driving assessment capabilities forward. New types of items and new methods of analysis were recommended for NAEP reading assessments.
- Every effort must be made in developing the assessment to represent a variety of opinions, perspectives, and emphases among professionals, as well as state and local school districts.

14.4 FRAMEWORK FOR THE 1998 READING ASSESSMENT

The framework adopted for the 1998 reading assessment, which also served as the framework for the 1992 and 1994 assessments, was organized according to a four-by-three matrix of reading *stances* by reading *purposes*. The stances include:

- Initial Understanding;
- Developing an Interpretation;
- Personal Reflection and Response; and
- Demonstrating a Critical Stance.

These stances were assessed across three global purposes defined as:

- Reading for Literary Experience;
- Reading to Gain Information; and
- Reading to Perform a Task.

Different types of texts were used to assess the various purposes for reading. Students' reading abilities were evaluated in terms of a single purpose for each type of text. At grade 4, only Reading for Literary Experience and Reading to Gain Information were assessed, while all three global purposes were assessed at grades 8 and 12. Figure 14-1 and 14-2 describe the four reading stances and three reading purposes that guided the development of NAEP's 1992, 1994, and 1998 reading assessments.

The Planning Committee was interested in creating an assessment that would be forward-thinking and reflect quality instruction. In recognition that the demands made of readers change as they mature and move through school, it was recommended that the proportion of items have some relation to reading purpose (i.e., for literary experience, to gain information, to perform a task). The distribution of items by reading purpose across grade levels recommended in the assessment framework is provided in Table 14-1.

Readers use a range of cognitive abilities and assume various stances that should be assessed within each of the reading purposes. While reading, students form an initial understanding of the text and connect ideas within the text to generate interpretations. In addition, they extend and elaborate their understanding by responding to the text personally and critically and by relating ideas in the text to prior knowledge.

For more detail on the development and specifications of the Reading Framework, refer to *Reading Framework for the National Assessment of Educational Progress, 1992-1998* (NAGB, 1990).

Figure 14-1
Description of Reading Stances

Readers interact with text in various ways as they use background knowledge and understanding of text to construct, extend, and examine meaning. The NAEP reading assessment framework specified four reading stances to be assessed that represent various interactions between readers and texts. These stances are not meant to describe a hierarchy of skills or abilities. Rather, they are intended to describe behaviors that readers at all developmental levels should exhibit.

Initial Understanding

Initial understanding requires a broad, preliminary construction of an understanding of the text. Questions testing this aspect ask the reader to provide an initial impression or unreflected understanding of what was read. The first question following a passage was usually one testing initial understanding.

Developing an Interpretation

Developing an interpretation requires the reader to go beyond the initial impression to develop a more complete understanding of what was read. Questions testing this aspect require a more specific understanding of the text and involve linking information across parts of the text as well as focusing on specific information.

Personal Reflection and Response

Personal reflection and response requires the reader to connect knowledge from the text more extensively with his or her own personal background knowledge and experience. The focus is on how the text relates to personal experience; questions on this aspect ask the readers to reflect and respond from a personal perspective. Personal reflection and response questions were typically formatted as constructed-response items to allow for individual possibilities and varied responses.

Demonstrating a Critical Stance

Demonstrating a critical stance requires the reader to stand apart from the text, consider it, and judge it objectively. Questions on this aspect require the reader to perform a variety of tasks such as critical evaluation, comparing and contrasting, application to practical tasks, and understanding the impact of such text features as irony, humor, and organization. These questions focus on the reader as critic and require reflection on and judgments about how the text is written.

Figure 14-2
Description of Purposes for Reading

Reading involves an interaction between a specific type of text or written material and a reader, who typically has a purpose for reading that is related to the type of text and the context of the reading situation. The reading assessment presented three types of text to students representing each of three reading purposes: literary text for literary experience, informational text to gain information, and documents to perform a task. Students' reading skills were evaluated in terms of a single purpose for each type of text.

Reading for Literary Experience

Reading for literary experience involves reading literary text to explore the human condition, to relate narrative events with personal experiences, and to consider the interplay in the selection among emotions, events, and possibilities. Students in the NAEP reading assessment were provided with a wide variety of literary text, such as short stories, poems, fables, historical fiction, science fiction, and mysteries.

Reading to Gain Information

Reading to gain information involves reading informative passages in order to obtain some general or specific information. This often requires a more utilitarian approach to reading that requires the use of certain reading/thinking strategies different from those used for other purposes. In addition, reading to gain information often involves reading and interpreting adjunct aids such as charts, graphs, maps, and tables that provide supplemental or tangential data. Informational passages in the NAEP reading assessment included biographies, science articles, encyclopedia entries, primary and secondary historical accounts, and newspaper editorials.

Reading to Perform a Task

Reading to perform a task involves reading various types of materials for the purpose of applying the information or directions in completing a specific task. The reader's purpose for gaining meaning extends beyond understanding the text to include the accomplishment of a certain activity. Documents requiring students in the NAEP reading assessment to perform a task included directions for creating a time capsule, a bus schedule, a tax form, and instructions on how to write a letter to a senator. Reading to perform a task was assessed only at grades 8 and 12.

Table 14-1
Percentage Distribution of Items by Reading Purpose as Specified in the NAEP Reading Framework

Grade	Purpose for Reading		
	Reading for Literary Experience	Reading to Gain Information	Reading to Perform a Task
4	55%	45%	(Not Assessed)
8	40%	40%	20%
12	35%	45%	20%

Table 14-2 shows the distribution of items by reading stance, as specified in the reading framework, for all three grade levels.

Table 14-2
Percentage Distribution of Items by Reading Stance as Specified in the NAEP Reading Framework

Reading Stance	Grades 4, 8, and 12
Initial Understanding/Developing an Interpretation	33%
Personal Reflection and Response	33%
Demonstrating a Critical Stance	33%

14.5 DEVELOPING THE READING COGNITIVE ITEMS

In developing the new portion of the 1998 NAEP reading assessment, the same framework and procedures used in 1992, and again in 1994, were followed. After careful review of the objectives, reading materials were selected and questions were developed that were appropriate to the objectives. All questions were extensively reviewed by specialists in reading, measurement, and bias/sensitivity, as well as by state representatives.

The development of cognitive items began with a careful selection of grade-appropriate passages for the assessment. Passages were selected from a pool of reading selections contributed by teachers from across the country. The framework states that the assessment passages should represent authentic, naturally occurring reading material that students may encounter in and out of school. Furthermore, these passages were to be reproduced in test booklets as they had appeared in their original publications. In some cases, materials (such as bus schedules) were provided to students separate from the printed assessment booklet. Final passage selections were made by the Reading Instrument Development Committee. In order to guide the development of items, passages were outlined or mapped to identify essential elements of the text.

The assessment included constructed-response (short and extended) and multiple-choice items. The decision to use a specific item type was based on a consideration of the most appropriate format for assessing the particular objective. Both types of constructed-response items were designed to provide an in-depth view of students' ability to read thoughtfully and to respond appropriately to what they read. Short constructed-response questions were used when students needed to respond in only one or two sentences in order to demonstrate full comprehension. Extended constructed-response questions were used when the task required more thoughtful consideration of the text and engagement in more complex

reading processes. Multiple-choice items were used whenever a reading outcome could be measured through use of these items.

A carefully developed and proven series of steps was used to create the assessment items. These steps are described in Chapter 2.

The assessment included 25-minute and 50-minute “blocks," each consisting of one or more passages and a set of multiple-choice and constructed-response items to assess students’ comprehension of the written material. At grade 8 and 12 students were asked to respond to either two 25-minute blocks or one 50-minute block. The grade-4 assessment included eight 25-minute blocks (four blocks measuring each of the two global purposes for reading assessed at this grade). The instruments at grades 8 and 12 each included nine 25-minute blocks (three blocks measuring each of the global purposes for reading). In addition, the grade 8 assessment included one 50-minute block and the grade-12 assessment included two 50-minute blocks.

14.6 DEVELOPING THE READING OPERATIONAL FORMS

A reading field test was conducted in March 1997 to test new reading questions that were developed to replace the few 1994 items that had been publicly released and were, therefore, no longer able to be used in an operational assessment. The field test was given to national samples of fourth-, eighth-, and twelfth-grade students. The field test data were collected, scored, and analyzed in preparation for meetings with the Reading Instrument Development Committee. Using item analysis, which provided the mean percentage of correct responses, the polyserial correlations, and the difficulty level for each item in the field test, committee members, ETS test development staff, and NAEP/ETS staff reviewed the materials. The objectives that guided these reviews included:

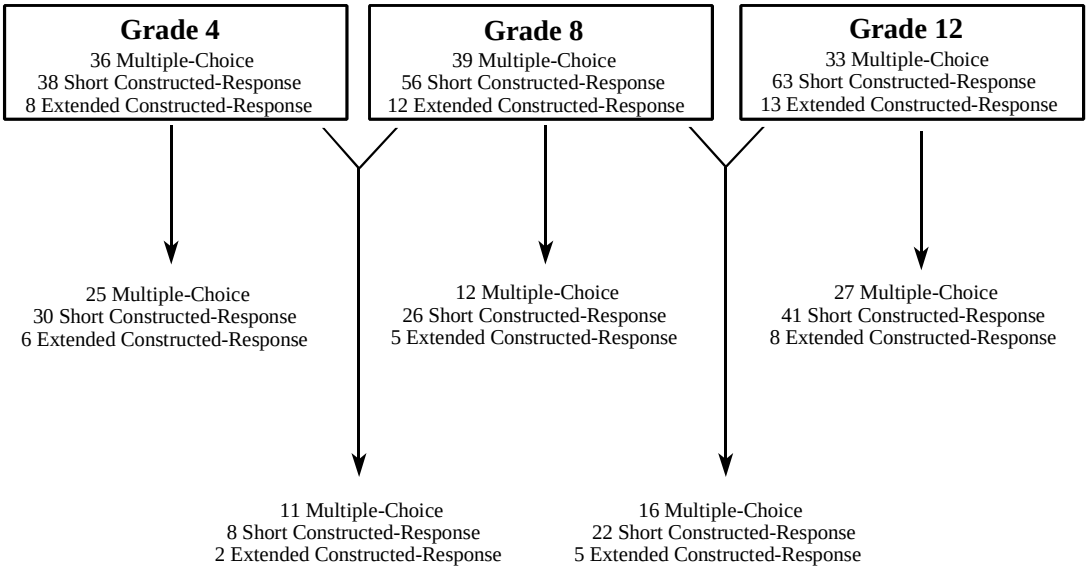
- determining which items were most related to overall student achievement,
- determining the need for revisions of items that lacked clarity or had ineffective item formats,
- prioritizing items to be included in the assessment, and
- determining appropriate timing for assessment items.

Once the committees had selected the items, all items were rechecked for content, measurement, and sensitivity concerns. The federal clearance process was initiated in June 1997 with the submission of draft materials to NCES. The package containing the final set of cognitive items assembled into blocks and questionnaires was submitted in June 1997. Throughout the clearance process, revisions were made in accordance with changes required by the government. Upon approval, the blocks (assembled into booklets) and questionnaires were prepared for printing.

14.7 DISTRIBUTION OF READING ASSESSMENT ITEMS

Figure 14-3 lists the total number of items at each grade level in the 1998 assessment. Of the total of 247 items, there are 93 unique multiple-choice items and 154 unique constructed-response questions that make up the 1998 reading assessment. Some of these items are used at more than one grade level. As a result, the sum of the items that appear at each grade level is greater than the total number of unique items.

Figure 14-3
Distribution of Items for the 1998 Reading Assessment



In the development process, every effort was made to meet the content and process targets specified in the assessment framework. Table 14-3 shows the approximate percentage of aggregate assessment time devoted to each purpose for reading at each grade level. Percentages are based on the classifications agreed upon by NAEP's 1998 Instrument Development Committee. Note that the numbers presented in Table 14-3 differ from Table 14-1 in that Table 14-1 shows the distribution of assessment items as specified in the reading framework.

Table 14-3
Percentage Distribution of Assessment Time by Grade and Reading Purpose for the NAEP 1998 Reading Assessment

Reading Purpose	Grade 4	Grade 8	Grade 12
Reading for Literary Experience	50%	38%	33%
Reading to Gain Information	50%	38%	47%
Reading to Perform a Task	N/A	23%	20%

Table 14-4 shows the approximate percentage of assessment time devoted to each reading stance. Unlike the purposes for reading, in which individual students did not receive questions in all areas, every student completed tasks involving each of the reading stances. It is recognized that making discrete classifications is difficult for these categories and that independent efforts to classify NAEP questions have led to different results (National Academy of Education, 1992). Also, it has been found that developing personal response questions that are considered equitable across students' different backgrounds and experiences is difficult. Note that the numbers presented in Table 14-4 differ from Table 14-2, in that Table 14-2 shows the distribution of items as specified in the reading framework.

Table 14-4
Percentage Distribution of Assessment Time by Grade and Reading Stance for the NAEP 1998 Reading Assessment

Reading Stance	Grade 4	Grade 8	Grade 12
Initial Understanding/ Developing an Interpretation	56%	49%	52%
Personal Reflection and Response	21%	19%	16%
Demonstrating a Critical Stance	23%	32%	32%

14.8 BACKGROUND QUESTIONNAIRES FOR THE 1998 READING ASSESSMENT

Research indicates that school, home, and attitudinal variables affect students' reading comprehension and literacy. Therefore, in addition to assessing how well students read, it is important to understand the instructional context in which reading takes place, students' home support for literacy, and their reading habits and attitudes. To gather contextual information, NAEP assessments include background questions designed to provide insight into the factors that may influence reading scale scores in the literary, informational, and document categories assessed.

NAEP includes both general background questionnaires given to participants in all subjects and subject-specific questionnaires for both students and their teachers. The development of the general background questionnaires is discussed below. It is worth noting that members of the Reading Instrument Development Committee were consulted on the appropriateness of the issues addressed in all questionnaires that may relate to reading instruction and achievement. Like the cognitive items, all background questions were submitted for extensive review and field testing. Recognizing the reliability problems inherent in self-reported data, particular attention was given to developing questions that were meaningful and unambiguous and that would encourage accurate reporting.

In addition to the cognitive questions, the 1998 assessment included one five-minute set each of general and reading background questions designed to gather contextual information about students, their instructional and recreational experiences in reading, and their attitudes toward reading. Students in the fourth grade were given additional time because the items in the general questionnaire were read aloud for them. A one-minute questionnaire was also given to students at the end of each booklet to measure students' motivation in completing the assessment and their familiarity with assessment tasks.

14.8.1 Student Reading Questionnaires

Three sets of multiple-choice background questions were included as separate sections in each student booklet:

General Background: The general background questions collected demographic information about race/ethnicity, language spoken at home, mother's and father's level of education, reading materials in the home, homework, school attendance, which parents live at home, and which parents work outside the home.

Reading Background: Students were asked to report their instructional experiences related to reading in the classroom, including group work, special projects, and writing in response to reading. In addition, they were asked about the instructional practices of their reading teachers and the extent to which the students themselves discussed what they read in class and demonstrated use of skills and strategies.

Motivation: Students were asked five questions about their attitudes and perceptions about reading and self-evaluation of their performance on the NAEP assessment.

Table 14-5 shows the number of questions per background section and the placement of each within student booklets.

Table 14-5
NAEP 1998 Background Sections of Student Reading Booklets

	Number of Questions	Placement in Student Booklet
Grade 4		
General Background	21	Section 1
Reading Background	22	Section 4
Motivation	5	Section 5
Grade 8		
General Background	17	Section 1
Reading Background	24	Section 4
Motivation	5	Section 5
Grade 12		
General Background	18	Section 1
Reading Background	25	Section 4
Motivation	5	Section 5

14.8.2 Language Arts Teacher Questionnaire

To supplement the information on instruction reported by students, the reading teachers of the fourth and eighth graders participating in the NAEP reading assessment were asked to complete a questionnaire about their educational background, content-area preparation, and classroom practices. The teacher questionnaire contained two parts. The first part pertained to the teachers' background and general training. The second part pertained to specific training in teaching reading and the procedures the teacher used for *each class* containing an assessed student.

The **Teacher Questionnaire, Part I: Background, Education, and Resources** (49 questions at grade 4 and 48 questions at grade 8) included questions pertaining to:

- gender;
- race/ethnicity;
- years of teaching experience;
- certification, degrees, major and minor fields of study;
- coursework in education;
- coursework in specific subject areas;
- amount of in-service training;
- extent of control over instructional issues; and
- availability of resources for their classroom.

The **Teacher Questionnaire, Part IIA: Reading/Writing Preparation** (12 questions at grade 4 and 12 at grade 8) included questions on the teacher's professional development in reading theory and instruction.

The **Teacher Questionnaire, Part IIB: Reading/Writing Instructional Information** (84 questions at grade 4 and 85 questions at grade 8) included questions pertaining to:

- ability level of students in the class;
- whether students were assigned to the class by ability level;
- time on task;
- homework assignments;
- frequency of instructional activities used in class;
- methods of assessing student progress in reading;
- instructional emphasis given to the reading abilities covered in the assessment; and
- use of particular resources.

14.9 STUDENT BOOKLETS FOR THE 1998 READING ASSESSMENT

The assembly of reading blocks into booklets and their subsequent assignment to sampled students was determined by a partially balanced incomplete block (PBIB) design with spiraled administration. The 25-minute blocks were assembled into 52 booklets such that two different blocks were assigned to each booklet and each block appeared in four booklets. Each 25-minute block was paired with another block measuring the same purpose for reading (i.e., reading for literary experience, reading to gain information, reading to perform a task) approximately 75 percent of the time at grade 4 and approximately 50 percent of the time at grades 8 and 12. This was the *partially balanced* part of the PBIB design.

The focused PBIB design also balances the order of presentation of the blocks—every block appears as the first cognitive block in two booklets and as the second cognitive block in two other booklets. This design allows for some control of context and fatigue effects.

At grade 4, the blocks were assembled into 16 booklets. At grade 8, the 25-minute blocks were assembled into 18 booklets, and the 50-minute block appeared in a single booklet. At grade 12, the 25-minute blocks were assembled into 18 booklets, and each 50-minute block appeared in a separate booklet. The assessment booklets were then spiraled and bundled. Spiraling involves interweaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in a position in a bundle.

As in the other subjects, the final step in the BIB or PBIB spiraling procedure was the assigning of booklets to the assessed students. The students in the assessment session were assigned booklets in the order in which the booklets were bundled. Thus, most students in an assessment session received different booklets. Tables 14-6, 14-7, and 14-8 detail the configuration of booklets administered in the 1998 national and state reading assessment.

Table 14-6

NAEP 1998 Reading Grade 4 Booklet Configuration

Booklet Number	Common Core Background	Question Block 1	Question Block 2	Reading Background	Motivation
1	CR	R4	R3	RB	RA
2	CR	R3	R5	RB	RA
3	CR	R5	R9	RB	RA
4	CR	R9	R4	RB	RA
5	CR	R4	R5	RB	RA
6	CR	R3	R9	RB	RA
7	CR	R6	R10	RB	RA
8	CR	R10	R7	RB	RA
9	CR	R7	R8	RB	RA
10	CR	R8	R6	RB	RA
11	CR	R6	R7	RB	RA
12	CR	R10	R8	RB	RA
13	CR	R7	R4	RB	RA
14	CR	R8	R3	RB	RA
15	CR	R5	R6	RB	RA
16	CR	R9	R10	RB	RA

Table 14-7

NAEP 1998 Reading Grade 8 Booklet Configuration

Booklet Number	Common Core Background	Question Block 1	Question Block 2	Reading Background	Motivation
1	CR	R3	R4	RB	RA
2	CR	R4	R5	RB	RA
3	CR	R5	R3	RB	RA
4	CR	R6	R8	RB	RA
5	CR	R8	R7	RB	RA
6	CR	R7	R6	RB	RA
7	CR	R10	R9	RB	RA
8	CR	R9	R11	RB	RA
9	CR	R11	R10	RB	RA
10	CR	R3	R8	RB	RA
11	CR	R7	R4	RB	RA
12	CR	R5	R6	RB	RA
13	CR	R6	R9	RB	RA
14	CR	R8	R11	RB	RA
15	CR	R10	R7	RB	RA
16	CR	R4	R10	RB	RA
17	CR	R9	R5	RB	RA
18	CR	R11	R3	RB	RA
21	CR	———— R13*————		RB	RA

* Block R13 contained one 50-minute task.

Table 14-8

NAEP 1998 Reading Grade 12 Booklet Configuration

Booklet Number	Common Core Background	Question Block 1	Question Block 2	Reading Background	Motivation
1	CR	R3	R4	RB	RA
2	CR	R4	R5	RB	RA
3	CR	R5	R3	RB	RA
4	CR	R6	R7	RB	RA
5	CR	R7	R8	RB	RA
6	CR	R8	R6	RB	RA
7	CR	R10	R9	RB	RA
8	CR	R9	R11	RB	RA
9	CR	R11	R10	RB	RA
10	CR	R3	R7	RB	RA
11	CR	R8	R4	RB	RA
12	CR	R5	R6	RB	RA
13	CR	R6	R9	RB	RA
14	CR	R7	R11	RB	RA
15	CR	R10	R8	RB	RA
16	CR	R4	R10	RB	RA
17	CR	R9	R5	RB	RA
18	CR	R11	R3	RB	RA
21	CR	———— R13* ————		RB	RA
22	CR	———— R14* ————		RB	RA

* Blocks R13 and R14 contained one 50-minute task each.

Chapter 15

INTRODUCTION TO THE DATA ANALYSIS FOR THE NATIONAL AND STATE READING ASSESSMENTS[1]

Jinming Zhang, Jiahe Qian, and Steven P. Isham
Educational Testing Service

15.1 INTRODUCTION

This chapter introduces the analyses performed on the responses to the cognitive and background items in the 1998 assessment of reading. The results of these analyses are presented in the *NAEP 1998 Reading: A Report Card for the Nation and the States* (Donahue et al., 1999). The emphasis of this chapter is on the description of student samples, items, assessment booklets, administrative procedures, scoring constructed-response items, and student weights, and on the methods and results of DIF analyses. The major analysis components are discussed in Chapter 16 for the national assessment and Chapter 17 for the state assessment.

The objectives of the reading analyses were to:

- prepare scale values and estimate subgroup scale score distributions for national and state samples of students who were administered reading items from the main assessment,
- link the 1998 main focused PBIB samples to the 1994 reading scale,
- perform all analyses necessary to produce a short-term trend report in reading (The reading short-term trend results include the years 1992, 1994 and 1998),
- link the 1998 state assessment scales to the corresponding scales from the 1998 national assessment.

15.2 DESCRIPTION OF STUDENT SAMPLES, ITEMS, ASSESSMENT BOOKLETS, AND ADMINISTRATIVE PROCEDURES

The student samples that were administered reading items in the 1998 assessment are shown in Table 15-1. The data from the national main focused PBIB assessment of reading (4 [Reading–Main], 8 [Reading–Main], and 12 [Reading–Main]) were used for national main analyses comparing the levels of reading achievement for various subgroups of the 1998 target populations. Chapters 1 and 3 contain descriptions of the target populations and the sample design used for the assessment. The target populations were grade 4, grade 8, and grade 12 students in the United States. Unlike previous reading NAEP assessments, only grade-defined cohorts were assessed in the 1998 NAEP. The sampled students in these three cohorts were assessed in the winter (January to March with final makeup sessions held

[1] Jinming Zhang was the primary person responsible for the planning, specification, and coordination of the national reading analyses. Jiahe Qian was the primary person responsible for the planning, specification, and coordination of the state reading analyses. Computing activities for all reading scaling and data analyses were directed by Steven P. Isham and completed by Lois H. Worthington. Others contributing to the analysis of reading data were David S. Freund, Bruce A. Kaplan, and Katharine E. Pashley.

from March 30 to April 3). As described in Chapter 3, the reporting sample for the national reading assessment consisted of students in the S2 sample and the S3 sample, excluding the SD/LEP students.

Table 15-1
NAEP 1998 Reading Student Samples

Sample	Booklet ID Number	Cohort Assessed	Time of Testing*	Reporting Sample Size
4 [Reading–Main]	R1–R16	Grade 4	1/5/98 – 3/27/98	7,672
8 [Reading–Main]	R1–R18, R21	Grade 8	1/5/98 – 3/27/98	11,051
12 [Reading–Main]	R1–R18, R21–R22	Grade 12	1/5/98 – 3/27/98	12,675
4 [Reading–State]	R1–R16	Grade 4	1/5/98 – 3/27/98	112,138
8 [Reading–State]	R1–R18, R21	Grade 8	1/5/98 – 3/27/98	94,429

* Final makeup sessions were held March 30–April 3, 1998.

LEGEND: Main NAEP national main assessment
State NAEP state assessment

The data from the state focused PBIB assessment of reading (4[Reading–State] and 8[Reading–State]) were used for the state analyses. The 1998 state reading assessment included the assessment of both public- and nonpublic-school students for many jurisdictions. The state results reported in the *NAEP 1998 Reading: Report Card for the Nation and the States* (Donahue et al., 1999) are based on public-school students. The state results for both public and nonpublic schools are presented separately in Chapter 17. The procedures used were similar to those of previous state assessments.

The items in the assessment were based on the curriculum framework described in *Reading Framework for the National Assessment of Educational Progress, 1992–1998* (NAGB, 1990). The 1998 reading assessment is based on the same objectives as the 1994 reading assessment. Compared to earlier NAEP assessments, the current assessment contains longer reading passages that are intended to be more authentic examples of the reading tasks encountered in and out of school. As described in the reading framework, these blocks are organized into three subscales, corresponding to three purposes for reading: reading for literary experience, reading to gain information, and reading to perform a task. At grade 4, only the first two purposes are represented. Scales were produced for each of the purposes of reading. In addition, a composite scale for reading was created as a weighted sum of the purposes-for-reading scales (see Table 14-1).

In the main samples, each student was administered a booklet containing either two separately timed 25-minute blocks of cognitive reading items or one 50-minute reading block (in lieu of the two 25-minute blocks). In addition, each student was administered a block of background questions, a block of reading-related background questions, and a block of questions concerning the student's motivation and his or her perception of the difficulty of the cognitive items. The background and motivational blocks were common to all reading booklets for a particular grade level. Eight (grade 4) or nine (grade 8 and grade 12) 25-minute blocks of reading items were administered at each grade level. As described in Chapter 2, the 25-minute blocks were combined into booklets according to a partially balanced incomplete block (PBIB) design. See Chapter 14 for more information about the blocks and booklets. Fifty-minute reading blocks were presented to the older students, one at grade 8 and two at grade 12. The

50-minute blocks were closely examined to ensure the appropriateness of including them with the shorter blocks in the scaling.[2]

For each grade, more than 80 percent of the items in the main assessment were identical to items in the 1994 main assessment. These items occurred in intact blocks, and provided the common information needed to establish the short-term trend. Table 15-2 gives the blocks and numbers of items common across assessment years.

Table 15-2
1998 Reading Blocks and Items Common to the 1992 and 1994 Assessments

Sample	New Blocks	Common Blocks to 1994 (Number of Common Items)	Common Blocks to 1992 and 1994 (Number of Common Items)
4 [Reading–Main] and 4 [Reading–State]	R3	R4, R5, R6, R7, R8, R9, R10; (73)	R4, R5, R6, R7, R10; (55)
8 [Reading–Main] and 8 [Reading–State]	R3, R8	R4, R5, R6, R7, R9, R10, R11, R13*; (90)	R5, R6, R7, R10, R11; (60)
12 [Reading–Main]	R3	R4, R5, R6, R7, R8, R9, R10, R11, R13*, R14*; (111)	R4, R6, R7, R10, R11, R13*; (78)

* 50-minute block

The total number of scaled items was 82, 110, and 118, respectively, for grades 4, 8, and 12. Note that some items overlap across grade. Table 15-3 shows the numbers of items within reading purpose subscales for each grade. The numbers presented in Table 15-3 show item counts both for the original item pool, and after the necessary adjustments were made during scaling (see Section 16.3.2.1).

