Chapter 18

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

Elissa A. Greenwald and Terry L. Schoeps
Educational Testing Service

18.1 INTRODUCTION

The framework that was used for the 1998 NAEP writing assessment detailed the structure of the assessment to be given at grades 4, 8, and 12 at the national level and at grade 8 at the state level. The framework was developed under contract by the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and American College Testing (ACT) for the National Assessment Governing Board (NAGB) in 1996. The framework for the writing assessment is available on the National Assessment Governing Board (NAGB) web site at *http://www.nagb.org*.

Sections 18.2 through 18.5 explain the development of the framework, objectives, and items for the 1998 NAEP writing assessment. Section 18.8 also describes the student background questionnaires and the writing teacher questionnaire. Additional information on the structure and content of assessment booklets can be found in Section 18.9. Various committees worked on the development of the framework, objectives, and items for the writing assessment. The list of committee members and consultants 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 Writing Report Card for the Nation and the States* (Greenwald, Persky, Campbell, & Mazzeo, 1999).

18.2 DEVELOPING THE WRITING ASSESSMENT FRAMEWORK

NAGB is responsible for setting policy for NAEP; this policy-making 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.

NAGB began the development process for the 1998 writing objectives by convening a writing framework panel. The panel solicited recommendations from members of the academic and business communities, from state and local government representatives, from members of the press, and from the general public. After reviewing the responses, the panel designed the framework.

For more detail on the development and specifications of the writing framework, refer to the *Writing Framework and Specifications for the 1998 National Assessment of Educational Progress, 1992–1998* (NAGB, 1996b).

[1] Elissa A. Greenwald managed the item-development process for the 1998 NAEP writing assessment. Terry L. Schoeps coordinates the production of NAEP technical reports.

18.3 WRITING FRAMEWORK AND ASSESSMENT DESIGN PRINCIPLES

The writing framework was designed to focus on writing processes and outcomes, rather than to reflect a particular instructional or theoretical approach. The framework focuses not on the specific writing skills that lead to outcomes, but rather on the quality of the outcomes themselves. The framework was intended to embody a broad view of writing by addressing the increasingly higher level of literacy needed for employment, personal development, and good citizenship. The people who designed the framework also relied on contemporary writing research and sought to use nontraditional assessment formats that resemble desired classroom activities to the extent possible within the constraints of a timed assessment.

The development of the framework objectives was guided by the consideration that the assessment should reflect many of the curricular emphases and objectives in various states, localities, and school districts, as well as what various scholars, practitioners, and interested citizens believed should be included in the assessment. Under contract to NAGB, ACT developed the test specifications to address overarching objectives of the 1998 writing assessment framework:

- Write for a variety of purposes—narrative, informative, and persuasive
- Write on a variety of tasks and for many different audiences
- Write from a variety of stimulus materials and within various time constraints
- Generate, draft, revise, and edit ideas and forms of expression in their writing
- Display effective choices in the organization of their writing
- Value writing as a communicative activity

18.4 FRAMEWORK FOR THE 1998 WRITING ASSESSMENT

The 1998 writing assessment framework was organized according to three *purposes for writing*:

- Narrative
- Informative
- Persuasive

Narrative writing tasks require students to produce a story or personal essay. Informative writing tasks focus primarily on the subject-matter element in communication. Informative writing is used to share knowledge and to convey messages, instructions, and ideas. In persuasive writing, the primary aim is to influence others to take some action or to bring about change. This type of writing involves a clear awareness of what arguments might most affect the audience being addressed. Further explanation of the purposes is contained in Figure 18-1.

The cognitive portion of the writing assessment included only constructed-response exercises. These tasks were designed to measure students' abilities to write for a variety of purposes and to a diverse set of audiences. To accomplish these goals, a wide variety of stimulus materials were used in the assessment. The first step in the development effort was the identification of appropriate stimulus materials that would allow the construction of tasks that would, in aggregate, measure the range of writing outcomes described in the framework.

Figure 18-1
*Description of NAEP 1998 Writing Purposes**

Narrative

Narrative writing involves the production of stories or personal essays. Practice with these forms helps writers to develop an ear for language. Also, informative and persuasive writing can benefit from many of the strategies used in narrative writing. For example, there must be an effective ordering of events when relating an incident as part of a report. Sometimes narrative writing contributes to an awareness of the world as the writer creates, manipulates, and interprets reality. Such writing—whether fact or fiction, poem, play, or personal essay—requires close observation of people, objects, and places. Further, this type of writing fosters creativity, imagination, and speculation by allowing the writer to express thoughts and then stand back, as a more detached observer might, and grasp more fully what is being felt and why. Thus, narrative writing offers a special opportunity to analyze and understand emotions and actions.

Informative

Informative writing focuses primarily on the subject-matter element in communication. This type of writing is used to share knowledge and to convey messages, instructions, and ideas. Like all writing, informative writing may be filtered through the writer's impressions, understanding, and feelings. Used as a means of exploration, informative writing helps both the writer and the reader to learn new ideas and to reexamine old conclusions. Informative writing may also involve reporting on events or experiences, or analyzing concepts and relationships, including developing hypotheses and generalizations. Any of these types of informative writing can be based on the writer's personal knowledge and experience or on information newly presented to the writer that must be understood in order to complete a task. Usually, informative writing involves a mix of the familiar and the new, and both are clarified in the process or writing. Depending on the task, writing based on either personal experience or secondary information may span the range of thinking skills from recall to analysis to evaluation.

Persuasive

Persuasive writing emphasizes the reader. Its primary aim is to influence others to take some action or bring about change. Persuasive writing may contain great amounts of information—facts, details, examples, comparisons, statistics, or anecdotes—but its main purpose is not simply to inform but to persuade. This type of writing involves a clear awareness of what arguments might most affect the audience being addressed. Writing persuasively also requires use of critical thinking skills such as analysis, inference, synthesis, and evaluation.

Persuasive writing is called for in a variety of situations. It may involve responding to a request for advice by giving an opinion and providing sound reasons to support it. It may also involve presenting an argument in such a way that a particular audience will find it convincing. When there is opposition, persuasive writing may entail refuting arguments that are contrary to the writer's point of view.

In all persuasive writing, authors must choose the approach they will use. They may, for instance, use emotional or logical appeals or an accommodating or demanding tone. Regardless of the situation or approach, persuasive writers must be concerned with having a particular desired effect on their readers, beyond merely adding to knowledge of the topic presented.

* The text in Figure 18-1 is from the *Writing Framework and Specifications for the 1998 National Assessment of Educational Progress, 1992–1998* (NAGB, 1996b), developed under contract by the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and American College Testing (ACT) for the National Assessment Governing Board (NAGB) in 1996.

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

The distribution of items by writing purpose across grade levels recommended in the assessment framework is provided in Table 18-1.

Table 18-1
Percentage Distribution of Items by Purpose for Writing as Specified in the NAEP Writing Framework

	Purposes for Writing		
Grade	**Narrative**	**Informative**	**Persuasive**
4	40%	35%	25%
8*	33%	33%	33%
12	25%	35%	40%

* The grade 8 percentages shown in this table do not total 100% because the numbers have been rounded.

The writing framework also discusses the ways in which the assessment tasks should be scored. Students' responses to each writing task were evaluated by trained raters who used scoring guides that emphasized development, organization, and control of language.

18.5 DEVELOPING THE WRITING COGNITIVE ITEMS

The assessment included 25-minute and 50-minute writing tasks (referred to as "blocks" in test development). Students were asked to respond to either two 25-minute writing tasks or one 50-minute writing task (for some students at grades 8 and 12). In accordance with the framework objective to include writing on a variety of tasks and for many different audiences, students were asked to write in a variety of forms. Some of the forms in which students were asked to write (across the tasks in the assessment) are listed in Figure 18-2.

Figure 18-2
NAEP 1998 Forms of Writing

Story
Essay
Letter to Authority
Letter to a Friend
Article
Report
Speech

18.6 DEVELOPING THE WRITING OPERATIONAL FORMS

Writing field tests were conducted in October and November of 1997 and involved national samples of fourth-, eighth-, and twelfth-grade students. More than 100 items were field tested across the three grades.

The field-test data were collected, scored, and analyzed in preparation for meetings with the Writing Instrument Development Committee. Committee members, ETS test-development staff, and NAEP/ETS staff reviewed the materials and chose the 66 writing tasks used in the operational assessment. The objectives that guided these reviews included determining

- which tasks were most related to overall student achievement;
- the need for revisions of tasks that lacked clarity or had ineffective formats; and
- which tasks could be scored with the highest levels of interrater reliability.

The tasks were chosen according to the distributions of narrative, informative, and persuasive writing tasks specified in the framework. Once the committees had selected the tasks, all tasks were rechecked for content, measurement, and sensitivity concerns. Finally, a clearance package was submitted to NCES. Throughout the clearance process, revisions were made in accordance with changes required by the government. Upon approval, the tasks (assembled into booklets) and questionnaires were ready for printing.

The 50-minute tasks that were administered at grades 8 and 12 were not administered as part of the state assessment.

18.7 DISTRIBUTION OF WRITING ASSESSMENT ITEMS

At grade 4, all tasks were 25-minute writing tasks; eight measured narrative writing, seven measured informative writing, and six measured persuasive writing. Of the 25-minute tasks administered at grade 8, seven measured narrative writing, seven measured informative writing, and six measured persuasive writing. At grade 12, of the 25-minute tasks, five measured narrative writing, seven measured informative writing, and eight measured persuasive writing. At grades 8 and 12, three 50-minute tasks were given—one for each writing purpose. The 50-minute tasks were administered in the national assessment but were not given in the state assessment.

Tables 18-2 through 18-4 provide the title and writing purpose of each writing task administered.

Table 18-2

NAEP 1998 Writing Grade 4 Blocks by Title and Writing Purpose

Writing Block Title	Block	Purpose
Aunt Dot	W3	Narrative
Cartoon Story	W4	Narrative
Very Unusual Day	W5	Narrative
Castle	W6	Narrative
Casey and Duke	W7	Narrative
Old Tree	W8	Narrative
Secret Door	W9	Narrative
Mr. Tooms	W10	Narrative
Letter from TX8	W11	Informative
Letter from MZ3	W12	Informative
Letter from Lilex	W13	Informative
Animal Lesson	W14	Informative
City Scenes	W15	Informative
Unusual Animal	W16	Informative
Favorite Object *	W17	Informative
Invisible Friend	W18	Persuasive
Day Trip*	W19	Persuasive
Class Pet	W20	Persuasive
Library Book	W21	Persuasive
Child or Adult	W22	Persuasive

* This block appeared in booklets administered to students requiring accommodations.

Table 18-3

NAEP 1998 Writing Grade 8 Blocks by Title and Writing Purpose

Writing Block Title	Block	Purpose
Cartoon Story	W3	Narrative
President for a Day	W4	Narrative
Plums	W5	Narrative
Tower	W6	Narrative
Principal for a Day*	W7	Narrative
Pioneer Journal	W8	Narrative
Space Visitor	W9	Narrative
Ancient Tree†	W10	Narrative
Performance Review	W11	Informative
New Park	W12	Informative
Dream Weekend	W13	Informative
Backpack	W14	Informative
Designing a TV Show	W15	Informative
Save a Book	W16	Informative
Life's Lessons	W17	Informative
Vandalism†	W18	Informative
Lengthening the School Year*	W19	Persuasive
School Schedule	W20	Persuasive
Fast Food	W21	Persuasive
Class Trip	W22	Persuasive
Driving Age	W23	Persuasive
Teens in Malls	W24	Persuasive
Student of the Year†	W25	Persuasive

* This block appeared in booklets administered to students requiring accommodations.

† This was a 50-minute block and was not part of the main national reporting sample.

Table 18-4
NAEP 1998 Writing Grade 12 Blocks by Title and Writing Purpose

Writing Block Title	Block	Purpose
Tall Tale	W3	Narrative
Plums	W4	Narrative
Special Object	W5	Narrative
The Arch	W6	Narrative
Pioneer Journal	W7	Narrative
Ancient Tree*	W8	Narrative
Cafeteria	W9	Informative
Writing Mentor	W10	Informative
Movie Review	W11	Informative
Technology	W12	Informative
Handbook	W13	Informative
Save a Book	W14	Informative
Life's Lessons	W15	Informative
Vandalism†	W16	Informative
Summer Job	W17	Persuasive
Big or Small Inventions	W18	Persuasive
Work Less/Study More	W19	Persuasive
Heroes	W20	Persuasive
One Vote*	W21	Persuasive
Teens in Malls	W22	Persuasive
Driving Age	W23	Persuasive
Person of the Year	W24	Persuasive
Campaign Speech*	W25	Persuasive

* This was a 50-minute block and was not part of the main reporting sample.
† This block appeared in booklets administered to students requiring accommodations.

Each student received an assessment booklet containing a either 25-minute exercises or one 50-minute exercise. Following the exercise or exercises in each booklet were a set of general background questions, a set of subject-specific background questions, and a set of questions about his or her motivation and familiarity with the assessment materials.

In the development process, every effort was made to meet the content targets specified in the assessment framework. Table 18-5 shows the approximate percentage of aggregate assessment time devoted to each purpose for writing, at each grade level. Percentages are based on the classifications agreed on by the Writing Instrument Development Committee. Note that the numbers presented in Table 18-5 differ slightly from those in Table 18-1 in that Table 18-1 (at grade 8 only) shows the distribution of assessment items as specified in the writing framework.

Table 18-5
*Percentage Distribution of Assessment Time by Grade and Purpose for Writing for the NAEP 1998 Writing Assessment**

Grade	Purposes for Writing		
	Narrative	Informative	Persuasive
4	40%	35%	25%
8	35%	35%	30%
12	25%	35%	40%

18.8 BACKGROUND QUESTIONNAIRES FOR THE 1998 WRITING ASSESSMENT

In addition to assessing how well students read, it is important to understand the instructional context in which writing takes place, students' home support for literacy, and students' writing habits and attitudes. To gather contextual information, NAEP assessments include background questions designed to provide insight into factors that may influence writing performance.

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. Members of the Writing Instrument Development Committee were consulted on the appropriateness of the issues addressed in all questionnaires that relate to writing instruction and achievement. Like the writing tasks, all background questions were submitted for extensive review and field testing. Recognizing the validity 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 of general and one five-minute set of subject-specific background questions designed to gather contextual information about students, their instructional and recreational experiences in writing, and their attitudes toward writing. 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 determine students' motivation in completing the assessment and their familiarity with assessment tasks.

18.8.1 Student Writing 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.

Writing Background: Students were asked to report their instructional experiences related to writing in the classroom, including how often their teachers asked them to write more than one draft of a paper and whether or not they or their teachers saved their written work in a folder or portfolio.

Motivation: Students were asked five questions about how hard they tried on the test and about friends' attitudes toward writing.

Table 18-6 gives the number of questions per background section and notes the placement of each within student booklets.

Table 18-6
NAEP 1998 Background Sections of Student Writing Booklets

	Number of Questions	Placement in Student Booklet
Grade 4		
General Background	21	Section 3
Writing Background	17	Section 4
Motivation	5	Section 5
Grade 8		
General Background	22	Section 3
Writing Background	28	Section 4
Motivation	5	Section 5
Grade 12		
General Background	24	Section 3
Writing Background	28	Section 4
Motivation	5	Section 5

18.8.2 Language Arts Teacher Questionnaire

To supplement the information on instruction reported by students, writing teachers of the fourth- and eighth-graders participating in the NAEP writing assessment were asked to complete a questionnaire about characteristics such as their gender, teaching backgrounds, and instructional 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 writing 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 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.

This component of the questionnaire was completed by teachers whose students participated in any subject assessed in NAEP.

The **Teacher Questionnaire, Part IIA: Reading/Writing Preparation** (12 questions at grade 4 and 12 at grade 8) included questions on the teachers' exposure to various issues related to writing instruction through college or university courses or professional-development workshops.

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

- the ability level of students in the class;
- whether students were assigned to the class by ability level;
- time spent weekly on teaching writing and helping students with their writing;
- writing homework assignments;
- frequency of various instructional activities in class;
- methods of assessing student progress in writing;
- instructional emphasis given to the writing abilities covered in the assessment; and
- use of particular resources.

18.9 STUDENT BOOKLETS FOR THE 1998 WRITING ASSESSMENT

At each grade in the assessment, the 25-minute tasks were assembled into 18 booklets. At grades 8 and 12, there were 3 additional booklets containing 50-minute tasks. 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.

The assembly of writing blocks (with one task per block) into booklets and their subsequent assignment to sampled students was determined by a partially balanced incomplete block (PBIB) design with spiraled administration (see Section 1.5). At each grade, the 25-minute tasks were assembled into 40 booklets such that two different blocks were assigned to each booklet and each block appeared in four booklets. Tables 18-6 through 18-8 show this configuration. At all grades, every 25-minute task appears in four booklets. This is the partially balanced part of the balanced incomplete block design. Every 50-minute task appears only in one booklet (although booklets containing the 50-minute tasks are included in the main national assessment, they cannot be assembled in the PBIB fashion).

The focused PBIB design also balances the order of presentation of the 25-minute blocks—every 25-minute block appears as the first cognitive task in two booklets and as the second cognitive task in two other booklets. This design allows for some control of context and fatigue effects.

As in the other subjects, the final step in the 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 18-7, 18-8, and 18-9 detail the configuration of booklets administered in the 1998 writing assessment.

18.10 WRITING CLASSROOM-BASED STUDY IN 1998

In 1998, NAEP conducted a special study designed to explore methods of assessing students' writing abilities by using written assignments that students had completed as part of their school curriculum. A full report on this study is due to be published in the year 2000.

Table 18-7

NAEP 1998 National and State Writing Grade 4 Booklet Configuration

Booklet Number	Question Block 1	Question Block 2	Common Core Background	Writing Background	Motivation
201	W4	W16	CW	WB	WA
202	W16	W11	CW	WB	WA
203	W11	W3	CW	WB	WA
204	W3	W18	CW	WB	WA
205	W18	W19	CW	WB	WA
206	W19	W20	CW	WB	WA
207	W20	W12	CW	WB	WA
208	W12	W7	CW	WB	WA
209	W7	W21	CW	WB	WA
210	W21	W22	CW	WB	WA
211	W22	W18	CW	WB	WA
212	W18	W14	CW	WB	WA
213	W14	W5	CW	WB	WA
214	W5	W19	CW	WB	WA
215*	W19	W17	CW	WB	WA
216	W17	W6	CW	WB	WA
217	W6	W20	CW	WB	WA
218	W20	W21	CW	WB	WA
219	W21	W15	CW	WB	WA
220	W15	W8	CW	WB	WA
221	W8	W22	CW	WB	WA
222	W22	W13	CW	WB	WA
223	W13	W9	CW	WB	WA
224	W9	W4	CW	WB	WA
225	W4	W3	CW	WB	WA
226	W3	W5	CW	WB	WA
227	W5	W6	CW	WB	WA
228	W6	W7	CW	WB	WA
229	W7	W8	CW	WB	WA
230	W8	W9	CW	WB	WA
231	W9	W10	CW	WB	WA
232	W10	W11	CW	WB	WA
233	W11	W14	CW	WB	WA
234	W14	W17	CW	WB	WA
235	W17	W12	CW	WB	WA
236	W12	W15	CW	WB	WA
237	W15	W13	CW	WB	WA
238	W13	W16	CW	WB	WA
239	W16	W10	CW	WB	WA
240	W10	W4	CW	WB	WA

* Booklet number 215 was an accommodations booklet. Accommodations booklets contain type that is larger than the type used in other booklets; they are given to participating students who have a visual disability.

Table 18-8

NAEP 1998 National and State Writing Grade 8 Booklet Configuration

Booklet Number	Question Block 1	Question Block 2	Common Core Background	Writing Background	Motivation
201	W3	W4	CW	WB	WA
202	W4	W5	CW	WB	WA
203	W5	W6	CW	WB	WA
204	W6	W7	CW	WB	WA
205	W7	W8	CW	WB	WA
206	W8	W9	CW	WB	WA
207	W9	W13	CW	WB	WA
208	W13	W19	CW	WB	WA
209*	W19	W7	CW	WB	WA
210	W7	W14	CW	WB	WA
211	W14	W21	CW	WB	WA
212	W21	W5	CW	WB	WA
213	W5	W12	CW	WB	WA
214	W12	W17	CW	WB	WA
215	W17	W23	CW	WB	WA
216	W23	W20	CW	WB	WA
217	W20	W21	CW	WB	WA
218	W21	W22	CW	WB	WA
219	W22	W19	CW	WB	WA
220	W19	W24	CW	WB	WA
221	W24	W8	CW	WB	WA
222	W8	W15	CW	WB	WA
223	W15	W22	CW	WB	WA
224	W22	W6	CW	WB	WA
225	W6	W16	CW	WB	WA
226	W16	W20	CW	WB	WA
227	W20	W4	CW	WB	WA
228	W4	W11	CW	WB	WA
229	W11	W12	CW	WB	WA
230	W12	W16	CW	WB	WA
231	W16	W14	CW	WB	WA
232	W14	W15	CW	WB	WA
233	W15	W13	CW	WB	WA
234	W13	W17	CW	WB	WA
235	W17	W11	CW	WB	WA
236	W11	W9	CW	WB	WA
237	W9	W3	CW	WB	WA
238	W3	W24	CW	WB	WA
239	W24	W23	CW	WB	WA
240	W23	W3	CW	WB	WA
241	——————W10†——————		CW	WB	WA
242	——————W18†——————		CW	WB	WA
243	——————W25†——————		CW	WB	WA

* Booklet number 209 was an accommodations booklet. Accommodations booklets contain type that is larger than the type used in other booklets; they are given to participating students who have a visual disability.

† Booklets containing blocks W10, W18, and W25 were booklets that contained 50-minute tasks.