Table 15-3
Number of Items in Subscales in the Reading Main Assessment, by Reading Purposes

Grade		Literary Experience	Gain Information	Perform a Task	Total
4	Prescaling	41	41	—	82
	Postscaling	41	41	—	82
8	Prescaling	29	48	33	110
	Postscaling	29	48	33	110
12	Prescaling	27	56	36	119
	Postscaling	27	55	36	118

The composition of each block of items by item type is given in Tables 15-4, 15-6, and 15-8. Common labeling of these blocks across grade levels does not necessarily denote common items (e.g., Block R4 at grade 4 does not contain the same items as Block R4 at grade 12). During scaling, some items received specific treatment (for details see Section 16.3). As a result, the composition of each block

[2] These analyses were identical to those described in *Assessing Some of the Properties of Longer Blocks in the 1992 NAEP Reading Assessment* (Donoghue & Mazzeo, 1995). Additional comparisons based on bootstrap comparisons (Donoghue, 1995) further supported the comparability of the 25- and 50-minute reading blocks.

of items by item type might changed. Tables 15-5, 15-7, and 15-9 present the final block composition by item type as defined after scaling.

Table 15-4
1998 NAEP Reading Block Composition by Purpose for Reading and Item Type As Defined Before Scaling, Grade 4

Block	Purpose for Reading	Multiple-Choice Items	Constructed-Response Items			Total Items
			2-category*	3-category	4-category	
Total		36	27	11	8	82
R3	Literary	3	3	2	1	9
R4	Literary	5	6	0	1	12
R5	Literary	7	3	0	1	11
R6	Information	5	4	0	1	10
R7	Information	4	5	0	1	10
R8	Information	3	0	5	1	9
R9	Literary	3	1	4	1	9
R10	Information	6	5	0	1	12

* For a small number of constructed-response items, adjacent categories were combined.

Table 15-5
1998 NAEP Reading Block Composition by Purpose for Reading and Item Type As Defined After Scaling, Grade 4

Block	Purpose for Reading	Multiple-Choice Items	Constructed-Response Items			Total Items
			2-category*	3-category	4-category	
Total		36	27	13	6	82
R3	Literary	3	3	2	1	9
R4	Literary	5	6	1	0	12
R5	Literary	7	3	0	1	11
R6	Information	5	4	0	1	10
R7	Information	4	5	0	1	10
R8	Information	3	0	6	0	9
R9	Literary	3	1	4	1	9
R10	Information	6	5	0	1	12

* For a small number of constructed-response items, adjacent categories were combined.

Table 15-6

1998 NAEP Reading Block Composition by Purpose for Reading and Item Type
As Defined Before Scaling, Grade 8

Block	Purpose for Reading	Multiple-Choice Items	Constructed-Response Items			Total Items
			2-category*	3-category	4-category	
Total		41	32	25	12	110
R3	Literary	3	2	4	1	10
R4	Literary	1	1	5	1	8
R5	Literary	7	3	0	1	11
R6	Information	5	5	0	2	12
R7	Information	6	6	0	1	13
R8	Information	4	1	4	1	10
R9	Task	4	0	5	0	9
R10	Task	4	6	0	2	12
R11	Task	3	8	0	1	12
R13	Information	4	0	7	2	13

* For a small number of constructed-response items, adjacent categories were combined.

Table 15-7

1998 NAEP Reading Block Composition by Purpose for Reading and Item Type
As Defined After Scaling, Grade 8

Block	Purpose for Reading	Multiple-Choice Items	Constructed-Response Items			Total Items
			2-category*	3-category	4-category	
Total		41	35	25	9	110
R3	Literary	3	3	3	1	10
R4	Literary	1	1	5	1	8
R5	Literary	7	3	0	1	11
R6	Information	5	5	0	2	12
R7	Information	6	6	0	1	13
R8	Information	4	1	4	1	10
R9	Task	4	1	4	0	9
R10	Task	4	7	1	0	12
R11	Task	3	8	1	0	12
R13	Information	4	0	7	2	13

* For a small number of constructed-response items, adjacent categories were combined.

Table 15-8

1998 NAEP Reading Block Composition by Purpose for Reading and Item Type
As Defined Before Scaling, Grade 12

Block	Purpose for Reading	Multiple-Choice Items	Constructed-Response Items			Total Items
			2-category*	3-category	4-category	
Total		43	35	28	13	119
R3	Literary	3	2	4	1	10
R4	Literary	3	5	0	1	9
R5	Literary	1	0	6	1	8
R6	Information	5	5	0	2	12
R7	Information	5	6	0	1	12
R8	Information	1	0	6	1	8
R9	Task	4	0	5	0	9
R10	Task	4	6	0	2	12
R11	Task	7	7	0	1	15
R13	Information	10	4	0	2	16
R14	Information	0	0	7	1	8

* For a small number of constructed-response items, adjacent categories were combined.

Table 15-9

1998 NAEP Reading Block Composition by Purpose for Reading and Item Type
As Defined After Scaling, Grade 12

Block	Purpose for Reading	Multiple-Choice Items	Constructed-Response Items			Total Items
			2-category*	3-category	4-category	
Total		43	39	28	8	118
R3	Literary	3	3	3	1	10
R4	Literary	3	5	1	0	9
R5	Literary	1	0	6	1	8
R6	Information	5	5	0	2	12
R7	Information	5	7	0	0	12
R8	Information	1	0	6	1	8
R9	Task	4	1	4	0	9
R10	Task	4	7	1	0	12
R11	Task	7	7	1	0	15
R13	Information	10	4	0	2	16
R14	Information	0	0	6	1	7

* For a small number of constructed-response items, adjacent categories were combined.

To ensure the quality of the administration in the state assessment, the sampling contractor Westat monitored some of the sampled schools. As described in Chapter 5, a randomly selected portion of the administration sessions within each jurisdiction were observed by Westat-trained quality control monitors. Thus, within and across jurisdictions, randomly equivalent samples of students received each block of items under monitored and unmonitored administration conditions. For most jurisdictions the monitored rate was about 25 percent of the schools. Since Kansas was new to the state assessment, 50 percent of the sessions were monitored.

15.3 SCORING CONSTRUCTED-RESPONSE ITEMS

A block consisted of one or two reading passages, each followed by several items. In addition to multiple-choice items, each block contained a number of constructed-response items, accounting for well over half of the testing time. Constructed-response items were scored by specially trained readers (described in Chapter 7). Some of the constructed-response items required only a few sentences or a paragraph response. These short constructed-response items were scored dichotomously as correct or incorrect. Other constructed-response items required somewhat more elaborated responses, and were scored polytomously on a 3-point (0–2) scale:

0 = Unsatisfactory (and omit)
1 = Partial
2 = Complete

In addition, most blocks (except one) contained at least one constructed-response item that required a more in-depth, elaborated response. These items were scored polytomously on a 4-point (0-3) scale:

0 = Unsatisfactory (and omit)
1 = Partial
2 = Essential
3 = Extensive, which demonstrates more in-depth understanding

Originally, the scoring guides for 3-point constructed-response items and 4-point constructed-response items separated the “unsatisfactory” from the “omit” responses, with omits and off-task responses forming a category below the “unsatisfactory” responses (the treatment of items that were not reached is discussed below in Section 16.2.1). During the 1992 scaling process, it was discovered that this scoring rule resulted in unexpectedly poor fit to the IRT model. After much investigation, the 0 category (omitted and off-task responses) was recoded. Off-task responses were treated as “not administered” for each of the items, and omitted responses were combined with the next lowest category, “unsatisfactory.” For new items (administered for the first time in 1998), decisions concerning the treatment of omit and off-task responses were reexamined and found to be appropriate for these new items.

In addition, adjacent categories of a small number of constructed-response items were combined (collapsed). These changes were made so that the scaling model used for these items fit the data more closely, and are described more fully in Section 16.3.2.2. Some of the short-term trend items had been collapsed in the original 1994 scaling. These items were collapsed in an identical manner for the 1998 assessment. New items (unique to 1998) were also examined, and where necessary, adjacent categories were collapsed.

Reliability of constructed-response scoring was calculated within year (1998) and across years (1994 and 1998). Interrater and trend scoring reliability information is provided in Appendix C.

15.4 DIF ANALYSIS

A differential item functioning (DIF) analysis of new items (administered for the first time in 1998) was done to identify potentially biased items that were differentially difficult for members of various subgroups with comparable overall scores. Sample sizes were large enough to compare male and female students, White and Black students, and White and Hispanic students. Appendix A specifies the sample size for each of these groups (see Table A-7). The purpose of these analyses was to identify items that should be examined more closely by a committee of trained test developers and subject-matter specialists for possible bias and consequent exclusion from the assessment. The presence of DIF in an item means that the item is differentially harder for one group of students than another, while controlling for the ability level of the students. DIF analyses were conducted separately by grade for national samples.

A similar DIF analysis was not conducted on the state data, since the results of the national DIF analysis were assumed to hold for the state sample. However, DIF analyses were carried out on 1998 state reading samples at both grade 4 and grade 8 to check items that were not differentially difficult for students between public and nonpublic schools with comparable overall scores. (The nonpublic-school population that was sampled included students from Catholic schools, private religious schools, and private nonreligious schools [all referred to by the term “nonpublic schools”].) Since the participation of nonpublic schools was less than public schools, the data included in the scaling process were only those from public schools. The results of DIF analyses were used to examine the appropriateness of the parameters of IRT models, based on public-school data, for the nonpublic-school data.

For dichotomous items, the Mantel-Haenszel procedure as adapted by Holland and Thayer (1988) was used as a test of DIF (this is described in Chapter 9). The Mantel procedure (Mantel, 1963) as described by Zwick, Donoghue, and Grima (1993) was used for detection of DIF in polytomous items. This procedure assumes that item scores are appropriately treated as ordered categories. SIBTEST (Shealy & Stout, 1993) was also used in the DIF analyses for the first time in NAEP.

For dichotomous items, the DIF index generated by the Mantel-Haenszel procedure is used to place items into one of three categories: “A,” “B,” or “C”. “A” items exhibit little or no evidence of DIF, while “C” items exhibit a strong indication of DIF and should be examined more closely. Positive values of the index indicate items that are differentially easier for the “focal” group (female, Black, or Hispanic students) than for the “reference” group (male or White students). Similarly, negative values indicate items that are differentially harder for the focal group than for the reference group. An item that was classified as a “C” item in *any* analysis was considered to be a “C” item. For details, see Section 9.3.4.

For polytomous items (regular constructed-response items and extended constructed-response items), the Mantel statistic provides a statistical test of the hypothesis of no DIF. A categorization similar to that described for dichotomous items was developed to classify items (this is discussed in detail in Donoghue, 2000). Polytomous items were placed into one of three categories: “AA”, “BB”, or “CC” similar to dichotomous items. “AA” items exhibit no DIF, while “CC” items exhibit a strong indication of DIF and should be examined more closely. The classification criterion for polytomous items is presented in Donoghue (2000). As with dichotomous items, positive values of the index indicate items that are differentially easier for the “focal” group (female, Black, or Hispanic students) than for the reference group (male or White students). Similarly, negative values indicate items that are differentially harder for the focal group than for the reference group. An item that was classified as a “CC” item in *any* analysis was considered to be a “CC” item.

For the national samples, Table 15-10 summarizes the results of DIF analyses for dichotomously scored items in the new blocks. One "C" item as showing significant DIF in favor of male students was identified in grade 8 by the Mantel-Haenszel procedure.

Table 15-10
DIF Category for National Samples by Grade for Dichotomous Items

Grade	DIF Category*	Analysis		
		Male/Female	White/Black	White/Hispanic
4	C-	0	0	0
	B-	0	0	0
	A-	5	4	4
	A+	1	1	1
	B+	0	1	1
	C+	0	0	0
8	C-	1	0	0
	B-	0	0	0
	A-	5	5	6
	A+	4	5	4
	B+	0	0	0
	C+	0	0	0
12	C-	0	0	0
	B-	0	1	0
	A-	5	1	1
	A+	0	2	4
	B+	0	1	0
	C+	0	0	0

* Positive values of the index indicate items that are differentially easier for the focal group (female, Black, or Hispanic students) than for the reference groups (male or White students). "A+" or "A-" means no indication of DIF, "B+" means a weak indication of DIF in favor of the focal group, "B-" means a weak indication of DIF in favor of the reference group, and "C+" or "C-" means a strong indication of DIF.

Table 15-11 summarizes the results of DIF analyses for polytomously scored items. No "CC" item was identified in the new blocks by the Mantel procedure. The only item that SIBTEST flagged as showing significant DIF is *exactly* the "C" item identified by the MH procedure. An independent reviewer examined the "C" item whose DIF statistics indicate that it favors males. The reviewer found no reason for its being biased for or against any group. Therefore, this item was not removed from scaling due to DIF.

In the analysis of DIF between public and nonpublic schools for the state assessment, Table 15-12 summarizes the results for dichotomous items. The focal group consists of students from nonpublic schools. Positive values indicate items that were differentially easier for the focal group. Table 15-13 summarizes the results for polytomous items. As for dichotomous items, the focal group consists of students from nonpublic schools and positive values indicate that the item was differentially easier for the focal group. To aid in interpreting the results for polytomous items, the standardized mean difference between focal and reference groups was produced. This statistic was rescaled by dividing the standardized mean differences by the standard deviation of the respective item. The description of this procedure can be found in Chapter 12. For polytomous items, a standardized mean difference ratio of .25 or greater (coupled with a significant Mantel statistic) was considered a strong indication of DIF. It can

be shown that standardized mean difference ratios of .25 are at least as extreme as Mantel-Haenszel statistics corresponding to "C" items (Donoghue, 1998a).

Table 15-11
DIF Category for National Samples by Grade for Polytomous Items

Grade	DIF Category*	Analysis		
		Male/Female	White/Black	White/Hispanic
4	CC-	0	0	0
	BB-	0	0	0
	AA-	2	2	0
	AA+	1	1	3
	BB+	0	0	0
	CC+	0	0	0
8	CC-	0	0	0
	BB-	0	0	1
	AA-	5	3	2
	AA+	5	6	7
	BB+	0	1	0
	CC+	0	0	0
12	CC-	0	0	0
	BB-	0	0	1
	AA-	2	3	2
	AA+	3	1	2
	BB+	0	1	0
	CC+	0	0	0

* Positive values of the index indicate items that are differentially easier for the focal group (female, Black, or Hispanic students) than for the reference groups (male or White students). "AA+" or "AA-" means no indication of DIF, "BB+" means a weak indication of DIF in favor of the focal group, "BB-" means a weak indication of DIF in favor of the reference group, and "CC+" or "CC-" means a strong indication of DIF.

For the dichotomous items, at grade 4, there were 82 items analyzed from two scales and, at grade 8, there were 110 items from three scales. Table 15-12 gives the number of items in each of six categories (C+, B+, A+, A-, B-, C-) for the comparison. No dichotomous items were classified as "C" items for any of the analyses for both fourth- and eighth-grade state reading assessment data. All the dichotomous items were classified as A+ or A- in the comparisons.

Table 15-12

The Category of DIF between Public and Nonpublic Schools for State Samples, by Grade for Dichotomous Items

Grade	DIF Category*	Analysis Public/Nonpublic
4	C-	0
	B-	0
	A-	33
	A+	30
	B+	0
	C+	0
8	C-	0
	B-	0
	A-	33
	A+	40
	B+	0
	C+	0

* Positive values of the index indicate items that are differentially easier for the focal group (nonpublic) than for the reference groups (public). "A+" or "A-" means no indication of DIF, "B+" means a weak indication of DIF in favor of the focal group, "B-" means a weak indication of DIF in favor of the reference group, and "C+" or "C-" means a strong indication of DIF.

For the polytomous items, there were 19 polytomous from grade 4 and 37 items from grade 8. Table 15-13 is in a format similar to that of Table 15-12, showing items in six categories (CC+, BB+, AA+, AA-, BB-, CC-). All the polytomous items were classified as "AA" for the analyses for both fourth- and eighth-grade state reading assessment data; no polytomous items were classified as "BB" or "CC" items.

Because no DIF items were found in the public and nonpublic comparisons for both fourth- and eighth-grade data, the results of IRT scaling, based on public-school data, were applied to nonpublic-school data.

Table 15-13
The Category of DIF between Public and Nonpublic Schools for State Samples, by Grade for Polytomous Items

Grade	DIF Category*	Analysis Public/Nonpublic
4	CC-	0
	BB-	0
	AA-	9
	AA+	10
	BB+	0
	CC+	0
8	CC-	0
	BB-	0
	AA-	25
	AA+	12
	BB+	0
	CC+	0

* Positive values of the index indicate items that are differentially easier for the focal group (nonpublic) than for the reference groups (public). "AA+" or "AA-" means no indication of DIF, "BB+" means a weak indication of DIF in favor of the focal group, "BB-" means a weak indication of DIF in favor of the reference group, and "CC+" or "CC-" means a strong indication of DIF.

15.5 THE WEIGHT FILES

For the 1998 reading assessments, Westat produced files of final student and school weights and corresponding replicate weights for both national and state samples. Information for the creation of the weight files was supplied by National Computer Systems (NCS) under the direction of Educational Testing Service (ETS). Because both the national and state samples were split into two subsamples, one using the revised inclusion rules for SD/LEP students (S2) and one using the revised inclusion rules and accommodations for SD/LEP students (S3), the weighting process was more complex than in previous assessments. Westat provided student files and school files to ETS for the assessments.

The student weight files contained one record for every student who was not classified as SD or LEP; the weight files contained two records for every student who was classified as SD or LEP. Each record had a full set of weights, including replicate weights. The first set of weights for the SD and LEP students is to be used when estimating results for either S2 or S3 alone. The second set of weights provided for those students is to be used when estimating results for students from both S2 and S3 together. (See Chapters 3 and 10 for more information about the sampling and weighting procedures for the S2 and S3 samples.)

From the student weight files, ETS constructed three sets of student weights, called modular weights, reporting weights, and all-inclusive weights. The modular weights were used when examining S2 and S3 separately, or for comparing S2 to S3. The reporting weights, used for most reports, were used when reporting results for the students in reading who were not classified as being SD or LEP in both S2 and S3 and the students classified as SD or LEP from S2 only. The reporting sample was formed so that unbiased estimation and valid comparisons with previous NAEP assessments could be made. The SD/LEP students were divided into two types, those who were assessed and those who could not be assessed (called excluded students). The all-inclusive weights were used for estimating results for both S2 and S3 together.

The reporting weights were formed from the student weight files by taking the records for students not classified as SD or LEP, the first record in the weight file for students in S2 classified as SD or LEP, and a record containing a missing value code for the students in S3 classified as SD or LEP. In this way, the old inclusion rules used with the students classified as SD or LEP in S3 would not affect the reading results of the 1998 state assessment. For the modular weights, all students approximately from that sample (S2 or S3) not classified as SD or LEP had their final and replicate weights proportionally increased (doubled), while the first record in the weight file for each SD/LEP student from the appropriate sample (S2 or S3) was selected directly from the student weight files. It is important to note that the samples should be separated into the S2 and S3 subsamples when using weights generated in this way. To analyze data from S2 and S3 together, the all-inclusive weights should be used. They were created from the student weight files by taking the records for the students not classified as SD or LEP, and the second records for all students classified as SD or LEP.

For the reporting sample for the state assessments, two other weights were created. These are called "house weights" and "senate weights." As with the respective branches of Congress, these weights represent jurisdictions in two different ways. The house weights weight the student records within a jurisdiction so that the sum of the weights for each jurisdiction is proportional to the fraction of the national in-grade enrollment in that jurisdiction. The senate weights weight the student records within a jurisdiction so that the sums of the weights for each jurisdiction are approximately equal to each other. In other words, a jurisdiction like California, with many eighth-grade students, and a jurisdiction like Rhode Island, with fewer eighth-grade students, would have equal weight when all of the state assessment data are combined. Both of these sets of weights are constructed only for the reporting sample. The reporting sample and either the house or senate weights are used during scaling, conditioning, and all major reporting.

The house weight is the student's reporting weight times a factor, which is the number of public-school students sampled over the sum of the reporting weights of the public-school students in all the jurisdictions. The senate weight is calculated for each jurisdiction separately. Within each jurisdiction a factor, which is 2,500 divided by the sum of the reporting weights of the jurisdiction's public-school students, is computed. (In previous state assessments, 2,000 was used.) The reporting weights for students in both public and nonpublic schools are multiplied by this factor to create the senate weights. For DoDEA/DDESS[3] and DoDEA/DoDDS[4] jurisdictions, all schools were considered public in the calculation of these factors.

Accordingly, there are three sets of weights (modular, reporting, and all-inclusive weights) for the national assessments and, for the state assessments, there are five sets of weights (modular, reporting, house, senate, and all-inclusive weights). Each set of weights has replicate weights associated with it. Replicate weights are used to estimate jackknife standard errors for each statistic estimated.

In addition to student weights, school weights are available for use in school-level analyses. These weights are modular weights for use when examining S2 and S3 separately or for comparing S2 to S3. No other school weights are available. School-level statistics should be calculated on the basis of S2 or S3 subsamples, as opposed to the reporting sample. If school-level statistics are calculated for the reporting sample, biases might occur.

[3] Department of Defense Education Activity /Department of Defense Elementary and Secondary Schools (DoDEA/DDESS) comprise the NAEP jurisdiction for domestic Department of Defense schools.

[4] Department of Defense Education Activity /Department of Defense Dependents Schools (DoDEA/DoDDS) comprise the NAEP jurisdiction for overseas Department of Defense schools.

Chapter 16

DATA ANALYSIS OF THE NATIONAL READING ASSESSMENT[1]

Jinming Zhang, Steven P. Isham, and Lois H. Worthington
Educational Testing Service

16.1 INTRODUCTION

This chapter describes the analyses performed on the responses to the cognitive and background items in the 1998 national assessment of reading. These analyses led to the results presented in Chapters 1 through 4 of the *NAEP 1998 Reading: Report Card for the Nation and the States* (Donahue et al., 1999). The emphasis of this chapter is on the methods and results of procedures used to develop the IRT-based scale scores that formed the basis of these chapters in that report. However, some attention is given to the analysis of constructed-response items as reported in the *NAEP 1998 Reading: Report Card for the Nation and the States*. The theoretical underpinnings of the IRT and plausible values methodology described in this chapter are given in Chapter 12, and several of the statistics are described in Chapter 9.

The major analysis components are discussed in turn. Some aspects of the analysis, such as procedures for item analysis, scoring of constructed-response items, and methods of scaling, are described in previous chapters and are therefore not detailed here. There were five major steps in the analysis of the reading data, each of which is described in a separate section:

1. Conventional item and test analyses (Section 16.2.1)
2. Item response theory (IRT) scaling (Section 16.3)
3. Estimation of national and subgroup scale score distributions based on the "plausible values" methodology (Section 16.4)
4. Transformation of the purposes-for-reading scales to the 1994 scale score metric (Section 16.5)
5. Creation of the reading composite scale (Section 16.5.2)

Section 16.6 describes the results of partitioning the error variance; 16.7 discusses the matching of student responses to those of their teachers.

16.2 NATIONAL ITEM ANALYSES

16.2.1 Conventional Item and Test Analyses

This section contains a detailed description of the conventional item analysis performed on the national reading data. This analysis was done within block so that a student's score is the sum of item scores in a block. In forming the block total score, dichotomous items (multiple-choice and 2-category constructed-response items) were scored as right or wrong; polytomous items were not scored as right or wrong but were scored with three or more categories reflecting several degrees of knowledge.

[1] Jinming Zhang was the primary person responsible for the planning, specification, and coordination of the national reading analyses. Computing activities for all reading scaling and data analyses were directed by Steven P. Isham and completed by Lois H. Worthington. Others contributing to the analysis of reading data were David S. Freund, Bruce A. Kaplan, Norma A. Norris, and Katharine E. Pashley.

Tables 16-1, 16-2, and 16-3 show the number of items in the block, the average weighted item score, average weighted polyserial correlation, and the weighted alpha reliability for each block administered. These statistics are described in Chapter 9. These values were calculated for the items within each block used in the scaling process. The tables also give the number of students who were administered the block and the percentage of students not reaching the last item in the block. These numbers include only those students who contributed to the summary statistics provided in the *NAEP 1998 Reading: Report Card for the Nation and the States*, Chapter 1 through Chapter 4. Student weights were used for all statistics, except for the sample sizes. The results for the blocks administered to each grade level indicate that the blocks differ in number of items, average difficulty, reliability, and percent not reaching the last item, and so are not parallel to each other. Preliminary item analyses for all items within a block were completed before scaling; however, the results shown here indicate the characteristics of the items that contributed to the final scale, and reflect decisions made in scaling to combine adjacent categories (collapse) for a small number of items.

As described in Chapter 12, in NAEP analyses (both conventional and IRT-based), a distinction is made between missing responses at the end of each block (not reached) and missing responses prior to the last observed response (omitted). Items that were not reached were treated as if they had not been presented to the examinee, while omitted items were regarded as incorrect. The proportion of students attempting the last item of a block (or, equivalently, one minus the proportion not reaching the last item) is often used as an index of the degree of speededness of the block of items.

Standard practice at ETS is to treat all nonrespondents to the last item as if they had not reached the item. For multiple-choice items, short constructed-response items, and regular constructed-response items (3-category), this convention produced a reasonable pattern of results, in that the proportion reaching the last item does not differ markedly from the proportion attempting the next-to-last item. However, for the blocks that ended with extended constructed-response items (4-category), this convention resulted in an implausibly large drop in the number of students attempting the final item. Therefore, for blocks that ended with an extended constructed-response item, students who attempted the next-to-last item but did not respond to the last item were classified as having intentionally omitted that item. Therefore, this item was regarded as incorrect.

The results in Tables 16-1 to 16-3 indicate that the difficulty and internal consistency of the blocks varied. Such variability is expected, because the blocks were not constructed to be parallel. Based on the proportion of students attempting the last item, all of the blocks appear to be somewhat speeded. This effect is larger for grade 4 than for the other grades.

Small but consistent differences were noted based on whether a block appeared first or second within a booklet. When the block appeared first in the booklet, the average item score tended to be higher and the average polyserial correlation tended to be lower. The largest differences were noted in the proportion of students not attempting the last item in the block; more students attempted the last item when the block appeared in the second position. It appears that students learned to pace themselves through the second block, based on their experience with the first block. Recall that the design of the reading assessment is not completely balanced. Thus, when these serial position effects were first noticed, it was feared that they might adversely affect the results of the IRT scaling. As part of the analysis of the 1992 reading assessment, a special study was completed to examine the effects of the serial position differences. The serial position effects were found to have minimal results on the scaling, most likely due to the balance of the partial BIB design of the booklets. The effects portrayed in Tables 16-1 through 16-3 are similar in size to the effects observed in the 1992 reading assessment, and were therefore unlikely to produce adverse effects on the final IRT scaling.

Table 16-1

Descriptive Statistics for Item Blocks by Position Within Test Booklet and Overall Occurrences for the National Main Reading Sample, Grade 4, As Defined After Scaling

Statistic	Position	R3	R4	R5	R6	R7	R8	R9	R10
Number of Scaled Items		9	12	11	10	10	9	9	12
Unweighted Sample Size	First	952	949	960	961	942	962	964	927
	Second	971	945	929	959	933	944	942	977
	Both	1,923	1,894	1,889	1,920	1,875	1,906	1,906	1,904
Weighted Average Item Score	First	.49	.64	.48	.59	.45	.52	.62	.66
	Second	.47	.63	.43	.57	.41	.49	.61	.63
	Both	.48	.64	.45	.58	.43	.51	.61	.64
Weighted Average R-Polyserial	First	.64	.68	.63	.60	.68	.63	.62	.65
	Second	.65	.68	.63	.62	.69	.65	.67	.65
	Both	.64	.68	.63	.61	.68	.64	.64	.65
Weighted Alpha Reliability	First	.69	.80	.76	.71	.74	.72	.76	.78
	Second	.69	.79	.73	.71	.74	.74	.76	.76
	Both	.69	.80	.75	.71	.74	.73	.76	.77
Weighted Proportion of Students Attempting Last Item	First	.67	.61	.76	.72	.60	.71	.65	.79
	Second	.82	.73	.82	.84	.75	.79	.82	.89
	Both	.75	.67	.79	.78	.67	.75	.74	.84

Table 16-2

Descriptive Statistics for Item Blocks by Position Within Test Booklet and Overall Occurrences for the National Main Reading Sample, Grade 8, As Defined After Scaling

Statistic	Position	R3	R4	R5	R6	R7	R8	R9	R10	R11	R13*
Number of Scaled Items		10	8	11	12	13	10	9	12	12	13
Unweighted Sample Size	First	986	968	1,035	1,034	996	1,016	989	1,016	977	—
	Second	999	1,006	1,000	994	1,004	991	1,037	961	999	—
	Both	1,985	1,974	2,035	2,028	2,000	2,007	2,026	1,977	1,976	2,012
Weighted Average Item Score	First	.43	.45	.67	.57	.69	.49	.61	.61	.69	—
	Second	.41	.41	.67	.54	.66	.47	.60	.59	.68	—
	Both	.42	.43	.67	.55	.68	.48	.61	.60	.68	.66
Weighted Average R-Polyserial	First	.68	.61	.73	.65	.70	.59	.69	.61	.72	—
	Second	.69	.64	.70	.64	.72	65	.69	.62	.74	—
	Both	.68	.63	.71	.65	.71	.62	.69	.62	73	.60
Weighted Alpha Reliability	First	.76	.67	.77	.72	.79	.66	.70	.73	.81	—
	Second	.76	.71	.75	.72	.80	.74	.73	.71	.81	—
	Both	.76	.70	.76	.72	.79	.70	.72	.72	.81	.73
Weighted Proportion of Students Attempting Last Item	First	.79	.65	.94	.85	.85	.84	.94	.79	.84	—
	Second	.83	.72	.95	.87	.87	.89	.94	.86	.89	—
	Both	.81	.68	.95	.86	.86	.87	.94	.82	.86	.95

* A 50-minute block that comprised an entire booklet.

Table 16-3

Descriptive Statistics for Item Blocks by Position Within Test Booklet and Overall Occurrences for the National Main Reading Sample, Grade 12, As Defined After Scaling

Statistic	Position	R3	R4	R5	R6	R7	R8	R9	R10	R11	R13*	R14*
Number of Scaled Items		10	9	8	12	12	8	9	12	15	16	7
Unweighted Sample Size	First	967	943	940	965	993	949	965	997	989	—	—
	Second	961	940	949	949	918	973	986	953	965	—	—
	Both	1,928	1,883	1,889	1,914	1,911	1,922	1,951	1,950	1,954	1,923	1,968
Weighted Average Item Score	First	.58	.54	.46	.68	.52	.59	.75	.72	.55	—	—
	Second	.56	.51	.43	.67	.52	.56	.74	.71	.53	—	—
	Both	.57	.52	.44	.68	.52	.58	.75	.72	.54	.64	.42
Weighted Average R-Polyserial	First	.69	.67	.63	.66	.54	.61	.73	.63	.55	—	—
	Second	.70	.69	.66	.70	.59	.63	.76	.66	.60	—	—
	Both	.70	.68	.64	.68	.57	.62	.74	.64	.57	.63	.66
Weighted Alpha Reliability	First	.76	.66	.69	.66	.54	.69	.66	.71	.66	—	—
	Second	.78	.67	.72	.69	.62	.70	.72	.73	.73	—	—
	Both	.77	.66	.71	.67	.58	.70	.69	.72	.70	.79	.66
Weighted Proportion of Students Attempting Last Item	First	.86	.65	.81	.92	.79	.87	.96	.82	.85	—	—
	Second	.90	.74	.83	.91	.86	.91	.95	.89	.83	—	—
	Both	.88	.70	.81	.91	.82	.89	.96	.85	.84	.92	.95

* A 50-minute block that comprised an entire booklet.

16.2.2 Scoring the Constructed-Response Items

As indicated earlier, the reading assessment included constructed-response items. Responses to these items were included in the scaling process. In addition, detailed analyses of the constructed-response items were also conducted, and are summarized in the *NAEP 1998 Reading: Report Card for the Nation and the States*. Chapter 7 provides the ranges for percent agreement between raters for the items as they were originally scored. The percent agreement for the raters and Cohen's (1968) Kappa are given in Appendix C.

16.3 NATIONAL IRT SCALING

16.3.1 Overview of Item Parameter Estimation

In 1992, separate IRT-based scales were developed for each of the purposes for reading identified in the reading framework. As described in Chapter 12, multiple-choice items were fit using a 3PL model. Short constructed-response items were fit using a 2PL model. Regular and extended constructed-response items were fit using a generalized partial-credit model.

For calibration, all items that were not reached were treated as if they had not been presented to the examinees.[2] Recall that responses to regular and extended constructed-response items that were off-task were also treated as if they had not been presented. The treatment of omitted responses differed according to the item type. Omitted responses to multiple-choice items were treated as fractionally correct (see Chapter 9 and Mislevy & Wu, 1988, for a discussion of these conversions). Omitted responses to short constructed-response items were treated as incorrect, and omitted responses to regular and extended constructed-response items were assigned to the lowest category.

For each purpose of reading, three separate scalings, one for each grade sample, were conducted. The analyses were conducted on the following samples:

- The 1998 grade 4 national main sample with the 1994 grade 4 only national sample
- The 1998 grade 8 national main sample with the 1994 grade 8 only national sample
- The 1998 grade 12 national main sample with the 1994 grade 12 only national sample

That is, item parameters were estimated using combined data from both assessment years. Items that were administered for more than one assessment (trend items) were constrained to have equal item response functions across assessment years. However, some items exhibited clear evidence of functioning differently across assessment years (see discussion in Section 16.3.2.3). These items were treated as separate items for each assessment year.

The calibration was performed using all the available examinees in the reporting sample. Student sampling weights were used for the analysis. For scaling, sampling weights were restandardized to ensure that each assessment year had a similar sum of weights, and so had approximately equal influence in the calibration. Each assessment year's data were treated as a sample from a separate subpopulation. Thus, separate scale score distributions were estimated for each assessment year.

Item responses were calibrated using the BILOG/PARSCALE program. Starting values were computed from item statistics based on the entire data set. BILOG/PARSCALE calibrations were done in

[2] An exception to this rule was the treatment of extended constructed-response items at the end of the block. See Section 16.2.1 for a discussion.

two stages. At stage one, the scale score distribution of each assessment year was constrained to be normally distributed, although the means and variances differed across assessments. The values of the item parameters from this normal solution were then used as starting values for a second-stage estimation run in which the scale score distribution (modeled as a separate multinomial distribution for each assessment) was estimated concurrently with item parameters. Calibration was concluded when changes in item parameter estimates became negligibly small.

A complexity introduced by the 50-minute blocks in reading is that those blocks of items must be linked in some way to the shorter blocks. This is complicated by the fact that no students received the shorter blocks in addition to the 50-minute blocks. Because the samples of students receiving each booklet are representative of the population as a whole, it was assumed that the distribution of student scale score was the same for the students receiving the 50-minute blocks as for the students receiving the booklets containing the shorter blocks.

16.3.2 Evaluation of Model Fit

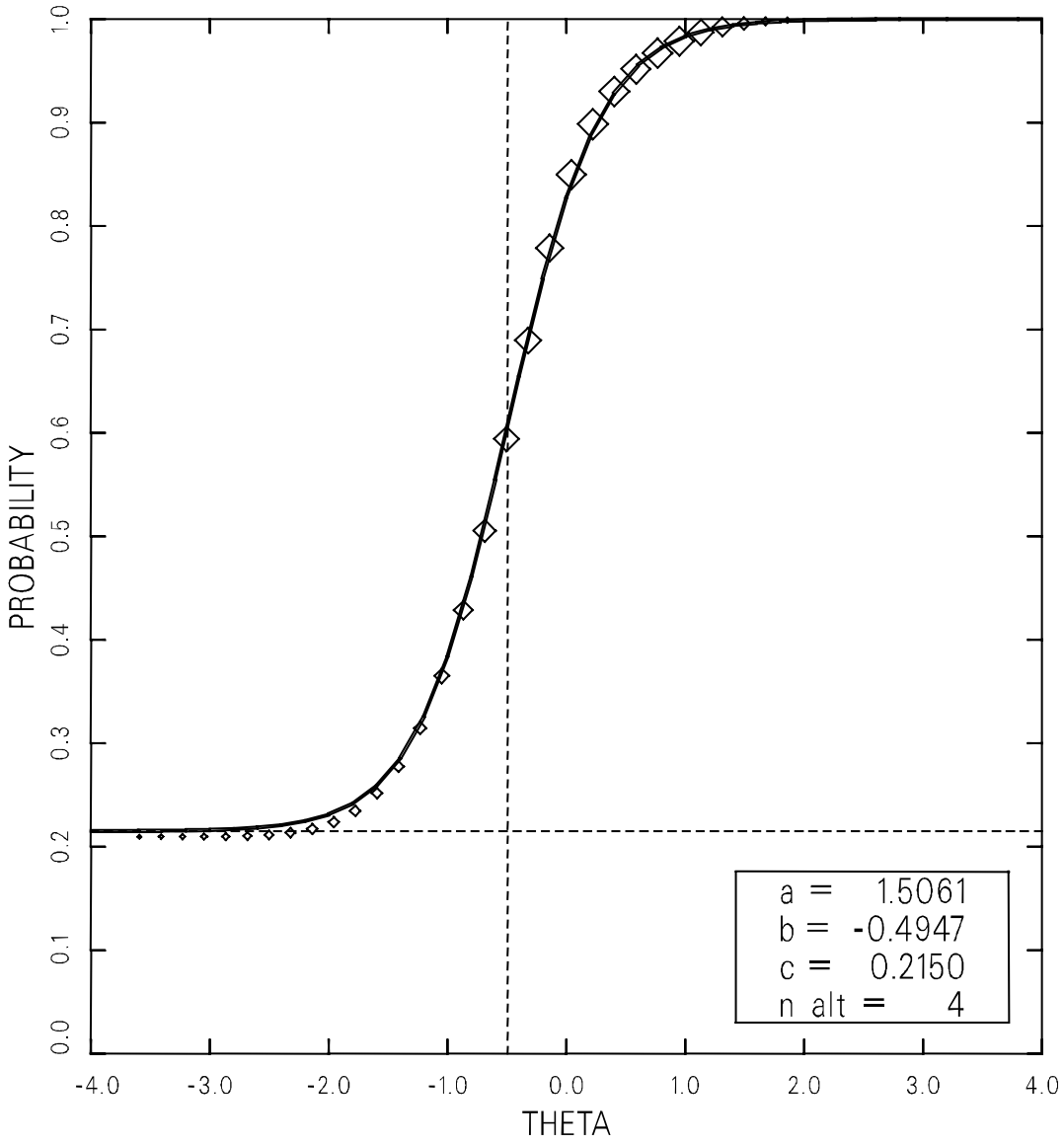
During and subsequent to item parameter estimation, evaluations of the fit of the IRT models were carried out for each of the items. These evaluations were based primarily on graphical analysis. First, model fit was evaluated by examining plots of nonmodel-based estimates of the expected proportion correct (conditional on scale score) versus the proportion correct predicted by the estimated item response function (see Chapter 12 and Mislevy & Sheehan, 1987, p. 302). Figure 16-1 gives an example plot of a multiple-choice item that demonstrates good model fit, R017002, from the Reading for Literary Experience scale at grade 4. For regular and extended constructed-response items, similar plots were produced for each item category response function (see Chapter 12). Figure 16-2 gives an example plot of a regular constructed-response item that demonstrates good model fit, R017104, from the Reading for Literary Experience scale at grade 8. Items that did not fit the model received some treatment (e.g., recoding), or were excluded from the final scales (see the next three subsections for details). Note that the remaining item plots in this section (Figures 16-3 through 16-7) were obtained from preliminary item parameter calibrations. They are presented to reflect the information used to make the decisions discussed in the text. Plots produced from the final item parameters (listed in Appendix E) were very similar to those presented and supported the decisions made.

16.3.2.1 Items Deleted from the Final Scale

In making decisions about excluding items from the final scales, a balance was sought between being too stringent, hence, deleting too many items and possibly damaging the content representativeness of the pool of scaled items, and being too lenient, hence including items with model fit poor enough to endanger the types of model-based inferences made from NAEP results. For the majority of the items, the model fit was extremely good. Items that clearly did not fit the model were not included in the final scales; however, a certain degree of misfit was tolerated for a number of items included in the final scales.

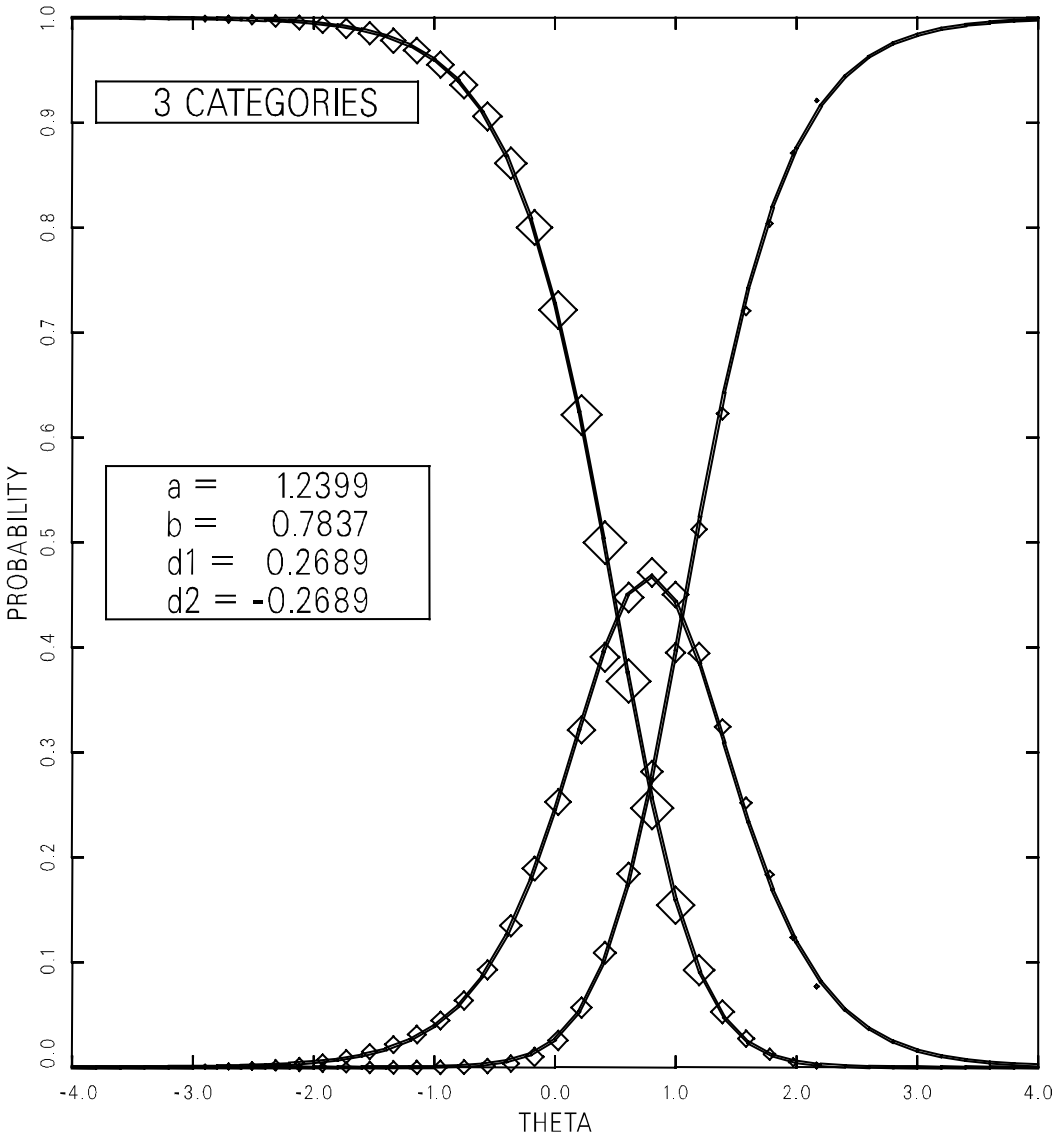
At grade 12, one item from the Reading to Gain Information scale, R016603, was dropped from the final scales due to poor fit to the IRT model in the 1994 reading assessment (See Chapter 12, *The NAEP 1994 Technical Report*, Allen, Kline, & Zelenak, 1997). In the 1998 data analysis, this item was reused to check whether it fitted a model or not, using the 1998 data. Figure 16-3 gives an IRT plot of this item. Category 1 provides virtually no discrimination; the empirical item category response function is essentially flat. Thus, the item was also deleted from the final scales in this analysis. As shown in Table 16-4, this is the only item that was deleted from the final scales in the 1998 reading national data analysis.

Figure 16-1
*Dichotomous Item (R017002) Exhibiting Good Model Fit**



** Diamonds represent 1998 grade 4 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item response function (IRF) assuming a logistic form.*

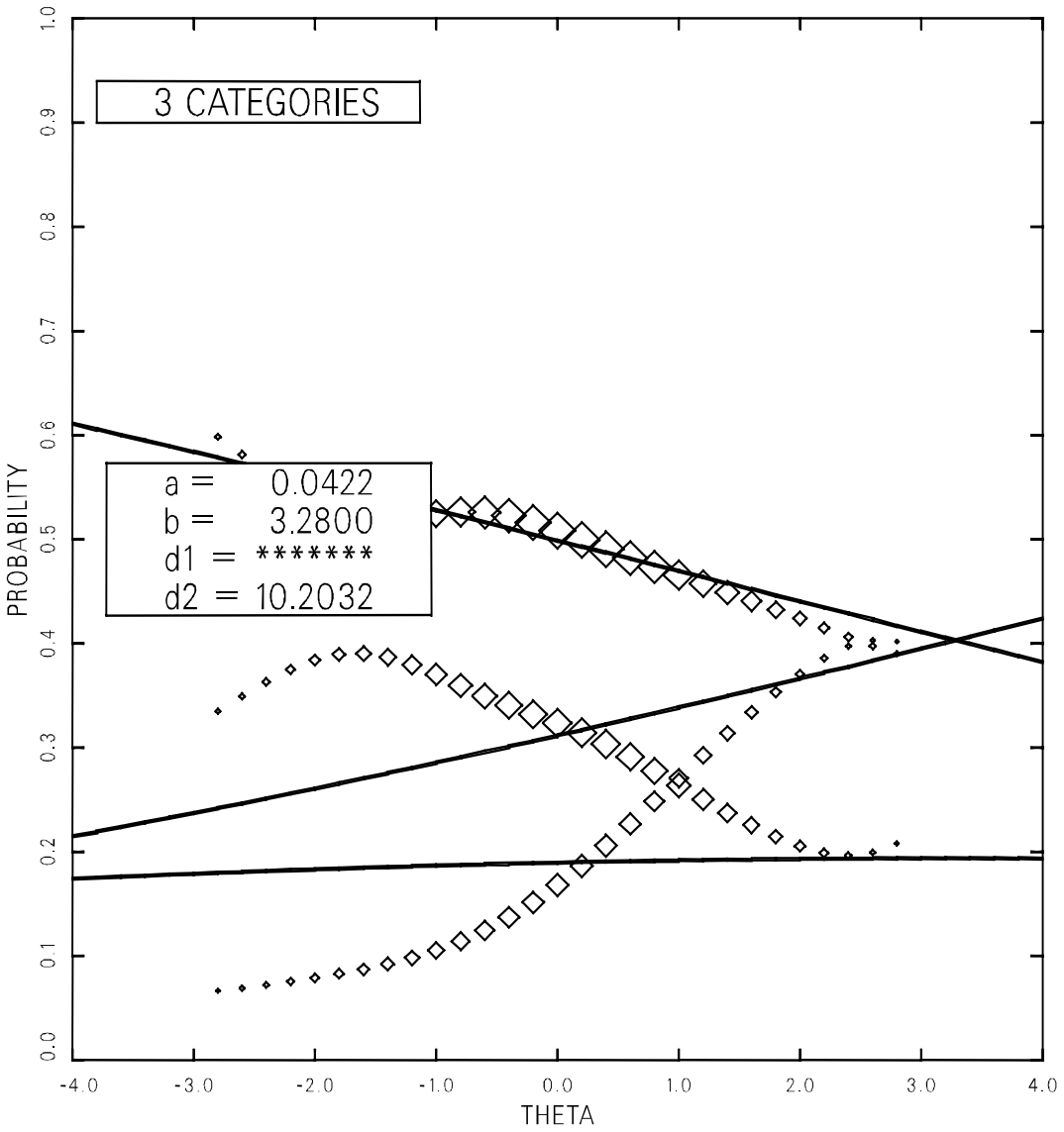
Figure 16-2
*Polytomous Item (R017104) Exhibiting Good Model Fit**



** Diamonds represent 1998 grade 8 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

Figure 16-3

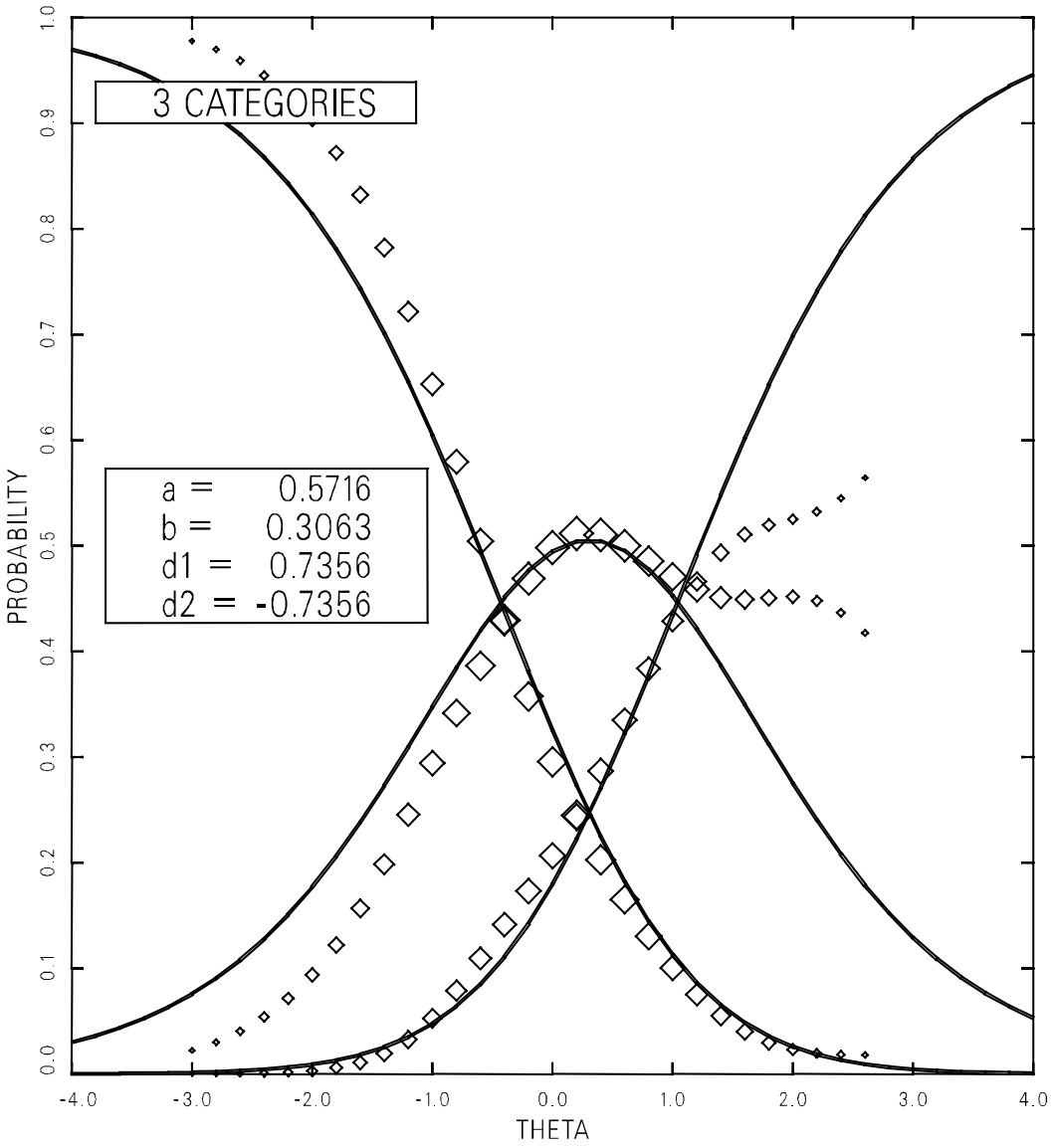
*Polytomous Item (R016603) Exhibiting Unacceptably Poor Model Fit**