Table 18-9

NAEP 1998 National and State Writing Grade 12 Booklet Configuration

Booklet Number	Question Block 1	Question Block 2	Common Core Background	Writing Background	Motivation
201	W3	W4	CW	WB	WA
202	W4	W5	CW	WB	WA
203	W5	W6	CW	WB	WA
204	W6	W7	CW	WB	WA
205	W7	W23	CW	WB	WA
206	W23	W15	CW	WB	WA
207	W15	W9	CW	WB	WA
208	W9	W10	CW	WB	WA
209	W10	W11	CW	WB	WA
210	W11	W12	CW	WB	WA
211	W12	W13	CW	WB	WA
212	W13	W14	CW	WB	WA
213	W14	W15	CW	WB	WA
214	W15	W17	CW	WB	WA
215	W17	W18	CW	WB	WA
216	W18	W19	CW	WB	WA
217	W19	W20	CW	WB	WA
218	W20	W21	CW	WB	WA
219	W21	W22	CW	WB	WA
220	W22	W23	CW	WB	WA
221	W23	W24	CW	WB	WA
222	W24	W9	CW	WB	WA
223	W9	W17	CW	WB	WA
224	W17	W24	CW	WB	WA
225	W24	W18	CW	WB	WA
226	W18	W10	CW	WB	WA
227	W10	W3	CW	WB	WA
228	W3	W19	CW	WB	WA
229	W19	W11	CW	WB	WA
230	W11	W4	CW	WB	WA
231	W4	W20	CW	WB	WA
232	W20	W12	CW	WB	WA
233	W12	W5	CW	WB	WA
234*	W5	W21	CW	WB	WA
235	W21	W13	CW	WB	WA
236	W13	W6	CW	WB	WA
237	W6	W22	CW	WB	WA
238	W22	W14	CW	WB	WA
239	W14	W7	CW	WB	WA
240	W7	W3	CW	WB	WA
241	——————W8† ——————		CW	WB	WA
242	——————W16† ——————		CW	WB	WA
243	——————W25† ——————		CW	WB	WA

* Booklet number 234 was an accommodations booklet. Accommodations booklets contain type that is larger than the type used in other booklets; they are given to participating students who have a visual disability.

† Booklets containing blocks W8, W16, and W25 were booklets that contained 50-minute tasks.

Chapter 19

INTRODUCTION TO THE DATA ANALYSIS FOR THE NATIONAL AND STATE WRITING SAMPLES[1]

Frank Jenkins, Jiahe Qian, Hua-Hua Chang, and Bruce A. Kaplan
Educational Testing Service

19.1 INTRODUCTION

This chapter gives an introduction to the analyses performed on the responses to the cognitive and background items in the 1998 assessment of writing. These analyses led to the results presented in the *NAEP 1998 Writing Report Card for the Nation and the States* (Greenwald et al., 1999). The topics discussed in this chapter center on issues such as the description of student samples, student weights, items, assessment booklet, administrative procedures, scoring of the constructed-response items and student weights. Reasons why a formal analysis of differential item functioning (DIF) were not attempted will be presented. The major analysis components are discussed in Chapter 20 for the national assessment and Chapter 21 for the state assessment.

The objectives of the writing analyses were to prepare scale values, estimate subgroup scale score distributions for pertinent populations of students, and estimate the percent of students performing at or above various achievement-level cut points. The 1998 state assessment scales were linked to the corresponding scales from the 1998 national assessment. All analyses used data from students participating in the 1998 national and state writing assessments.

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

The student samples that were administered writing items in the 1998 assessment are shown in Table 19-1. The data from the national main focused partially balanced incomplete block (PBIB) assessment of writing (4 [Writing–Main], 8 [Writing–Main], and 12 [Writing–Main]) were used for national main analyses comparing the levels of writing achievement for various subgroups of the 1998 target populations. See Section 1.5 for an explanation of the focused partially balanced incomplete block (PBIB). Chapters 3 and 4 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 writing NAEP assessments, only grade-defined cohorts were assessed in the 1998 NAEP. The students were sampled in the winter (January to March with final makeup sessions held from March 30 to April 3). As described in Chapter 3, the reporting sample for the national writing assessment has students with disabilities (SD) and limited English proficient students (LEP) who were included under new inclusion rules and who were given appropriate accommodations as available.

The sample designated as 8 [Writing–State] was used for the grade 8 state writing analysis. This sample included the assessment of both public- and nonpublic-school students for most jurisdictions. The procedures used were similar to those of previous state assessments.

[1] Frank Jenkins was the primary person responsible for coordinating the national writing analysis. Hua-Hua Chang and Jiahe Qian were responsible for coordinating the state writing analysis. Computing activities for all writing analyses were directed by Bruce A. Kaplan and assisted by Youn-Hee Lim. Others contributing to the analysis were David S. Freund and Katherine Pashley.

Table 19-1
NAEP 1998 Writing Student Samples[*]

Sample	Booklet Number	Cohort Assessed	Time of Testing[†]	Reporting Sample Size
4 [Writing–Main]	W201–W240	Grade 4	1/5/98 – 3/27/98	19,816
8 [Writing–Main]	W201–W240	Grade 8	1/5/98 – 3/27/98	20,586
12 [Writing–Main]	W201–W237	Grade 12	1/5/98 – 3/27/98	19,505
8 [Writing–50 Min]	W241–W243	Grade 8	1/5/98 – 3/27/98	6,009
12 [Writing–50 Min]	W241–W243	Grade 12	1/5/98 – 3/27/98	5,804
8 [Writing–State]	W201–W240	Grade 8	1/5/98 – 3/27/98	97,589

[*] All sessions were administered in a printed format.
[†] Final makeup sessions were held March 30–April 3, 1998.

The major analysis components are discussed below. Some aspects of the analysis, such as procedures for item analysis, scoring of constructed-response items, and methods of scaling, are described in Chapters 9 and 12 and are therefore not detailed here. There were four major steps in the analysis of the writing data, each of which is described in a separate section:

- Conventional item and test analyses (Section 20.2)
- Item response theory (IRT) scaling (Section 20.3)
- Estimation of subgroup scale score distributions based on the plausible values methodology (Section 20.4)
- Transforming the 1998 assessment scales to the final reporting metric (Section 20.5)

Section 20.6 describes the results of partitioning the error variance, 20.7 discusses the matching of student responses to those of their teachers, and 19.6 provides a brief explanation of sampling weights. Analysis of the state writing assessment consisted of similar steps and is detailed in Chapter 21.

To set the context within which to describe the methods and results of scaling procedures, a brief review of the assessment instruments and administration procedures is provided.

The 1998 NAEP national main writing assessment differed from the long-term trend assessment in the sample age definition, the time of testing, the objectives that define the emphasis of the assessment, and the items used. It also differed from the 1992 national main NAEP writing assessment in that (1) the framework was revised, (2) most of the prompts (the exercises administered to the students) were new, and (3) for those prompts that were also administered in 1992, different rubrics (the rules for assigning scores to responses) were used to score responses. Because of these differences, equating or linking to the earlier main and the long-term trend assessments was not appropriate. The 1998 national main writing assessment can be used to start a new baseline for measuring trends in the nation.

The prompts used in the 1998 writing assessment consisted of two types of six-point constructed-response items: those allowing for a 25-minute response and those allowing for a 50-minute response. The items in the assessment were based on the curriculum framework described in *Writing Framework and Specifications for the 1998 National Assessment of Educational Progress* (NAGB, 1996b). The 1998 framework resulted from augmenting the 1992 framework with new exercise specifications. This lead to the development of new writing prompts and scoring guides. As described in the writing framework, the prompts represented three purposes of writing: narrative, informative, and persuasive. All three item types were used to measure a single scale of writing performance. Table 19-2 gives the number of 25-minute writing prompts in each grade that were used in the national main assessment. There were a

total of 20 25-minute prompts per grade in the main assessment. In grade 4, there was an emphasis on narrative items (8 of 20), whereas at grade 12 the emphasis was on persuasive prompts (8 of 20).

Table 19-2
Number of 25-Minute Items in the National Main Writing Assessment Within the Three Purposes of Writing

Grade	Narrative	Informative	Persuasive	Total
4	8	7	5	20
8	7	7	6	20
12	5	7	8	20

Three 50-minute prompts were administered at grades 8 and 12, one for each purpose of writing, as shown in Table 19-3. Administering these items provided an opportunity to study how students responded to longer writing exercises that were more like regular classroom assignments. These items were not included as part of the main writing scale, however, because only one such prompt was administered per person. It was thought that a single item per person yielded too unreliable a measure of writing skill. Therefore, only 25-minute prompts were used in calculating scale score results. Data from the 50-minute prompts were not included.

Table 19-3
Number of 50-Minute Items in the National Writing Assessment Within the Three Purposes of Writing

Grade	Narrative	Informative	Persuasive	Total
8	1	1	1	3
12	1	1	1	3

In the main samples, each student was administered a booklet containing two separately timed 25-minute blocks. Each block contained a single writing prompt. In addition, each student was administered a block of background questions, a block of writing-related background questions, and a block of questions concerning the student's motivation and his or her perception of the difficulty of the NAEP writing items. The background and motivational blocks were common to all writing booklets for a particular grade level. Twenty 25-minute blocks of writing prompts were administered at each grade level. As described in Chapter 18, the 25-minute blocks were combined into booklets according to a partially balanced incomplete block (PBIB) design. See Chapter 18 for more information about the blocks and booklets. In addition, the 50-minute writing prompts were given to some students at grades 8 and 12 in lieu of two 25-minute prompts. In these cases, the single prompt given a student composed the block and the book. As mentioned before, these prompts were not included in the writing scale.

Tables 19-4 through 19-6 give the correspondence between writing prompts and the respective blocks they define. As mentioned above, the 50-minute prompts were the only writing task in a book. The 25-minute prompts, however, are arranged into 40 books. Tables 19-7 through 19-9 gives the correspondence between prompts (which are also blocks) and books. It also indicates in which books a block (or item) was ordered first and in which book a block (or item) was ordered second.

Table 19-4

Grade 4: Prompt, Block, and Purpose Correspondence

Prompt	Description	Block	Purpose
W004002	Aunt Dot	W3	Narrative
W004102	Cartoon Story	W4	Narrative
W004202	Very Unusual Day	W5	Narrative
W004302	Castle	W6	Narrative
W004402	Casey and Duke	W7	Narrative
W004502	Old Tree	W8	Narrative
W004602	Secret Door	W9	Narrative
W004702	Mr. Tooms	W10	Narrative
W004802	Letter from TX8	W11	Informative
W004902	Letter from MZ3	W12	Informative
W005002	Letter from Lilex	W13	Informative
W005102	Animal Lesson	W14	Informative
W005202	City Scenes	W15	Informative
W005302	Unusual Animal	W16	Informative
W005402	Favorite Object	W17*	Informative
W005502	Invisible Friend	W18	Persuasive
W005602	Day Trip	W19*	Persuasive
W005702	Class Pet	W20	Persuasive
W005802	Library Book	W21	Persuasive
W005902	Child or Adult	W22	Persuasive

* This block appears in booklets administered to students requiring accommodations.

Table 19-5

Grade 8: Prompt, Block, and Purpose Correspondence

Prompt	Description	Block	Purpose
W006002	Cartoon Story	W3	Narrative
W006102	President for a Day	W4	Narrative
W006202	Plums	W5	Narrative
W006302	Tower	W6	Narrative
W006402	Principal for a Day	W7*	Narrative
W006502	Pioneer Journal	W8	Narrative
W006602	Space Visitor	W9	Narrative
W006702	Ancient Tree	W10†	Narrative
W006802	Performance Review	W11	Informative
W006902	New Park	W12	Informative
W007002	Dream Weekend	W13	Informative
W007102	Backpack	W14	Informative
W007202	Designing a TV Show	W15	Informative
W007302	Save a Book	W16	Informative
W007402	Life's Lessons	W17	Informative
W007502	Vandalism	W18†	Informative
W007602	Lengthening the School Year	W19*	Persuasive
W007702	School Schedule	W20	Persuasive
W007802	Fast Food	W21	Persuasive
W007902	Class Trip	W22	Persuasive
W008002	Driving Age	W23	Persuasive
W008102	Teens in Malls	W24	Persuasive
W008202	Student of the Year	W25†	Persuasive

* This block appeared in booklets administered to students requiring accommodations.

† This was a 50-minute block and was not part of the main spiral.

Table 19-6

Grade 12: Prompt, Block, and Purpose Correspondence

Prompt	Description	Block	Purpose
W008302	Tall Tale	W3	Narrative
W008402	Plums	W4	Narrative
W008502	Special Object	W5*	Narrative
W008602	The Arch	W6	Narrative
W008702	Pioneer Journal	W7	Narrative
W008802	Ancient Tree	W8†	Narrative
W008902	Cafeteria	W9	Informative
W009002	Writing Mentor	W10	Informative
W009102	Movie Review	W11	Informative
W009202	Technology	W12	Informative
W009302	Handbook	W13	Informative
W009402	Save a Book	W14	Informative
W009502	Life's Lessons	W15	Informative
W009602	Vandalism	W16†	Informative
W009702	Summer Job	W17	Persuasive
W009802	Big or Small Inventions	W18	Persuasive
W009902	Work Less/Study More	W19	Persuasive
W010002	Heroes	W20	Persuasive
W010102	One Vote	W21*	Persuasive
W010202	Teens in Malls	W22	Persuasive
W010302	Driving Age	W23	Persuasive
W010402	Person of the Year	W24	Persuasive
W010502	Campaign Speech	W25†	Persuasive

* This block appeared in booklets administered to students requiring accommodations.

† This was a 50-minute block and was not part of the main spiral.

Table 19-7

Correspondence of Prompts, Blocks, and Books: Grade 4

Item	Block	Books Where Item Occurs in 1st Position		Books Where Item Occurs in 2nd Position	
W004002	W3	204	226	203	225
W004102	W4	201	225	224	240
W004202	W5	214	227	213	226
W004302	W6	217	228	216	227
W004402	W7	209	229	208	228
W004502	W8	221	230	220	229
W004602	W9	224	231	223	230
W004702	W10	232	240	231	239
W004802	W11	203	233	202	232
W004902	W12	208	236	207	235
W005002	W13	223	238	222	237
W005102	W14	213	234	212	233
W005202	W15	220	237	219	236
W005302	W16	202	239	201	238
W005402	W17	216	235	215	234
W005502	W18	205	212	204	211
W005602	W19	206	215	205	214
W005702	W20	207	218	206	217
W005802	W21	210	219	209	218
W005902	W22	211	222	210	221

Table 19-8

Correspondence of Prompts, Blocks, and Books: Grade 8

Item	Block	Books Where Item Occurs in 1st Position		Books Where Item Occurs in 2nd Position	
W006002	W3	201	238	237	240
W006102	W4	202	228	201	227
W006202	W5	203	213	202	212
W006302	W6	204	225	203	224
W006402	W7	205	210	204	209
W006502	W8	206	222	205	221
W006602	W9	207	237	206	236
W006702	W10*	241	—	—	—
W006802	W11	229	236	228	235
W006902	W12	214	230	213	229
W007002	W13	208	234	207	233
W007102	W14	211	232	210	231
W007202	W15	223	233	222	232
W007302	W16	226	231	225	230
W007402	W17	215	235	214	234
W007502	W18*	242	—	—	—
W007602	W19	209	220	208	219
W007702	W20	217	227	216	226
W007802	W21	212	218	211	217
W007902	W22	219	224	218	223
W008002	W23	216	240	215	239
W008102	W24	221	239	220	238
W008202	W25*	243	—	—	—

* Booklets containing 50-minute blocks included only one block.

Table 19-9
Correspondence of Prompts, Blocks, and Books: Grade 12

Item	Block	Books Where Item Occurs in 1st Position		Books Where Item Occurs in 2nd Position	
W008302	W1	201	228	227	240
W008402	W2	202	231	201	230
W008502	W3	203	234	202	233
W008602	W4	204	237	203	236
W008702	W5	205	240	204	239
W008802	W6*	241	—	—	—
W008902	W7	208	223	207	222
W009002	W8	209	227	208	226
W009102	W9	210	230	209	229
W009202	W10	211	233	210	232
W009302	W11	212	236	211	235
W009402	W12	213	239	212	238
W009502	W13	207	214	206	213
W009602	W14*	242	—	—	—
W009702	W15	215	224	214	223
W009802	W16	216	226	215	225
W009902	W17	217	229	216	228
W010002	W18	218	232	217	231
W010102	W19	219	235	218	234
W010202	W20	220	238	219	237
W010302	W21	206	221	205	220
W010402	W22	222	225	221	224
W010502	W23*	243	—	—	—

* Booklets containing 50-minute blocks included only one block.

Some writing prompts were common with the 1992 assessment. However, because the scoring rubrics differed from those used in the 1992 assessment, all items were treated as if they were new. As a result, there was no trend with the 1992 assessment. Also, there was no overlap of items across grades. Thus, a separate writing scale was defined for each grade.

19.3 SCORING CONSTRUCTED-RESPONSE ITEMS

Responses to each writing prompt were scored holistically using a six-category rubric. The six categories defined six levels of partial credit and are referred to by the following descriptors:

0 = Unsatisfactory
1 = Insufficient Response
2 = Uneven Response
3 = Sufficient Response
4 = Skillful Response
5 = Excellent Response

"Missing" responses (students did not write a response to the task, or provided an off-task response) were treated as if the item had not been presented to the student (see Section 12.3.1 or Mislevy & Wu [1988]).

Teams of trained raters scored the written student responses according to scoring guides that defined particular features for the score points appropriate to the grade and purpose of writing. This means that there were nine scoring guides: one for narrative, informative, and persuasive purposes for each grade. See the upcoming *NAEP 1998 Writing Report Card for the Nation and the States* (Greenwald et al., 1999) for details of the scoring rubrics.

In order to determine interrater reliability of scoring, a percentage of responses was scored twice: for the 25-minute prompts, 25 percent of the responses at grades 4 and 12, and 10 percent of the responses at grade 8 (the only grade at which the state-by-state assessment was given) were scored by two raters. In addition, 25 percent of responses to the 50-minute prompts were scored by a second rater.

For the national and state writing assessments, approximately 370,000 responses to writing prompts were scored. This number includes rescoring to monitor interrater reliability. The average within-year percentages of agreement on the six-level scale for the 1998 reliability samples were 77 percent at grade 4, 71 percent at grade 8, and 74 percent at grade 12. The reliabilities for each writing prompt can be found in Appendix C.

19.4 DIFFERENTIAL ITEM FUNCTIONING

A differential item functioning (DIF) analysis is customarily done to identify potentially biased items. In standard DIF analyses such as Mantel-Haenszel and SIBTEST, it is well established that a moderately long matching test is required for the procedures to be valid (i.e., identify DIF in items unconfounded by other irrelevant factors [e.g., Donoghue, Holland, & Thayer, 1993]). In the 1998 NAEP writing assessment, the booklets contain two 25-minute blocks, with one writing prompt per block. Thus, each examinee has (at most) two responses on six-category prompts. This is too little information for the test statistics associated with Mantel (1963) or SIBTEST (Shealy & Stout, 1993) procedures to function effectively. Thus, standard DIF approaches based on statistical tests of items are likely to function poorly, and so were not used in the 1998 writing assessment.

In the writing assessment the standardization method of Dorans and Kulick (1986) was used to produce descriptive statistics. The matching variable was the total score on the booklet (see Section 9.3.4). As in other NAEP DIF analyses, the statistics were computed based on pooled booklet matching; the results are accumulated over the booklets in which a given item appears (e.g., Allen & Donoghue, 1996). This analysis was accomplished using the standard NAEP DIF program NDIF. The statistic of interest appears under the label SMD for "standardized mean DIF." (First, differences in the item score between the two comparison groups are calculated for each level of the booklet score. Then, the standardized mean DIF for the item is the average of these differences divided by their standard deviation.

Significance testing was not performed, due to the low reliability of the matching variable. Instead, the standardized mean difference values were used descriptively, to identify those items that demonstrate the most evidence of DIF. A rough criterion used in the past to describe DIF for polytomous items has been to create the ratio of the SMD to the item's standard deviation and flag any item with a ratio of at least .25. In the writing data no items approached that level. If, as a rule of thumb we use as a criterion for flagging DIF, that the absolute SMD was at least .1, six prompts are flagged. These are listed in Table 19-10. This ad hoc descriptive analysis of DIF did not lead to the rejection of any items as biased.

Table 19-10
Items With Absolute SMD (Standardized Mean DIF) > .10

Group	Grade	SMD	ID
NonAcc/Acc	4	-.106	W005402
B/W	4	-.108	W005302
B/W	12	-.129	W009802
B/W	12	.127	W010402
H/W	4	-.101	W004602
H/W	12	-.112	W009202

LEGEND

NonAcc/Acc	Nonaccommodated versus accommodated students
B/W	Black versus White students
H/W	Hispanic versus White students

Tables A-6 and A-8 in Appendix A provide sample sizes for each of the race/ethnicity and accommodated/nonaccommodated groups noted in the table above.

ETS NAEP staff examined these items, although no formal DIF committee for writing was convened. As a result of this informal analysis of DIF it was decided that there was insufficient evidence of DIF to delete any items. It should be noted that this descriptive procedure was not a formal DIF analysis. Since there were only two items per book, standard DIF procedures wear not appropriate. The descriptive procedure used (standardized mean DIF) did not rule out the possibility of DIF in writing items.