** Diamonds represent 1998 grade 12 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

Figure 16-4

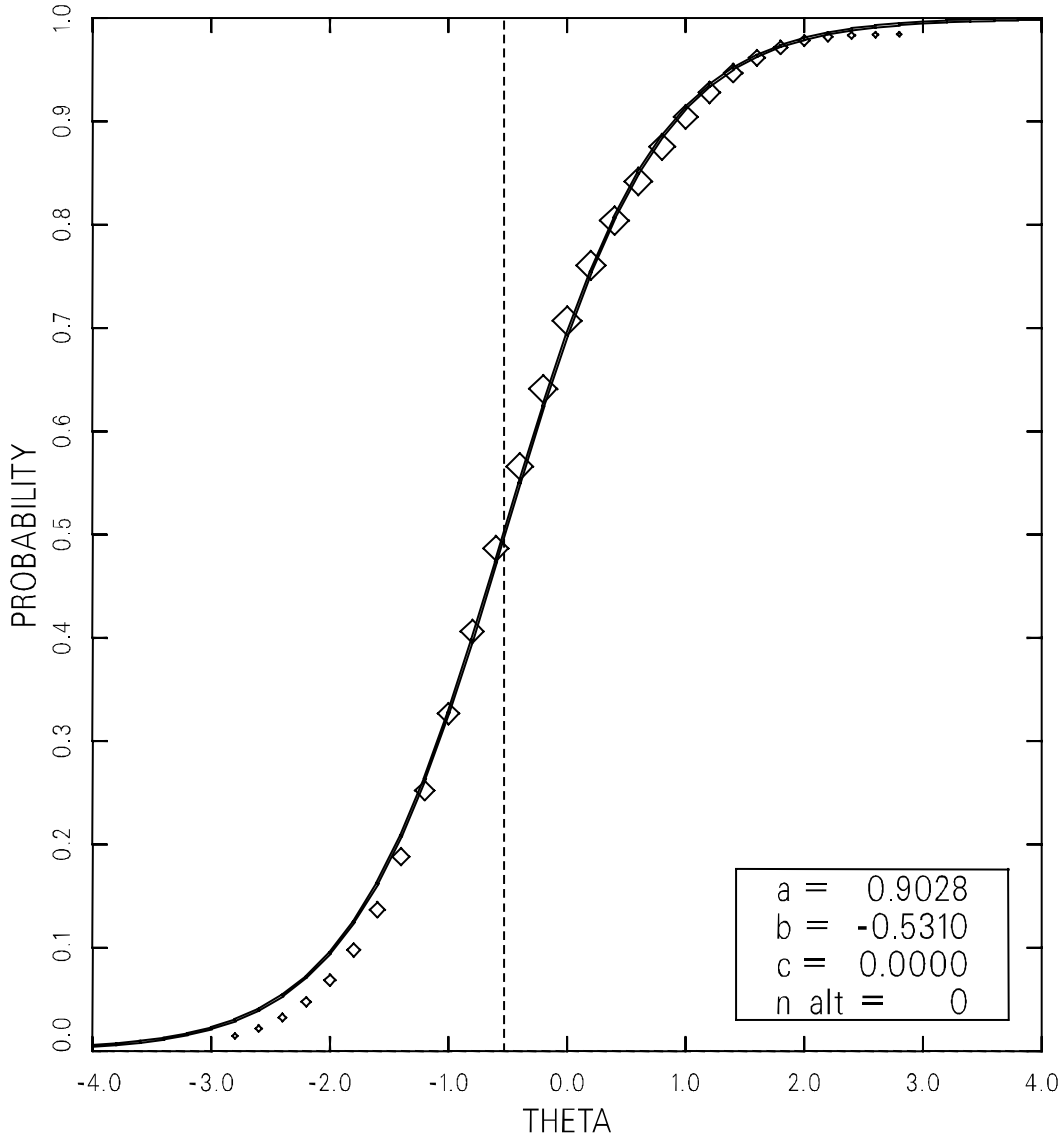
*Polytomous Item (R017110) Exhibiting Poor Model Fit**



** Diamonds represent 1998 grade 12 reading assessment data They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

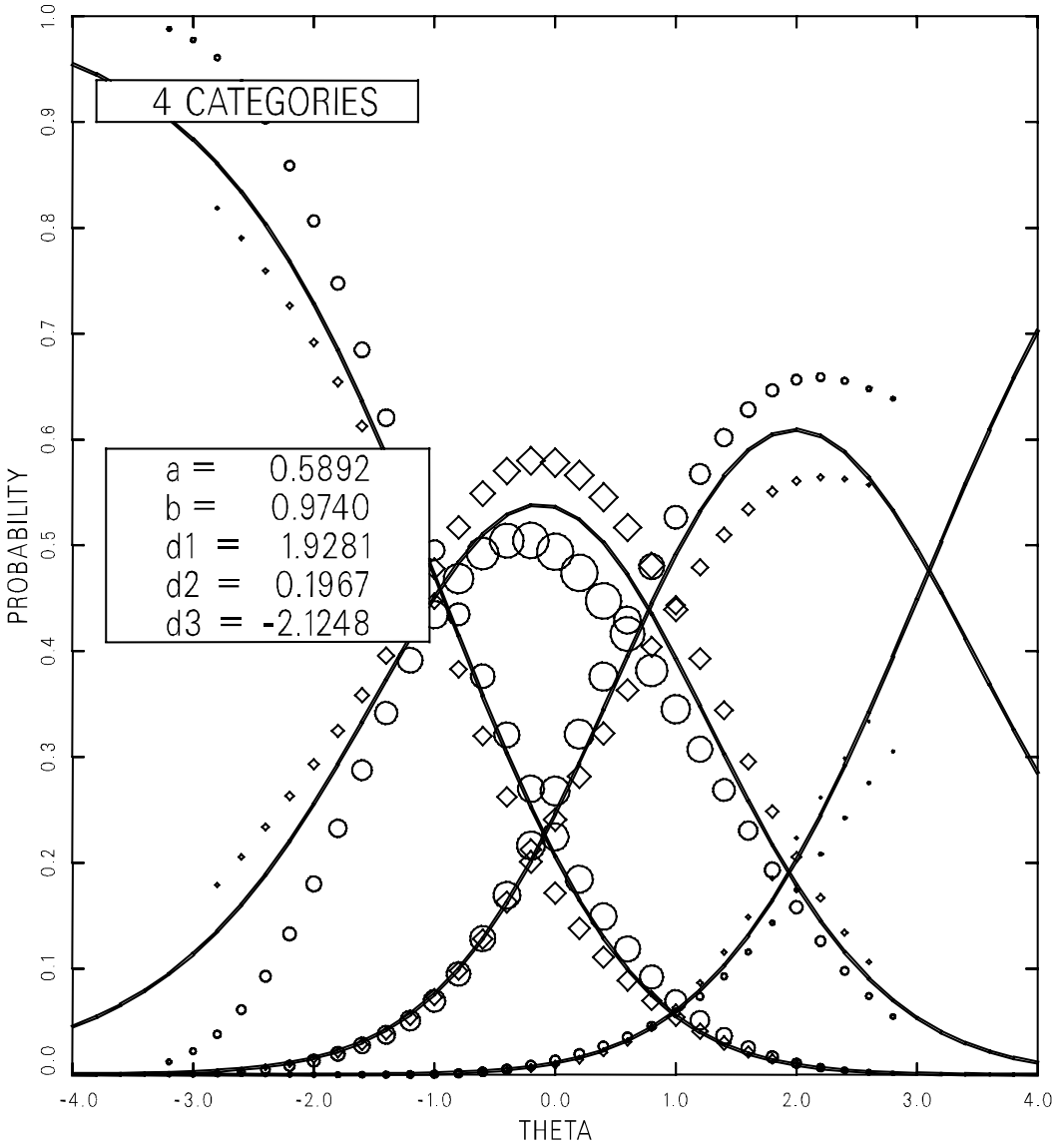
Figure 16-5

*Dichotomous Item (R017110) After Collapsing Categories 1 and 2**



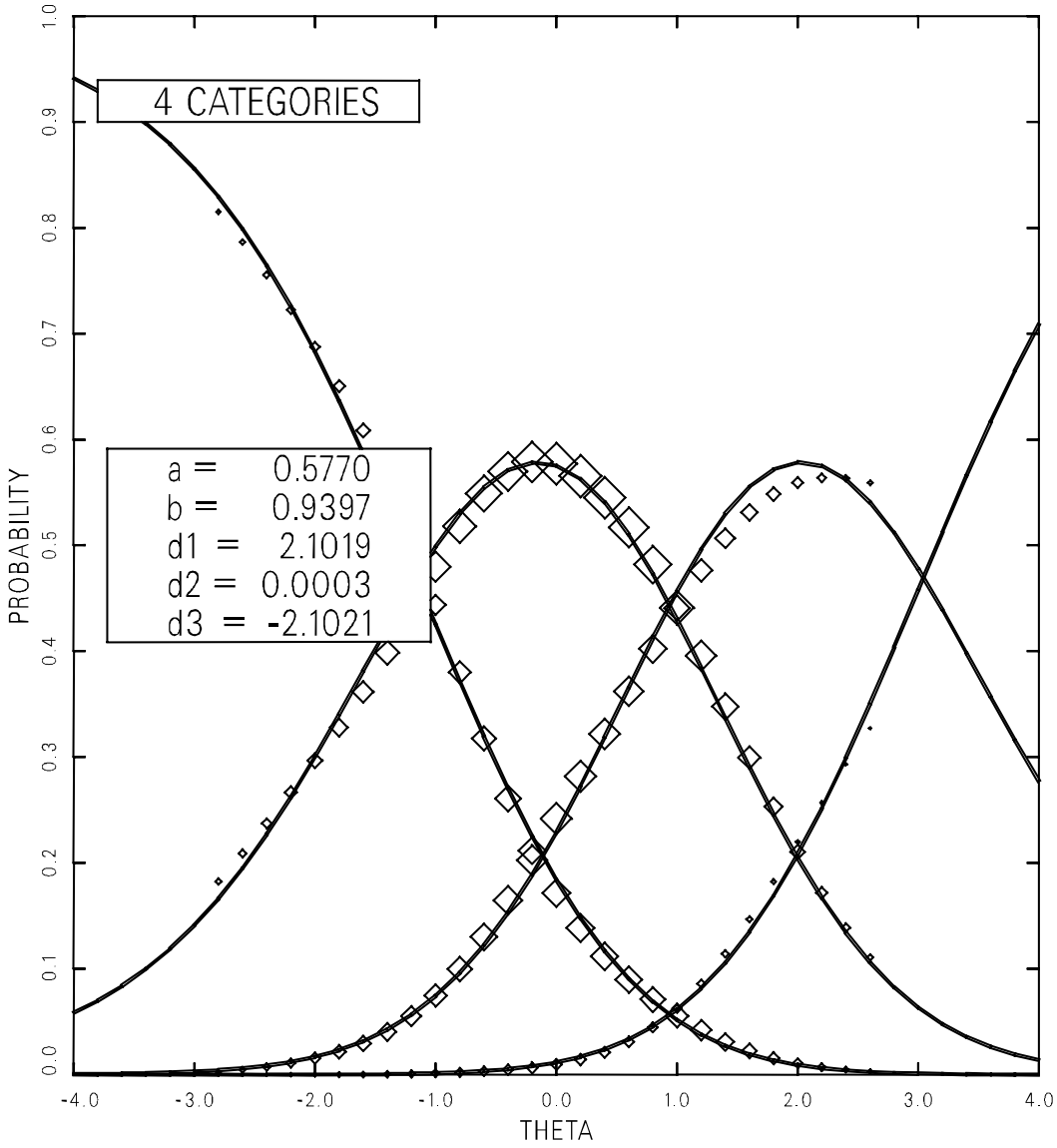
** Diamonds represent 1998 grade 12 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item response function (IRF) assuming a logistic form.*

Figure 16-6
Short-Term Trend Polytomous Item (R016210)
*Demonstrating Differential Item Functioning Across Assessment Years 1994 and 1998**



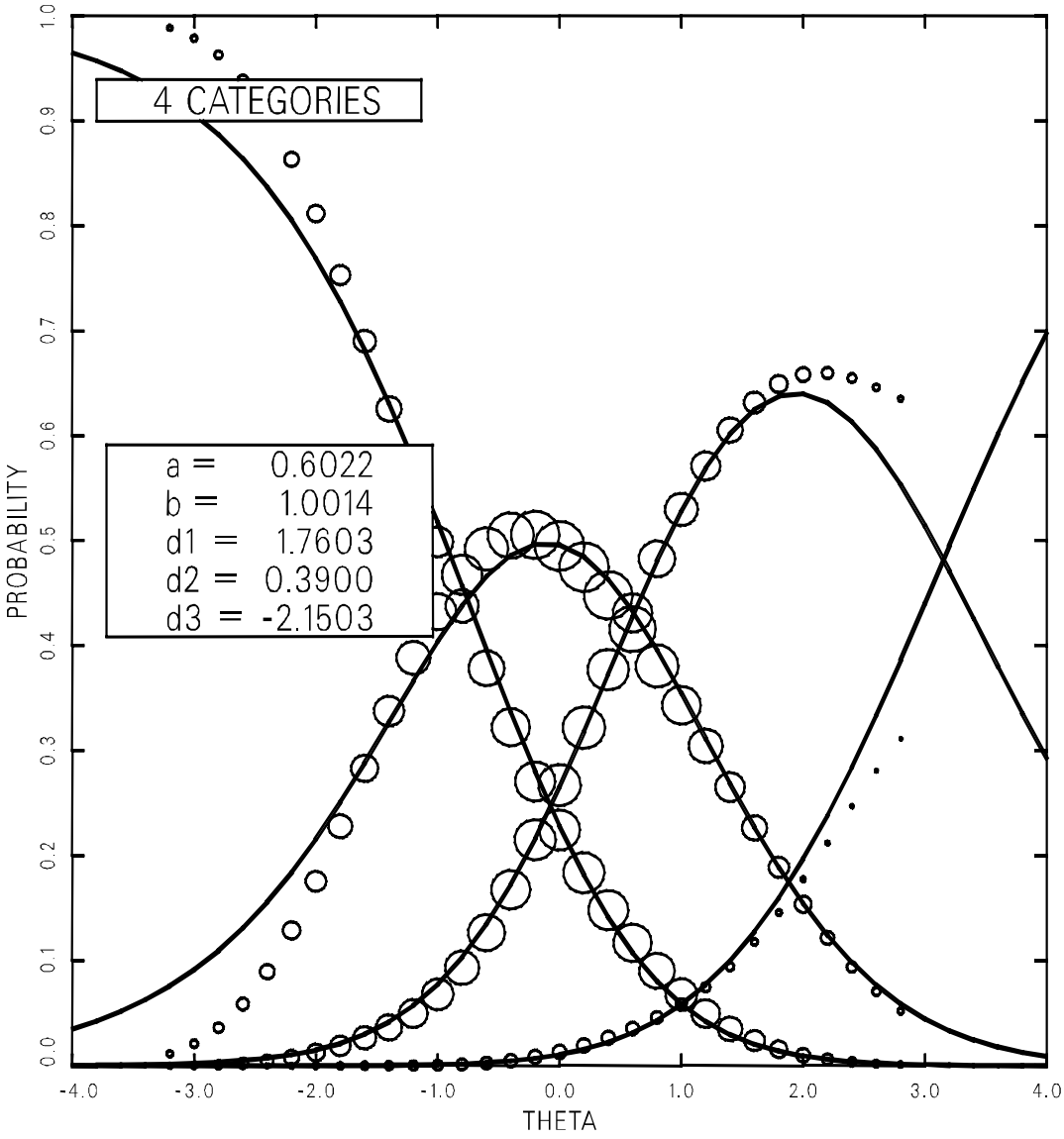
** Diamonds represent 1998 grade 8 reading assessment data; circles represent 1994 grade 8 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

Figure 16-7a
Short-Term Trend Polytomous Item (R016210)
*Fitting Separate Item Response Functions for Each Assessment Year**



** Diamonds represent 1998 grade 8 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

Figure 16-7b
Short-Term Trend Polytomous Item (R016210)
*Fitting Separate Item Response Functions for Each Assessment Year**



** Circles represent 1994 grade 8 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

Table 16-4
Items Deleted from the Final Scaling

Scale	NAEP ID	Block	Grade Affected	Reason for Decision
Reading to Gain Information	R016603	R14	12	Poor fit in 1994 and 1998

16.3.2.2 Recoded Polytomous Items

Polytomous items received special treatment (i.e., recoding) for one of two reasons. First, some of the short-term trend items were recoded in the original 1994 scaling. These items were recoded again for the 1998 assessment. Second, two of the new (unique to 1998) polytomous items received this treatment in the scaling. Figure 16-4 shows one such item, R017110, from the Reading for Literary Experience scale at grade 12.

There is a lack of fit for both the unsatisfactory and partial categories for low scale score ($\theta < -1.0$) values. There is also a marked misfit for categories 1 and 2 in high scale score ($\theta > 1.0$) values. Categories 1 and 2 of this item were collapsed:

0 = Unsatisfactory
1 = Partial
2 = Complete

Figure 16-5 shows the recoded version of R017110 from the final scaling. The fit is substantially improved.

Table 16-5 lists polytomous items that were recoded for scaling in 1998.

Table 16-5
Recoding of Polytomous Items for Scaling

Scale	NAEP ID	Block	Grade(s) Affected	Reason for Decision	Disposition
Reading for Literary Experience	R012111	R4	4	Recoded in 1992 and 1994	Combine categories 0 + 1
	R013506	R4	12	Recoded in 1992 and 1994	Combine categories 0 + 1
	R017110	R3	8, 12	Poor fit in 1998	Combine categories 1 + 2 (dichotomize)
Reading to Gain Information	R015707	R8	4	Recoded in 1994	Combine categories 2 + 3
	R013706	R7	12	Recoded in 1992 and 1994	Combine categories 0 + 1, 2 + 3 (dichotomize)
Reading to Perform a Task	R013004	R11	8	Recoded in 1992 and 1994	Combine categories 0 + 1
	R013403	R10	8, 12	Recoded in 1992 and 1994	Combine categories 0 + 1
	R013406	R10	8, 12	Recoded in 1992 and 1994	Combine categories 0 + 1, 2 + 3 (dichotomize)
	R013915	R11	12	Poor fit in 1998	Combine categories 0 + 1
	R016104	R9	8, 12	Recoded in 1994	Combine categories 1 + 2 (dichotomize)

16.3.2.3 Item Category Response Functions (ICRFs) Common Across Assessment Years

The adequacy of the assumption of a common item (category) response function across assessment years was also evaluated. For dichotomous items, this was evaluated by comparing the nonmodel-based expected proportions for each assessment year to the single, model-based item response function fit by BILOG/PARSCALE. For polytomously scored items, similar plots were produced for each item category response function (ICRF, see Chapter 12). Plots showing each assessment year's data separately and the common item (category) response function were then examined. Items that showed clear evidence of functioning differently across assessments were treated as separate items for each assessment year. As was the case with deleting items, in making decisions about scaling items separately by assessment year, a balance was sought between being too stringent, hence, splitting too many items and possibly damaging the common item link between the assessment years, and being too lenient, hence, including items with model fit poor enough to endanger the model-based trend inferences.

For each short-term trend constructed-response item, a sample of approximately 600–1,000 of the 1994 responses was rescored in 1998. Most items showed an acceptably high level of exact agreement. However, several items showed a clear trend in the disagreements. Special attention was paid to these items in the process of scaling.

Figure 16-6 gives an example plot for an item that was split early in the process, R016210 at grade 8. The circles represent data from the 1994 assessment, and the diamonds represent the data from the 1998 assessment. There is a marked separation between the two sets of symbols that indicate that the item functioned substantially differently across assessment years.

Figures 16-7a and 16-7b show the result of splitting this item. Figure 16-7a gives the ICRF fit using only the 1998 data, and Figure 16-7b gives the ICRF fit to the 1994 data. Within each assessment year, there is good or acceptable agreement between the curve and the plotted points.

At each grade, several items were calibrated separately for each assessment year, because these items functioned differently across assessment years according to item plots. In addition, these items are constructed-response items that either have relatively low rater agreement across assessment years (as revealed in rescoring) or have relatively low rater reliabilities in the 1998 scoring. Tables 16-6 through 16-8 list the short-term trend items that were calibrated separately across assessment years. A list of the items scaled for each of the grades, along with their final item parameter estimates, appears in Appendix E.

Table 16-6
Grade 4 Items Scaled Separately by Assessment Years

Scale	Block	NAEP ID	Type
Reading for Literary Experience	R9	R015802	Short constructed-response
		R015803	Regular constructed-response
		R015807	Regular constructed-response
Reading to Gain Information	R8	R015702	Regular constructed-response

Table 16-7
Grade 8 Items Scaled Separately by Assessment Years

Scale	Block	NAEP ID	Type
Reading for Literary Experience	R5	R012607	Extended constructed-response
		R012611	Short constructed-response
Reading to Gain Information	R6	R013212	Extended constructed-response
	R7	R012711	Short constructed-response
	R13	R016210	Extended constructed-response
Reading to Perform a Task	R11	R013004	Extended constructed-response

Table 16-8
Grade 12 Items Scaled Separately by Assessment Years

Scale	Block	NAEP ID	Type
Reading for Literary Experience	R5	R016301	Regular constructed-response
		R016302	Regular constructed-response
		R016305	Regular constructed-response
Reading to Gain Information	R6	R013207	Short constructed-response
		R013211	Short constructed-response
	R7	R013704	Short constructed-response
	R8	R016401	Regular constructed-response
		R016402	Regular constructed-response
		R016405	Regular constructed-response
	R13	R015514	Extended constructed-response
	R14	R016602	Regular constructed-response
Reading to Perform a Task	R11	R013913	Short constructed-response

16.4 GENERATION OF PLAUSIBLE VALUES

Multivariate plausible values were generated for each grade group separately using the CGROUP program. Final student weights were used in this analysis. Reporting plans required analyses that examined the relationships between proficiencies and a large number of background variables. The background variables included student demographic characteristics (e.g., race/ethnicity of the student, highest level of education attained by parents), students' perceptions about reading, student behavior both in and out of school (e.g., amount of television watched daily, amount of homework done each day), and a variety of other aspects of the educational, social, and financial environment of the schools they attended. For grade 4 and grade 8, information was also collected from students' teachers concerning teachers' background, education, and instructional practices in the classroom (see Section 3.4.9).

To avoid bias in reporting results and to minimize biases in secondary analyses, it was desirable to incorporate a large number of independent variables in the conditioning model. When expressed in terms of contrast-coded main effects and interactions, the number of variables to be included totaled

1,081 for grade 4, 1,059 for age grade 8, and 568 for grade 12. The much larger numbers for grade 4 and grade 8 reflect the number of contrasts from the teacher questionnaires.

Some of these contrasts involved relatively small numbers of individuals and some were highly correlated with other contrasts or sets of contrasts. Given the large number of contrasts, an effort was made to reduce the dimensionality of the predictor variables. Consistent with what was done for the 1994 reading assessment, the original background variable contrasts were standardized and transformed into a set of linearly independent variables by extracting separate sets of principal components at each grade level. The principal components, rather than the original variables, were used as the independent variables in the conditioning model. The number of principal components was the number required to account for at least 90 percent of the variance in the original contrast variables. Research based on data from the 1990 trial state assessment in mathematics suggests that results obtained using such a subset of components will differ only slightly from those obtained using the full set (Mazzeo, Johnson, Bowker, & Fong, 1992). Table 16-9 contains a list of the number of principal components included in conditioning, as well as the proportion of variance accounted for by the conditioning model for each grade.

Table 16-9
Proportion of Scale Score Variance Accounted for by the Conditioning Model for the National Main Reading Assessment

			Proportion of Scale Score Variance		
Grade	**Number of Conditioning Contrasts***	**Number of Principal Components***	**Reading for Literary Experience**	**Reading to Gain Information**	**Reading to Perform a Task**
4	1,081	381	.600	.610	NA
8	1,059	380	.599	.608	.662
12	568	235	.600	.565	.589

* Excluding the constant term

For each grade, Table 16-10 provides an estimated residual variance for each purpose-for-reading scale and the residual correlation matrix between the reading scales. The values, taken directly from the output of the CGROUP program, are estimates of relationships between the subscales conditional on the set of principal components included in the conditioning model. The marginal correlations between the purpose-for-reading scales are presented in Table 16-11.

Table 16-10
Conditional Correlations and Variances from Conditioning (CGROUP)

Grade	Scale	Reading for Literary Experience	Reading to Gain Information	Reading to Perform a Task
4	Reading for Literary Experience	1.000	—	NA
	Reading to Gain Information	0.853	1.000	NA
	Residual Variance	0.327	0.337	NA
8	Reading for Literary Experience	1.000	—	—
	Reading to Gain Information	0.863	1.000	—
	Reading to Perform a Task	0.827	0.868	1.000
	Residual Variance	0.353	0.357	0.341
12	Reading for Literary Experience	1.000	—	—
	Reading to Gain Information	0.807	1.000	—
	Reading to Perform a Task	0.688	0.758	1.000
	Residual Variance	0.404	0.428	0.393

Table 16-11
*Marginal Correlations of Reading Scales**

Grade	Scale	Reading for Literary Experience	Reading to Gain Information	Reading to Perform a Task
4	Reading for Literary Experience	1.000	—	NA
	Reading to Gain Information	0.851	1.000	NA
8	Reading for Literary Experience	1.000	—	—
	Reading to Gain Information	0.858	1.000	—
	Reading to Perform a Task	0.837	0.866	1.000
12	Reading for Literary Experience	1.000	—	—
	Reading to Gain Information	0.861	1.000	—
	Reading to Perform a Task	0.797	0.827	1.000

* Tabled values were obtained by computing a separate Pearson correlation coefficient for each plausible value, computing Fisher's *z*-transformation for each value, computing the average of the transformed values, and computing the inverse transformation of the average.

16.5 THE FINAL READING SCALES

16.5.1 Purpose-for-Reading Scales

The linear indeterminacy of the reading scale was resolved by linking the 1998 reading short-term trend scales to previous scales. For each grade, the item parameters from the joint calibration based on data from 1994 and 1998 were used with the 1994 data to find plausible values for the 1994 data. The mean and standard deviation of all of the plausible values were calculated and matched to the mean and standard deviation of all of the plausible values based on the original analysis of the 1994 data, as given

in earlier reports. This linking was performed separately for each of the purpose-for-reading scales using the transformation:

$$\theta_{scale\ score} = A \bullet \theta_{calibrated} + B$$

where $\theta_{scale\ score}$ denotes values on the final transformed scale and $\theta_{calibrated}$ denotes values on the original calibration scale from BILOG/PARSCALE. The constants for the linear transformation for each scale are given in Table 16-12.

Table 16-12
Coefficients of Linear Transformations of the Purpose-for-Reading Scales from the Calibrating Scale Units to the Units of the Reporting Scale

Grade	Scale	A	B
4	Reading for Literary Experience	43.17	217.25
	Reading to Gain Information	42.23	213.71
8	Reading for Literary Experience	36.27	260.82
	Reading to Gain Information	38.05	261.17
	Reading to Perform a Task	41.37	262.68
12	Reading for Literary Experience	48.04	285.44
	Reading to Gain Information	33.81	291.87
	Reading to Perform a Task	39.65	286.17

16.5.2 The Composite Reading Scale

For the national assessment, a composite scale was created as an overall measure of reading proficiency. The composite was a weighted average of plausible values on the purpose-for-reading scales (Reading for Literary Experience, Reading to Gain Information, and, at grade 8 and grade 12, Reading to Perform a Task). The weights for the scales were proportional to the importance assigned to each reading purpose contained in the assessment specifications given in the *Reading Framework*. The percentages of assessed time are given in Table 16-13. Weights for each reading purpose are similar to the actual proportion of assessment time devoted to that purpose. In developing the composite scale, the weights were applied to the plausible values for each reading purpose as expressed in terms of the final scale (i.e., after transformation from the provisional θ scales). Overall summary statistics for the composite scale are given in Tables 16-14.

Table 16-13
Weighting of the Purpose-for-Reading Scales on the Reading Composite Scale

Grade	Reading for Literary Experience	Reading to Gain Information	Reading to Perform a Task
4	55%	45%	Not assessed
8	40%	40%	20%
12	35%	45%	20%

Table 16-14
*Means and Standard Deviations on the Reading Composite Scale**

Grade	Year	Mean	S. D.
4	1998	217.32	37.61
	1994	214.26	40.58
	1992	216.74	35.57
8	1998	263.63	34.65
	1994	259.64	36.75
	1992	260.04	35.89
12	1998	290.79	37.63
	1994	287.35	36.66
	1992	292.15	32.81

* Tabled values were computed separately for each plausible value. The mean is the mean of the individual means. The standard deviation is computed as the square root of the average of the individual variances.