19.5 50-MINUTE WRITING STUDY

It was previously mentioned that there were three 50-minute writing prompts at grade 8 as well as grade 12. For those assigned such prompts, the writing portion of the book consisted of the single 50-minute prompt. Response to these items were not put on the main writing scale. The single response per student was thought to yield inadequate information about students' writing abilities to put their scores on the writing scale. The 50-minute prompts were administered in order to provide a writing experience that more closely reflects actual classroom assignment. It was also an attempt to see if students would do more pre-writing (e.g., outlining) if given more time. Indeed, as the result of an analysis of pre-writing behavior, it was determined that there was more pre-writing with the 50-minute prompts. Details of the responses to 50-minute prompts will be given in the item release materials.

19.6 THE WEIGHT FILES

The sampling contractor Westat produced the final student and school weights and the corresponding replicate weights for the 1998 writing assessment. Information for the creation of the weight files was supplied by NCS under the direction of ETS. Details of the general weighting scheme for the 1998 assessments is given in Chapters 10 and 11. Some features of the weighting procedure peculiar to the 1998 writing assessment will be discussed here.

Students designated as SD or LEP were included in the assessment under new inclusion rules. SD and LEP students who customarily received accommodations were offered those same accommodations in NAEP (i.e., writing used an S3 sample only). At each grade, all accommodated

students took the same booklet, which consisted of two 25-minute blocks. The weighting of accommodated students was handled somewhat differently in different phases of the analysis.

The first stage of a NAEP analysis is an item analysis (IA), which yields information such as item-level frequencies, item means, and item-to-block score correlations. For the IA, the weights were normalized so that the sum of the weights equaled the sample size of the reporting sample (all students taking 25-minute items).

In order to understand the effect that the accommodated students had on the responses for the two items in the "accommodation" book, the item analysis was run three ways:

1. With accommodated students deleted. In this way the responses to items in the "accommodated" book were directly comparable with the responses to other items.

2. With the accommodated students included and using the weights provided by Westat. When compared with the first IA analysis, this showed the full effect that accommodated students had on item responses.

3. Finally, IA was run with accommodated students included, but weighted down by a factor of 4/40. This showed the effect accommodated students would have on items, if the responses for those items were a representative sample from the population. The 4/40 factor was derived from the fact that there are 40 booklets and each item appears in 4 booklets. If evenly distributed, only 4/40s of the entire sample takes each item.

The two items in the accommodated book are "downweighted" in the final IA analysis because there were more accommodated students taking these items than would be expected from a simple random sample. This is because all accommodated students initially assigned to other books were reassigned to the accommodated book. The 4/40 factor comes from the fact that there are 40 books funneling accommodated students into this one book, but an item occurs in 4 books. So we downweight by 1/40 and weight up by 4, which is the same as weighting by 4/40.

The "downweighting" of the accommodated students was also used in the IRT scaling analysis.

For estimation of imputed values (using NSWEEP and CGROUP, see Section 20.4), the accommodated students were not downweighted and the weights were used as they were provided by Westat, as they were in the second IA analysis mentioned above. This was done to assure that statistics based on weighted proficiencies would be representative of the entire population.

Chapter 20

DATA ANALYSIS FOR THE NATIONAL WRITING SAMPLES[1]

Frank Jenkins, Bruce A. Kaplan, and Youn-Hee Lim
Educational Testing Service

20.1 INTRODUCTION

The purpose of the national writing analysis was to produce estimates of subgroup means and standard deviations on the 1998 writing achievement scale and to estimate the percentage of students scoring within each of the achievement level ranges (basic, proficient and advanced) as defined by the National Assessment Governing Board (NAGB) achievement level cut points. To accomplish these goals, data from the 1998 national writing assessment was analyzed through the stages detailed in the following sections. Standard item analyses (e.g., estimation of item means) were performed. Next, an IRT scaling was done to create a writing achievement scale at each grade. Third, estimated (plausible) values on a latent writing trait were estimated in order to get unbiased estimates of subgroup achievement distributions, and finally estimates were put in a convenient metric to facilitate interpretation and prevent confusion with other assessments.

20.2 NATIONAL ITEM ANALYSIS

This section contains a detailed description of the conventional item analysis performed on the writing data. Since there was only one item per block, this analysis could not be done within block as is usual in NAEP assessments. Item to total correlations are meaningless with one item per block. Instead, item analysis was run within grade as if all twenty 25-minute blocks (items) came from one large block. Frequencies of responses at each score point and item averages were the only meaningful statistics that could be reported. Tables 20-1 through 20-3 give the item statistics for the 25-minute items in the three grades. These tables show the number of students taking each item, the percentage of those taking the item that scored in each category, the overall average item score, the average score for the item when it appeared first in a booklet and the average item score when it appeared second in a booklet. The means by block order show a small but consistent order effect advantaging the item when it is in the first position. Fortunately, order effects were balanced over all subsamples through the partially balanced incomplete block (PBIB) design for assigning blocks to books. Books were then assigned to students through a spiral procedure, which results in an equivalent sample of students being assigned to each book (see Chapter 9, Section 9.2). The item means do not vary greatly, ranging from 3.3 to 4.0 at grade 4, 3.4 to 3.9 at grade 8, and 3.3 to 4.2 at grade 12. The reader is cautioned that average item means cannot be compared across grades since there is not a cross-grade scale.

[1] Frank Jenkins was the primary person responsible for the coordination of the National writing analysis. Computing activities for all writing analyses were directed by Bruce A. Kaplan and assisted by Youn-Hee Lim. Others contributing to the analysis were David S. Freund and Katherine E. Pashley.

Table 20-1

Descriptive Statistics for 25-Minute Writing Prompts: Grade 4

			Percentage of Students in Each Category									
Item ID	**Description**	**n**	**Missing**	**0**	**1**	**2**	**3**	**4**	**5**	**Total Item Mean**	**1st Position Item Mean**	**2nd Position Item Mean**
W004002	Aunt Dot	1,680	8.6	1.3	8.8	34.5	40.4	10.0	5.1	3.64	3.67	3.62
W004102	Cartoon Story	1,805	5.6	3.0	17.0	42.9	24.1	10.8	2.2	3.29	3.36	3.23
W004202	Very Unusual Day	1,698	10.8	4.9	12.8	36.2	28.3	13.9	3.9	3.45	3.49	3.42
W004302	Castle	1,730	8.5	2.0	12.1	30.7	38.4	14.0	2.8	3.59	3.65	3.53
W004402	Casey And Duke	1,831	3.2	1.9	6.7	22.8	43.2	20.9	4.4	3.88	3.96	3.80
W004502	Old Tree	1,740	8.3	2.4	7.8	21.3	47.6	16.9	4.0	3.81	3.82	3.80
W004602	Secret Door	1,733	8.3	1.1	6.0	19.4	44.0	23.0	6.5	4.01	4.05	3.98
W004702	Mr. Tooms	1,740	8.3	3.3	7.1	22.7	41.8	20.6	4.5	3.83	3.87	3.80
W004802	Letter from TX8	1,791	3.5	6.4	11.6	36.2	29.8	12.6	3.3	3.40	3.42	3.39
W004902	Letter from MZ3	1,841	4.2	4.4	8.3	45.5	32.6	7.9	1.4	3.36	3.38	3.33
W005002	Letter from Lilex	1,846	3.3	4.1	14.7	43.2	29.2	7.9	1.0	3.25	3.30	3.21
W005102	Animal Lesson	1,893	2.2	1.4	7.9	31.1	47.4	10.5	1.7	3.63	3.68	3.58
W005202	City Scenes	1,747	7.5	4.4	13.7	36.9	35.9	7.8	1.4	3.33	3.39	3.28
W005302	Unusual Animal	1,848	2.9	1.7	5.3	38.3	42.7	9.3	2.8	3.61	3.65	3.57
W005402	Favorite Object	1,827	7.9	1.7	8.7	37.5	41.0	9.4	1.7	3.53	3.59	3.48
W005502	Invisible Friend	1,746	6.3	1.8	8.1	25.2	46.9	15.2	2.8	3.74	3.80	3.68
W005602	Day Trip	1,790	6.9	5.5	13.5	28.3	39.0	11.4	2.3	3.44	3.59	3.29
W005702	Class Pet	1,712	8.4	4.6	9.9	30.0	43.6	9.1	2.7	3.51	3.53	3.49
W005802	Library Book	1,721	7.8	2.8	7.9	31.7	48.2	7.6	1.7	3.55	3.60	3.50
W005902	Child or Adult	1,721	8.6	4.6	7.5	33.7	44.1	9.0	1.1	3.49	3.54	3.44
Average		**1,772**								**3.57**	**3.62**	**3.52**

LEGEND

n = Unweighted sample size
0 = Unsatisfactory
1 = Insufficient
2 = Uneven
3 = Sufficient
4 = Skilled
5 = Excellent

Table 20-2

Descriptive Statistics for 25-Minute Writing Prompts: Grade 8

			Percentage of Students in Each Category									
Item ID	**Description**	**n**	**Missing**	**0**	**1**	**2**	**3**	**4**	**5**	**Total Item Mean**	**1st Position Item Mean**	**2nd Position Item Mean**
W006002	Cartoon Story	1,940	3.3	1.4	13.4	29.7	33.6	16.1	5.9	3.67	3.78	3.56
W006102	President For a Day	1,943	2.3	1.2	12.6	31.0	37.6	12.7	4.8	3.62	3.73	3.52
W006202	Plums	1,988	2.3	2.0	16.2	34.1	32.6	11.7	3.3	3.46	3.57	3.34
W006302	Tower	1,932	1.5	6.0	6.4	21.2	39.3	23.1	4.0	3.79	3.84	3.74
W006402	Principal For a Day	1,921	2.6	3.3	9.3	20.5	39.4	20.4	7.2	3.86	3.97	3.75
W006502	Pioneer Journal	1,935	2.5	1.4	6.9	21.3	46.6	21.5	2.4	3.87	3.96	3.78
W006602	Space Visitor	1,928	3.0	1.5	11.0	20.8	46.2	15.2	5.4	3.79	3.91	3.67
W006802	Performance Review	1,927	2.3	1.4	8.5	30.9	42.4	13.8	3.1	3.68	3.77	3.60
W006902	New Park	1,971	2.1	1.8	8.6	28.1	51.2	8.7	1.6	3.62	3.68	3.55
W007002	Dream Weekend	1,950	1.7	1.8	7.5	26.6	50.3	10.4	3.4	3.70	3.81	3.60
W007102	Backpack	1,936	1.6	2.7	6.4	24.5	49.1	15.2	2.1	3.74	3.79	3.69
W007202	Designing a TV Show	1,929	2.3	3.2	12.7	39.9	33.8	8.5	1.8	3.37	3.44	3.31
W007302	Save a Book	1,915	3.7	4.0	9.4	29.4	47.3	7.1	2.8	3.53	3.68	3.37
W007402	Life's Lessons	1,964	2.2	3.2	8.1	25.8	43.6	15.5	3.9	3.72	3.88	3.56
W007602	Lengthening School Year	1,949	1.8	4.0	9.6	34.1	35.2	14.0	3.0	3.55	3.64	3.45
W007702	School Schedule	1,921	2.3	3.6	11.6	33.8	40.2	9.5	1.3	3.44	3.54	3.36
W007802	Fast Food	1,976	1.2	5.2	9.4	28.3	38.5	15.3	3.3	3.59	3.71	3.47
W007902	Class Trip	1,940	1.9	2.4	8.6	35.9	43.8	6.7	2.5	3.51	3.59	3.44
W008002	Driving Age	1,969	2.5	1.8	11.2	34.2	40.8	10.4	1.7	3.52	3.59	3.44
W008102	Teens in Malls	1,966	1.8	4.7	10.6	24.3	42.4	15.3	2.7	3.61	3.69	3.54
Average		**1,945**								**3.63**	**3.73**	**3.54**

LEGEND

n = Unweighted sample size
0 = Unsatisfactory
1 = Insufficient
2 = Uneven
3 = Sufficient
4 = Skilled
5 = Excellent

Table 20-3

Descriptive Statistics for 25-Minute Writing Prompts: Grade 12

			Percentage of Students in Each Category									
Item ID	**Description**	**n**	**Missing**	**0**	**1**	**2**	**3**	**4**	**5**	**Total Item Mean**	**1st Position Item Mean**	**2nd Position Item Mean**
W008302	Tall Tale	1,838	3.0	6.7	3.6	17.3	49.1	21.6	1.8	3.81	3.87	3.74
W008402	Plums	1,863	3.5	3.1	3.9	11.9	44.5	34.8	1.8	4.10	4.15	4.04
W008502	Special Object	1,889	3.2	1.2	4.3	14.2	36.7	42.0	1.7	4.19	4.28	4.09
W008602	The Arch	1,945	1.8	0.3	4.0	18.0	49.2	26.6	2.0	4.04	4.11	3.97
W008702	Pioneer Journal	1,932	2.0	0.8	6.7	21.5	45.7	21.0	4.3	3.92	4.03	3.81
W008902	Cafeteria	1,878	2.7	0.5	4.3	13.8	46.6	29.9	4.9	4.16	4.23	4.08
W009002	Writing Mentor	1,841	2.4	3.0	4.3	21.4	40.3	25.5	5.4	3.97	4.13	3.83
W009102	Movie Review	1,761	5.1	2.1	7.0	19.7	53.3	13.1	4.8	3.83	3.92	3.73
W009202	Technology	1,815	3.1	3.2	7.8	18.1	38.8	30.3	1.8	3.90	3.98	3.82
W009302	Handbook	1,850	2.5	2.0	5.9	14.5	39.4	26.8	11.5	4.17	4.31	4.04
W009402	Save a Book	1,826	3.1	4.3	8.9	19.6	39.7	25.8	1.6	3.79	3.92	3.64
W009502	Life's Lessons	1,805	5.2	3.5	6.2	14.3	44.8	27.2	4.1	3.98	4.07	3.89
W009702	Summer Job	1,892	2.3	3.2	8.5	28.1	39.3	16.3	4.5	3.70	3.76	3.65
W009802	Big or Small Inventions	1,874	2.9	2.9	8.5	18.3	48.5	15.8	5.9	3.84	3.88	3.79
W009902	Work Less/Study More	1,842	2.0	3.7	9.9	26.1	43.6	10.9	5.7	3.65	3.73	3.57
W010002	Heroes	1,884	2.3	2.3	8.5	17.2	45.7	21.4	4.9	3.90	4.00	3.80
W010102	One Vote	1,892	2.2	4.2	21.3	30.1	31.8	10.1	2.6	3.30	3.40	3.20
W010202	Teens in Malls	1,876	2.5	3.3	9.7	23.6	41.0	18.0	4.4	3.74	3.84	3.63
W010302	Driving Age	1,907	2.6	3.4	11.9	24.6	36.9	18.1	5.1	3.70	3.82	3.58
W010402	Person of the Year	1,882	2.5	2.3	7.0	21.7	37.1	22.3	9.6	3.99	4.11	3.87
Average		**1,865**								**3.88**	**3.98**	**3.79**

LEGEND

n = Unweighted sample size
0 = Unsatisfactory
1 = Insufficient
2 = Uneven
3 = Sufficient
4 = Skilled
5 = Excellent

A few details about the tables need to be explained. Item means were calculated using weights. The denominator for calculating means and percents in responses 1 through 6 were the weighted total number giving legitimate responses (1 through 6). "Missing" responses (i.e., students did not write a response to the task, or provided an off-task response) were treated as "not presented," (i.e., were not given a score and were not used in IRT calibration [see Section 12.3.1 or Mislevy & Wu, 1988]). The denominator for calculating percent missing was the sum of total missing and legitimate responses for the item. The column labeled "n" in the tables shows the unweighted number of students presented with the item who gave a legitimate response. In order to facilitate comparisons among items, the accommodated students were not included in these item analysis tables. At each grade, accommodated students were given the same two items and including this data would make the responses on these two items noncomparable to responses of other items.

20.3 ITEM RESPONSE THEORY (IRT) SCALING

In 1993, the National Assessment Governing Board (NAGB) determined that future NAEP assessments should be developed using within-grade frameworks. Within-grade scaling removes the constraint that the trait being measured is cumulative across the grade levels of the assessment. It also means that there is no need for overlap items across grades. Consistent with this view, NAGB also declared that scaling be performed within-grade. Any items that happened to be the same across grades in the assessment were scaled separately for each grade, thus making it possible for common items to function differently in the separate grades. Therefore, the writing framework specifies that the 1998 writing assessment be developed within-grade. Likewise, all IRT scaling was performed within-grade. Within each grade, a single writing scale was defined that summarizes student performance on the 25-minute items.

20.3.1 Item Parameter Estimation

Item parameter estimates were obtained for the univariate writing achievement scale by 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 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) (see Chapter 12). In the writing assessment, only the partial-credit model was used. Although only two prompts are present in any booklet, each booklet is administered to a randomly equivalent sample of students by employing a spiral procedure of assigning books to students (see Section 20.2).

The accommodated students were weighted down in the scaling analysis. This is because all accommodated students were assigned to the same book. With 40 books and each item occurring in 4 books, this implies that accommodated students were oversampled for these items by a factor of 40/4, (i.e., there were 10 times too many accommodated students). As a result, the accommodated students were weighted down by a factor of 4/40 (1/10) to make their influence on the items the same as would occur in a representative sample. As with the item analysis, weights were normalized (multiplied by a constant) so that the sum of the weights was equal to the sample size.

BILOG/PARSCALE was run with model assumptions to more accurately account for the influence of accommodated students. Two subgroups were defined, one for accommodated and the other for nonaccommodated students. Separate prior achievement scale distributions were estimated for the two subgroups. The subgroup priors were defined as normal with combined mean equal to zero and the combined standard deviation equal to one. The means and standard deviations of the subsamples were

free to vary. As it turned out, the accommodated group mean was always lower than the nonaccommodated group, and the subgroup variances were less than one. The scale was transformed to the reporting metric with an overall mean of 150 and overall standard deviation of 35, in a later stage of the analysis (see Section 20.5).

As with the item analysis, “missing” responses (i.e., students did not reach the task, or provided an off-task response) were treated as “not presented,” (i.e., were not given a score and were not used in IRT calibration).

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) of the item parameter prior distributions were updated at each program-estimation cycle in accordance with provisional estimates of the item parameters. Starting values were computed from item statistics. Item parameters are listed in Appendix E.

20.3.2 Evaluation of Model Fit

During and subsequent to item parameter estimation, an evaluation of the fit of the IRT models was carried out for each of the items in the item pool. These evaluations were conducted to determine if any items had to be dropped or have categories collapsed. Evaluations of model fit were based primarily on graphical analyses. The 6-category polytomous items are depicted by graphs that display response curves for each item category (see Chapter 12). The model-based (theoretical) item category curves were compared with empirical response plots derived from the observed responses. An item’s fit was assessed by comparing the theoretical curves with the empirical ones. The closer they coincide, the better the fit.

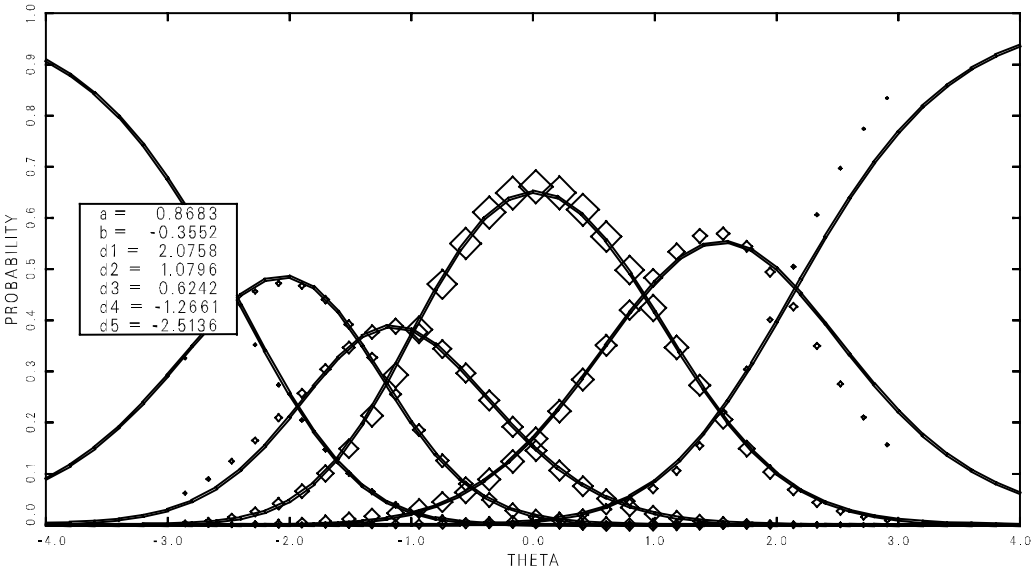
As with most procedures that involve evaluating plots of data versus model predictions, a certain degree of subjectivity was involved in determining the degree of fit necessary to justify use of the model. The seemingly objective procedures of assessing model fit based on goodness-of-fit indices such as the "pseudo chi-squares" produced in BILOG (Mislevy & Bock, 1982) cannot be used as an absolute gauge of fit. The exact sampling distributions of these indices when the model fits 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 were used only as rough guides in interpreting the severity of model departures.

In the case of the writing assessment, there was not much information with which to evaluate model fit. Since there were only, at most, two items administered to each respondent, about half of the achievement scale was determined by the item being evaluated for fit. The IRT model fits well if higher levels of the scale are associated with higher score levels on an item. Since much of a person’s scale score was determined by the item in question, items almost always fit. Without an independent measure of achievement, with only two items per person, item fit will usually be (trivially) good.

As expected, the fit of the model to the item responses was good for all items. Figure 20-1 provides an example of a particularly good-fitting item. In the plot, the *y*-axis indicates the probability of a correct response and the *x*-axis indicates scale score level (theta). The diamonds show empirical

estimates of item category responses. The sizes of the diamonds are proportional to the estimated sample size at the indicated value. The solid curve shows the estimated theoretical item response function. The item response function provides estimates of the probability of a correct response at each scale point (θ) when a logistic response function is assumed.[2] Also shown in the plot are the values of the item parameter estimates (in the box on the left side). As is evident from the plot, the empirical item category traces are in extremely close agreement with the model-based item response function curves.