16.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE

For each grade, the variance of the final, transformed scale mean was partitioned into two parts. This analysis yielded estimates of the proportion of error variance due to sampling students and the proportion due to the latent nature of θ. These estimates are given in Table 16-15 for each purpose-for-reading scale and the composite scale (for stability, the estimates of the between-imputation variance B in Equation 12.12 are based on 100 plausible values). Additional results, including those by gender and race/ethnicity, are presented in Appendix H.

Table 16-15
Estimation Error Variance and Related Coefficients for the National Main Reading Assessment

Grade	Scale	Total Estimation Error Variance	Proportion of Variance Due to ...	
			Student Sampling	Latency of θ
4	Reading for Literary Experience	0.72	0.84	0.16
	Reading to Gain Information	0.88	0.85	0.15
	Composite	0.64	0.89	0.11
8	Reading for Literary Experience	0.75	0.85	0.15
	Reading to Gain Information	0.77	0.91	0.09
	Reading to Perform a Task	0.89	0.87	0.13
	Composite	0.62	0.93	0.07
12	Reading for Literary Experience	1.07	0.79	0.21
	Reading to Gain Information	0.44	0.80	0.20
	Reading to Perform a Task	0.62	0.75	0.25
	Composite	0.51	0.88	0.12

16.7 READING TEACHER QUESTIONNAIRES

Teachers of fourth- and eighth-grade students were surveyed about their educational background and teaching practices. Each student's records were matched first with his or her reading teacher, and then with the specific classroom period. Variables derived from the questionnaire were used in the conditioning models. An additional conditioning variable was included that indicated whether the student had been matched with a teacher record. This contrast controlled estimates of subgroup means for differences that exist between matched and nonmatched students. Of the 7,672 fourth-grade students in the sample, 6,741 (88%, unweighted) were matched with teachers who answered both parts of the teacher questionnaire, and 334 (4%, unweighted) of the students had teachers who answered only the teacher background section of the questionnaire. For the eighth-grade sample, 8,935 of the 11,051 students (81%, unweighted) were matched to both sections of the teacher questionnaire. An additional 935 students (8%, unweighted) were matched with the first part of the teacher questionnaire, but could not be matched to the appropriate classroom period. Thus, 92 percent of the fourth-graders and 89 percent of the eighth-graders were matched with at least the background information about their reading teacher.

Chapter 17

DATA ANALYSIS OF THE STATE READING ASSESSMENT[1]

Jiahe Qian, Steven P. Isham, Lois H. Worthington, and Jo-Lin Liang
Educational Testing Service

17.1 INTRODUCTION

This chapter describes the analyses used in developing the reading scales for the 1998 state assessment of reading that was carried out at grades 4 and 8. The procedures used were similar to those employed in the analysis of the 1992 and 1994 state assessments in reading (Allen, Mazzeo, Ip, Swinton, Isham, & Worthington, 1995; Allen, Mazzeo, Isham, Fong, & Bowker, 1994) and are based on the philosophical and theoretical rationale given in the previous chapter. For 1998, the NAEP reading assessment framework incorporated a balance of knowledge and skills based on current reform reports, exemplary curriculum guides, and research on the teaching and learning of reading. The 1998 state assessment included the assessment of both public- and nonpublic-school students for most jurisdictions. The NAEP report card for state assessments only presents average scale scores and achievement-level results for public-school students selected using the 1996 inclusion rules and provided no accommodations. The inclusion rules used are discussed in more detail in Section 1.1.

There were five major steps in the analysis of the state assessment reading data, each of which is described in a separate section:

- Conventional item and test analyses (Section 17.2)
- Item response theory (IRT) scaling (Section 17.3)
- Estimation of state and subgroup scale score distributions based on the "plausible values" methodology (Section 17.4)
- Linking of the 1998 state assessment scales to the corresponding scales from the 1998 national assessment (Section 17.5)
- Creation of the state assessment reading composite scale (Section 17.5)

For the context of the assessment instruments and administration procedures of the reading assessments, see Chapters 5 and 14.

17.2 STATE ITEM AND TEST ANALYSES

For grades 4 and 8, Tables 17-1 through 17-4 contain summary statistics for each block of items for public- and nonpublic-school sessions, respectively. (The nonpublic-school population that was sampled included students from Catholic schools, private religious schools, and private nonreligious schools [all referred to by the term "nonpublic schools"].) Block-level statistics are provided both overall and by serial position of the block within booklet. To produce the tables for grade 4, data from all 44

[1] Jiahe Qian was the primary person responsible for the planning, specification, and coordination of the state reading analyses. Computing activities for all reading scaling and data analyses were directed by Steven P. Isham and completed by Lois H. Worthington. Others contributing to the analysis of reading data were David S. Freund, Bruce A. Kaplan, Jo-Lin Liang, and Katharine E. Pashley.

jurisdictions were aggregated and statistics were calculated using rescaled versions of the final (reporting sample) sampling weights provided by Westat. The same processes employed the data from all 41 jurisdictions in the grade 8 assessment. The senate weights were used in item analysis and scaling procedure (see Section 15.5). Use of the senate weights does nothing to alter the value of statistics calculated separately within each jurisdiction. However, for statistics obtained from samples that combine students from different jurisdictions, use of the senate weights results in a roughly equal contribution of each jurisdiction's data to the final value of the estimate. As discussed in Mazzeo (1991), equal contribution of each jurisdiction's data to the results of the IRT scaling was viewed as a desirable outcome and the same rescaled weights were only adjusted slightly in carrying out the scaling. Hence, the item analysis statistics for each grade shown in Tables 17-1 through 17-4 are approximately consistent with the weighting used in scaling.

Table 17-1

Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall Public Schools, Grade 4*

Statistic	Position	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted Sample Size	First	12,349	12,296	12,136	12,233	12,272	12,440	12,307	12,335
	Second	12,414	12,390	12,158	12,265	12,228	12,227	12,224	12,283
	Both	24,763	24,686	24,294	24,498	24,500	24,667	24,531	24,618
Average Item Score	First	.49	.65	.46	.59	.43	.53	.62	.67
	Second	.47	.63	.44	.56	.42	.50	.60	.64
	Both	.48	.64	.45	.58	.42	.51	.61	.65
Weighted Alpha Reliability	First	.68	.79	.73	.71	.73	.71	.75	.78
	Second	.70	.80	.73	.70	.74	.73	.75	.77
	Both	.69	.79	.72	.70	.73	.72	.75	.77
Average R-Polyserial	First	.63	.67	.61	.60	.67	.61	.60	.65
	Second	.66	.70	.63	.62	.70	.64	.65	.67
	Both	.65	.68	.62	.61	.68	.63	.62	.66
Proportion of Students Attempting Last Item	First	.70	.60	.71	.67	.59	.69	.63	.79
	Second	.82	.74	.84	.84	.74	.82	.78	.88
	Both	.76	.67	.78	.75	.66	.75	.71	.85

* The number and types of items contained in each block are shown in Table 15-4.

Tables 17-1 through 17-4 show the number of students assigned each block of items, the average item score, the weighted alpha reliability, the average polyserial correlation, and the proportion of students attempting the last item in the block for each grade. The average item score for the block is the average, over items, of the score means for each of the individual items in the block. For binary-scored multiple-choice and constructed-response items, these score means correspond to the proportion of students who correctly answered each item. For the extended constructed-response items, the score means were calculated as item score mean divided by the maximum number of points possible.

In NAEP analyses (both conventional and IRT-based), a distinction is made between missing responses at the end of each block (i.e., missing responses subsequent to the last item the student answered) and missing responses prior to the last observed response. Missing responses before the last observed response are considered intentional omissions. Intentional omissions were considered "omitted" and were treated as incorrect responses. In calculating the average score for each item, only students classified as having been presented the item were included in the denominator of the statistic. Missing responses at the end of the block are considered "not-reached," and treated as if they had not been

presented to the student. The proportion of students attempting the last item of a block (or, equivalently, one minus the proportion of students not reaching the last item) is often used as an index of the degree of speededness associated with the administration of that block of items. Mislevy and Wu (1988) discussed these conversions.

Table 17-2

Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall Nonpublic Schools, Grade 4*

Statistic	Position	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted Sample Size	First	942	945	950	958	973	974	946	969
	Second	965	954	941	951	965	968	944	957
	Both	1,907	1,899	1,891	1,909	1,938	1,942	1,890	1,926
Average Item Score	First	.57	.73	.53	.67	.52	.59	.68	.74
	Second	.56	.71	.54	.64	.52	.58	.66	.72
	Both	.56	.72	.53	.66	.52	.58	.67	.73
Weighted Alpha Reliability	First	.57	.69	.72	.65	.71	.64	.70	.69
	Second	.62	.69	.69	.64	.72	.67	.67	.72
	Both	.59	.69	.70	.64	.71	.65	.68	.70
Average R-Polyserial	First	.57	.63	.60	.56	.65	.57	.54	.60
	Second	.60	.64	.61	.59	.67	.61	.61	.66
	Both	.59	.64	.60	.57	.66	.59	.58	.63
Proportion of Students Attempting Last Item	First	.81	.70	.80	.78	.66	.77	.73	.89
	Second	.88	.83	.92	.90	.83	.88	.86	.92
	Both	.84	.77	.86	.84	.74	.82	.80	.90

* The number and types of items contained in each block are shown in Table 15-4.

The average polyserial correlation is the average, over items, of the item-level polyserial correlations (*r*-biserial for dichotomous items) between the item and the number-correct block score. For each item-level *r*-polyserial, total block number-correct score (including the item in question, and with students receiving zero points for all not-reached items) was used as the criterion variable for the correlation. The number-correct score was the sum of the item scores where correct dichotomous items are assigned 1 and correct polytomous (or multiple-category) items are assigned the score category for the response. Data from students classified as not reaching the item were omitted from the calculation of the statistic. As is evident from Tables 17-1 through 17-4, the difficulty and the average item-to-total correlations of the blocks varied somewhat for each grade. Such variability was expected, since these blocks were not created to be parallel in either difficulty or content. In general, the proportion of nonpublic-school students reaching the last item in blocks was higher. For public-school students, only 67 percent of the fourth-graders and 69 percent of the eighth-graders receiving block R4 reached the last item in the block. For nonpublic-school students, 77 percent of fourth-graders and 82 percent of eighth-graders receiving block R4 reached the last item in the block.

Table 17-3
*Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall**
Public Schools, Grade 8

Statistic	Position	R3	R4	R5	R6	R7	R8	R9	R10	R11
Unweighted Sample Size	First	7,781	7,882	7,836	7,741	7,792	7,683	7,850	7,760	7,917
	Second	7,864	7,586	7,788	7,942	7,796	7,860	7,638	7,833	7,726
	Both	15,645	15,468	15,624	15,683	15,588	15,543	15,488	15,593	15,643
Average Item Score	First	.42	.44	.68	.57	.70	.49	.61	.60	.68
	Second	.40	.42	.66	.55	.67	.47	.60	.61	.67
	Both	.41	.43	.67	.56	.69	.48	.60	.60	.68
Weighted Alpha Reliability	First	.77	.67	.74	.68	.77	.66	.69	.70	.79
	Second	.77	.70	.77	.71	.79	.69	.70	.72	.79
	Both	.77	.69	.75	.70	.78	.68	.70	.71	.79
Average R-Polyserial	First	.69	.61	.69	.61	.70	.59	.68	.59	.70
	Second	.70	.64	.72	.64	.71	.61	.68	.61	.71
	Both	.70	.63	.71	.63	.70	.60	.68	.60	.70
Proportion of Students Attempting Last Item	First	.79	.67	.95	.86	.83	.85	.95	.77	.81
	Second	.85	.72	.95	.86	.88	.90	.95	.84	.90
	Both	.82	.69	.95	.86	.85	.88	.95	.81	.86

* The number and types of items contained in each block are shown in Table 15-6.
Block R13 did not appear with any other cognitive block, so no information on positions is available.

These tables also indicate that there was little variability in average item scores or average polyserial correlations for each block by serial position within the assessment booklet. The differences in item statistics were small for items appearing in blocks in the first position and in the second position. However, differences were consistent in their direction. Average item scores were almost always highest when each block was presented in the first position. Average polyserial correlations were usually higher when each block was presented in the second position. An aspect of block-level performance that did differ noticeably by block position was the proportion of students attempting the last item in the block. As shown in Tables 17-1 through 17-4, the percentage of the students attempting the last item increased in the second block position. Students may have learned to pace themselves through the later block after they had experienced the format of the first block they received. This was similar to what occurred in the previous state reading assessments. For the 1992 state assessment, a study was completed to examine the effect of the block position differences on scaling. Due to the partial BIB design of the booklets, those effects were minimal.

As mentioned earlier, in an attempt to maintain rigorous standardized administration procedures across the jurisdictions, a randomly selected 50 percent of all sessions within each jurisdiction that had never participated in a state assessment were observed by a Westat-trained quality control monitor. In the 1998 state reading assessment, Kansas was the only new participant, and 50 percent of those sessions were monitored. A randomly selected 25 percent of the sessions within other jurisdictions were monitored. Observations from the monitored sessions provided information about the quality of administration procedures and the frequency of departures from standardized procedures in the monitored sessions (see Chapter 5 for a discussion of the substance of these observations).

Table 17-4
Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall Nonpublic Schools, Grade 8*

Statistic	Position	R3	R4	R5	R6	R7	R8	R9	R10	R11
Unweighted Sample Size	First	482	491	466	461	482	458	479	483	484
	Second	473	471	486	493	483	468	463	479	459
	Both	955	962	952	954	965	926	942	962	943
Average Item Score	First	.51	.50	.75	.65	.80	.57	.72	.69	.80
	Second	.50	.50	.76	.64	.79	.55	.71	.70	.79
	Both	.51	.50	.75	.65	.79	.56	.71	.70	.79
Weighted Alpha Reliability	First	.71	.60	.75	.58	.65	.55	.62	.63	.71
	Second	.75	.60	.68	.55	.71	.59	.62	.60	.63
	Both	.73	.60	.72	.56	.68	.58	.62	.62	.67
Average R-Polyserial	First	.64	.59	.74	.56	.68	.55	.64	.55	.66
	Second	.68	.58	.70	.55	.73	.57	.65	.54	.66
	Both	.66	.58	.72	.55	.70	.56	.65	.54	.66
Proportion of Students Attempting Last Item	First	.83	.78	.96	.94	.92	.91	.97	.80	.90
	Second	.89	.85	.98	.94	.96	.94	.96	.88	.92
	Both	.86	.82	.97	.94	.94	.92	.96	.84	.91

* The number and types of items contained in each block are shown in Table15-6.
Block R13 did not appear with any other cognitive block, so no information on positions is available.

Tables 17-5 through 17-8 provide the block-level descriptive statistics for the monitored and unmonitored sessions. When results were aggregated over all participating jurisdictions, there was little difference between the performance of students who attended monitored or unmonitored sessions. When data were classified by school type, there was also little difference between the performance of students who attended monitored or unmonitored sessions. For grade 4, the average item score over all 8 blocks and over all 44 participating jurisdictions was 0.54 for both monitored and unmonitored public-school sessions. The average item score was 0.62 for monitored nonpublic-school sessions and 0.62 for unmonitored nonpublic-school sessions. For grade 8, the average item score over all 10 blocks and over all 41 participating jurisdictions was 0.577 and 0.582 for monitored and unmonitored public-school sessions, respectively. The average item score was 0.67 for both monitored and unmonitored nonpublic-school sessions.

Table 17-5

Block-Level Descriptive Statistics for Monitored and Unmonitored Public-School Sessions, Grade 4*

Statistic	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted Sample Size								
Unmonitored	18,540	18,473	18,159	18,322	18,359	18,500	18,325	18,386
Monitored	6,223	6,213	6,135	6,176	6,141	6,167	6,206	6,232
Average Item Score								
Unmonitored	.48	.64	.45	.58	.42	.51	.61	.66
Monitored	.48	.64	.45	.57	.42	.51	.61	.65
Weighted Alpha Reliability								
Unmonitored	.69	.79	.73	.70	.73	.72	.75	.77
Monitored	.68	.80	.74	.70	.73	.73	.75	.78
Average R-Polyserial								
Unmonitored	.65	.68	.62	.61	.69	.63	.62	.66
Monitored	.64	.69	.63	.62	.68	.63	.62	.66
Proportion of Students Attempting Last Item								
Unmonitored	.77	.67	.78	.76	.67	.76	.71	.84
Monitored	.74	.66	.77	.75	.65	.74	.69	.83

* The number and types of items contained in each block are shown in Table 15-4.

Table 17-6

Block-Level Descriptive Statistics for Monitored and Unmonitored Nonpublic-School Sessions, Grade 4*

Statistic	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted Sample Size								
Unmonitored	1,372	1,361	1,345	1,365	1,382	1,381	1,342	1,370
Monitored	535	538	546	544	556	561	548	556
Average Item Score								
Unmonitored	.57	.72	.54	.66	.52	.58	.67	.73
Monitored	.56	.72	.51	.65	.52	.59	.68	.74
Weighted Alpha Reliability								
Unmonitored	.59	.68	.70	.64	.70	.65	.67	.70
Monitored	.60	.71	.71	.63	.75	.64	.70	.70
Average R-Polyserial								
Unmonitored	.58	.64	.60	.57	.64	.59	.58	.64
Monitored	.60	.63	.62	.57	.70	.59	.58	.63
Proportion of Students Attempting Last Item								
Unmonitored	.82	.78	.87	.84	.75	.82	.81	.91
Monitored	.84	.74	.83	.82	.73	.84	.76	.90

* The number and types of items contained in each block are shown in Table 15-4.

Table 17-7

Block-Level Descriptive Statistics for Monitored and Unmonitored Public-School Sessions, Grade 8*

Statistic	R3	R4	R5	R6	R7	R8	R9	R10	R11	R13
Unweighted Sample Size										
Unmonitored	11,803	11,618	11,732	11,798	11,681	11,691	11,609	11,695	11,720	11,823
Monitored	3,842	3,850	3,892	3,885	3,907	3,852	3,879	3,898	3,923	3,914
Average Item Score										
Unmonitored	.41	.43	.67	.55	.69	.48	.60	.60	.67	.67
Monitored	.42	.43	.67	.56	.69	.49	.61	.61	.69	.67
Weighted Alpha Reliability										
Unmonitored	.77	.69	.76	.70	.78	.68	.70	.71	.79	.74
Monitored	.77	.67	.75	.70	.78	.67	.69	.71	.78	.73
Average R-Polyserial										
Unmonitored	.70	.63	.71	.63	.71	.60	.68	.60	.70	.62
Monitored	.71	.62	.71	.63	.70	.60	.68	.60	.69	.60
Proportion of Students Attempting Last Item										
Unmonitored	.82	.69	.95	.86	.85	.87	.94	.81	.86	.95
Monitored	.83	.70	.95	.86	.86	.88	.96	.81	.85	.95

* The number and types of items contained in each block are shown in Table 15-6.

Table 17-8

Block-Level Descriptive Statistics for Monitored and Unmonitored Nonpublic-School Sessions Grade 8*

Statistic	R3	R4	R5	R6	R7	R8	R9	R10	R11	R13
Unweighted Sample Size										
Unmonitored	645	651	649	655	652	631	637	646	641	673
Monitored	310	311	303	299	313	295	305	316	302	299
Average Item Score										
Unmonitored	.51	.49	.75	.64	.79	.56	.72	.70	.80	.73
Monitored	.50	.52	.76	.66	.80	.58	.69	.70	.79	.74
Weighted Alpha Reliability										
Unmonitored	.74	.60	.72	.57	.70	.58	.64	.62	.65	.57
Monitored	.70	.59	.72	.54	.63	.55	.59	.62	.72	.53
Average R-Polyserial										
Unmonitored	.67	.59	.71	.56	.73	.56	.65	.55	.65	.53
Monitored	.63	.56	.76	.54	.64	.55	.65	.54	.67	.46
Proportion of Students Attempting Last Item										
Unmonitored	.87	.81	.97	.94	.95	.92	.96	.82	.92	.97
Monitored	.83	.83	.97	.94	.92	.93	.98	.87	.89	.94

* The number and types of items contained in each block are shown in Table 15-6.

Table 17-9 for grade 4 and Table 17-10 for grade 8 summarize the differences between monitored and unmonitored average item scores for the jurisdictions. These are mean differences within a jurisdiction averaged over all items in all blocks. The results in the tables are from combined samples of public- and nonpublic-school data. The mean difference and median difference were close to zero. For grade 4, 26 jurisdictions had negative differences (i.e., students from unmonitored sessions scored higher than students from monitored sessions). None was larger in absolute magnitude than 0.029. For grade 8, 17 jurisdictions had negative differences. The largest in absolute magnitude is 0.052. The results indicate that across jurisdictions, the differences between monitored and unmonitored sessions are relatively small for both grades. While these tables list differences, no significance tests were done. This is true for all the descriptive statistics in Tables 17-5 to 17-12.

As has been the case since the 1994 trial state assessment in reading, the 1998 state assessment in reading included students sampled from nonpublic schools. Tables 17-11 and 17-12 show the difference between public and nonpublic schools with respect to sample size, average item scores, alpha reliability, average *r*-polyserial correlation, and proportion of students attempting the last item in a block. As with the monitored/unmonitored comparisons, results were aggregated over all participating jurisdictions. For grade 4, 43 of the 44 jurisdictions that participated in the state assessment in reading had public-school samples and 29 of the 44 jurisdictions had nonpublic-school samples that met reporting requirements. For grade 8, 40 of the 41 jurisdictions had public-school samples and 23 of the 41 jurisdictions had nonpublic-school samples that met reporting requirements.

Consistent differences are evident between the public- and nonpublic-school groups. Table 17-11, for grade 4, indicates that the difference in average item score between public- and nonpublic-school students (i.e., public block mean minus nonpublic block mean) ranged from -.095 to -.061, with an average of -.079, indicating that public-school students were generally lower in average item score.

Table 17-9

Effect of Monitoring Sessions by Jurisdiction:
Average Jurisdiction Item Scores for Monitored and Unmonitored Sessions, Grade 4

	Monitored	Unmonitored	Monitored – Unmonitored
Alabama	0.506	0.489	0.017
Arizona	0.467	0.494	-0.027
Arkansas	0.512	0.491	0.022
California	0.459	0.473	-0.014
Colorado	0.548	0.553	-0.005
Connecticut	0.609	0.592	0.017
Delaware	0.490	0.500	-0.009
Florida	0.517	0.493	0.024
Georgia	0.495	0.501	-0.006
Hawaii	0.483	0.473	0.010
Iowa	0.553	0.557	-0.004
Kansas	0.549	0.548	0.001
Kentucky	0.519	0.527	-0.008
Louisiana	0.490	0.488	0.002
Maine	0.571	0.561	0.010
Maryland	0.539	0.538	0.001
Massachusetts	0.584	0.569	0.015
Michigan	0.541	0.535	0.006
Minnesota	0.560	0.558	0.002
Mississippi	0.468	0.473	-0.005
Missouri	0.554	0.525	0.029
Montana	0.550	0.571	-0.021
Nebraska	0.561	0.608	-0.047
Nevada	0.493	0.489	0.004
New Hampshire	0.538	0.575	-0.036
New Mexico	0.475	0.488	-0.013
New York	0.523	0.533	-0.010
North Carolina	0.505	0.535	-0.030
Oklahoma	0.520	0.533	-0.013
Oregon	0.517	0.515	0.002
Rhode Island	0.546	0.545	0.001
South Carolina	0.499	0.502	-0.002
Tennessee	0.499	0.503	-0.004
Texas	0.538	0.525	0.013
Utah	0.515	0.518	-0.002
Virginia	0.525	0.532	-0.007
Washington	0.525	0.544	-0.019
West Virginia	0.511	0.530	-0.019
Wisconsin	0.551	0.566	-0.014
Wyoming	0.529	0.539	-0.010
District of Columbia	0.365	0.373	-0.008
DoDEA/DDESS	0.538	0.535	0.002
DoDEA/DoDDS	0.539	0.554	-0.016
Virgin Islands	0.348	0.399	-0.051
Mean			**-0.005**
Median			**-0.005**
Minimum			**-0.051**
1st Quartile			**-0.013**
3rd Quartile			**0.003**
Maximum			**0.029**

Table 17-10

Effect of Monitoring Sessions by Jurisdiction:
Average Jurisdiction Item Scores for Monitored and Unmonitored Sessions, Grade 8

	Monitored	Unmonitored	Monitored - Unmonitored
Alabama	0.499	0.514	-0.014
Arizona	0.545	0.541	0.004
Arkansas	0.533	0.516	0.017
California	0.527	0.514	0.012
Colorado	0.567	0.559	0.008
Connecticut	0.606	0.600	0.006
Delaware	0.559	0.507	0.052
Florida	0.540	0.513	0.027
Georgia	0.533	0.534	-0.002
Hawaii	0.510	0.480	0.031
Kansas	0.590	0.569	0.021
Kentucky	0.568	0.546	0.022
Louisiana	0.513	0.521	-0.008
Maine	0.601	0.607	-0.006
Maryland	0.555	0.569	-0.014
Massachusetts	0.594	0.583	0.010
Minnesota	0.596	0.576	0.020
Mississippi	0.509	0.487	0.022
Missouri	0.558	0.560	-0.002
Montana	0.584	0.594	-0.010
Nebraska	0.640	0.627	0.014
Nevada	0.532	0.527	0.005
New Mexico	0.535	0.532	0.004
New York	0.573	0.582	-0.009
North Carolina	0.567	0.559	0.008
Oklahoma	0.564	0.560	0.004
Oregon	0.559	0.572	-0.012
Rhode Island	0.588	0.560	0.028
South Carolina	0.508	0.510	-0.002
Tennessee	0.522	0.537	-0.014
Texas	0.533	0.547	-0.015
Utah	0.576	0.553	0.023
Virginia	0.588	0.564	0.024
Washington	0.565	0.566	-0.002
West Virginia	0.548	0.545	0.003
Wisconsin	0.580	0.566	0.014
Wyoming	0.517	0.559	-0.043
District of Columbia	0.414	0.436	-0.022
DoDEA/DDESS	0.607	0.562	0.045
DoDEA/DoDDS	0.567	0.583	-0.016
Virgin Islands	0.436	0.447	-0.011
Mean			**0.005**
Median			**0.004**
Minimum			**-0.043**
1st Quartile			**-0.009**
3rd Quartile			**0.020**
Maximum			**0.052**

The public/nonpublic difference in average item-to-total block correlation (the average *r*-polyserial) ranged from 0.017 to 0.059, with an average of 0.037, indicating that public-school students generally had a somewhat higher item-to-total correlation. As for the proportion of students attempting the last item, public minus nonpublic differences ranged from -.097 to -.06, with an average of -.080, indicating that somewhat fewer students in public schools attempted the last item.

Table 17-11

Block-Level Descriptive Statistics for Overall Public- and Nonpublic-School Sessions
Grade 4

Statistic	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted Sample Size								
Public	24,763	24,686	24,294	24,498	24,500	24,667	24,531	24,618
Nonpublic	1,907	1,899	1,891	1,909	1,938	1,942	1,890	1,926
Weighted Average Item Score								
Public	.48	.64	.45	.58	.42	.51	.61	.65
Nonpublic	.56	.72	.53	.66	.52	.58	.67	.73
Weighted Alpha Reliability								
Public	.69	.79	.72	.70	.73	.72	.75	.77
Nonpublic	.59	.69	.70	.64	.71	.65	.68	.70
Weighted Average R-Polyserial								
Public	.65	.68	.62	.61	.68	.63	.62	.66
Nonpublic	.59	.64	.60	.57	.66	.59	.58	.63
Weighted Proportion of Students Attempting Last Item								
Public	.76	.67	.78	.75	.66	.75	.71	.85
Nonpublic	.84	.77	.86	.84	.74	.82	.80	.90

Table 17-12

Block-Level Descriptive Statistics for Overall Public- and Nonpublic-School Sessions
Grade 8

Statistic	R3	R4	R5	R6	R7	R8	R9	R10	R11	R13
Unweighted Sample Size										
Public	15,645	15,468	15,624	15,683	15,588	15,543	15,488	15,593	15,643	15,737
Nonpublic	955	962	952	954	965	926	.942	962	943	972
Weighted Average Item Score										
Public	.41	.43	.67	.56	.69	.48	.60	.60	.68	.67
Nonpublic	.51	.50	.75	.65	.79	.56	.71	.70	.79	.74
Weighted Alpha Reliability										
Public	.77	.69	.75	.70	.78	.68	.70	.71	.79	.74
Nonpublic	.73	.60	.72	.56	.68	.58	.62	.62	.67	.56
Weighted Average R-Polyserial										
Public	.70	.63	.71	.63	.70	.60	.68	.60	.70	.61
Nonpublic	.51	.50	.75	.65	.79	.56	.71	.70	.79	.51
Weighted Proportion of Students Attempting Last Item										
Public	.82	.69	.95	.86	.85	.88	.95	.81	.86	.95
Nonpublic	.86	.82	.97	.94	.93	.92	.96	.84	.91	.96

17.3 STATE IRT SCALING

As described in Chapter 12, separate IRT-based scales were developed using the scaling models. For grade 4, two scales were produced by separately calibrating the sets of items classified in each of the two content areas. For grade 8, three scales were produced in each of the three content areas.

For the reasons discussed in Mazzeo (1991), for each scale, a single set of item parameters for each item was estimated and used for all jurisdictions. Item-parameter estimation was carried out using a 25 percent systematic random sample of the students participating in the 1998 state assessment and included equal numbers of students from each participating jurisdiction, half from monitored sessions and half from unmonitored sessions whenever possible. All students in the scaling sample were public-school students. The grade 4 sample consisted of 98,873 students, with 590 students being sampled from each of the 42 participating jurisdictions (excluding DoDEA/DDESS[2] and DoDEA/DoDDS[3] schools). Of the 590 records sampled from each jurisdiction, 295 were drawn from the monitored sessions and 295 were drawn from the unmonitored sessions. The grade 8 sample consisted of 86,210 students, with 554 students being sampled from each of the 39 participating jurisdictions. Of the 554 records sampled from each jurisdiction, 277 were drawn from the monitored sessions and 277 were drawn from the unmonitored sessions. In grade 8, there were less than 277 monitored students in the District of Columbia and Virgin Islands; therefore, all the monitored students in these two jurisdictions were included. The rescaled weights for the 25 percent sample of students used in item calibration were adjusted slightly to ensure that (1) each jurisdiction's data contributed equally to the estimation process, and (2) data from monitored and unmonitored sessions contributed equally. All calibrations were carried out using the rescaled sampling weights described in Section 11.2 in an effort to ensure that each jurisdiction's data contributed equally to the determination of the item-parameter estimates.

To the extent that items may have functioned differently in monitored and unmonitored sessions, the single set of item parameters obtained defines a set of item characteristic curves "averaged over" the two types of sessions. Tables 17-5 through 17-8 (shown earlier) presented block-level item statistics that suggested little, if any, difference in item functioning by session type.

Only public-school data were used in the scaling models for the state assessments, since no DIF items were found in the public versus nonpublic comparisons for both fourth- and eighth-grade data. For details on DIF analysis, see Chapter 15, Section 15.4.

17.3.1 Item Parameter Estimation

For each content-area scale, item parameter estimates were obtained using the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs. The program uses marginal maximum likelihood estimation procedures to estimate the parameters of the one-, two-, and three-parameter logistic models, and the generalized partial-credit model described by Muraki (1992).

Multiple-choice items were dichotomously scored and were scaled using the three-parameter logistic model. Omitted responses to multiple-choice items were treated as fractionally correct, with the fraction being set to 1 over the number of response options. Short constructed-response items that were also in the 1992 assessment were dichotomously scored and scaled using the two-parameter logistic model. New short (regular) constructed-response items were scored on a three-point generalized partial-

[2] DoDEA/DDESS is the Department of Defense Education Activity Department of Defense Domestic Dependent Elementary and Secondary Schools.
[3] DoDEA/DoDDS is the Department of Defense Education Activity Department of Defense Dependents Schools.

credit scale. These items appear in block 3 for grade 4, and in blocks 3 and 8 for grade 8. Omitted responses to short constructed-response items were treated as incorrect.

There were a total of eight extended constructed-response items. Each of these items was also scaled using the generalized partial-credit model. Four scoring levels were defined:

0 = Unsatisfactory response or omitted
1 = Partial response
2 = Essential response
3 = Extensive response

Note that omitted responses were treated as the lowest possible score level. As stated earlier, not-reached and off-task responses were treated as if the item were not administered to the student. Table 17-13 provides a listing of the blocks, positions within the block, content-area classifications, and NAEP identification numbers for all extended constructed-response items included in the 1998 assessment for grade 4 and grade 8 data.

Table 17-13
Extended Constructed-Response Items, 1998 State Assessment in Reading

Grade	Block	Position in Block	Content Area Classifications	NAEP ID
4	R3	6	Literary Experience	R017007
	R4	11	Literary Experience	R012111
	R5	7	Literary Experience	R012607
	R6	4	Gain Information	R012204
	R7	8	Gain Information	R012708
	R8	7	Gain Information	R015707
	R9	4	Literary Experience	R015804
	R10	12	Gain Information	R012512
8	R3	5	Literary Experience	R017105
	R4	6	Literary Experience	R015906
	R5	7	Literary Experience	R012607
	R6	1	Gain Information	R013201
	R6	12	Gain Information	R013212
	R7	8	Gain Information	R012708
	R8	5	Gain Information	R017205
	R13	4	Gain Information	R016204

Empirical Bayes modal estimates of all item parameters were obtained from the BILOG/PARSCALE program. Prior distributions were imposed on item parameters with the following starting values: thresholds, normal [0,2]; slopes, log-normal [0,.5]; and asymptotes, two-parameter beta with parameter values determined as functions of the number of response options for an item and a weight factor of 50. The locations (but not the dispersions) were updated at each program-estimation cycle in accordance with provisional estimates of the item parameters.

Item parameter estimation proceeded in two phases. First, the subject ability distribution was assumed fixed (normal [0,1]) and a stable solution was obtained. Starting values for the item parameters were provided by item analysis routines. The parameter estimates from this initial solution were then

used as starting values for a subsequent set of runs in which the subject ability distribution was freed and estimated concurrently with item parameter estimates. After each estimation cycle, the subject ability distribution was standardized to have a mean of zero and standard deviation of one. Correspondingly, parameter estimates for that cycle were also linearly standardized.

During and subsequent to item parameter estimation, evaluations of the fit of the IRT models were carried out for each of the items in the item pool. These evaluations were conducted to determine the final composition of the item pool making up the scales by identifying misfitting items that should not be included. Evaluations of model fit were based primarily on graphical analyses. For dichotomously scored multiple-choice and two-category response items, model fit was evaluated by examining plots of estimates of the expected conditional (on theta) probability of a correct response that do not assume a two-parameter or three-parameter logistic model versus the probability predicted by the estimated item-characteristic curve (see Mislevy & Sheehan, 1987, p. 302). For the extended constructed-response items, similar plots were produced for each item-category characteristic curve.

As with most procedures that involve evaluating plots of data versus model predictions, a certain degree of subjectivity is involved in determining the degree of fit necessary to justify use of the model. There are a number of reasons why evaluation of model fit relied primarily on analyses of plots rather than seemingly more objective procedures based on goodness-of-fit indices such as the “pseudo chi-squares” produced in BILOG (Mislevy & Bock, 1982). First, when the model fits, the exact sampling distributions of these indices are not well understood, even for fairly long tests. Mislevy and Stocking (1989) point out that the usefulness of these indices appears particularly limited in situations like NAEP, where examinees have been administered relatively short tests. A study by Stone, Mislevy, and Mazzeo (1994) using simulated data suggests that the correct reference chi-square distributions for these indices have considerably fewer degrees of freedom than the value indicated by the BILOG/PARSCALE program, and require additional adjustments of scale. However, it is not yet clear how to estimate the correct number of degrees of freedom and necessary scale factor adjustment factors. Consequently, pseudo chi-square goodness-of-fit indices are used only as rough guides in interpreting the severity of model departures.

Second, as discussed in Chapter 12, it is almost certainly the case that, for most items, item response models hold only to a certain degree of approximation. Given the large sample sizes used in NAEP and the state assessment, there will be sets of items for which one is almost certain to reject the hypothesis that the model fits the data, even though departures are minimal in nature or involve kinds of misfit unlikely to impact on important model-based inferences. In practice, one is almost always forced to temper statistical decisions with judgments about the severity of model misfit and the potential impact of such misfit on final results.

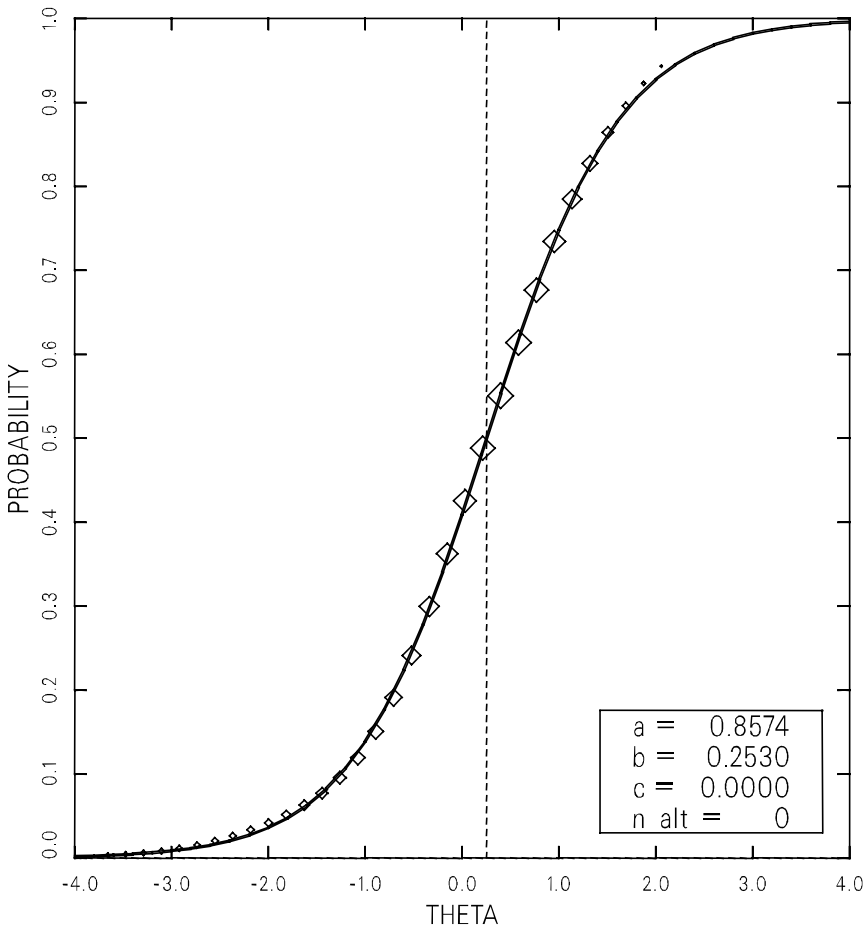
To maximize the agreement between the state analysis and national analysis, the 1998 state assessment incorporated most adjustments and deletions resulting from the analysis of the 1998 national assessment in reading.

For the large majority of the items for grade 4 and grade 8 data, the fit of the model was extremely good. Figure 17-1 provides typical examples of what the plots look like for this class of items. Item R012106 for grade 4 is a binary-scored constructed-response item. Item R012711 for grade 8, at the top of Figure 17-1 (continued), is a multiple-choice item; item R013405 for grade 8, at the bottom of Figure 17-1 (continued), is a binary-scored constructed-response item. In each plot, the *x*-axis indicates scale score level (theta) and the *y*-axis indicates the probability of a correct response. The diamonds show estimates of the conditional (on theta) probability of a correct response that do not assume a logistic form (referred to subsequently as nonlogistic-based estimates). The sizes of the diamonds are proportional to the number of students categorized as having thetas at or close to the indicated value. The solid curve shows the estimated item response function. The item response function provides estimates of the

conditional probability of a correct response based on an assumed logistic form. The vertical dashed line indicates the estimated location parameter (*b*) for the item and the horizontal dashed line (e.g., item R012711) indicates the estimated lower asymptote (*c*). Also shown in the plot are the values of the item parameter estimates. As is evident from the plots, the nonlogistic-based estimates of conditional (diamonds) probabilities are in extremely close agreement with those given by the estimated item response function (the solid curves).

Figure 17-1

*Dichotomous Items (R012106, R012711, and R013405) Exhibiting Good Model Fit**

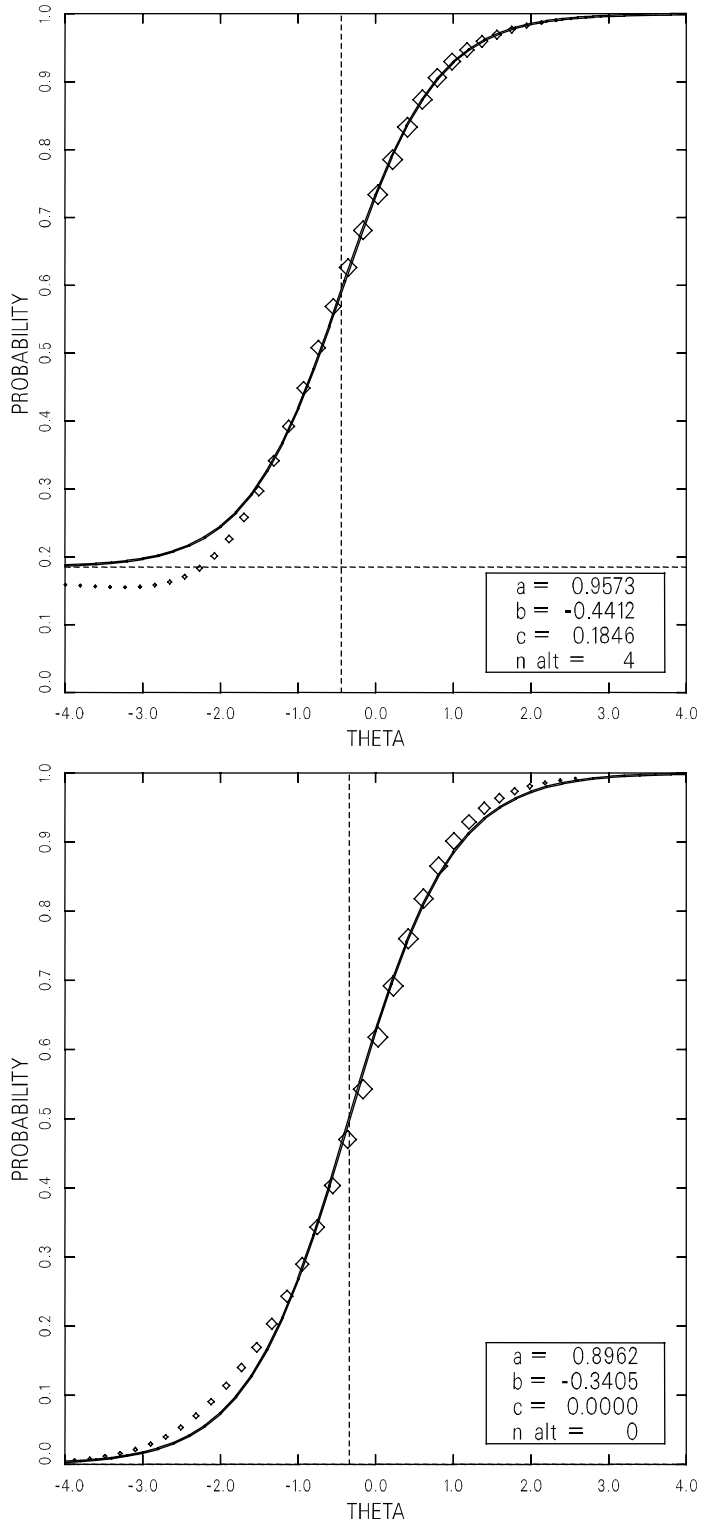


** Diamonds represent 1998 grade 4 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item response function (IRF) assuming a logistic form.*

(continued)

Figure 17-1 (continued)

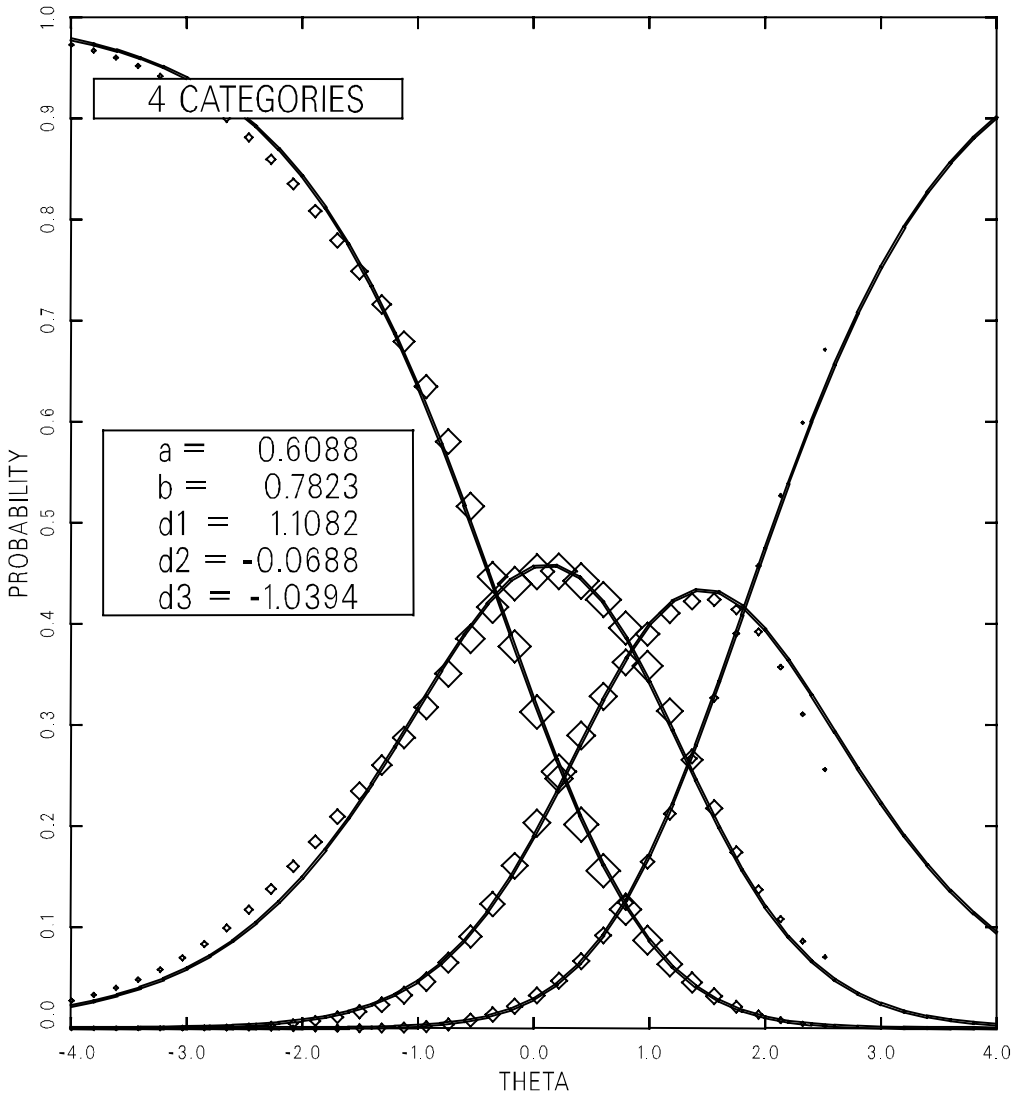
*Dichotomous Items (R012106, R012711, and R013405) Exhibiting Good Model Fit**



** Diamonds represent 1998 grade 8 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item response function (IRF) assuming a logistic form.*

Figure 17-2 provides an example of a plot for a four-category extended constructed-response item (R013201, grade 8) exhibiting good model fit. Like the plots for the binary items, this plot shows two estimates of each item category characteristic curve, one set that does not assume the partial-credit model (shown as diamonds) and one that does (the solid curves). The estimates for all parameters for the item in question are also indicated on the plot. As shown by the figure, there is strong agreement and only slight differences between the item category characteristic curve and the curve of diamonds at the high categories. Although few student responses were scored in the highest category, there were adequate data to calculate the model-based estimates for those categories (the solid curves). Such results were typical for the extended constructed-response items.

Figure 17-2
*Polytomous Item (R013201) Exhibiting Good Model Fit**



** Diamonds represent 1998 grade 8 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

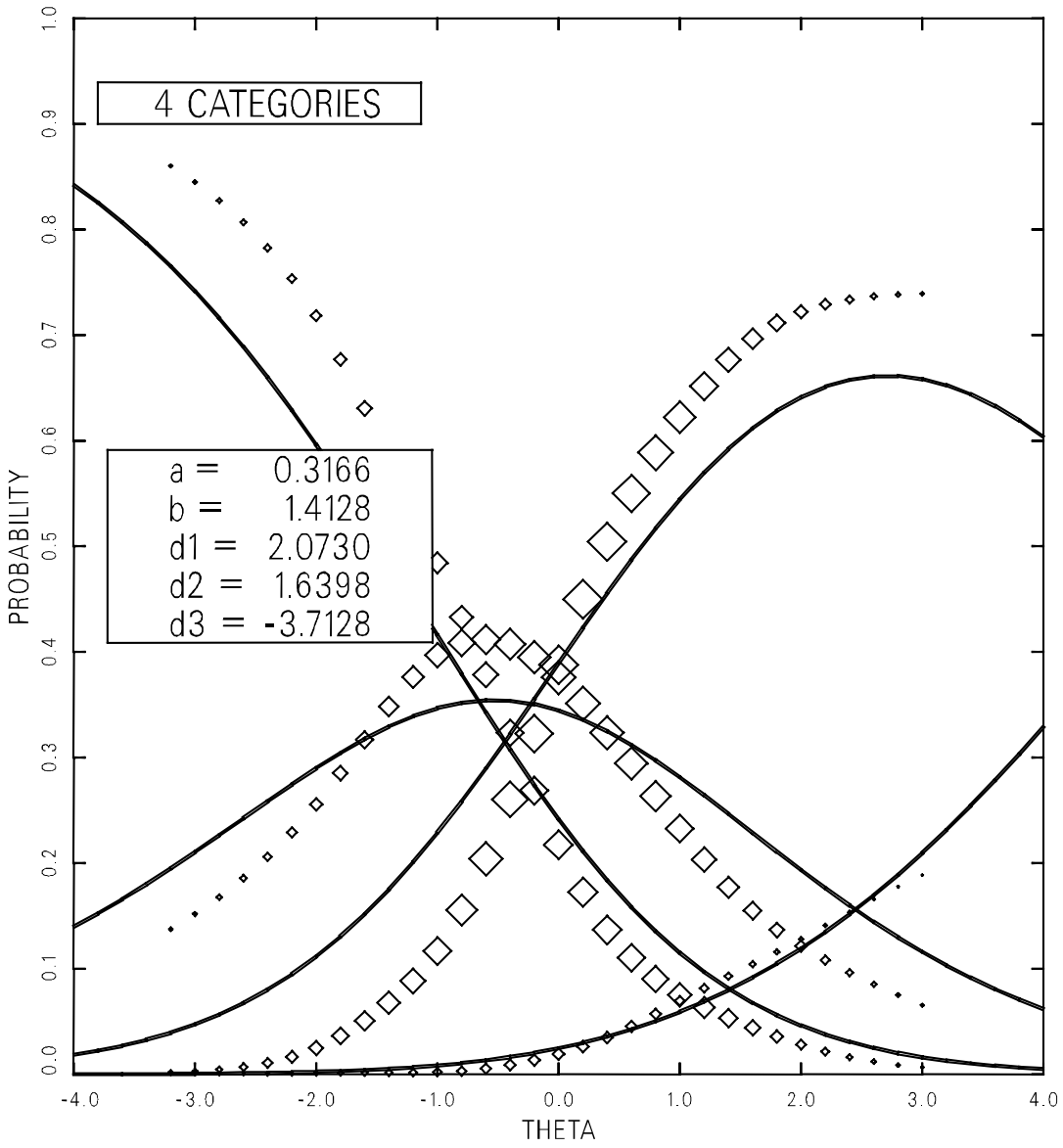
17.3.2 Recoded Extended Constructed-Response Items

As discussed above, some of the items retained for the final scales display some degree of model misfit. In general, good agreement between nonlogistic and logistic estimates of conditional probabilities was found in the regions of the theta scale that includes most of the examinees. Misfit was confined to conditional probabilities associated with theta values in the tails of the subject ability distributions.

For grade 4 data, item R012111, an item of Literary Experience in the eleventh position in block R4, received special treatment in the scaling process in the 1992, 1994, and 1998 assessments. Figure 17-3 shows the plot of item R012111 before collapsing unsatisfactory and partial-response categories using 1998 assessment data.

Figure 17-3

*Polytomous Item (R012111) Before Collapsing Unsatisfactory and Partial-Response Categories**



** Diamonds represent 1998 grade 4 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

To obtain a good fit of the generalized partial-credit model to the extended constructed-response items in 1998 assessment, the categories 0 and 1 were combined and the other categories were relabeled as in previous assessments. Therefore, the codings for the three scoring levels were defined:

0 = Unsatisfactory, partial response, or omitted
1 = Essential response
2 = Extensive response

The plot for this item for the 1998 data after collapsing the unsatisfactory and partial-response categories is given in Figure 17-4. The figure shows good model fit, except that the nonlogistic-based estimates tend to be somewhat different from the model-based estimates for theta values greater than 1. Note that this item is functioning essentially as a dichotomous item due to the small frequencies in the top category. There were enough data, however, to calculate the model-based estimates of the category-characteristic curve for this category (shown as the rightmost solid curve in both figures).

Another fourth-grade item, R015707, an item of Gain Information in the seventh position in block R8, also received special treatment in the 1994 and 1998 assessments. As with item R012111, the general partial-credit model did not fit the response to the extended constructed-response item R015707 well. This Reading to Gain Information item was treated the same way as was item R012111, and good model-data fit was obtained.