Figure 20-1
*Polytomous Item (W010002) Exhibiting Good Model Fit**



** Diamonds represent 1998 grade 12 writing 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 20-2 shows an item with poorer fit. This is especially true for the lower end of the achievement distribution, where the empirical plots for two category functions (diamonds) are quite far from the theoretical item category function (solid line). Fortunately, this misfit represents a very small portion of the respondents, as is evidenced by the small size of the diamonds. This is the poorest fitting item even though the figure shows quite good fit. As a result, it was not necessary to delete or collapse categories for any items to improve the fit of the model.

20.4 GENERATION OF PLAUSIBLE VALUES

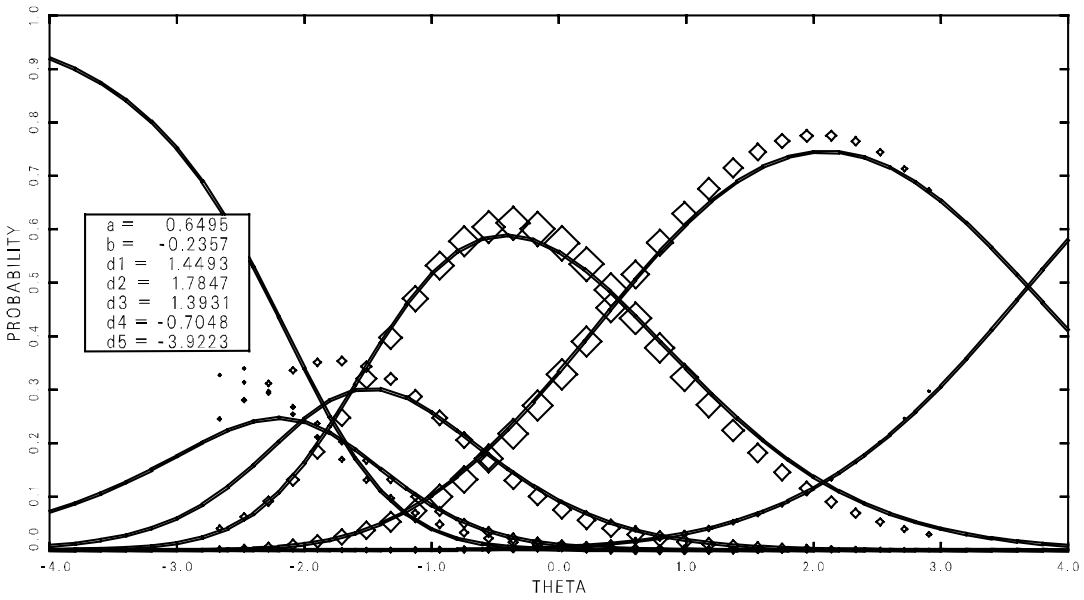
20.4.1 Principal Components (NSWEEP Program)

Univariate plausible values were generated for each sample using the univariate conditioning program BGROUP as written by Thomas (1993b). This procedure employed student weights. Prior to the 1990 assessment, selected background variables were used for conditioning. However, from 1990 to the present, principal components of the background variables have been used as conditioning variables. Almost all of the background variables were coded as 0-1 contrasts, so no standardization took place.

[2] Note that in the generalized partial-credit model, the displayed theoretical curves are not logistic. Rather, logistic curves represent the conditional probabilities given adjacent values, so that $P(x=k|x=(k-1)\ or\ x=k,\theta)=\frac{P(x=k|\theta)}{P(x=(k-1)|\theta)+P(x=k|\theta)}$ is logistic.

Principal components of these contrasts were employed to remedy problems of extreme collinearity among some of the original conditioning variables. The principal components used accounted for at least 90 percent of the variance of the original conditioning variables.

Figure 20-2
*Polytomous Item (W008402) Exhibiting Less Than Optimal Model Fit**



** Diamonds represent 1998 grade 12 writing 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.*

Results from research on the 1990 trial state assessment in mathematics suggests that using a large subset of principal components will yield estimates that differ only slightly from those obtained using the full set (Mazzeo et al., 1992). Table 20-4 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 20-4
Proportion of Scale Score Variance Accounted for by the Conditioning Model for the 1998 National Main Writing Assessment

Grade	Number of Conditioning Contrasts*	Number of Principal Components*	Proportion of Scale Score Variance Accounted For
4	1,095	416	.53
8	1,123	405	.62
12	633	255	.59

* Excluding the constant term

20.4.2 Conditioning (BGROUP Program)

The codings of the original writing-specific conditioning variables, before principal components were calculated, are presented in Appendix F. NAEP BGROUP (described in Chapter 12) creates posterior distributions of scale scores by combining information from item responses of individuals and information from linear regression of scale score on conditioning variables. For each individual, five plausible values were randomly drawn from their posterior scale distribution.

The values of the conditioning effects were expressed in the metrics of the original calibration scale. Definitions of derived conditioning variables are given in Appendix G.

20.5 FINAL REPORTING SCALES

Like all IRT scales, the writing scales have a linear indeterminacy that may be resolved by an arbitrary choice of origin and unit size. The 1998 writing assessment was developed using a new definition of the content domain of the items (see Section 18.2). Because it was not appropriate to compare results from the 1998 assessment with those of previous NAEP writing assessments, no attempt was made to link or align scores on the new assessment to those of previous assessments. Therefore, it was necessary to establish a new scale for reporting. The NAGB has decided that all NAEP scales will be defined within-grade. As a result, the univariate writing achievement scales at each grade were transformed to a reporting metric with scale points ranging from 0 to 300, with an overall mean of 150 and with a standard deviation of 35. Because of the arbitrary nature of the metric, cross-grade comparisons are meaningless.

At each grade the writing scale was transformed from the original scaling metric (mean 0, SD=1) to the reporting metric (mean 150, SD=35) using the transformation:

$$\theta_{reporting} = A \bullet \theta_{scaling} + B.$$

with $\theta_{scaling}$ being the scale score in the scaling metric (approximately mean=0, SD=1), and $\theta_{reporting}$ being the scale the scale score in the reporting metric (mean=150, SD=35). Calculation of the constants for this linear transformation, "A" and "B", is described in Chapter 9. These linear transformation constants are given for each grade in Table 20-5. As previously mentioned, the scaling metric is roughly standardized with mean about 0 and standard deviation about 1 and the scale score metric has mean 150 and standard deviation 35. As a result, one would expect all A's to be 35 and all B's to be 150. As Table 20-5 shows, this is not the case. The reason is that accommodated students were weighted differently in the scaling and conditioning phases of analysis.

Table 20-5
Coefficients of Linear Transformations of the Writing Scales from the Scaling Metric to the Reporting Metric

Sample	A	B
Grade 4	34.01	152.24
Grade 8	34.06	151.50
Grade 12	34.54	151.11

20.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE

For each grade, the error variance of the final, transformed scale mean was partitioned as described in Chapter 12. The variance was partitioned into two parts: the proportion of error variance due to sampling students (sampling variance) and the proportion of variance due to the fact that the scale score, θ, is a latent variable that was estimated rather than observed. Table 20-6 contains estimates of the total error variance, the proportion due to sampling of students, and the proportion due to the latent nature of scale scores. To get greater stability of the variance estimates, they are based on drawing 100 imputations from the posterior achievement distribution of each student. More detailed information of proportion of variance by gender and race/ethnicity is presented in Appendix H.

Table 20-6
Estimation Error Variance and Related Coefficients for the National Main Writing Assessment

	Proportion of Variance Due to . . .	
Grade	**Student Sampling**	**Latency of θ**
4	.90	.10
8	.94	.06
12	.93	.07

20.7 WRITING TEACHER QUESTIONNAIRES

Teachers of fourth- and eighth-grade students were surveyed about their educational background and teaching practices. Each student's records were matched with his or her teacher's survey information. Variables derived from the questionnaire were used in the conditioning models, along with a variable that indicated whether a student record had been matched with a teacher record, which controls estimates of subgroup means for differences that exist between the matching and nonmatching students. Of the 19,816 fourth-grade students in the sample, 89 percent were matched with both parts of the teacher questionnaire and 4 percent were matched with only the first, teacher background, part of the questionnaire. Of the 20,586 eighth-grade students sampled, 72 percent were matched with both parts of the teacher questionnaire and 8 percent were matched with only the first part (the demographic background section) of the questionnaire. The lower match rate for both parts of the questionnaire for eighth-grade students was due in part to the fact that in grade 8 students were matched to the particular class that the teacher taught. Class membership information was often missing or ambiguous. For grade 4, students only had to be matched to the main teacher, resulting in higher match rates. Thus, 93 percent of the fourth-graders and 79 percent of the eighth-graders were matched with at least the background information about their writing teachers.

Chapter 21

DATA ANALYSIS OF THE STATE WRITING ASSESSMENT[1]

Jiahe Qian, Hua-Hua Chang, Bruce A. Kaplan, Jo-Lin Liang, and Youn-Hee Lim
Educational Testing Service

21.1 INTRODUCTION

This chapter describes the analyses used in developing the 1998 state assessment writing scale. The 1998 state writing assessment was administered to eighth-grade public- and nonpublic-school students for 40 jurisdictions. This was the first state assessment in writing. The procedures used were similar to those employed in the analysis of the 1990, 1992, and 1996 state assessments in mathematics (Jenkins, Kulick, Kaplan, Wang, Qian, Wang, 1997; Mazzeo, 1991; Mazzeo, Chang, Kulick, Fong, & Grima, 1993), 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 Chapter 12. For 1998, the NAEP writing 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 writing. The NAEP report card for state assessments presents average scale scores and achievement-level results for public-school students. In the 1998 state assessment, an attempt was made to include more students with disabilities (SD) and students with limited English proficiency (LEP) by liberalizing inclusion rules allowing for accommodations. Although the 1998 state writing analysis is the first state writing assessment, comparisons of writing results for state and national assessments are essential. The sample of students used for analysis and reporting was formed so that comparable inclusion rules were used.

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

- Conventional item and test analyses (Section 21.2)
- Item response theory (IRT) scaling (Section 21.3)
- Estimation of state and subgroup scale score distributions based on the "plausible values" methodology (Section 21.4)
- Linking of the 1998 state assessment scales to the corresponding scales from the 1998 national assessment (Section 21.5)

For the context of the assessment instruments and administration procedures of the writing assessments, see Section 19.2.

[1] Jiahe Qian was the primary person responsible for the planning, specification, and coordination of the state writing analyses in collaboration with Hua-Hua Chang. Computing activities for all writing scaling and data analyses were directed by Bruce A. Kaplan and completed by Youn-Hee Lim and Ting Lu. Others contributing to the analysis of writing data were David S. Freund, Jo-Lin Liang, and Katharine E. Pashley.

21.2 STATE ITEM ANALYSES

21.2.1 Conventional Item and Test Analyses

This section contains a detailed description of the item analysis performed on the state writing data. As was discussed in Chapter 20, only the 25-minute writing blocks were included in the writing scale. Because there is only one item per block, all twenty 25-minute blocks (items) were treated together as one large block in the item analysis. The main statistics analyzed are mean item scores and frequencies of responses at each score point. Table 21-1 contains summary statistics for overall samples and by the order of the block within booklet, based on the data from all 40 jurisdictions. The senate weights were used in item analysis and scaling procedure (see Sections 15.5 and 17.5). Use of the senate weights does nothing to alter the value of statistics calculated separately within each jurisdiction. Items W006402 and W007602 were presented to accommodated students in the writing assessment. To make the statistics comparable with those of other items, the accommodated students were not included in the item analysis calculation.

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 that scaling. Hence, the item analysis statistics shown in Table 21-1 is approximately consistent with the weighting used in scaling.

Table 21-1 shows the number of students assigned each item, the average item scores and the percentage of students in each category of an item. For the constructed-response items in the writing assessment, the score means were calculated as item score mean. As is evident from Table 21-1, the difficulty of the items did not vary greatly.

This table also indicates that there was little variability in average item scores by block 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. The average item scores were higher when each block was presented in the first position.

In an attempt to maintain rigorous standardized administration procedures across the jurisdictions, a Westat-trained quality control monitor would observe randomly selected sessions within each jurisdiction. If a jurisdiction had never participated in a state assessment, a randomly selected 50 percent of the sessions within jurisdictions were monitored; otherwise, a 25 percent of sampled sessions would be monitored within jurisdictions. Because all jurisdictions in the 1998 state writing assessment had participated in previous state assessments, 25 percent of sessions were monitored in each jurisdiction. 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.

The 1998 state assessment in writing included students sampled from nonpublic schools. 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 school"). Table 21-2 contains the item descriptive statistics for total, public-school sessions, and nonpublic-school sessions, respectively. Of the 40 jurisdictions that reported in the state assessment in writing, 39 had public-school samples, while 18 of the 40 jurisdictions had nonpublic-school samples that met reporting requirements.

Table 21-1

Descriptive Statistics Writing Prompts, Writing 25-Minute State Samples, Grade 8

			Percentage of Students in Each Category							Total Item	1st Position	2nd Position
Item ID	**Description**	**n**	**Off-task**	**0**	**1**	**2**	**3**	**4**	**5**	**Mean**	**Item Mean***	**Item Mean***
W006002	Cartoon Story	9,190	3.70	0.81	12.14	30.79	33.76	17.47	5.04	3.70	3.80	3.57
W006102	President for a Day	9,272	1.67	0.66	12.30	30.43	38.37	14.37	3.87	3.65	3.70	3.58
W006202	Plums	9,274	1.82	0.88	14.26	38.39	31.34	12.21	2.92	3.48	3.60	3.40
W006302	Tower	9,300	1.64	6.20	6.18	23.63	38.62	21.50	3.87	3.75	3.81	3.68
W006402†	Principal for a Day	9,337	1.87	3.04	8.09	20.87	40.45	19.70	7.84	3.89	4.03	3.83
W006502	Pioneer Journal	9,316	2.37	0.91	5.46	21.99	45.99	22.35	3.29	3.93	4.00	3.86
W006602	Space Visitor	9,376	2.43	0.96	10.52	20.51	49.25	14.48	4.29	3.79	3.87	3.73
W006802	Performance Review	9,261	2.18	1.07	8.05	32.14	43.56	12.12	3.05	3.67	3.74	3.62
W006902	New Park	9,392	1.22	1.05	8.61	29.01	48.80	10.31	2.23	3.65	3.79	3.56
W007002	Dream Weekend	9,428	1.43	1.01	6.76	28.58	48.04	12.72	2.88	3.73	3.85	3.66
W007102	Backpack	9,262	1.61	1.89	5.47	24.04	54.24	12.38	1.98	3.76	3.86	3.67
W007202	Designing a TV Show	9,260	1.78	2.65	12.56	44.97	31.62	6.61	1.60	3.32	3.38	3.26
W007302	Save a Book	9,286	2.38	3.10	9.55	32.05	47.23	6.20	1.87	3.49	3.60	3.41
W007402	Life's Lessons	9,291	2.14	3.14	7.90	26.78	41.83	15.98	4.36	3.73	3.84	3.65
W007602†	Lengthening School Year	9,430	1.47	2.67	11.04	36.07	34.86	12.15	3.22	3.52	3.62	3.52
W007702	School Schedule	9,344	1.79	2.73	11.55	38.20	40.08	6.76	0.68	3.39	3.47	3.32
W007802	Fast Food	9,335	1.57	4.99	7.63	27.52	40.82	15.16	3.87	3.65	3.78	3.58
W007902	Class Trip	9,370	1.21	2.02	10.32	38.28	41.53	6.21	1.64	3.44	3.61	3.38
W008002	Driving Age	9,315	1.87	1.40	10.72	34.24	41.94	10.91	0.78	3.53	3.59	3.49
W008102	Teens in Malls	9,326	1.51	4.23	11.10	29.33	40.35	12.07	2.91	3.54	3.66	3.47

* The means were calculated by coding responses from 1 to 6, according to standard IA procedures.

† This item was presented to the accommodated students in the writing assessment. To make the comparisons of statistics comparable with those of other items, the accommodated students were not included in the item analysis calculation.

Key:

n = Unweighted sample size
0 = Unsatisfactory
1 = Insufficient
2 = Uneven
3 = Sufficient
4 = Skilled
5 = Excellent

Table 21-2

Descriptive Statistics for Each Item of the Writing State Assessment Using Senate Weights (Scaled from 0 to 5) , Grade 8

	Public and Private						Public						Private					
	n			Mean*			n			Mean*			n			Mean*		
Item ID	Overall	1st Position	2nd Position	Overall	1st Position	2nd Position	Overall	Mon.	Unmon.	Overall	Mon.	Unmon.	Overall	Mon.	Unmon.	Overall	Mon.	Unmon.
W006002	9,190	4,556	4,634	3.70	3.80	3.57	8,399	2,094	6,305	3.66	3.68	3.66	791	242	549	4.03	3.94	4.08
W006102	9,272	4,615	4,657	3.65	3.70	3.58	8,445	2,093	6,352	3.60	3.65	3.58	827	252	575	4.11	3.83	4.24
W006202	9,274	4,635	4,639	3.48	3.60	3.40	8,453	2,136	6,317	3.46	3.45	3.47	821	248	573	4.00	4.08	3.97
W006302	9,300	4,654	4,646	3.75	3.81	3.68	8,474	2,074	6,400	3.71	3.65	3.73	826	252	574	4.21	4.06	4.27
W006402†	9,337	4,603	4,734	3.89	4.03	3.83	8,539	2,168	6,371	3.91	3.92	3.90	798	255	543	4.17	4.03	4.24
W006502	9,316	4,694	4,622	3.93	4.00	3.86	8,466	2,108	6,358	3.91	3.91	3.91	850	260	590	4.26	4.40	4.20
W006602	9,376	4,694	4,682	3.79	3.87	3.73	8,567	2,116	6,451	3.78	3.81	3.77	809	253	556	4.08	4.10	4.07
W006802	9,261	4,643	4,618	3.67	3.74	3.62	8,458	2,031	6,427	3.66	3.65	3.66	803	233	570	4.03	4.15	3.99
W006902	9,392	4,663	4,729	3.65	3.79	3.56	8,590	2,102	6,488	3.65	3.67	3.64	802	235	567	4.09	4.19	4.04
W007002	9,428	4,722	4,706	3.73	3.85	3.66	8,601	2,146	6,455	3.72	3.78	3.70	827	251	576	4.14	4.04	4.18
W007102	9,262	4,611	4,651	3.76	3.86	3.67	8,485	2,090	6,395	3.75	3.71	3.76	777	240	537	4.06	4.01	4.08
W007202	9,260	4,624	4,636	3.32	3.38	3.26	8,443	2,041	6,402	3.29	3.29	3.28	817	244	573	3.74	3.71	3.75
W007302	9,286	4,606	4,680	3.49	3.60	3.41	8,474	2,021	6,453	3.47	3.46	3.48	812	242	570	3.94	3.99	3.92
W007402	9,291	4,638	4,653	3.73	3.84	3.65	8,465	2,066	6,399	3.72	3.72	3.72	826	244	582	4.10	4.13	4.09
W007602†	9,430	4,715	4,715	3.52	3.62	3.52	8,573	2,111	6,462	3.54	3.50	3.55	857	248	609	3.96	3.88	3.99
W007702	9,344	4,670	4,674	3.39	3.47	3.32	8,491	2,030	6,461	3.36	3.37	3.36	853	246	607	3.77	3.80	3.75
W007802	9,335	4,650	4,685	3.65	3.78	3.58	8,513	2,044	6,469	3.66	3.61	3.67	822	239	583	4.04	4.01	4.06
W007902	9,370	4,699	4,671	3.44	3.61	3.38	8,531	2,025	6,506	3.47	3.46	3.47	839	246	593	3.82	3.94	3.78
W008002	9,315	4,639	4,676	3.53	3.59	3.49	8,477	2,102	6,375	3.51	3.54	3.50	838	244	594	3.89	3.93	3.87
W008102	9,326	4,662	4,664	3.54	3.66	3.47	8,464	2,044	6,420	3.53	3.54	3.53	862	247	615	3.95	3.93	3.96

Mon. = Monitored Unmon. = Unmonitored

* The means were calculated by coding responses from 1 to 6, according to standard IA procedures.

† This item was presented to the accommodated students in the writing assessment. To make the comparisons of statistics comparable with those of other items, the accommodated students were not included in the item analysis calculation.

Consistent differences were evident between the public- and nonpublic-school students. The difference in average item score between public- and nonpublic-school students (i.e., public item mean minus nonpublic item mean) range from -0.54 to -0.26 with an average of -0.40, indicating that public-school students were generally lower in average item scores.

Within each school type session, Table 21-2 also provides the item descriptive statistics for the monitored or 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 public-school results were aggregated over all participating jurisdictions, there was little difference between the performance of students who attended monitored or unmonitored sessions. For nonpublic-school data, the difference was also very small. The average item score was 3.62 for both monitored public-school sessions and unmonitored public-school sessions. The average item score was 4.01 for monitored nonpublic-school sessions and 4.03 for unmonitored nonpublic-school sessions.

Table 21-3 summarizes 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 the booklets. The information in the table combines public- and nonpublic-school data. The mean difference and median difference were close to zero. There are 15 jurisdictions with negative differences (i.e., students from unmonitored sessions scored higher than students from monitored sessions). None were larger in absolute magnitude than 0.083. The results indicate that across jurisdictions, the differences between monitored and unmonitored sessions were relatively small.

21.3 STATE IRT SCALING

21.3.1 Samples Used in State IRT Scaling

As in other state assessments, a single set of item parameters for each item was estimated and used for all jurisdictions (Mazzeo, 1991). Item parameter estimation was carried out using a 25 percent systematic random sample of the public-school 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 sample consisted of 89,164 students, with 590 students being sampled from each of the 39 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. There were not enough monitored students in the District of Columbia and Virgin Islands to sample these two jurisdictions. All the monitored students were taken in these two jurisdictions. 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 11.3 in an effort to ensure that each jurisdiction's data contributed equally to the determination of the item parameter estimates.

[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.

Table 21-3

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

Jurisdiction	Monitored Mean	Unmonitored Mean	Monitored – Unmonitored
Alabama	0.488	0.495	-0.007
Arizona	0.498	0.502	-0.004
Arkansas	0.488	0.475	0.013
California	0.508	0.494	0.014
Colorado	0.539	0.536	0.002
Connecticut	0.610	0.593	0.016
Delaware	0.556	0.487	0.069
Florida	0.497	0.491	0.006
Georgia	0.515	0.517	-0.002
Hawaii	0.483	0.452	0.031
Kentucky	0.525	0.512	0.014
Louisiana	0.472	0.486	-0.014
Maine	0.547	0.558	-0.012
Maryland	0.548	0.528	0.021
Massachusetts	0.562	0.563	-0.001
Minnesota	0.520	0.519	0.002
Mississippi	0.469	0.450	0.019
Missouri	0.527	0.512	0.015
Montana	0.511	0.538	-0.027
Nevada	0.497	0.477	0.021
New Mexico	0.502	0.496	0.006
New York	0.509	0.519	-0.010
North Carolina	0.552	0.541	0.011
Oklahoma	0.535	0.536	-0.001
Oregon	0.512	0.527	-0.015
Rhode Island	0.552	0.527	0.025
South Carolina	0.492	0.480	0.013
Tennessee	0.509	0.519	-0.010
Texas	0.533	0.550	-0.017
Utah	0.508	0.488	0.020
Virginia	0.560	0.537	0.024
Washington	0.512	0.526	-0.014
West Virginia	0.516	0.511	0.004
Wisconsin	0.564	0.536	0.028
Wyoming	0.510	0.509	0.000
District of Columbia	0.430	0.412	0.018
DoDEA/DDESS	0.598	0.564	0.034
DoDEA/DoDDS	0.550	0.558	-0.008
Virgin Islands	0.355	0.438	-0.083
Mean			**0.007**
Median			**0.006**
Minimum			**-0.027**
1st Quartile			**-0.006**
3rd Quartile			**0.019**
Maximum			**0.069**

Only public-school data were used in the scaling models for the state assessments. Based on the analysis of item response function plots for the public/nonpublic comparisons, the public/nonpublic data have similar item response functions for the state writing sample. The plots of empirical and model-based estimates of the item response function were used to study the appropriateness. Each plot contained three estimates of each item category characteristic curve: two sets of empirical estimates that represented public- and nonpublic-school samples, respectively, were compared with a third set that assumed the partial-credit model, which was estimated from public-school data only. The plots for all the items showed reasonable closeness between two empirical curves and the theoretical curve.

21.3.2 Item Parameter Estimation

For the 1998 state assessment, a writing IRT-based scale was developed using the generalized partial-credit model described in Chapter 12. The 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 (Muraki, 1992).

All the items in writing assessments were extended constructed-response items. Each of these items was also scaled using the generalized partial-credit model. Six scoring levels were defined:

0 = Unsatisfactory
1 = Insufficient Response
2 = Uneven Response
3 = Sufficient Response
4 = Skilled Response
5 = Excellent Response

As was done in previous assessments of writing, "missing" responses (i.e., students did not reach the task, or provided an off-task response) were treated as if the item had not been presented to the student. (See Section 12.3.1 for more information on this topic.)

Empirical Bayes modal estimates of all item parameters were obtained from the BILOG/PARSCALE program. Item parameter estimation proceeded as follows. 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. After each estimation cycle, the subject ability distribution was restandardized to have a mean of 0 and standard deviation of 1. Correspondingly, parameter estimates for that cycle were also linearly standardized. Two items, W006402 and W007602, were presented to the accommodated students in the state assessment. The data of accommodated students were calibrated as a separate population in the scaling procedure. Their weights were appropriately reduced to the proportion of the students in the student group who took the items in the test.

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.

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, the exact sampling distributions of these indices when the model fits 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. Studies by Stone, Ankenmann, Lane, and Liu (1993), and by Stone, Mislevy, and Mazzeo (1994) using simulated data suggest 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 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 decisions based on hypothesis testing with judgments about the severity of model misfit and the potential impact of such misfit on final results.

For all of the items of the state writing assessment, the fit of the model was extremely good. Figure 21-1 and Figure 21-2 provide typical examples of what the plots look like for this class of items. The item W006502 in Figure 21-1, an extended constructed-response item, has a good fit. This plot shows two estimates of each item category characteristic curve, one set that does not assume the generalized 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, the estimates agree quite well, although some diamonds on the empirical curve lie above the theoretical curve in the lowest category. They contain just a few students. The sizes of the diamonds are proportional to the number of students categorized as having thetas at or close to the indicated value. Although few student responses were categorized in the highest category, there were adequate data to estimate the model-based estimates for those categories (the solid curves). Such results were typical for the extended constructed-response items.

The plot of item W007602 in Figure 21-2 shows three estimates of each item category characteristic curve, one that assumes the partial-credit model (the solid curves) that was fit on the accommodated and nonaccommodated cases together, and two sets that do not assume the generalized partial-credit model (shown as diamonds for nonaccommodated cases and circles for accommodated cases). The figure also shows a very good fit, except for some accommodated cases lying above theoretical curve in the third category.

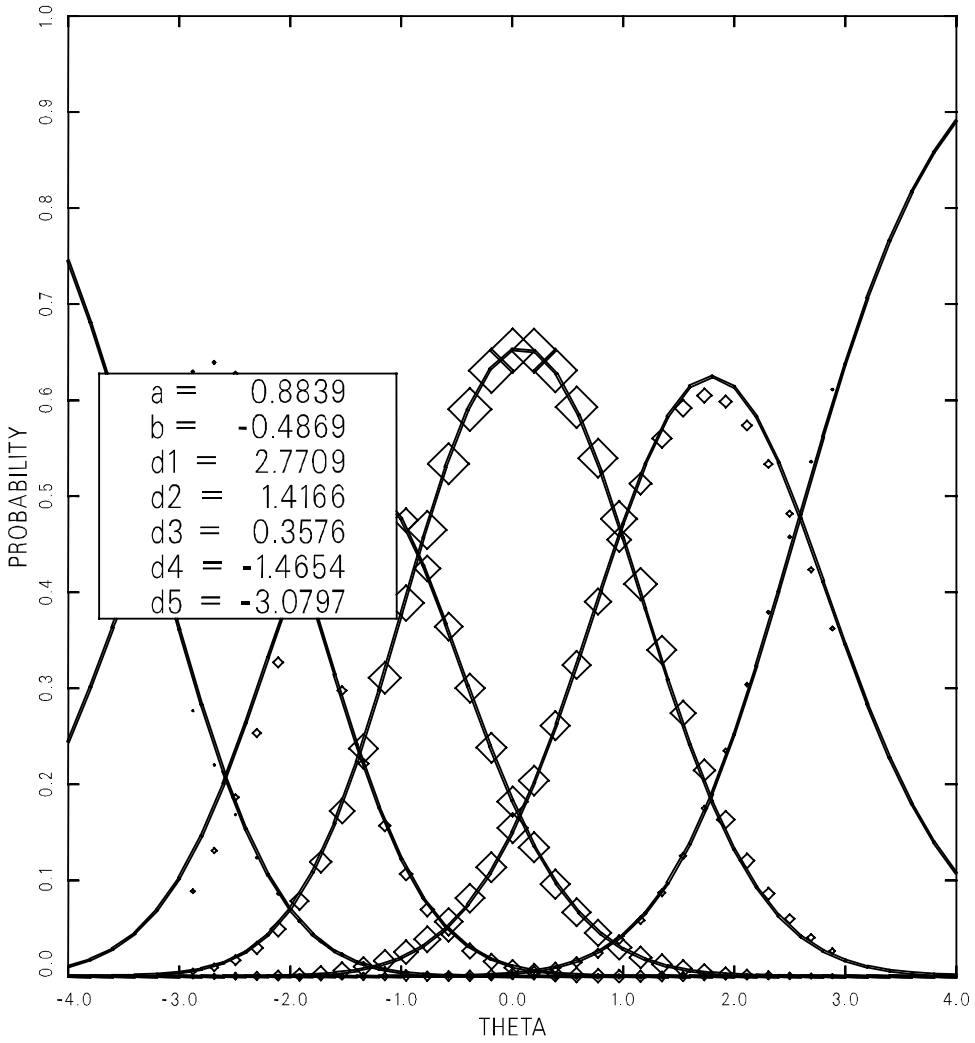
As discussed above, all of the items retained for the final scaling display good model fit. No item needed to be recoded for the state writing assessment. The IRT parameters for the items included in the state assessment are listed in Appendix E.

21.4 GENERATION OF PLAUSIBLE VALUES

The scale score distributions in each jurisdiction (and for some demographic subgroups within each jurisdiction) were estimated by using the univariate plausible values methodology and the corresponding BGROUP computer program. As described in Chapter 12, the BGROUP 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 needed to 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 41 distinct conditioning models for the eighth-grade 1998 state writing assessment.

Figure 21-1
*Polytomous Item (W006502) Exhibiting Good Model Fit**

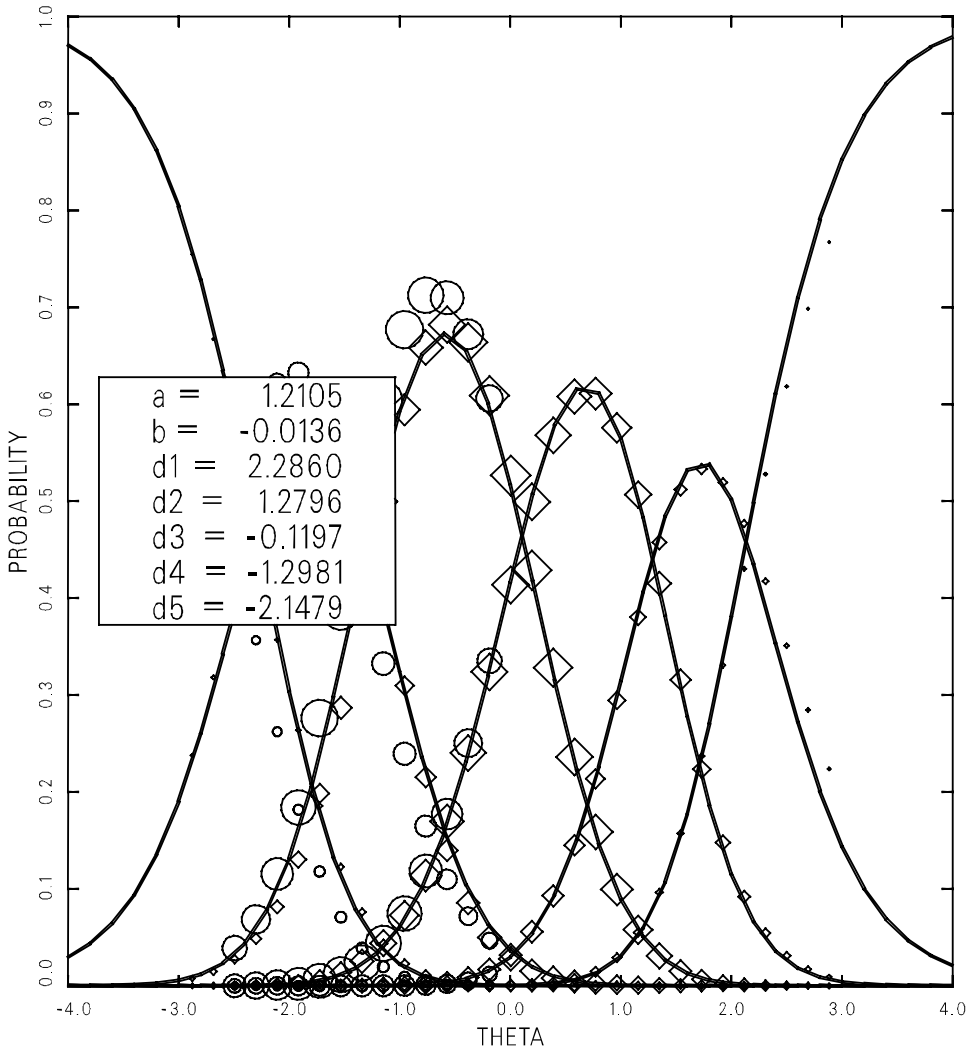


** Diamonds represent 1998 grade 8 writing assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

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, status of test accommodation), students' perceptions about writing, student behavior both in and out of school (e.g., amount of TV watched daily, amount of writing homework done each day), the type of writing 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 also was collected from students' teachers about the types of educational practice, such as the amount of classroom emphasis on various topics included in the assessment provided by the students' teachers, the background and preparation of their teachers.

Figure 21-2
*Polytomous Item (W007602) Exhibiting Good Model Fit**



** Diamonds represent 1998 grade 8 writing 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.*

As described in the Chapter 12, 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. When expressed in terms of contrast-coded main effects and interactions, the number of variables to be included totaled 1,129. 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 BGROUP models. As was done for the 1990, 1992, and 1996 state assessments in mathematics and the 1992, 1994, and 1998 state assessment 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 40 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 suggests 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 21-4 lists the number of principal components included in and the proportion of scale score variance accounted for by the conditioning model for each participating jurisdictions.

It is important to note that the proportion of variance accounted for by the conditioning model differs across jurisdictions. 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, the restriction of the range in the predictor variables will attenuate relationship.

Table 21-4
Proportion of Scale Score Variance Accounted by Conditioning Model for the Writing State Assessment, Grade 8

Jurisdiction	Number of Principal Components	Proportion of Scale Score Variance*
Alabama	242	0.670
Arizona	264	0.704
Arkansas	249	0.731
California	270	0.752
Colorado	259	0.698
Connecticut	276	0.712
Delaware	198	0.775
Florida	284	0.647

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

(continued)

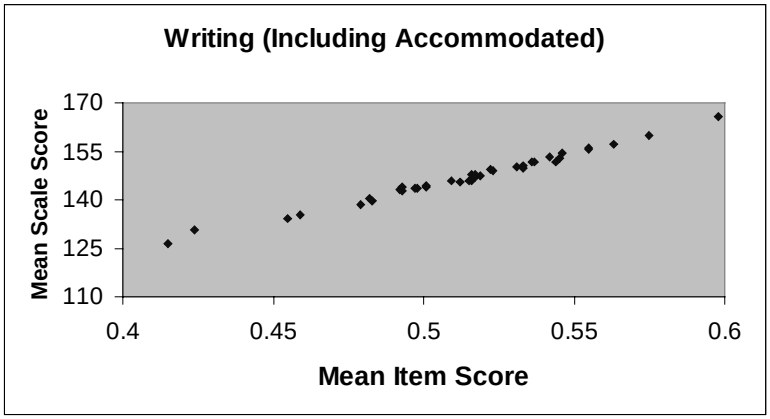
Table 21-4 (continued)
Proportion of Scale Score Variance Accounted by Conditioning Model for the Writing State Assessment, Grade 8

Jurisdiction	Number of Principal Components	Proportion of Scale Score Variance*
Georgia	293	0.732
Hawaii	213	0.665
Kentucky	240	0.699
Louisiana	274	0.696
Maine	228	0.657
Maryland	257	0.719
Massachusetts	256	0.714
Minnesota	219	0.705
Mississippi	241	0.663
Missouri	255	0.705
Montana	194	0.647
Nevada	229	0.685
New Mexico	260	0.709
New York	240	0.714
North Carolina	287	0.690
Oklahoma	232	0.680
Oregon	246	0.667
Rhode Island	225	0.718
South Carolina	290	0.766
Tennessee	234	0.711
Texas	263	0.664
Utah	267	0.621
Virginia	291	0.733
Washington	267	0.705
West Virginia	249	0.731
Wisconsin	214	0.672
Wyoming	200	0.641
District of Columbia	163	0.730
DoDEA/DDESS	142	0.834
DoDEA/DoDDS	173	0.667
Virgin Islands	138	0.841

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

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 scaled scores and in terms of the average proportion correct for the set of items. High agreement was found in all of these analyses. One set of such analyses is presented in Figure 21-3.

Figure 21-3
Plot of Mean Scale Score Versus Mean Item Score by Jurisdiction, Grade 8



The figure contains scatterplots of the state scaled score mean versus the state item score means, for the writing scale. In calculating the statistics for both metrics, the accommodated students are included. 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.

21.5 FINAL SCORE SCALES

21.5.1 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 writing assessment were administered the same test booklets as the eighth-graders in the national assessment, separate state and national scalings were carried out (for reasons explained in Mazzeo, 1991, and Yamamoto & Mazzeo, 1992). 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, 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 size was 89,164. A subsample of the eighth-grade national sample, consisting of grade-eligible public-school students from any of the 40 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

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

national data is referred to as the national linking (NL)[5] sample, and appropriate NL weights were obtained from Westat. Again, 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 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 BGROUP program. The weights used were the final sampling weights provided by Westat (see Section 11.7). For each scale, the arithmetic mean of the five sets of plausible values was taken as the overall estimated mean and the square root of arithmetic average of the variances 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 BGROUP 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 writing results

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

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

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

[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.

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 21-5. All state assessment results are reported in terms of the θ^* metric.

Table 21-5
Coefficients of Linear Transformations for the 1998 State Writing Assessment

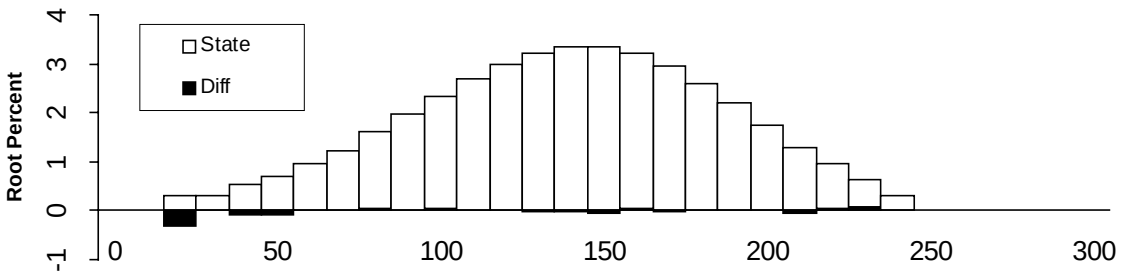
Grade	Writing Scale	A	B
8	State Writing	33.70	147.13

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 to 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. 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 Figure 21-4. The height of each of the unshaded bar 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.

[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 50 and above 250) are dramatically smaller than the counts obtained near the middle of the scale.

Figure 21-4
Rootogram Comparing Scale Score Distributions
for the State Assessment Aggregate Sample
and the National Linking Sample for the Composite Scale, Grade 8



The shaded bars show the differences in root percents between the NL and SA estimates. Positive differences indicate intervals in which the estimated percentages from the SA sample are lower than those obtained from the NL. Conversely, negative differences indicate intervals in which the estimated percentages from the SA sample are higher. Differences in root percents are quite small, suggesting that the shapes of the two estimated distributions are quite similar (i.e., unimodal with slight negative skewness). There is some evidence that the estimates produced using the SA data are slightly heavier in the extreme lower tails, below 50. However, even these differences at the extremes are small in magnitude (0.3 in the root percent metric) and have little impact on estimates of reported statistics such as percentages of students below the achievement levels.

21.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE

For each grade in state writing assessments, the error variance of the final transformed scale score mean was partitioned as described in Chapter 10. 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. Table 21-6 contains 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 the national assessment, the calculations for the state samples are based on 5 plausible values.

Table 21-6
Estimation Error Variance and Related Coefficients for the Writing State Assessment, Grade 8

State	Total Estimation Error Variance	Proportion of Variance due to …	
		Student Sampling	Latency of θ
Alabama	1.958	0.95	0.05
Arizona	2.331	0.86	0.14
Arkansas	1.470	0.89	0.11
California	3.162	0.93	0.07
Colorado	1.719	0.91	0.09
Connecticut	1.843	0.83	0.17
Delaware	2.077	0.32	0.68
Florida	1.534	0.83	0.17
Georgia	1.822	0.83	0.17
Hawaii	1.019	0.37	0.63
Kentucky	2.320	0.92	0.08
Louisiana	1.902	0.93	0.07
Maine	2.110	0.47	0.53
Maryland	2.270	0.89	0.11
Massachusetts	2.814	0.94	0.06
Minnesota	3.492	0.81	0.19
Mississippi	1.689	0.71	0.29
Missouri	2.087	0.87	0.13
Montana	2.107	0.64	0.36
Nevada	0.750	0.48	0.52
New Mexico	0.663	0.80	0.20
New York	2.209	0.94	0.06
North Carolina	2.111	0.77	0.23
Oklahoma	1.603	0.90	0.10
Oregon	2.317	0.87	0.13
Rhode Island	0.431	0.84	0.16
South Carolina	1.196	0.82	0.18
Tennessee	3.121	0.94	0.06
Texas	2.246	0.88	0.12
Utah	1.522	0.63	0.37
Virginia	1.424	0.76	0.24
Washington	2.371	0.80	0.20
West Virginia	2.692	0.43	0.57
Wisconsin	1.746	0.96	0.04
Wyoming	2.043	0.28	0.72
District of Columbia	1.413	0.52	0.48
DoDEA/DDESS	6.695	0.40	0.60
DoDEA/DoDDS	1.476	0.47	0.53
Virgin Islands	14.194	0.14	0.86

21.7 WRITING TEACHER QUESTIONNAIRES

Teachers of the eighth-grade students were surveyed about their educational background and teaching practices. The students were matched first with their writing 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 97,589 eighth-grade students in the sample, 84,605 (86.7%, unweighted) were matched with teachers who answered both parts of the teacher questionnaire, and 6,920 (7.1%, unweighted) of the students had teachers who answered only the teacher background section of the questionnaire.