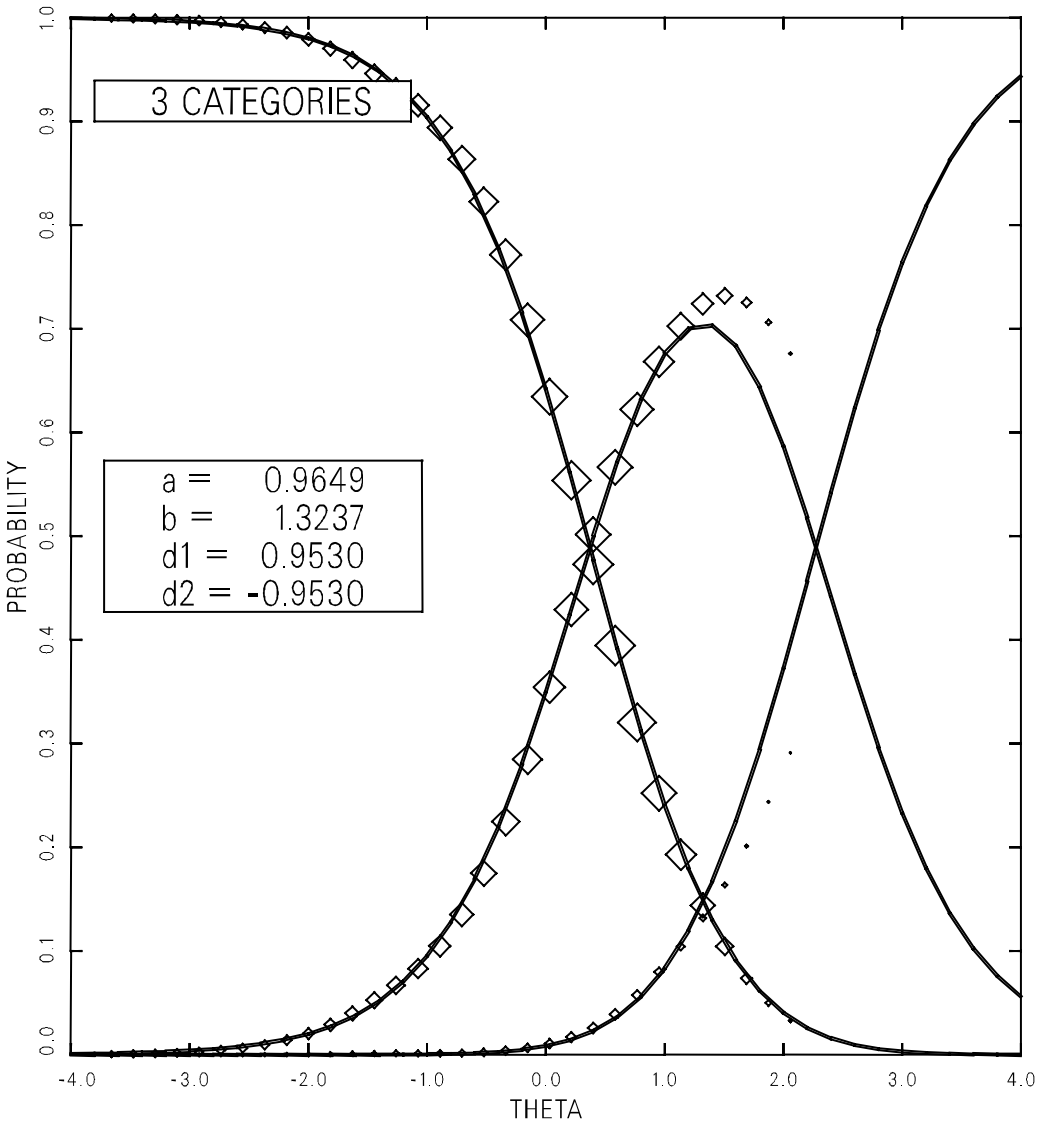
To be consistent with the scaling of the 1998 national reading assessment for grade 8 data, item R017110, an item of Literary Experience in the tenth position in block R3, received special treatment. The categories 0 and 1 were combined as 0 and the other categories were relabeled as 1. Therefore R017110 was defined as a dichotomous item. A plot for this item after collapsing the categories is displayed in Figure 17-5.

To be consistent with the previous assessments, for grade 8 data, item R017102, an item of Literary Experience in the second position in block R3, received special treatment. It was recoded as a dichotomous item: the categories 0 and 1 were combined as 0 and the other categories were relabeled as 1. Item R016212, an item of Gain Information in the twelfth position in block R13, was recoded in the state assessment as it was recoded in the national assessment: The categories 0 and 1 were combined as 0 and the other categories were relabeled as 1. A plot for this item after collapsing the categories is displayed in Figure 17-6.

The IRT parameters for the items included in the state assessment are listed in Appendix E.

Figure 17-4

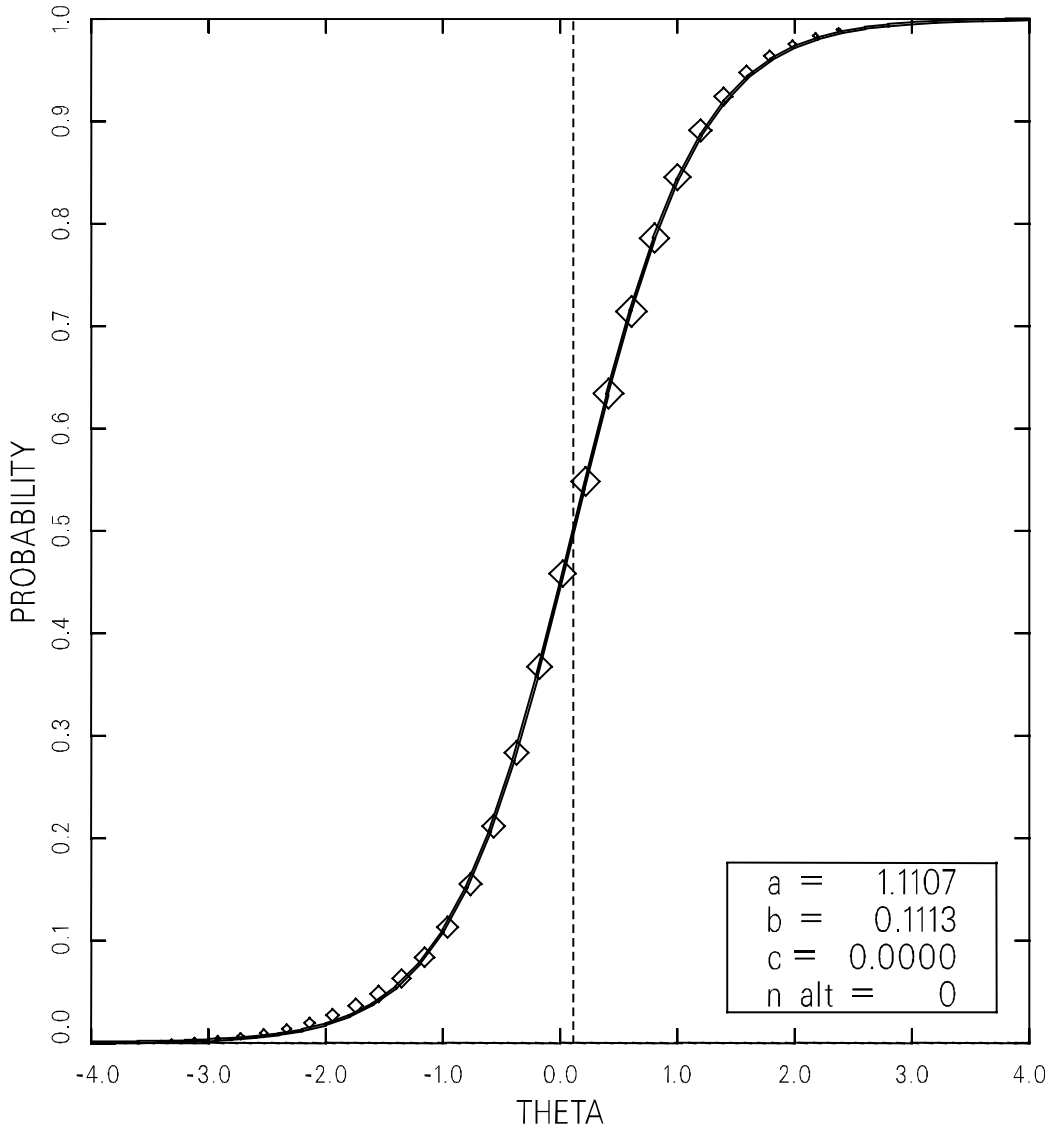
*Polytomous Item (R012111) After Collapsing Unsatisfactory and Partial-Response Categories**



** Diamonds represent 1998 grade 4 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

Figure 17-5

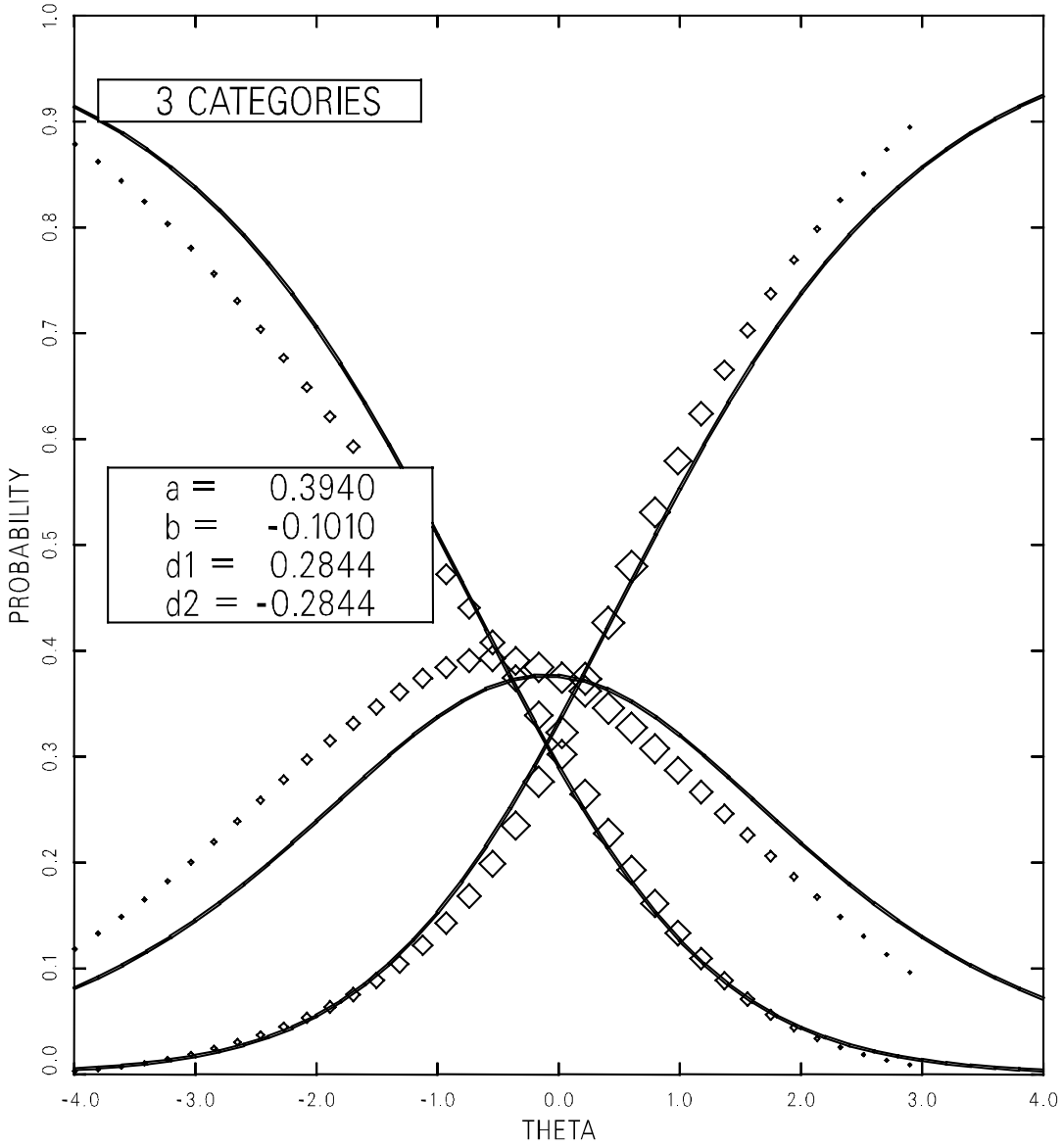
*Polytomous Item (R017110) After Collapsing Unsatisfactory and Partial-Response Categories**



** Diamonds represent 1998 grade 8 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

Figure 17-6

*Polytomous Item (R016212) After Collapsing Unsatisfactory and Partial-Response Categories**



** Diamonds represent 1998 grade 8 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

17.4 GENERATION OF PLAUSIBLE VALUES

The scale score distributions for each jurisdiction (and for subgroups of interest within each jurisdiction) were estimated using the multivariate plausible values methodology and the corresponding CGROUP computer program. As described in Chapter 12, the CGROUP program estimates scale score distributions using information from student item responses, measures of student background variables, and the item parameter estimates obtained from the BILOG/PARSCALE program.

Results from Mazzeo's research (1991) suggested that separate conditioning models be estimated for each jurisdiction because the parameters estimated by the conditioning model differed across jurisdictions. If a jurisdiction had a nonpublic-school sample, students from that sample were included in this part of the analysis, and a conditioning variable differentiating between public- and nonpublic-school students was included. This resulted in the estimation of 44 distinct conditioning models for grade 4, and 41 distinct conditioning models for grade 8.

Reporting each jurisdiction's results required analyses describing the relationships between scale scores and a large number of background variables. The background variables included in each jurisdiction's model were principal component scores derived from the within-jurisdiction correlation matrix of selected main-effects and two-way interactions associated with a wide range of student, teacher, school, and community variables. The background variables included student demographic characteristics (e.g., the race/ethnicity of the student, highest level of education attained by parents), students' perceptions about reading, student behavior both in and out of school (e.g., amount of TV watched daily, amount of reading homework done each day), the type of reading class being taken, and a variety of other aspects of the students' background and preparation, and the educational, social, and financial environment of the schools they attended. Information was also collected from students' teachers about their teaching practices, such as the amount of classroom emphasis on various topics included in the assessment, and their educational background and professional preparation.

As described in the previous chapter, to avoid biases in reporting results and to minimize biases in secondary analyses, it is desirable to incorporate measures of a large number of independent variables in the conditioning model. For grade 4, when expressed in terms of contrast-coded main effects and interactions, the number of variables to be included totaled 1,086; for grade 8, the number of variables to be included totaled 1,064. Appendix F provides a listing of the full set of contrasts defined. These contrasts were the common starting point in the development of the conditioning models for each of the participating jurisdictions.

Because of the large number of these contrasts and the fact that, within each jurisdiction, some contrasts had zero variance, some involved relatively small numbers of individuals, and some were highly correlated with other contrasts or sets of contrasts, an effort was made to reduce the dimensionality of the predictor variables in each jurisdiction's CGROUP models. As was done for the 1990 and 1992 state assessments in mathematics and the 1992 and 1994 state assessments in reading, the original background variable contrasts were standardized and transformed into a set of linearly independent variables by extracting separate sets of principal components (one set for each of the 44 jurisdictions) from the within-jurisdiction correlation matrices of the original contrast variables. The principal components, rather than the original variables, were used as the independent variables in the conditioning model. As was done for the previous assessments, the number of principal components included for each jurisdiction was the number required to account for approximately 90 percent of the variance in the original contrast variables. Research based on data from the 1990 state assessment in mathematics suggested that results obtained using such a subset of the components will differ only slightly from those obtained using the full set (Mazzeo et al., 1992).

Table 17-14

Summary Statistics for State Assessment Conditioning Models, Grade 4

Jurisdiction	Number of Principal Components	Proportion* of Scale Score Variance in the Reading Assessment for Literary Experience Scale Accounted for by the Conditioning Model	Proportion* of Scale Score Variance in the Reading Assessment to Gain Information Scale Accounted for by the Conditioning Model	Conditional Correlation Between Literary Experience and Gain Information
Alabama	240	0.68	0.69	0.86
Arizona	242	0.71	0.72	0.89
Arkansas	253	0.68	0.69	0.86
California	195	0.70	0.71	0.89
Colorado	236	0.61	0.65	0.86
Connecticut	262	0.71	0.69	0.78
Delaware	231	0.77	0.75	0.85
District of Columbia	186	0.64	0.69	0.87
Florida	278	0.69	0.67	0.90
Georgia	275	0.74	0.75	0.84
Hawaii	260	0.62	0.56	0.84
Iowa	202	0.66	0.65	0.77
Kansas	191	0.69	0.74	0.85
Kentucky	221	0.70	0.67	0.87
Louisiana	256	0.56	0.61	0.86
Maine	230	0.73	0.76	0.80
Maryland	218	0.58	0.48	0.91
Massachusetts	235	0.68	0.72	0.89
Michigan	229	0.69	0.71	0.86
Minnesota	243	0.72	0.66	0.89
Mississippi	247	0.54	0.70	0.90
Missouri	241	0.66	0.63	0.89
Montana	180	0.80	0.75	0.80
Nebraska	110	0.93	0.89	0.91
Nevada	256	0.56	0.71	0.92
New Hampshire	209	0.84	0.80	0.86
New Mexico	238	0.65	0.67	0.91
New York	238	0.67	0.68	0.75
North Carolina	258	0.58	0.59	0.84
Oklahoma	234	0.66	0.72	0.89
Oregon	226	0.70	0.72	0.84
Rhode Island	253	0.68	0.68	0.76
South Carolina	254	0.67	0.66	0.88
Tennessee	253	0.68	0.61	0.85
Texas	235	0.75	0.73	0.90
Utah	238	0.64	0.64	0.88
Virginia	259	0.71	0.67	0.93
Virgin Islands	160	0.49	0.62	0.90
Washington	233	0.55	0.58	0.91
West Virginia	217	0.64	0.66	0.80
Wisconsin	219	0.87	0.82	0.90
Wyoming	206	0.80	0.78	0.86
DoDEA/DDESS	184	0.65	0.69	0.90
DoDEA/DoDDS	207	0.88	0.86	0.77

* (Total Variance – Residual Variance)/Total Variance, where Total Variance consists of both sampling and measurement error variance.

Table 17-15

Summary Statistics for State Assessment Conditioning Models, Grade 8

Jurisdiction	Number of Principal Components	Proportion* of Scale Score Variance in the Reading for Literary Experience Scale Accounted for by the Conditioning Model	Proportion* of Scale Score Variance in the Reading to Gain Information Scale Accounted for by the Conditioning Model	Proportion* of Scale Score Variance in the Reading to Perform a Task Scale Accounted for by the Conditioning Model	Conditional Correlation Between Literary Experience and Gain Information	Conditional Correlation Between Literary Experience and Perform a Task	Conditional Correlation Between Gain Information and Perform a Task
Alabama	229	0.70	0.66	0.74	0.90	0.90	0.93
Arizona	244	0.69	0.72	0.82	0.87	0.85	0.85
Arkansas	233	0.72	0.68	0.76	0.79	0.76	0.88
California	245	0.76	0.72	0.82	0.82	0.87	0.82
Colorado	233	0.69	0.71	0.73	0.83	0.85	0.92
Connecticut	264	0.73	0.78	0.81	0.92	0.80	0.83
Delaware	179	0.78	0.72	0.84	0.92	0.89	0.91
District of Columbia	148	0.77	0.72	0.78	0.91	0.86	0.87
Florida	267	0.76	0.60	0.79	0.79	0.71	0.88
Georgia	283	0.77	0.78	0.83	0.89	0.90	0.90
Hawaii	194	0.58	0.59	0.70	0.82	0.78	0.83
Kansas	191	0.81	0.71	0.74	0.92	0.92	0.87
Kentucky	222	0.70	0.63	0.72	0.92	0.85	0.89
Louisiana	255	0.75	0.74	0.77	0.78	0.76	0.81
Maine	210	0.75	0.77	0.83	0.87	0.83	0.91
Maryland	234	0.66	0.67	0.67	0.86	0.89	0.91
Massachusetts	232	0.75	0.74	0.85	0.91	0.86	0.88
Minnesota	197	0.81	0.69	0.80	0.83	0.77	0.82
Mississippi	223	0.72	0.57	0.67	0.88	0.92	0.92
Missouri	236	0.67	0.69	0.75	0.85	0.88	0.89
Montana	172	0.88	0.76	0.89	0.91	0.86	0.93
Nebraska	99	1.00	0.96	1.00	0.55	0.33	0.58

* (Total Variance – Residual Variance)/Total Variance, where Total Variance consists of both sampling and measurement error variance.

(continued)

Table 17-15 (continued)

Summary Statistics for State Assessment Conditioning Models, Grade 8

Jurisdiction	Number of Principal Components	Proportion* of Scale Score Variance in the Reading for Literary Experience Scale Accounted for by the Conditioning Model	Proportion* of Scale Score Variance in the Reading to Gain Information Scale Accounted for by the Conditioning Model	Proportion* of Scale Score Variance in the Reading to Perform a Task Scale Accounted for by the Conditioning Model	Conditional Correlation Between Literary Experience and Gain Information	Conditional Correlation Between Literary Experience and Perform a Task	Conditional Correlation Between Gain Information and Perform a Task
Nevada	213	0.75	0.64	0.79	0.91	0.92	0.92
New Mexico	234	0.73	0.69	0.84	0.71	0.66	0.93
New York	221	0.78	0.75	0.77	0.83	0.84	0.89
North Carolina	271	0.64	0.60	0.71	0.81	0.72	0.82
Oklahoma	219	0.69	0.74	0.85	0.90	0.80	0.85
Oregon	225	0.82	0.76	0.82	0.87	0.90	0.91
Rhode Island	206	0.74	0.70	0.79	0.85	0.80	0.88
South Carolina	279	0.77	0.75	0.78	0.90	0.87	0.94
Tennessee	222	0.62	0.70	0.82	0.89	0.86	0.89
Texas	249	0.79	0.71	0.78	0.85	0.89	0.86
Utah	241	0.72	0.70	0.76	0.77	0.81	0.84
Virginia	273	0.78	0.72	0.81	0.82	0.76	0.84
Virgin Islands	129	0.75	0.64	0.81	0.96	0.95	0.94
Washington	247	0.74	0.70	0.75	0.91	0.87	0.91
West Virginia	229	0.78	0.76	0.77	0.92	0.92	0.90
Wisconsin	195	0.84	0.83	0.90	0.91	0.86	0.88
Wyoming	181	0.88	0.85	0.92	0.79	0.84	0.87
DoDEA/DDESS	130	0.98	0.92	0.97	0.87	0.87	0.88
DoDEA/DoDDS	160	0.89	0.86	0.90	0.83	0.83	0.90

* (Total Variance – Residual Variance)/Total Variance, where Total Variance consists of both sampling and measurement error variance

Tables 17-14 for grade 4 and 17-15 for grade 8 list the number of principal components included in and the proportion of scale score variance accounted for by the conditioning model for each participating jurisdiction. It is important to note that the proportion of variance accounted for by the conditioning model differs across scales within a jurisdiction, and across jurisdictions within a scale. Such variability is not unexpected for at least two reasons. First, there is no reason to expect the strength of the relationship between scale score and demographics to be identical across all jurisdictions. In fact, one of the reasons for fitting separate conditioning models is that the strength and nature of this relationship may differ across jurisdictions. Second, the homogeneity of the demographic profile also differs across jurisdictions. As with any correlation analysis, restriction of the range in the predictor variables will attenuate the relationship.

Table 17-16 provides a matrix of estimated within-state correlations among the three purpose for reading scales averaged over the 40 jurisdictions for grade 8. In parentheses are the lowest and the highest estimated correlation among the 40 jurisdictions. The listed values, taken directly from the CGROUP program, are estimates of the within-state correlations conditional on the set of principal components included in the conditioning model. For grade 4, the average correlation between Literary Experience and Gain Information is 0.86, with a range of (0.75, 0.93).

Table 17-16
Average Correlations and Ranges of Scale Correlations Among the Reading Scales for 40 Jurisdictions[*] for Grade 8

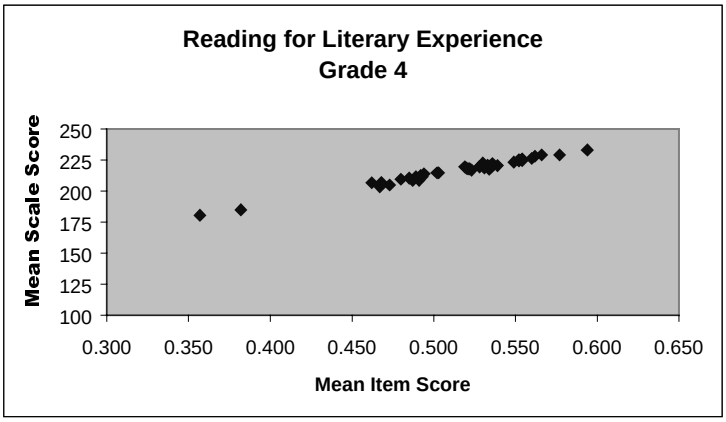
	Literary Experience	**Perform A Task**
Literary Experience	1.0 (1.0)	0.83 (0.66 - 0.95)
Gain Information	0.86 (0.71 - 0.96)	0.88 (0.81 - 0.94)

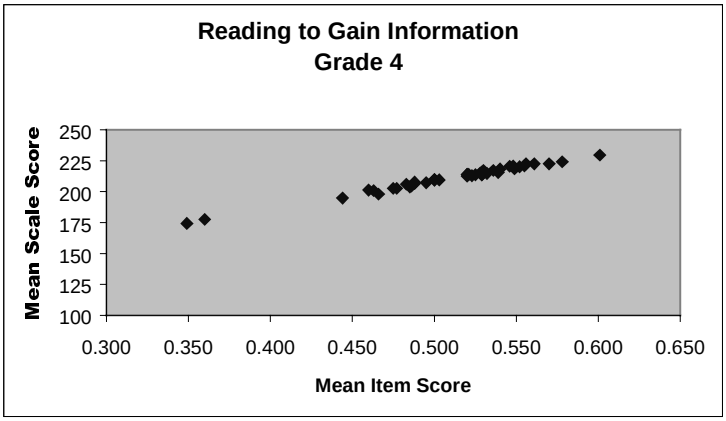
[*] Since Nebraska only had private schools participating, it was not included in the calculation of the average correlation.

As discussed in Chapter 12, NAEP scales are viewed as summaries of consistencies and regularities that are present in item-level data. Such summaries should agree with other reasonable summaries of the item-level data. In order to evaluate the reasonableness of the scaling and estimation results, a variety of analyses were conducted to compare state-level and subgroup-level performance in terms of the content-area scale scores and in terms of the average proportion correct for the set of items in a content area. High agreement was found in all of these analyses. One set of such analyses is presented in Figures 17-7 and 17-8. The figures contain scatterplots of the state scale score mean (mean scale score) versus the state item score means, for each of the two reading content areas and the composite scale for grade 4 and the three reading content areas and the composite scale for grade 8. As is evident from the figures, there is an extremely strong relationship between the estimates of state-level performance in the scale score and item score metrics for both figures.