Chapter 22

ASSESSMENT FRAMEWORKS AND INSTRUMENTS FOR THE 1998 CIVICS ASSESSMENT[1]

Andrew R. Weiss and Terry L. Schoeps
Educational Testing Service

22.1 INTRODUCTION

In 1998, NAEP conducted a national *main* civics assessment and national *special trend* civics assessment at grades 4, 8, and 12.[2] Chapters 22, 23, and 24 cover only the main assessment; a forthcoming report will detail the procedures and analyses of the special trend assessment.

The framework that was used for the 1998 NAEP civics assessment detailed the structure of the assessment to be given at grades 4, 8, and 12 at the national level. The framework for the civics assessment is available on the National Assessment Governing Board (NAGB) web site at *http://www.nagb.org*.

Sections 22.2 through 22.5 include a detailed description of the development of the framework, objectives, and items for the 1998 NAEP civics assessment. Sections 22.6 and 22.7 describe the final cognitive instruments. Section 22.8 describes the student background questionnaires and the civics teacher questionnaire. Additional information on the structure and content of assessment booklets can be found in Section 22.9. Section 22.10 mentions the special trend study in civics. Various committees worked on developing the framework, objectives, and items for the civics assessment. The list of committee members and consultants who participated in the 1998 development process is provided in Appendix K.

Samples of assessment questions and student responses are published in the *NAEP 1998 Civics Report Card for the Nation* (Lutkus, Weiss, Campbell, Mazzeo, & Lazer, 1999).

22.2 DEVELOPING THE CIVICS ASSESSMENT FRAMEWORK

NAGB is responsible for setting policy for NAEP; this policymaking role includes developing 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.

NAGB began the development process for the 1998 civics objectives by establishing the NAEP Civics Consensus Project in February 1995 with the award of the framework contract to the Council of Chief State School Officers (CCSSO). The project's committees gained input through public hearings, student forums, and written reviews of successive drafts of the framework.

[1] Andrew R. Weiss manages the item-development process for NAEP civics assessments. Terry L. Schoeps coordinates the production of NAEP technical reports.
[2] Civics was not part of the NAEP state asssesments in 1998.

For more detail on the development and specifications of the civics framework, refer to the *Civics Framework for the 1998 National Assessment of Educational Progress* (CCSSO, 1996).

Additional information on the NAEP Civics Framework can be found in three technical publications available through NAGB—*Civics Assessment and Exercise Specifications, Recommendations for Background Questions,* and *Reporting Recommendations.*

22.3 CIVICS FRAMEWORK AND ASSESSMENT DESIGN PRINCIPLES

The framework authors stated that given the extreme importance of competent citizenship and effective civic education for the well-being of our constitutional democracy, it is imperative that we have adequate information about what students know and are able to do with regard to civics and government. The aim of the 1998 NAEP Civics assessment was to indicate generally how much and how well students are learning essential knowledge and skills about democratic citizenship and government.

22.4 FRAMEWORK FOR THE 1998 CIVICS ASSESSMENT

The framework comprised three interrelated components: *knowledge, intellectual skills,* and *civic dispositions.* Of these, the *knowledge* component served as the core of the framework. The framework designers established five content areas of knowledge on which to base civics test questions:

- What are civic life, politics, and government?
- What are the foundations of the U.S. political system?
- How does the government established by the Constitution embody the purposes, values, and principles of U.S. democracy?
- What is the relationship of the United States to other nations and to world affairs?
- What are the roles of citizens in U.S. democracy?

The second component, *intellectual skills*, includes:

- identifying and describing,
- explaining and analyzing, and
- evaluating, taking, and defending a position.

The distribution of questions by intellectual skill across grade levels recommended in the assessment framework is provided in Table 22-1. Table 22-2 shows the actual distribution of these questions in the assessment.

Table 22-1
Percentage Distribution of Questions by Intellectual Skill as Recommended in the NAEP Civics Framework

	Intellectual Skill		
	Identifying and Describing	**Explaining and Analyzing**	**Evaluating, Taking, and Defending a Position**
Grade 4	40%	30%	30%
Grade 8	35%	35%	30%
Grade 12	25%	40%	35%

Table 22-2
Actual Percentage Distribution of Questions by Intellectual Skill

	Intellectual Skill		
	Identifying and Describing	**Explaining and Analyzing**	**Evaluating, Taking, and Defending a Position**
Grade 4	33%	37%	30%
Grade 8	29%	38%	33%
Grade 12	18%	33%	38%

Civic dispositions refers to those aspects of a person's character that drive him or her to contribute to the preservation and improvement of United States constitutional democracy.

All three components are summarized in the civics framework (CCSSO, 1996) as shown in Figure 22-1.

22.5 DEVELOPING THE CIVICS COGNITIVE ITEMS

Civics questions were developed by NAEP test developers and outside consultants to meet the requirements of the civics framework. In addition to matching the content and intellectual skills components, NAEP staff had to balance the question pool by question format. The question format included multiple-choice, short constructed-response, and extended constructed-response questions. Short constructed-response questions required answers ranging from a few words to a few sentences and were intended to be answered in up to two minutes. Extended constructed-response questions generally required longer written answers or more time for thinking and were intended to be answered in up to five minutes. The decision to use a specific question format was based on a consideration of how best to measure particular civics knowledge and skills.

Figure 22-1
Description of the NAEP 1998 Civics Framework Components

Knowledge

The *knowledge* component is embodied in the form of five significant and enduring questions: (1) What are civic life, politics, and government? (2) What are the foundations of the American political system? (3) How does the government established by the Constitution embody the purposes, values, and principles of American democracy? (4) What is the relationship of the United States to other nations and to world affairs? (5) What are the roles of citizens in American democracy?

Intellectual and Participatory Skills

The *intellectual and participatory skills* component involves the use of knowledge to think and act effectively in a constitutional democracy. Intellectual skills enable students to learn and apply civic knowledge in the many and varied roles of citizens. These skills help citizens identify, describe, explain, and analyze information and arguments as well as evaluate, take, and defend positions on public policies. Participatory skills enable citizens to monitor and influence public and civic life by working with others, clearly articulating ideas and interests, building coalitions, seeking consensus, negotiating compromise, and managing conflict.

Civic Dispositions

Civics dispositions refer to the inclination or "habits of the heart," as de Tocqueville called them, that pervade all aspects of citizenship. In a constitutional democracy, these dispositions pertain to the rights and responsibilities of individuals in society and to the advancement of ponsibilities of individuals in society and to the ideals of the polity. They include the dispositions to become an independent member of society; respect individual worth and responsibilities of a citizen; abide by the "rules of the game," such as accepting the legitimate decisions of the majority while protecting the rights of the minority; participate in civic affairs in an informed, thoughtful, and effective manner; and promote the healthy functioning of American constitutional democracy.

Table 22-3 contains the percent of assessment time for each question format as specified in the framework and as estimated for the questions selected for the assessment. Grades 8 and 12 estimated percents are closer to the target percent.

Table 22-3
NAEP 1998 Civics Assessment
Percentage of Student Assessment Time by Question Format

Question Type	Specified in Framework*	Actual Percentage of Time		
		Grade 4	Grade 8	Grade 12
Multiple Choice	60%	53%	61%	61%
Short Constructed-Response	30%	29%	27%	30%
Extended Constructed-Response	10%	18%	12%	9%

* These percentages were specified to be the same for all three grades.

Finally, the assessment framework directed test developers to ensure that 15 percent of the questions measured civic dispositions and participatory skills, and that a significant portion of questions were based on textual and visual stimulus material.

22.6 DEVELOPING THE CIVICS OPERATIONAL FORMS

In preparation for the 1998 operational assessment, questions were field-tested in 1997. The purpose of the field test was to administer a large pool of questions so that those with the best content and statistical properties could be selected for the 1998 operational assessment. The civics field test was conducted in January and February of 1997 and involved national samples of fourth-, eighth-, and twelfth-grade students. A total of 555 questions were developed for the field test. Two hundred questions were administered at grade 4, 224 at grade 8, and 244 at grade 12. The questions were organized in a series of 25-minute blocks, each containing multiple-choice, short constructed-response questions, and extended constructed-response questions. Each student received two blocks. Thirty blocks were administered as follows:

- Eight blocks at grade 4 only,
- Four blocks at grade 4 and grade 8,
- Six blocks at grade 8 only,
- Three blocks at grade 8 and grade 12, and
- Nine blocks at grade 12 only.

Field test results were used by ETS test developers to assemble the 1998 operational instruments. Approximately 500 responses were obtained for each question in the field test. Multiple-choice questions were machine scored and constructed-response questions were read and scored by staff at the National Computer Systems scoring center under the direction of NAEP/ETS staff. The raw field test data were subjected to statistical analyses by NAEP/ETS data analysts. The resulting question analyses yielded mean percentage correct, polyserial correlations, difficulty levels, and other information for each question. NAEP test developers reviewed the analyses to help determine:

- which items best measured civics knowledge and skills,
- the need for revisions of items that lacked clarity or had ineffective item formats, and
- the appropriate number of items to include in each operational assessment test book.

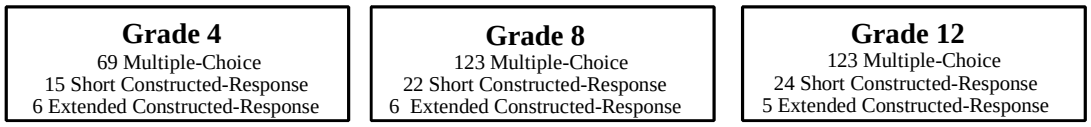
The items chosen for the operational assessment were revised as needed and assembled into new blocks. With the approval of the Civics Instrument Development Committee, cross-grade blocks were eliminated, because it was believed that few questions were successful measures of student knowledge at more than one grade. The blocks were reviewed by the committee in May 1997 for content and balance. Once approved by the committee, all items were subjected to content, measurement, fairness, and editorial reviews by appropriate ETS staff. The draft materials, including background questionnaires, were submitted to the Office of Management and Budget (OMB) in July 1997 for clearance. Changes requested by OMB were made in August 1997, and upon receiving approval, the assessment was sent to print.

Six blocks were assembled for grade 4 and eight blocks were assembled for each of grades 8 and 12. Each student participating in NAEP received two blocks of items. Grade 4 blocks included 15 items each, whereas the blocks at grade 8 and grade 12 included 19 items each.

22.7 DISTRIBUTION OF CIVICS ASSESSMENT ITEMS

Of the total of 393 items, there are 315 multiple-choice items, 61 short constructed-response items, and 17 extended constructed-response items that make up the 1998 civics assessment. A few 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 22-2
Distribution of Items for the 1998 Civics Assessment



22.8 BACKGROUND QUESTIONS FOR THE 1998 CIVICS ASSESSMENT

To gather contextual information, NAEP assessments include background questions designed to provide insight into the factors that may influence civics proficiency.

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 Civics Instrument Development Committee were consulted on the appropriateness of the issues addressed in all questionnaires that may relate to civics instruction and achievement. Like the civics questions, all background questions were submitted for extensive review and field testing. Recognizing the validity problems inherent in self-reported data, particular attention was given to developing questions that were meaningful and unambiguous and that would encourage accurate reporting.

The 1998 assessment included two five-minute sets of general and civics background questions designed to gather contextual information about students and their instructional experiences in civics. Students in the fourth grade were given additional time for these sections (up to fifteen minutes per section), 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 determine students' motivation in completing the assessment and their familiarity with assessment tasks.

22.8.1 Student Civics 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.

Civics Background: Students were asked to report their instructional experiences related to civics including the amount of civics instruction they received and the topics they studied. In addition, they were asked about the instructional practices of their civics teachers including, for example, how often they used textbooks, discussed current events, and took part in classroom activities that simulated civic participation.

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

Table 22-4 shows the number of questions per background section and notes the placement of each within student booklets.

Table 22-4
NAEP 1998 Background Sections of Student Civics Booklets

	Number of Questions	Placement in Student Booklet (of 5 Sections)
Grade 4		
General Background	21	Section 3
Civics Background	22	Section 4
Motivation	5	Section 5
Grade 8		
General Background	22	Section 3
Civics Background	24	Section 4
Motivation	5	Section 5
Grade 12		
General Background	24	Section 3
Civics Background	29	Section 4
Motivation	5	Section 5

22.8.2 Civics Teacher Questionnaire

To supplement the information on instruction reported by students, the civics teachers of the fourth and eighth graders participating in the NAEP civics assessment were asked to complete a questionnaire about their backgrounds, education, experience, and instructional practices. To make the

link between student data and teacher information as complete as possible, teachers were asked to provide information for *each class* containing an assessed student.

The **Teacher Questionnaire, Part I: Background, Education, and Resources** (49 questions at grade 4 and 47 at grade 8) included questions pertaining to:

- 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: Civics Preparation** (7 questions at grade four and 7 at grade eight) included questions on the teacher's preparedness in various areas related to civics education, for example:

- preparedness in social studies instruction;
- preparedness in use of community resources in instruction;
- preparedness in using national standards for civics; and
- preparedness in using software for social studies.

The **Teacher Questionnaire, Part IIB: Civics Classroom Information** (33 questions at grade four and 32 at grade eight) 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 civics;
- instructional emphasis given to the civics abilities covered in the assessment; and
- use of particular resources.

22.9 STUDENT BOOKLETS FOR THE 1998 CIVICS ASSESSMENT

Each student assessed in civics received a booklet containing two blocks of test questions, a five-minute section of general background questions, a five-minute section of civics background questions, a one-minute section of questions about his or her motivation and familiarity with the assessment materials, and content questions. The test questions were assembled into sections or blocks, each containing a range of questions covering the five knowledge categories.

The assembly of civics blocks into booklets and their subsequent assignment to sampled students was determined by a balanced incomplete block (BIB) design with spiraled administration. The civics blocks were assigned to booklets in such a way that every block was paired with every other block at least once. The BIB design balanced the order of presentation of the blocks of items so that every block appears as the first question block and as the second question block an equal number of times. This design allows for some reduction of the impact of context and fatigue effects to be measured and reported. The BIB design in Table 22-5 would call for 15 booklets to allow each of the six blocks to be paired with every other block. Three additional booklets (316-318) were added to ensure that each block appeared equally often in the first and second position. These booklets are the reverse of booklets 313-315.

Once assembled, the assessment booklets were then spiraled and packaged. Spiraling involves interweaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The packages were designed so that each booklet would appear equally often in each position in a package.

The final step in the BIB-spiraling procedure was the assigning of the booklets to the assessed students. The students within an 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 22-5 through 22-7 detail the configuration of booklets administered in the 1998 civics assessment.

Table 22-5

NAEP 1998 Civics Grade 4 Booklet Configuration

Booklet Number	Question Block 1	Question Block 2	Common Core Background	Civics Background	Motivation
301	C3	C4	CW	PB	PA
302	C4	C5	CW	PB	PA
303	C5	C6	CW	PB	PA
304	C6	C7	CW	PB	PA
305	C7	C8	CW	PB	PA
306	C8	C3	CW	PB	PA
307*	C3	C5	CW	PB	PA
308	C4	C6	CW	PB	PA
309	C5	C7	CW	PB	PA
310	C6	C8	CW	PB	PA
311	C7	C3	CW	PB	PA
312	C8	C4	CW	PB	PA
313	C3	C6	CW	PB	PA
314	C4	C7	CW	PB	PA
315	C5	C8	CW	PB	PA
316	C6	C3	CW	PB	PA
317	C7	C4	CW	PB	PA
318	C8	C5	CW	PB	PA

* A large-type version of this booklet was administered as an accommodation to students who had a visual disability.

Table 22-6

NAEP 1998 Civics Grade 8 Booklet Configuration

Booklet Number	Question Block 1	Question Block 2	Common Core Background	Civics Background	Motivation
301	C3	C4	CW	PB	PA
302	C4	C5	CW	PB	PA
303	C5	C6	CW	PB	PA
304	C6	C7	CW	PB	PA
305	C7	C8	CW	PB	PA
306	C8	C9	CW	PB	PA
307	C9	C10	CW	PB	PA
308	C10	C3	CW	PB	PA
309	C3	C5	CW	PB	PA
310*	C4	C6	CW	PB	PA
311	C5	C7	CW	PB	PA
312	C6	C8	CW	PB	PA
313	C7	C9	CW	PB	PA
314	C8	C10	CW	PB	PA
315	C9	C3	CW	PB	PA
316	C10	C4	CW	PB	PA
317	C3	C6	CW	PB	PA
318	C4	C7	CW	PB	PA
319	C5	C8	CW	PB	PA
320	C6	C9	CW	PB	PA
321	C7	C10	CW	PB	PA
322	C8	C3	CW	PB	PA
323	C9	C4	CW	PB	PA
324	C10	C5	CW	PB	PA
325	C3	C7	CW	PB	PA
326	C4	C8	CW	PB	PA
327	C5	C9	CW	PB	PA
328	C6	C10	CW	PB	PA
329	C7	C3	CW	PB	PA
330	C8	C4	CW	PB	PA
331	C9	C5	CW	PB	PA
332	C10	C6	CW	PB	PA

* A large-type version of this booklet was administered as an accommodation to students who had a visual disability.

Table 22-7

NAEP 1998 Civics Grade 12 Booklet Configuration

Booklet Number	Question Block 1	Question Block 2	Common Core Background	Civics Background	Motivation
301*	C3	C4	CW	PB	PA
302	C4	C5	CW	PB	PA
303	C5	C6	CW	PB	PA
304	C6	C7	CW	PB	PA
305	C7	C8	CW	PB	PA
306	C8	C9	CW	PB	PA
307	C9	C10	CW	PB	PA
308	C10	C3	CW	PB	PA
309	C3	C5	CW	PB	PA
310	C4	C6	CW	PB	PA
311	C5	C7	CW	PB	PA
312	C6	C8	CW	PB	PA
313	C7	C9	CW	PB	PA
314	C8	C10	CW	PB	PA
315	C9	C3	CW	PB	PA
316	C10	C4	CW	PB	PA
317	C3	C6	CW	PB	PA
318	C4	C7	CW	PB	PA
319	C5	C8	CW	PB	PA
320	C6	C9	CW	PB	PA
321	C7	C10	CW	PB	PA
322	C8	C3	CW	PB	PA
323	C9	C4	CW	PB	PA
324	C10	C5	CW	PB	PA
325	C3	C7	CW	PB	PA
326	C4	C8	CW	PB	PA
327	C5	C9	CW	PB	PA
328	C6	C10	CW	PB	PA
329	C7	C3	CW	PB	PA
330	C8	C4	CW	PB	PA
331	C9	C5	CW	PB	PA
332	C10	C6	CW	PB	PA

* A large-type version of this booklet was administered as an accommodation to students who had a visual disability.

22.10 CIVICS SPECIAL TREND STUDY IN 1998

In 1998, NAEP conducted a special study designed to compare trends in civics proficiency between 1988 and 1998. Students participating in this special trend study were given booklets from the 1988 NAEP civics assessment. Because the questions in the trend study were based on the 1988 framework, the results cannot be linked to 1998 national assessment results. At the fourth grade level, 2,087 student participated. For grades 8 and 12 the number of students participating totaled 2,053 and 2,181, respectively. Differences in mean item scores for the 1988 booklet were calculated. Results from this special trend study appear in a separate report (Weiss et al., 2000).

Chapter 23

INTRODUCTION TO THE DATA ANALYSIS FOR THE CIVICS ASSESSMENT[1]

Spencer S. Swinton and Edward Kulick
Educational Testing Service

23.1 INTRODUCTION

This chapter gives an introduction to the analyses performed on the responses to the cognitive and background items in the 1998 assessment of civics. These analyses led to the results presented in the *NAEP 1998 Civics Report Card for the Nation* (Lutkus et al., 1999). This chapter describes the student samples, items, assessment booklets, administrative procedures, student weights, and the process used in scoring constructed-response items, as well as the methods and results of differential item functioning (DIF) analyses. The major analysis components are discussed in Chapter 24.

The objectives of the civics analyses were to prepare scale values and estimate subgroup scale score distributions for samples of students who were administered civics items from the national main assessment.

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

The student samples that were administered civics items in the 1998 assessment are shown in Table 23-1. The data from the national main focused balanced in completed block (BIB) assessment (see Section 1.5) of civics (4 [Civics-Main], 8 [Civics-Main], and 12 [Civics-Main]) were used for national main analyses comparing the levels of civics 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. (See Appendix A for tables describing the students assessed and the reporting sample for each component of the civics assessment).

The items in the assessment were based on the framework described in *Civics Framework for the National Assessment of Educational Progress* (NAGB, 1996a). Five areas are described in the civics framework, and were used in developing the assessment questions. For purposes of scaling, all items were fit to a single scale.

In the national main samples, each student was administered a booklet containing two separately timed 25-minute blocks of cognitive civics items. In addition, each student was administered a block of background questions, a block of civics-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; these blocks were common to all civics booklets for a particular grade level. Eight 25-minute blocks of

[1] Spencer S. Swinton was the primary person responsible for the planning, specification, and coordination of the civics analyses. Computing activities for all civics scaling and data analyses were directed by Edward Kulick and completed by Venus Leung. Others contributing to the analysis of civics data were David S. Freund, Bruce A. Kaplan, and Katharine E. Pashley.

civics items were administered at grade 4, and 10 at each of grades 8 and 12. As described in Chapter 22, the 25-minute blocks were combined into booklets according to a BIB design. See Chapter 22 for more information about the blocks and booklets.

At each grade, two civics blocks were repeated from the 1988 assessment of citizenship and social studies to provide data for a special trend study. These items were not scaled with the national main civics assessment items, but were reported using a mean percent-correct metric. The results are reported in *The Next Generation of Citizens: NAEP Trends in Civics, 1988 to 1998* (Weiss, Lutkus, Grigg, & Niemi, 2000).

The mean percent-correct metric involves the percent of people who answered the item correctly. Since all students in the civics trend special study took all items, it was possible to report results for single items and subsets of items by demographic groups. In contrast, the main civics items were scaled using item response theory (IRT). IRT scaling provides parameters that describe the overall difficulty and discrimination of the item. The scale score metric defined by IRT makes comparisons possible across assessments, even if different students took different items.

Table 23-1
NAEP 1998 National Main Civics Assessment Student Samples

Sample	Booklet ID Number*	Cohort Assessed	Time of Testing†	Reporting Sample Size
4 [Civics–Main]	C301-C318	Grade 4	1/5/98 – 3/27/98	5,948
8 [Civics–Main]	C301-C332	Grade 8	1/5/98 – 3/27/98	8,212
12 [Civics–Main]	C301-C332	Grade 12	1/5/98 – 3/27/98	7,763

* Common labeling of booklet numbers across grade levels does not denote common items (e.g., Booklet C301 at grade 8 does not contain the same items as Booklet C301 at grade 12).

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

The total number of scaled items in the main civics assessments was 89, 149, and 151, respectively, for grades 4, 8, and 12. Note that some items overlap across grade. Table 23-2 shows the numbers of items within civics purpose subscales for each grade—both for the original item pool, and after the necessary adjustments were made during scaling.

The composition of each block of items by item type is given in Tables 23-3, 23-5, and 23-7. Common labeling of these blocks across grade levels does not necessarily denote common items (e.g., Block C3 at grade 4 does not contain the same items as Block C3 at grade 8). The numbers of items scaled in 1998 for each grade are presented in Tables 23-4, 23-6, and 23-8.

Table 23-2

Number of Items in the National Main Civics Assessment by Content Area

Grade		Content Areas 1	2	3	4	5	Total
4	Prescaling	19	17	16	8	30	90
	Postscaling	19	16	16	8	30	89
8	Prescaling	19	35	44	22	31	151
	Postscaling	18	35	43	22	31	149
12	Prescaling	14	29	43	30	37	152
	Postscaling	14	29	43	29	37	151

CONTENT-AREA LEGEND

1 What are civic life, politics, and government?
2 What are the foundations of the U.S. political system?
3 How does the government established by the Constitution embody the purposes, values, and principles of U.S. democracy?
4 What is the relationship of the United States to other nations and to world affairs?
5 What are the roles of citizens in U.S. democracy?

Table 23-3

1998 NAEP Civics Block Composition
As Defined Before Scaling, Grade 4

Block	Multiple-Choice Items	Constructed-Response Items Scored Polytomously 2-category*	3-category	4-category	Total Items
Total	69	0	15	6	90
C3	11	0	3	1	15
C4	11	0	4	0	15
C5	12	0	2	1	15
C6	12	0	3	0	15
C7	11	0	1	3	15
C8	12	0	2	1	15

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

Table 23-4

1998 NAEP Civics Block Composition
After Scaling, Grade 4

Block	Multiple-Choice Items	Constructed-Response Items Scored Polytomously 2-category*	3-category	4-category	Total Items
Total	68	1	15	5	89
C3	11	0	3	1	15
C4	11	1	3	0	15
C5	12	0	2	1	15
C6	12	0	3	0	15
C7	11	0	2	2	15
C8	11	0	2	1	14

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

Table 23-5

1998 NAEP Civics Block Composition
As Defined Before Scaling, Grade 8

Block	Multiple-Choice Items	Constructed-Response Items Scored Polytomously			Total Items
		2-category*	3-category	4-category	
Total	123	0	22	6	151
C3	15	0	4	0	19
C4	16	0	1	2	19
C5	15	0	4	0	19
C6	15	0	4	0	19
C7	15	0	3	1	19
C8	16	0	2	1	19
C9	16	0	2	1	19
C10	15	0	2	1	18

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

Table 23-6

1998 NAEP Civics Block Composition
After Scaling, Grade 8

Block	Multiple-Choice Items	Constructed-Response Items Scored Polytomously			Total Items
		2-category*	3-category	4-category	
Total	121	1	21	6	149
C3	15	0	4	0	19
C4	16	0	1	2	19
C5	14	0	4	0	18
C6	15	1	3	0	19
C7	15	0	3	1	19
C8	15	0	2	1	18
C9	16	0	2	1	19
C10	15	0	2	1	18

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

Table 23-7

1998 NAEP Civics Block Composition
As Defined Before Scaling, Grade 12

Block	Multiple-Choice Items	Constructed-Response Items Scored Polytomously			Total Items
		2-category*	3-category	4-category	
Total	123	0	23	6	152
C3	15	0	3	1	19
C4	16	0	3	0	19
C5	15	0	3	1	19
C6	16	0	3	0	19
C7	15	0	2	2	19
C8	15	0	4	0	19
C9	16	0	2	1	19
C10	15	0	3	1	19

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

Table 23-8
1998 NAEP Civics Block Composition
After Scaling, Grade 12

Block	Multiple-Choice Items	Constructed-Response Items Scored Polytomously			Total Items
		2-category*	3-category	4-category	
Total	122	1	22	6	151
C3	15	0	3	1	19
C4	16	0	3	0	19
C5	15	1	2	1	19
C6	15	0	3	0	18
C7	15	0	2	2	19
C8	15	0	4	0	19
C9	16	0	2	1	19
C10	15	0	3	1	19

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

23.3 SCORING CONSTRUCTED-RESPONSE ITEMS

In addition to multiple-choice items, each block contained a number of constructed-response items, accounting for 47 percent of testing time in grade 4 and 39 percent of testing time in grades 8 and 12. Constructed-response items were scored by specially trained readers. (Chapter 7 describes scoring procedures and ranges of interrater reliability for constructed-response items.) 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 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

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 Chapter 12.

23.4 DIF ANALYSIS

A differential item functioning (DIF) analysis of items 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. Table A-9 of Appendix A specifies the sample size for each of these groups. The purpose of the analysis 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.

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) was used for detection of DIF in polytomous items and also as described by Zwick, Donoghue, and Grima (1993). This procedure assumes ordered categories.

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 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 groups (male or White students). Similarly, negative values indicate items that are differentially harder for the focal group than the reference group. An item that was classified as a "C" item in *any* analysis was considered to be a "C" item.

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.

Table 23-9 summarizes the results of DIF analyses for dichotomously scored items. One C item was identified in grade 4, 2 in grade 8, and 3 in grade 12. The committee decided that only the C item in grade 8 showed evidence of bias. The item tested for understanding that the rights of United States citizens date back to the Constitution and Bill of Rights, but used a World War II poster as a stimulus. It was judged that the concept being tested did not require a military theme, making it unnecessarily more difficult for females. Note that if the concept in the framework being assessed had *required* a military context, the same performance differential would not necessarily have resulted in the dropping of the item.

Table 23-9

DIF Category by Grade for Dichotomous Civics Items

Grade	DIF Category*	Analysis		
		Male/Female	White/Black	White/Hispanic
4	C-	0	0	0
	B-	3	5	0
	A-	31	30	29
	A+	35	30	37
	B+	0	3	3
	C+	0	1	0
8	C-	1	0	0
	B-	5	4	2
	A-	70	51	52
	A+	46	58	65
	B+	1	10	3
	C+	0	0	1
12	C-	0	0	0
	B-	14	6	4
	A-	49	45	46
	A+	55	65	68
	B+	4	5	5
	C+	1	2	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 23-10
DIF Category by Grade for Polytomous Civics Items

Grade	DIF Category*	Analysis		
		Male/Female	White/Black	White/Hispanic
4	CC-	0	0	0
	BB-	0	2	0
	AA-	7	9	10
	AA+	14	9	11
	BB+	0	1	0
	CC+	0	0	0
8	CC-	0	0	0
	BB-	1	1	2
	AA-	5	13	11
	AA+	16	13	15
	BB+	5	1	0
	CC+	1	0	0
12	CC-	0	3	0
	BB-	0	2	1
	AA-	6	10	12
	AA+	21	13	14
	BB+	2	0	2
	CC+	0	1	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.

In addition to the Mantel-Haenszel DIF procedure, a second bias test was performed using a SIBTEST analysis (Shealy & Stout, 1993). This analysis identified essentially the same items as were flagged by the other DIF procedure.

23.5 THE WEIGHT FILES

To include special-needs students in its assessment, NAEP test developers established accommodations or adaptations of test forms for students with disabilities (SD) and those characterized as having limited English proficiency (LEP). Inclusion criteria for these students were developed by the Department of Education in consultation with a number of other federal government offices. Its goal was to achieve optimal inclusion of students with disabilities and increase the salience of subject-related instructional matters in inclusion decisions.

For the 1998 civics assessments, the sampling contractor Westat produced the final student and school weights and the corresponding replicate weights. Information for the creation of the weight files was supplied by National Computer Systems (NCS) under the direction of Educational Testing Service (ETS).

Chapter 24

DATA ANALYSIS FOR THE CIVICS ASSESSMENT[1]

Spencer S. Swinton, Edward Kulick, and Venus Leung
Educational Testing Service

24.1 INTRODUCTION

This chapter describes the analyses performed on the responses to the cognitive and background items in the 1998 assessment of civics. The focus of this chapter is on the methods and procedures used to estimate scale score distributions for subgroups of students. This includes a wide array of topics, such as the scoring of constructed-response items, classical item statistics, item response theory (IRT) analysis of civics scales, and estimation of subgroup means by the imputation of plausible values. The statistical bases of the IRT and plausible values methodology described in this chapter are given in Chapter 12. These analyses serve as a basis for the results presented in *NAEP 1998 Civics Report Card for the Nation* (Lutkus et al., 1999).

The student samples that were administered civics items in the 1998 national assessment were shown in Table 23-1. (See Chapters 1 and 3 for descriptions of the target populations and the sample design used for the assessment.). These samples were defined only by grade (4, 8, or 12) and not by age of the student. Data from the samples denoted (Civics–Main) comprised the spiraled partially balanced incomplete block design (spiral BIB design, described in Chapter 22) and the present chapter contains information about the scaling of data from these samples. The analyses for the special trend study of 1988–1998 civics will be published in a separate report through the National Center for Education Statistics (NCES).

24.2 ITEM ANALYSIS

This section contains a detailed description of the item analysis performed using sample data. The analysis examines items within blocks. In preparation for this step, constructed-response items were polytomously scored, and derived background variables were calculated. Item statistics such as mean percent correct, average score, item to total score correlations, and percent responding in each item category were calculated.

Tables 24-1, 24-2, and 24-3 show the number of scaled items, number of constructed-response items, unweighted sample size, weighted mean item score, weighted alpha reliability, weighted mean item to total score correlation, and the weighted proportion of students attempting the last item in the block for each block administered at each grade level for the national main assessment for grades 4, 8, and 12, respectively. These values were calculated within block only for those items used in the scaling process. For these item analyses, accommodated students were excluded, because they were not evenly distributed across items; all of the accommodated students in a grade received the same two blocks. Because of the concentration in these blocks of accommodated students, who are generally lower-scoring, inclusion of the accommodated students in the data for these blocks would have made these

[1] Spencer S. Swinton was the primary person responsible for the planning, specification, and coordination of the civics analyses. Computing activities for all civics scaling and data analyses were directed by Edward Kulick and completed by Venus Leung. Others contributing to the analysis of civics data were David S. Freund, Bruce A. Kaplan, and Katharine E. Pashley.

items appear more difficult than they would have in other blocks. Student weights were used, except for the sample sizes. The results for the blocks administered to each grade level indicated that despite nearly identical numbers of items, the blocks differ in average difficulty (i.e., weighted average item score [Block C4=.48 – Block C7=.55]), reliability (i.e., weighted alpha reliability [Block C8=.68 – Block C3=.74]), and proportion reaching the last item (Block C3=.84 – Block C6=.93]). Note that these tables are descriptive, since no significance tests of differences were done.

As described in Chapter 9, 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 completed response (omitted). Not-reached items are those occurring after the last item the student completed in a block. Items that were not reached are treated as if they had not been presented to the examinee, while omitted items are regarded as incorrect.

The r-polyserial is a generalization of the r-biserial statistic traditionally employed in item analysis. Like the alpha reliability, the r-biserial and r-polyserial statistic provides information about the reliability of the block of items. Smaller values are less desirable than large values. 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.

Tables 24-1 to 24-3 also contain information about the effect of the position of blocks within booklets on the average item score for items within each block presented to the national main samples for each grade. Because the special trend study 1988–1998 blocks appeared in only one position, they are not included in these tables. The averages for the national main samples show that the order of blocks within booklets has a small, but consistent, effect on mean item score in the national main civics assessment.

Table 24-1

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

Statistic	Position	C3	C4	C5	C6	C7	C8
Number of Scaled Items		15	15	15	15	15	14
Number Constructed-Response Items		4	4	3	3	4	3
Unweighted Sample Size	First	942	921	984	965	975	946
	Second	985	904	947	938	969	971
	Both	1,927	1,825	1,931	1,903	1,944	1,917
Weighted Average Item Score	First	.53	.50	.51	.55	.56	.51
	Second	.52	.47	.48	.53	.54	.50
	Both	.52	.48	.49	.54	.55	.50
Weighted Alpha Reliability	First	.73	.73	.72	.69	.69	.68
	Second	.75	.72	.71	.70	.71	.68
	Both	.74	.72	.71	.70	.70	.68
Weighted Average R-Polyserial*	First	.52	.54	.54	.54	.50	.48
	Second	.56	.55	.55	.55	.54	.51
	Both	.54	.55	.54	.55	.52	.50
Weighted Proportion of Students Attempting Last Item	First	.77	.86	.83	.93	.87	.82
	Second	.91	.91	.90	.94	.92	.93
	Both	.84	.88	.86	.93	.90	.88

Table 24-2

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

Statistic	Position	C3	C4	C5	C6	C7	C8	C9	C10
Number of Scaled Items		19	19	18	19	19	18	19	18
Number Constructed-Response Items		15	16	14	15	15	15	16	15
Unweighted Sample Size	First	1,000	980	981	1,002	993	1,021	994	1,009
	Second	1,003	1,012	992	1,009	974	975	1,000	997
	Both	2,003	1,992	1,973	2,011	1,967	1,996	1,994	2,006
Weighted Average Item Score	First	.50	.44	.47	.56	.49	.56	.53	.49
	Second	.47	.43	.46	.54	.47	.55	.51	.47
	Both	.48	.44	.47	.55	.48	.55	.52	.48
Weighted Alpha Reliability	First	.77	.78	.75	.77	.71	.72	.74	.69
	Second	.76	.77	.76	.77	.73	.73	.76	.71
	Both	.76	.77	.75	.77	.72	.73	.75	.70
Weighted Average R-Polyserial	First	.53	.57	.53	.55	.48	.51	.53	.48
	Second	.53	.55	.54	.54	.50	.52	.55	.50
	Both	.53	.56	.53	.55	.49	.52	.54	.49
Weighted Proportion of Students Attempting Last Item	First	.88	.94	.90	.95	.82	.93	.96	.91
	Second	.93	.94	.92	.96	.90	.96	.98	.94
	Both	.90	.94	.91	.95	.86	.94	.97	.92

Table 24-3

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

Statistic	Position	C3	C4	C5	C6	C7	C8	C9	C10
Number of Scaled Items		19	19	19	18	19	19	19	19
Number Constructed-Response Items		15	16	15	15	15	15	16	15
Unweighted Sample Size	First	988	970	929	940	922	957	951	974
	Second	931	976	924	996	947	928	955	944
	Both	1,919	1,946	1,853	1,936	1,869	1,885	1,906	1,918
Weighted Average Item Score	First	.54	.56	.53	.57	.50	.51	.54	.58
	Second	.51	.53	.51	.55	.49	.48	.52	.55
	Both	.53	.54	.52	.56	.50	.50	.53	.57
Weighted Alpha Reliability	First	.83	.75	.79	.75	.77	.72	.76	.79
	Second	.85	.77	.81	.76	.79	.75	.78	.79
	Both	.84	.76	.80	.76	.78	.74	77	.79
Weighted Average R-Polyserial	First	.61	.54	.54	.54	.54	.48	.55	.56
	Second	.63	.54	.57	.54	.55	.51	.55	.56
	Both	.62	.54	.56	.54	.55	.50	.55	.56
Weighted Proportion of Students Attempting Last Item	First	.87	.95	.76	.94	.88	.86	.96	.86
	Second	.91	.95	.85	.92	.90	.91	.94	.93
	Both	.89	.95	.80	.93	.89	.89	.95	.89

In grades 4 and 8, and in most grade 12 blocks, the proportion of students attempting the last item is higher for blocks in the second position. This suggests that students learn to pace themselves better as they go through the assessment. Since slower students are more likely to be somewhat lower-scoring, if more of them run out of time in the first block and do not attempt the final items, they will not contribute to those item statistics, which will be based on a group of relatively more able individuals. This will make the average item appear somewhat easier in the first position than in the second.

24.2.1 Constructed-Response Items

As indicated previously in Tables 23-3, 23-5, and 23-7, about 20 percent of the civics items were constructed-response. Constructed-response items were scored in 3 or 4 categories. The categories of responses for the items and the number of responses that were rescored for each item are indicated in Appendix C. The percent agreement for the raters and the intraclass correlation, a rater reliability estimate appropriate for items with several categories, are also given in the appendix. The sample sizes listed in the tables correspond to the samples used in calculating the rater reliability.

In general, the rater reliability of the scoring for dichotomized responses was reasonably high. Reliabilities ranged over items from 0.69 to 0.96 for grade 4, mean 0.82; from 0.50 to 0.94 for grade 8, mean 0.80; and from 0.61 to 0.90 for grade 12, mean 0.78. The item in grade 8 with unusually low scorer reliability, P040903, was a 3-category item requiring the student to explain characteristics of a good representative.

Chapter 7 discusses the definition of the item ratings and describes the process by which teams of raters scored the constructed-response items. This discussion includes the rating definitions for short and extended constructed-response items as well as the range of interrater reliabilities that occurred. Constructed-response items were scored on a scale from 1 to 4 or 1 to 3 to reflect degrees of knowledge. In scaling, this scale is shifted to 0 to 3 or 0 to 2, respectively. Rating information on constructed-response items can be found in Appendix C, which lists the sample sizes, percent agreement, and Cohen’s Kappa reliability index. No items were excluded because of low rater reliabilities.

24.3 ITEM RESPONSE THEORY (IRT) SCALING

For each grade, a separate univariate IRT scale was constructed. The BILOG/PARSCALE computer program was used to estimate the item parameters for the national main assessment. For dichotomous multiple-choice and dichotomized constructed-response items, a three-parameter IRT model was used. Three- and four-category items were polytomously scored and were analyzed with a generalized partial- credit model (Muraki, 1992).

Recall from Section 24.2 that for calibration, item responses that were missing prior to the last completed item in a block were considered omitted and scored as wrong. Also, items that were not reached were treated as if they were not presented to the examinees (and therefore, not counted as wrong). Omitted multiple-choice items were treated as fractionally [1 / (number of alternatives)] correct. Responses to constructed-response items that were classified by scorers as “off-task” (not responsive to the question) were treated as omitted and assigned to the lowest category (0 = omitted). For score-point descriptions, see Section 15.3; for details on scaling procedures, see Section 12.3.1.

The item parameter estimation was done separately within grade, with accommodated student responses included as a separate population. 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 (modeled as a multinomial distribution) 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.

In the final BILOG/PARSCALE run, the prior distributions of the population abilities were free to be estimated and the overall distribution was set to range from –6 to +4. The calibration was based on student weights that were rescaled so that the their sum equaled the unweighted sample size of the 1998 sample. The weights of accommodated students were further rescaled so that for a given item from the accommodation blocks, the proportion of responses from accommodated students was made similar to their proportion in the weighted sample. As a result, the sum of population weights for accommodated students is smaller than the sum of population weights for nonaccommodated students.

Items that received special treatment in the scaling procedure are listed in Table 24-4, along with the reason for special treatment. Items were either dropped or collapsed. If items had empirical item response functions that were severely nonmonotonic, they were dropped. If polytomous items had sparse or nonmonotonic responses in one or more categories, the items were collapsed so that some adjacent response categories were combined into a single category. Only eight of the total items were given special treatment.

Table 24-4
1998 Civics Items Receiving Special Treatment

Grade	NAEP ID	Block	Treatment
4	P040102	C4	Collapsed: (0,1,2) becomes (0,0,1)
	P040402	C7	Collapsed: (0,1,2,3) becomes (0,0,1,2)
	P040506	C8	Dropped due to lack of fit
8	P040905	C5	Dropped due to DIF
	P041003	C6	Collapsed: (0,1,2) becomes (0,0,1)
	P041204	C8	Dropped due to lack of fit
12	P041705	C5	Collapsed: (0,1,2) becomes (0,1,1)
	P041810	C6	Dropped due to lack of fit

24.3.1 Evaluating the Fit of the IRT Model

During the course of estimating an IRT model, individual items were evaluated to determine how well the item response model fit the data. This was done by visual inspection of plots comparing empirically based and theoretical item response functions. Specifically, for dichotomous items these plots consisted of empirically based estimates of the expected proportion correct for each level of civics performance compared to the proportion correct for each level of civics scale score as predicted by the theoretical item response function. For polytomous extended constructed-response items, similar plots

were produced for each item category response function. See Chapter 12 for a fuller explanation of these plots.