Figure 17-7

Plot of Mean Scale Score Versus Mean Item Score by Jurisdiction, Grade 4





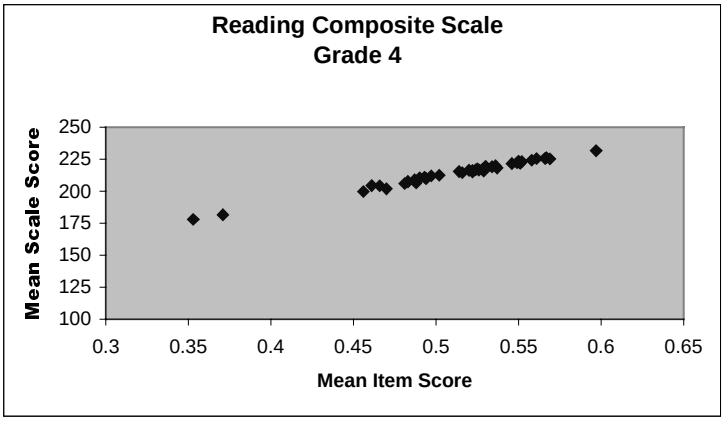
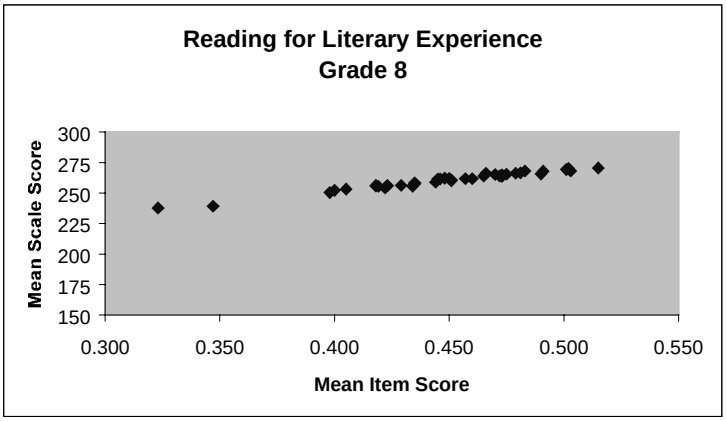
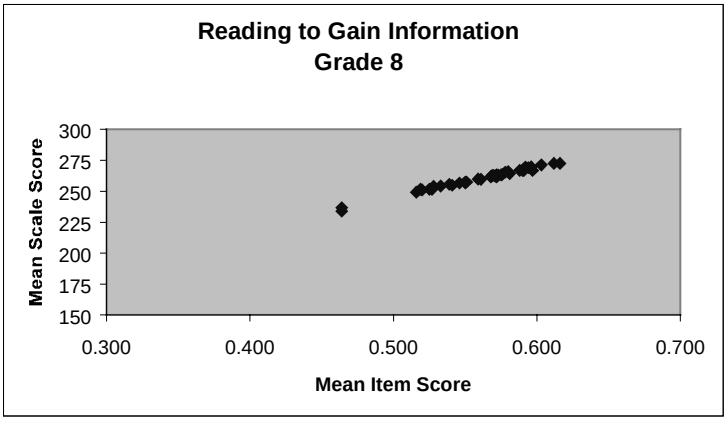
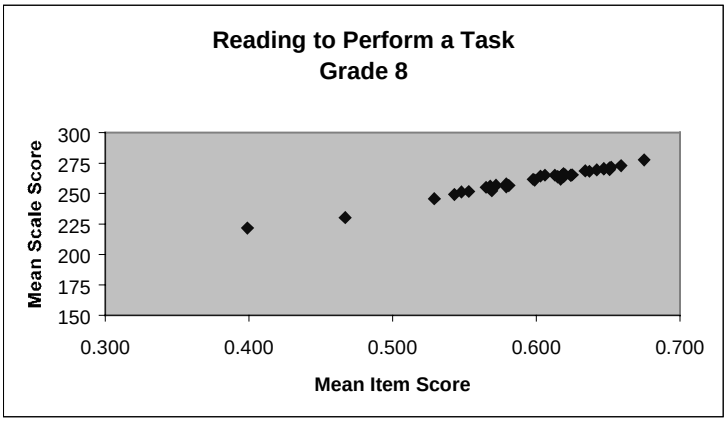


Figure 17-8

Plot of Mean Scale Score Versus Mean Item Score by Jurisdiction, Grade 8

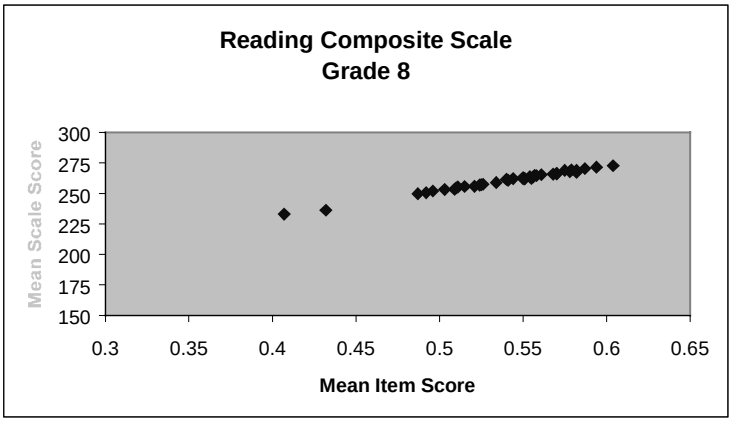






(continued)

Figure 17-8 (continued)
Plot of Mean Scale Score Versus Mean Item Score by Jurisdiction, Grade 8



17.5 THE FINAL SCORE SCALES

17.5.1 Linking State and National Scales

A major purpose of the state assessment program was to allow each participating jurisdiction to compare its 1998 results with the nation as a whole and with the region of the country in which that jurisdiction is located. Although the students in the 1998 state reading assessment were administered the same test booklets as the fourth- and eighth-graders in the national assessment, separate state and national scalings were carried out (for reasons explained in Mazzeo, 1991, and Yamamoto & Mazzeo, 1992). Again, to ensure a similar scale unit system for the state and national metrics, the scales had to be linked.

For meaningful comparisons to be made between each of the state assessment jurisdictions and the relevant national samples, results from these two assessments had to be expressed in terms of a similar system of scale units. The purpose of this section is to describe the procedures used to align the 1998 state assessment scales with their 1998 national counterparts. The procedures that were used represent an extension of the common population equating procedures employed to link the previous national and state scales (Mazzeo, 1991; Yamamoto & Mazzeo, 1992).

Using the house sampling weights provided by Westat (see Section 15.5), the combined sample of students from all participating jurisdictions was used to estimate the distribution of scale scores for the population of students enrolled in public schools that participated in the state assessment.[4] The total sample sizes were 104,129 for the fourth-graders, and 94,429 for the eighth-graders. A subsample of the fourth- grade national sample, consisting of grade-eligible public-school students from any of the 44 jurisdictions that participated in the 1998 state assessment, was used to obtain estimates of the distribution of scale scores for the same target population. A subsample of the eighth-grade national sample, consisting of the students from any of the 41 jurisdictions that participated in the 1998 state assessment, was used to obtain estimates of the distribution of scale scores for the same target population. This subsample of national data is referred to as the national linking sample (NL).[5] Again,

[4] Students from Virgin Islands, DoDEA/DDESS, and DoDEA/DoDDS schools were excluded from the state aggregate sample for purposes of linking.

[5] Note that in previous state assessments, the national linking sample was called the state aggregate comparison, or SAC, sample. Many people thought this was easy to confuse with state data, so the term “national linking” is used in this report.

appropriate weights provided by Westat were used. Thus, for each scale, two sets of scale score distributions were obtained and used in the linking process. One set, based on the sample of combined data from the state assessment (referred to as the state aggregate, or SA) and using item parameter estimates and conditioning results from that assessment, was in the metric of the 1998 state assessment. The other, based on the NL sample from the 1998 national assessment and obtained using item parameters and conditioning results from the national assessment, was in the reporting metric of the 1998 national assessment. The state assessment and national scales, two for grade 4 and three for grade 8, were made comparable by constraining the mean and standard deviation of the two sets of estimates to be equal.

More specifically, the following steps were followed to linearly link the scales of the two assessments:

1) For each scale, estimates of the scale score distribution for the SA sample was obtained using the full set of plausible values generated by the CGROUP program. The weights used were the final (reporting sample) sampling weights provided by Westat (see Section 15.5). For each scale, the arithmetic mean of the five sets of plausible values was taken as the overall estimated mean and the arithmetic average of the standard deviations of the five sets of plausible values was taken as the overall estimated standard deviation.

2) For each scale, the estimated scale score distribution of the NL sample was obtained, again using the full set of plausible values generated by the CGROUP program. The weights used were specially provided by Westat to allow for the estimation of scale score distributions for the same target population of students estimated by the jurisdiction data. The means and standard deviations of the distributions (in the 1998 national reporting metric) for each scale were obtained for this sample in the same manner as described in Step 1.

3) For each scale, a set of linear transformation coefficients was obtained to link the state scale to the corresponding national scale. The linking was of the form

$$\theta^* = A \bullet \theta + B$$

where

θ = a scale score level in terms of the system of units of the provisional BILOG/PARSCALE scale of the state assessment scaling

θ^* = a scale score level in terms of the system of units comparable to those used for reporting the 1998 national reading results

A = $[\text{Standard Deviation}_{NL}]/[\text{Standard Deviation}_{SA}]$

B = $\text{Mean}_{NL} - A \bullet [\text{Mean}_{SA}]$

where the subscripts refer to the NL sample and to the SA sample.

The final conversion parameters for transforming plausible values from the provisional BILOG/PARSCALE scales to the final state assessment reporting scales are given in Table 17-17. All state assessment results are reported in terms of the Y^* metric.

Table 17-17
Coefficients of Linear Transformations for the 1998 State Reading Assessment

Grade	Field of Reading Scale	A	B
4	Literary Experience	39.66	216.15
	Gain Information	38.88	211.09
8	Literary Experience	31.55	260.11
	Gain Information	35.89	259.25
	Perform a Task	38.33	261.11

As is evident from the discussion above, a linear method was used to link the scales from the state and national assessments. While these linear methods ensure equality of means and standard deviations for the SA (after transformation) and the NL samples, they do not guarantee the shapes of the estimated scale score distributions for the two samples will be the same. As these two samples are both from a common target population, estimates of the scale score distribution of that target population based on each of the samples should be quite similar in shape in order to justify strong claims of comparability for the state and national scales. Substantial differences in the shapes of the two estimated distributions would result in differing estimates of the percentages of students above achievement levels or of percentile locations depending on whether state or national scales were used—a clearly unacceptable result given claims about the comparability of the scales. In the face of such results, nonlinear linking methods would be required.

Analyses were carried out to verify the degree to which the linear linking process described above produced comparable scales for state and national results. Comparisons were made between two estimated scale score distributions, one based on the SA sample and one based on the NL sample, for each of the three fields of reading scales. The comparisons were carried out using slightly modified versions of what Wainer (1974) refers to as suspended rootograms. The final reporting scales for the state and national assessments were each divided into 10-point intervals. Two sets of estimates of the percentage of students in each interval were obtained, one based on the SA sample and one based on the NL sample. Following Tukey (1977), the square roots of these estimated percentages were compared.[6]

The comparisons are shown in Figures 17-9 through 17-13. The height of each of the unshaded bars corresponds to the square root of the percentage of students from the state assessment aggregate sample in each 10-point interval on the final reporting scale. The shaded bars show the differences in root percents between the SA and NL estimates. Positive differences indicate intervals in which the estimated percentages from the NL sample are lower than those obtained from the SA. Conversely, negative differences indicate intervals in which the estimated percentages from the NL sample are higher. For all three scales, differences in root percents are quite small, suggesting that the shapes of the two estimated distributions are quite similar (i.e., unimodal with small positive coefficient of skewness). There is some evidence that the estimates produced using the NL data are slightly heavier in the extreme upper tails (above 400 for Literary reading and Information reading for grade 4; above 350 for Literary reading, above 380 for Information reading, and above 400 for Perform a Task for grade 8). However, even these differences at the extremes are small in magnitude (0.2 in the root percent metric and 0.09 in the percent metric) and have little impact on estimates of reported statistics such as percentages of students above the achievement levels.

[6] The square root transformation allows for more effective comparisons for counts (or equivalently, percentages) when the expected number of counts in each interval is likely to vary greatly over the range of intervals, as is the case for the NAEP scales where the expected counts of individuals in intervals near the extremes of the scale (e.g., below 150 and above 350) are dramatically smaller than the counts obtained near the middle of the scale.

Figure 17-9
Rootogram Comparing Scale Score Distributions
for the State Assessment Aggregate Sample
and the National Linking Sample
for the Reading for Literary Experience Scale, Grade 4

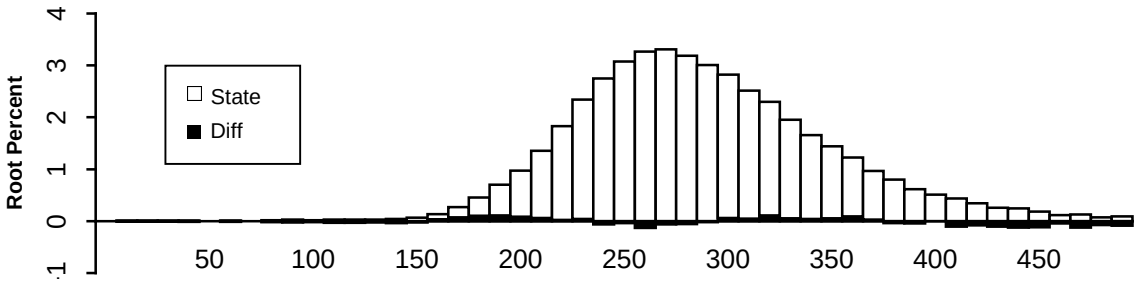


Figure 17-10
Rootogram Comparing Scale Score Distributions
for the State Assessment Aggregate Sample
and the National Linking Sample
for the Reading to Gain Information Scale, Grade 4

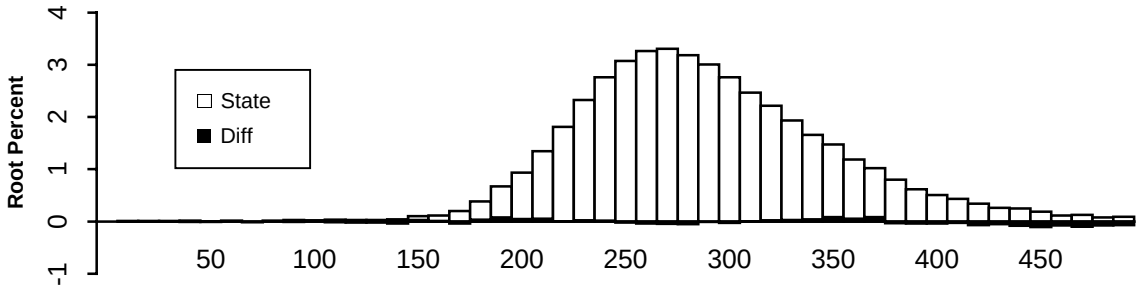


Figure 17-11
Rootogram Comparing Scale Score Distributions
for the State Assessment Aggregate Sample
and the National Linking Sample
for the Reading for Literary Experience Scale, Grade 8

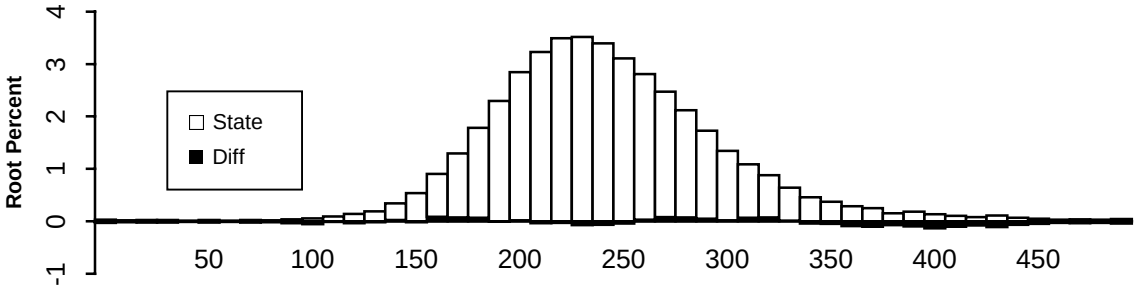


Figure 17-12
Rootogram Comparing Scale Score Distributions
for the State Assessment Aggregate Sample
and the National Linking Sample
for the Reading to Gain Information Scale, Grade 8

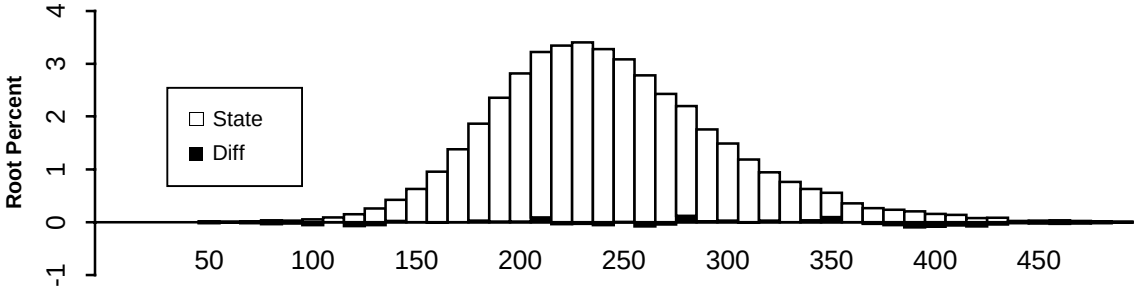
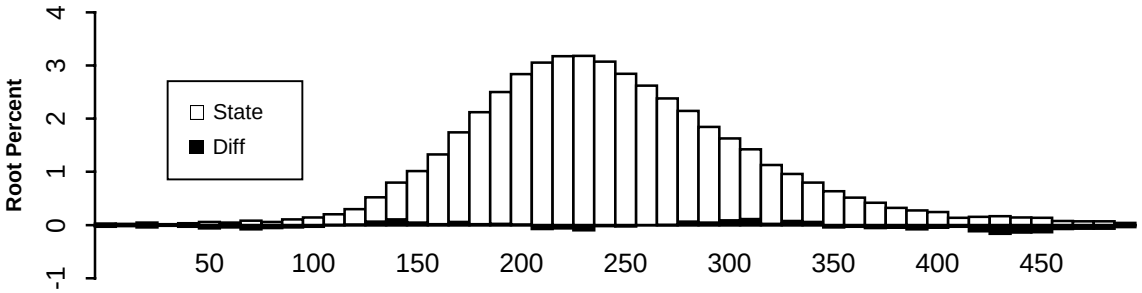


Figure 17-13
Rootogram Comparing Scale Score Distributions
for the State Assessment Aggregate Sample
and the National Linking Sample
for the Reading to Perform a Task Scale, Grade 8



17.5.2 Producing a Reading Composite Scale

For the national assessment, a composite scale was created for the fourth, eighth, and twelfth grades as an overall measure of reading scale scores for students at that grade. The composite was a weighted average of plausible values on the purpose-for-reading scales (Reading for Literary Experience, Reading to Gain Information, and at grades 8 and 12, Reading to Perform a Task). The weights for the national fields of reading scale scales were proportional to the relative importance assigned to each field of reading scale in each grade in the assessment specifications developed by the Reading Objectives Panel. Consequently, the weights for each of the fields of reading scales are similar to the actual proportion of items from that field of reading scale.

State assessment composite scales for grades 4 and 8 were developed using weights identical to those used to produce the composites for the 1998 national reading assessment. The weights are given in Table 16-14. In developing the state assessment composite, the weights were applied to the plausible values for each field of reading scale as expressed in terms of the final state assessment scales (i.e., after transformation from the provisional BILOG/PARSCALE scales.)

Figures 17-14 and 17-15 provide rootograms comparing the estimated scale score distributions based on the SA and NL samples for the grade 4 and grade 8 composites. Consistent with the results presented separately by scale, there is some evidence that the estimates produced using the NL are slightly heavier in the upper tails than the corresponding estimate based on the SA samples. Again however, these differences in root relative percents are small in magnitude.

Figure 17-14
Rootogram Comparing Scale Score Distributions
for the State Assessment Aggregate Sample
and the National Linking Sample
for the Reading Composite Scale, Grade 4

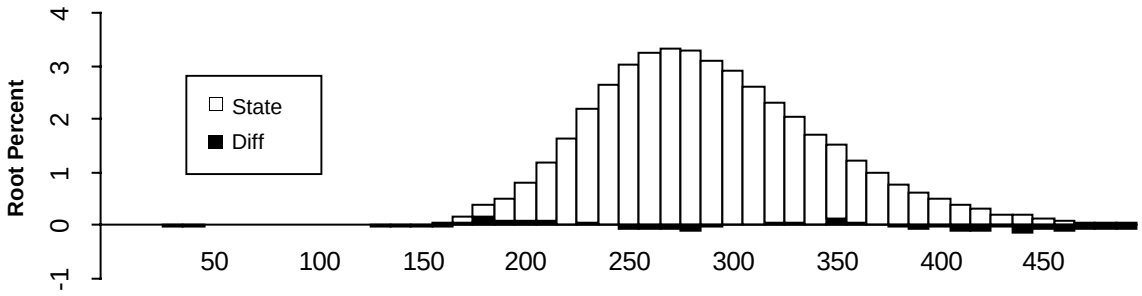
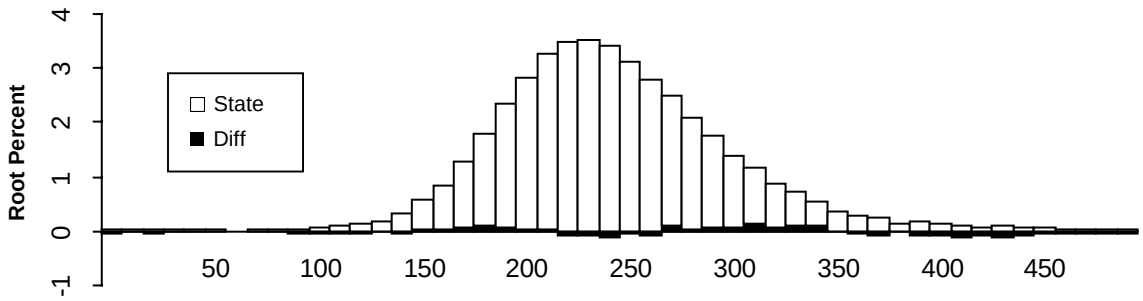


Figure 17-15
Rootogram Comparing Scale Score Distributions
for the State Assessment Aggregate Sample
and the National Linking Sample
for the Reading Composite Scale, Grade 8



17.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE

For each grade in state reading assessments, the error variance of the final transformed scale score mean was partitioned as described in Chapter 12. The partition of error variance consists of two parts: the proportion of error variance due to sampling students (sampling variance) and the proportion of error variance due to the fact that scale score, θ, is a latent variable that is estimated rather than observed. For grades 4 and 8, Tables 17-18 and 17-19 contain estimates of the total error variance, the proportion of error variance due to sampling students, and the proportion of error variance due to the latent nature of θ. Instead of using 100 plausible values as in national assessment, the calculations for the state samples are based on 5 plausible values. More detailed information is available for gender and race/ethnicity subgroups in Appendix H.

17.7 READING TEACHER QUESTIONNAIRES

Teachers of fourth- and eighth-grade students were surveyed about their educational background and teaching practices. The students were matched first with their reading teacher, and then the specific classroom period. Variables derived from the questionnaire were used in the conditioning models. An additional conditioning variable was included that indicated whether the student had been matched with a teacher record. This contrast controlled estimates of subgroup means for differences that exist between matched and nonmatched students. Of the 112,138 fourth-grade students in the sample, 105,026 (93.7%, unweighted) were matched with teachers who answered both parts of the teacher questionnaire, and 13 of the students had teachers who answered only the teacher background section of the questionnaire. For the eighth-grade sample, 82,118 of the 94,429 students (87%, unweighted) were matched to both sections of the teacher questionnaire. There were 6,575 students (7%, unweighted) who were matched with the first part of the teacher questionnaire, but could not be matched to the appropriate classroom period. Thus, 93.7 percent of the fourth-graders and 94 percent of the eighth-graders were matched with at least the background information about their reading teacher.

Table 17-18

Estimation Error Variance and Related Coefficients for the Reading State Assessment, Grade 4

State	Total Estimation Error Variance	Proportion of Variance due to …	
		Student Sampling	Latency of θ
Alabama	3.197	0.94	0.06
Arizona	4.062	0.97	0.03
Arkansas	2.208	0.93	0.07
California	10.325	0.96	0.04
Colorado	1.721	0.94	0.06
Connecticut	3.425	0.93	0.07
Delaware	1.637	0.57	0.43
Florida	2.128	0.96	0.04
Georgia	2.519	0.95	0.05
Hawaii	3.085	0.66	0.34
Iowa	1.397	0.97	0.03
Kansas	2.173	0.89	0.11
Kentucky	2.218	0.81	0.19
Louisiana	2.254	0.98	0.02
Maine	1.529	0.72	0.28
Maryland	2.656	0.97	0.03
Massachusetts	1.965	0.89	0.11
Michigan	2.755	0.94	0.06
Minnesota	2.195	0.89	0.11
Mississippi	2.123	0.98	0.02
Missouri	2.762	0.96	0.04
Montana	2.774	0.59	0.41
Nevada	1.855	0.93	0.07
New Hampshire	1.783	0.76	0.24
New Mexico	4.089	0.79	0.21
New York	2.639	0.89	0.11
North Carolina	1.804	0.89	0.11
Oklahoma	1.286	0.92	0.08
Oregon	2.644	0.94	0.06
Rhode Island	3.018	0.84	0.16
South Carolina	1.648	0.91	0.09
Tennessee	2.224	0.95	0.05
Texas	4.493	0.97	0.03
Utah	1.775	0.86	0.14
Virginia	1.777	0.97	0.03
Washington	1.791	0.97	0.03
West Virginia	2.205	0.96	0.04
Wisconsin	1.322	0.95	0.05
Wyoming	2.624	0.47	0.53
District of Columbia	1.971	0.38	0.62
DoDEA/DDESS	1.702	0.32	0.68
DoDEA/DoDDS	1.208	0.57	0.43
Virgin Islands	3.779	0.39	0.61

Table 17-19
Estimation Error Variance and Related Coefficients for the Reading State Assessment, Grade 8

State	Total Estimation Error Variance	Proportion of Variance due to …	
		Student Sampling	Latency of θ
Alabama	1.822	0.97	0.03
Arizona	1.394	0.95	0.05
Arkansas	1.753	0.79	0.21
California	2.726	0.96	0.04
Colorado	1.196	0.98	0.02
Connecticut	1.159	0.89	0.11
Delaware	1.626	0.72	0.28
Florida	2.890	0.91	0.09
Georgia	2.052	0.95	0.05
Hawaii	1.745	0.39	0.61
Kansas	1.437	0.94	0.06
Kentucky	1.664	0.98	0.02
Louisiana	2.157	0.95	0.05
Maine	1.389	0.92	0.08
Maryland	3.376	0.82	0.18
Massachusetts	2.435	0.92	0.08
Minnesota	1.672	0.93	0.07
Mississippi	2.054	0.79	0.21
Missouri	1.728	0.85	0.15
Montana	1.291	0.72	0.28
Nevada	1.301	0.95	0.05
New Mexico	1.524	0.79	0.21
New York	2.531	0.91	0.09
North Carolina	1.301	0.85	0.15
Oklahoma	1.631	0.71	0.29
Oregon	2.087	0.91	0.09
Rhode Island	0.925	0.89	0.11
South Carolina	1.756	0.93	0.07
Tennessee	1.679	0.91	0.09
Texas	2.142	0.99	0.01
Utah	1.123	0.78	0.22
Virginia	1.232	0.90	0.10
Washington	1.639	0.88	0.12
West Virginia	1.417	0.88	0.12
Wisconsin	2.466	0.91	0.09
Wyoming	1.734	0.58	0.42
District of Columbia	3.846	0.30	0.70
DoDEA/DDESS	10.719	0.24	0.76
DoDEA/DoDDS	1.054	0.44	0.56
Virgin Islands	8.264	0.26	0.74