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 so poor as to weaken the types of model-based inferences made from NAEP results. Items showing extreme misfit 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.

For most items, the model fit reasonably well in the scale score region containing most of the observations. In a few cases, poor fit with the data led to special treatment or deletion of the item. Figures 24-1, 24-3, and 24-5 give item response plots of dichotomous items. In the plots, the *x*-axis depicts scale score (theta), and the *y*-axis the probability of a correct response. The solid line is the logistic model prediction, and the symbols (diamonds) are the empirically based proportions. The size of the symbols are proportional to the estimated number of students at a particular scale score level. The item parameter values are also included in the plot.

Item response plots for polytomously scored items are given in Figures 24-2, 24-4, 24-6, and 24-7. These are similar to the plots for dichotomous items except that there are several solid lines, one for each item category, with each line indicating the probability of responding in the respective item category. As before, the diamonds indicate the empirical response function, with the size of the symbols proportional to the estimated number of students at a scale score level.

In the plots, good fit of the model to the data is indicated when the model-based functions (solid lines) coincide with the empirical functions (diamonds). When the empirical plot is far away from the model-based line, there is poor fit of the model to the data.

Four examples of fit are illustrated. First there is good model fit, which is shown by Figure 24-1 for a dichotomous item and Figure 24-2 for a polytomous item. In both cases empirical and theoretical lines nearly coincide.

Second are examples of items that displayed moderate lack of fit to the theoretical function. Figure 24-3 shows a dichotomous item and Figure 24-4 a polytomous item with moderate model misfit.

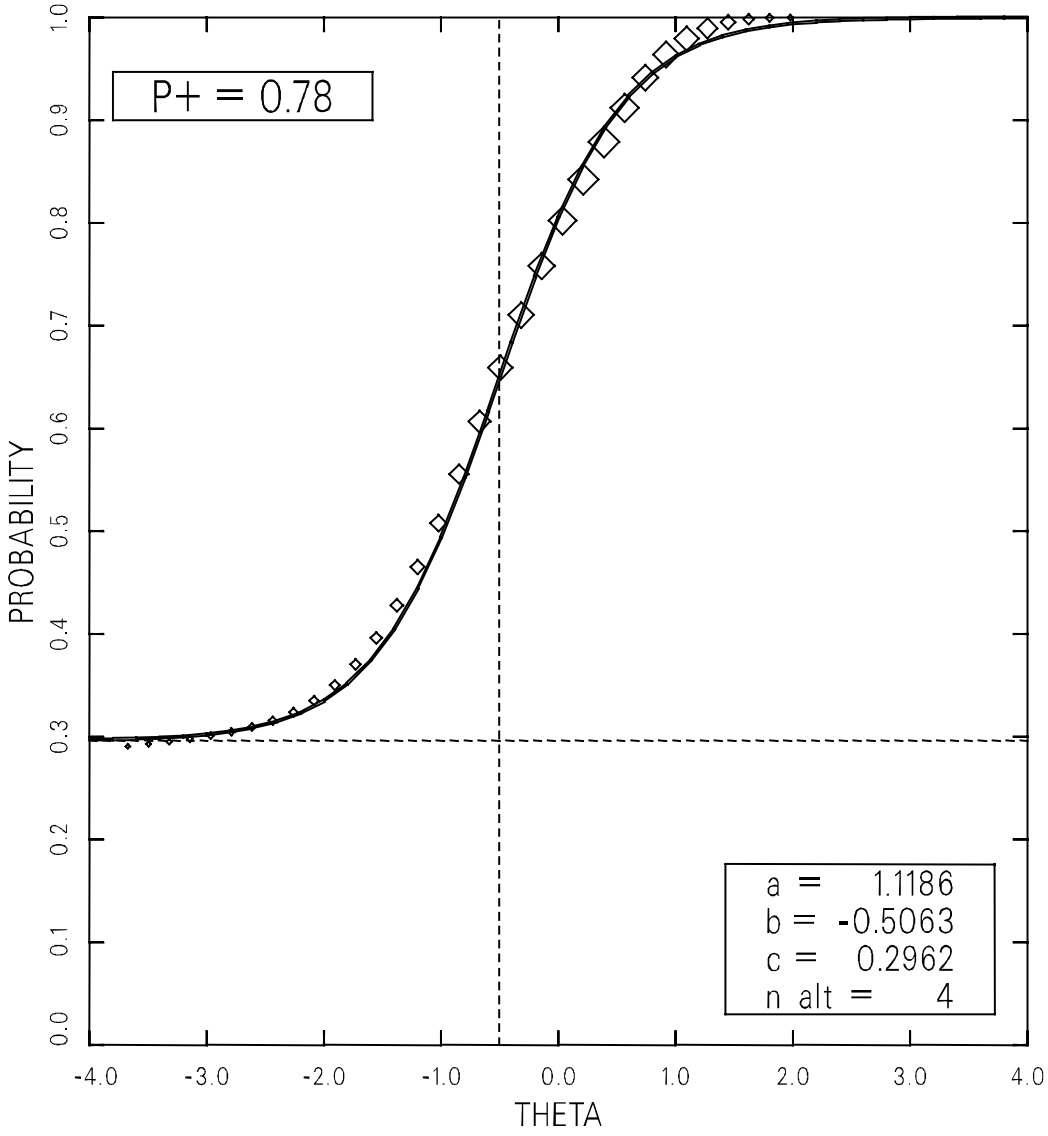
Third (Figure 24-5) is an example of a dichotomous item exhibiting unacceptably poor model fit. This item was dropped from the assessment. This item asked the student to identify a function of a nongovernmental organization.

The fourth example is of a poorly fitting polytomous item that was modified by collapsing categories. Figure 24-6 shows a 4-category item that evidences poor fit mostly in the lower categories. As a result, the lower two categories were collapsed, resulting in a 3-category item, as illustrated in Figure 24-7. This plot still exhibits some degree of misfit, but was judged to fit satisfactorily to be included in the scale. This item asked the student to write on the contrast between a rule and a law.

24.3.2 Derived Background Variables

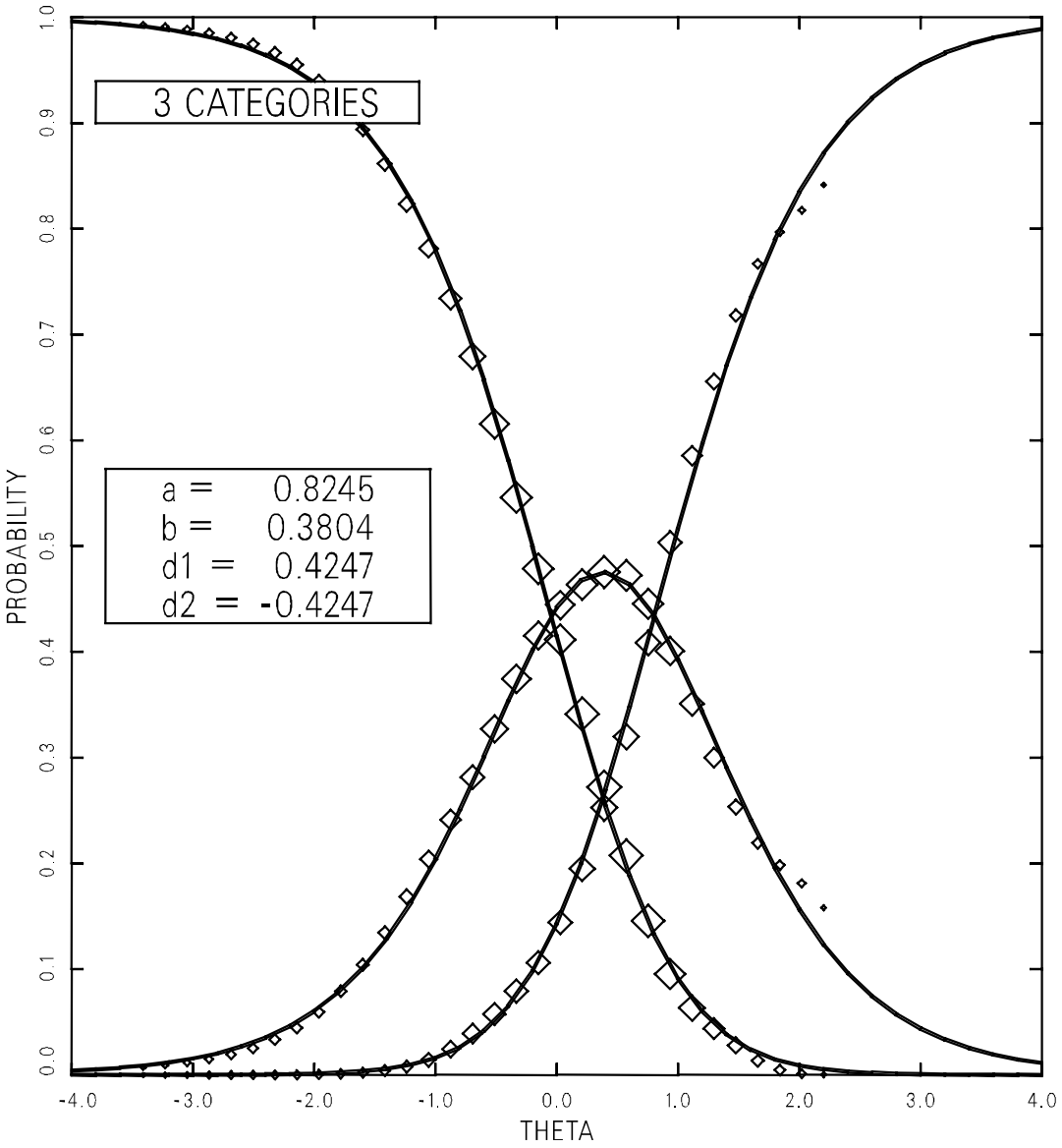
Derived variables are variables that use information from more than one background question. They were used for two purposes: as conditioning variables and as reporting variables used to define subgroups. Some of these variables are common to all the subject areas; others are specific to the 1998 civics assessment. Derived variables used for conditioning and reporting are described in Appendix G.

Figure 24-1
*Dichotomous Item (P040719) Exhibiting Good Model Fit**



** Diamonds represent 1998 grade 12 civics 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.*

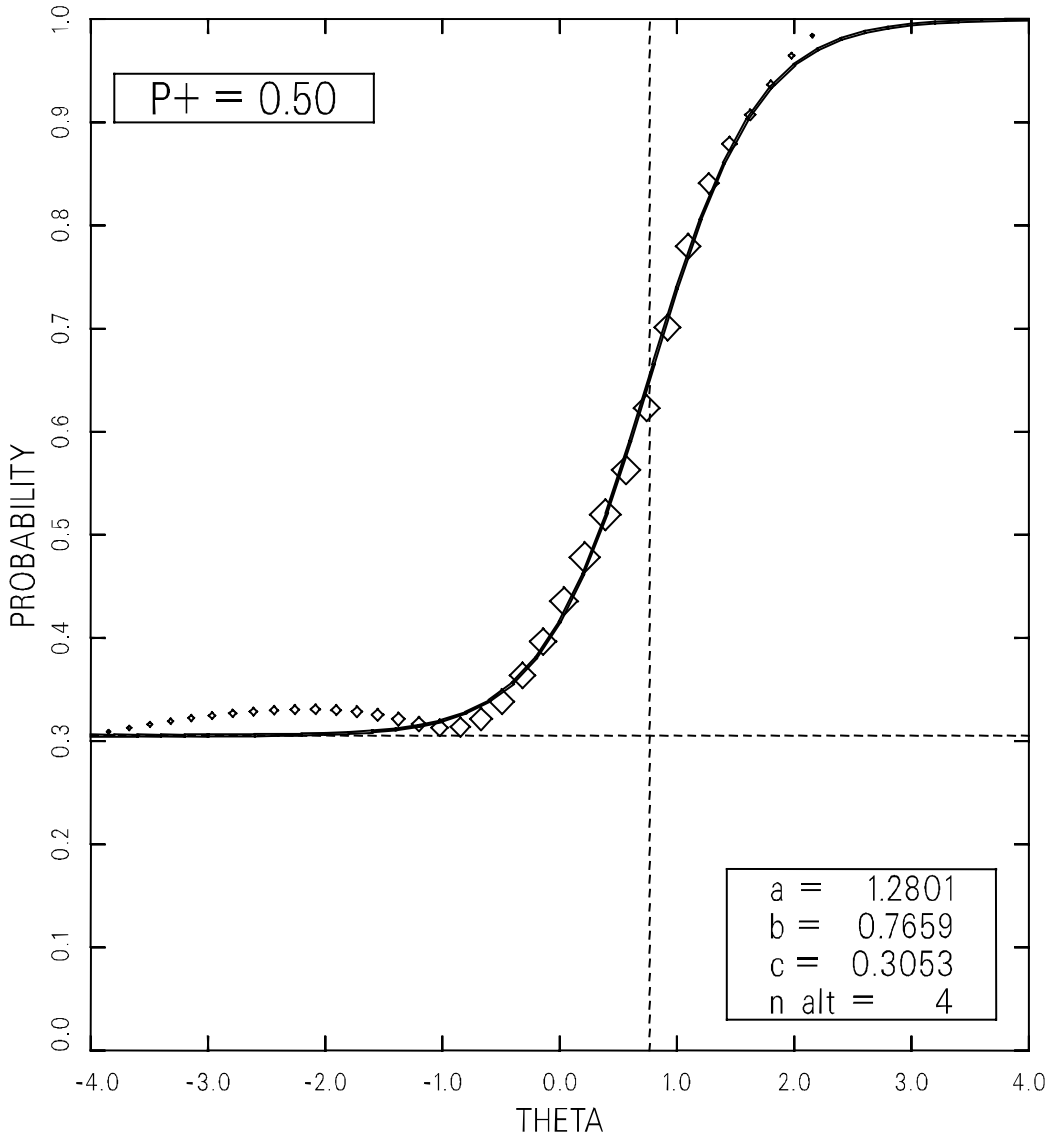
Figure 24-2
*Polytomous Item (P042008) Exhibiting Good Model Fit**



** Diamonds represent 1998 grade 12 civics 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 24-3

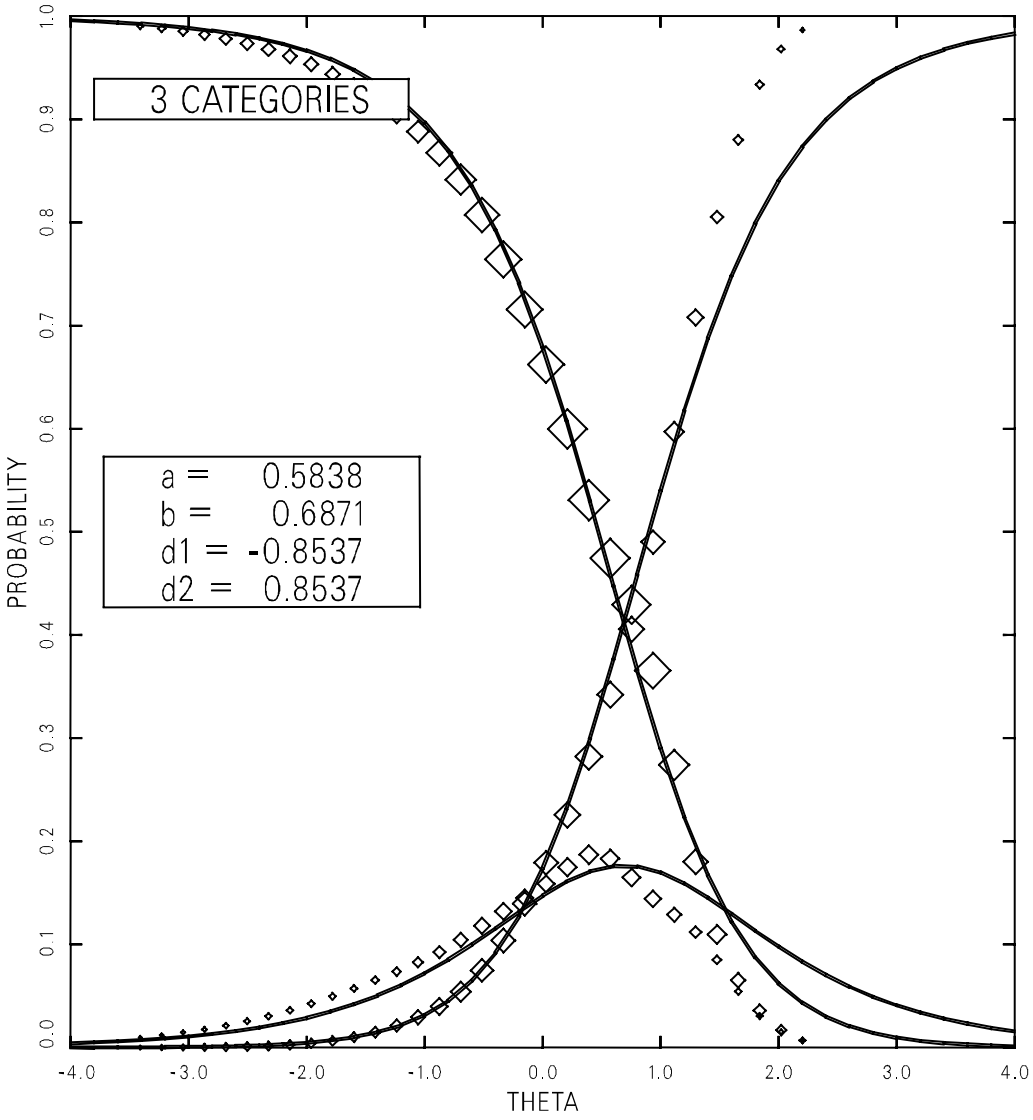
*Dichotomous Item (P041209) Exhibiting Moderate Model Misfit**



** Diamonds represent 1998 grade 12 civics 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.*

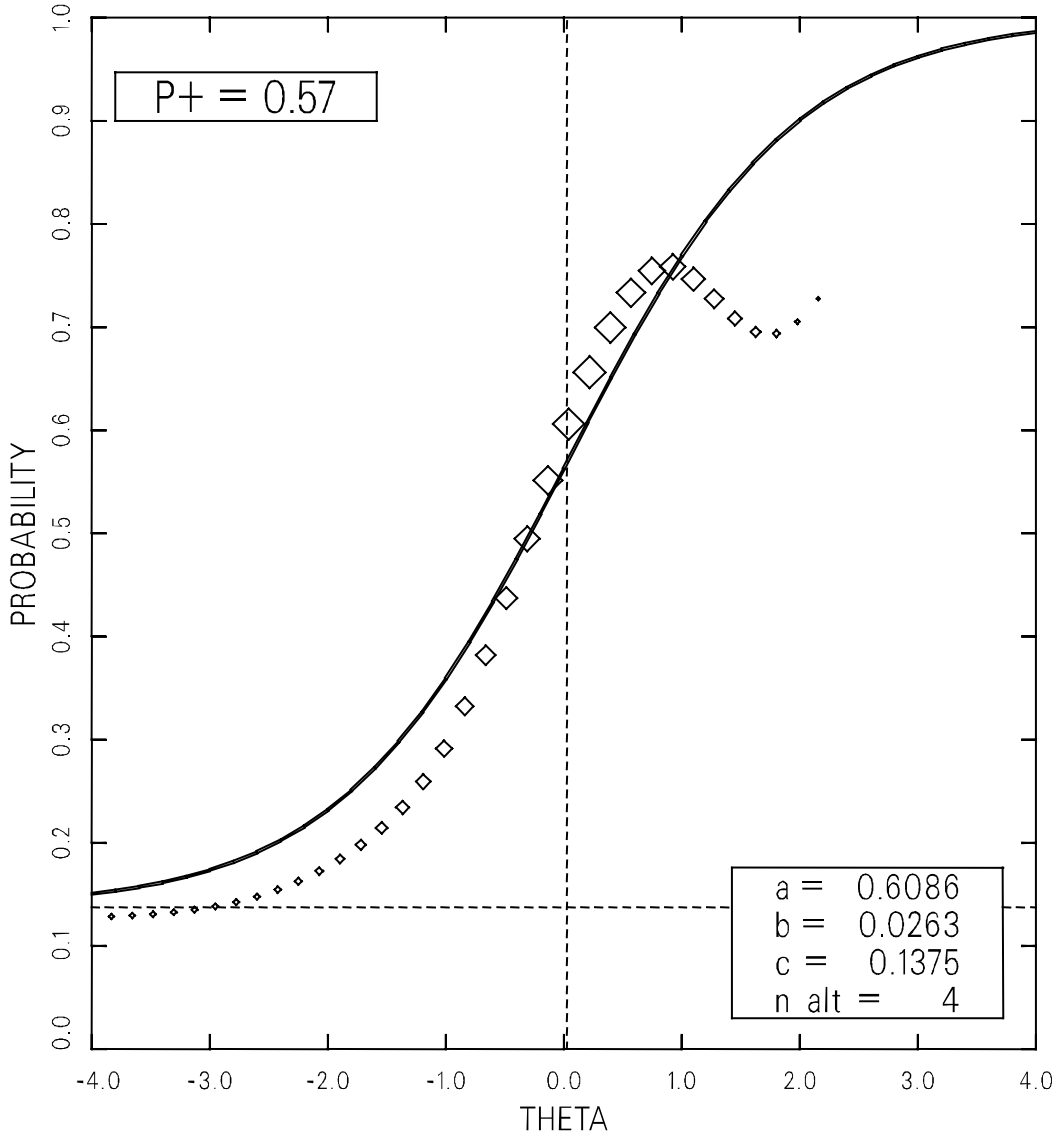
Figure 24-4

Polytomous Item (P041902) Exhibiting Moderate Model Misfit



** Diamonds represent 1998 grade 12 civics 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.*

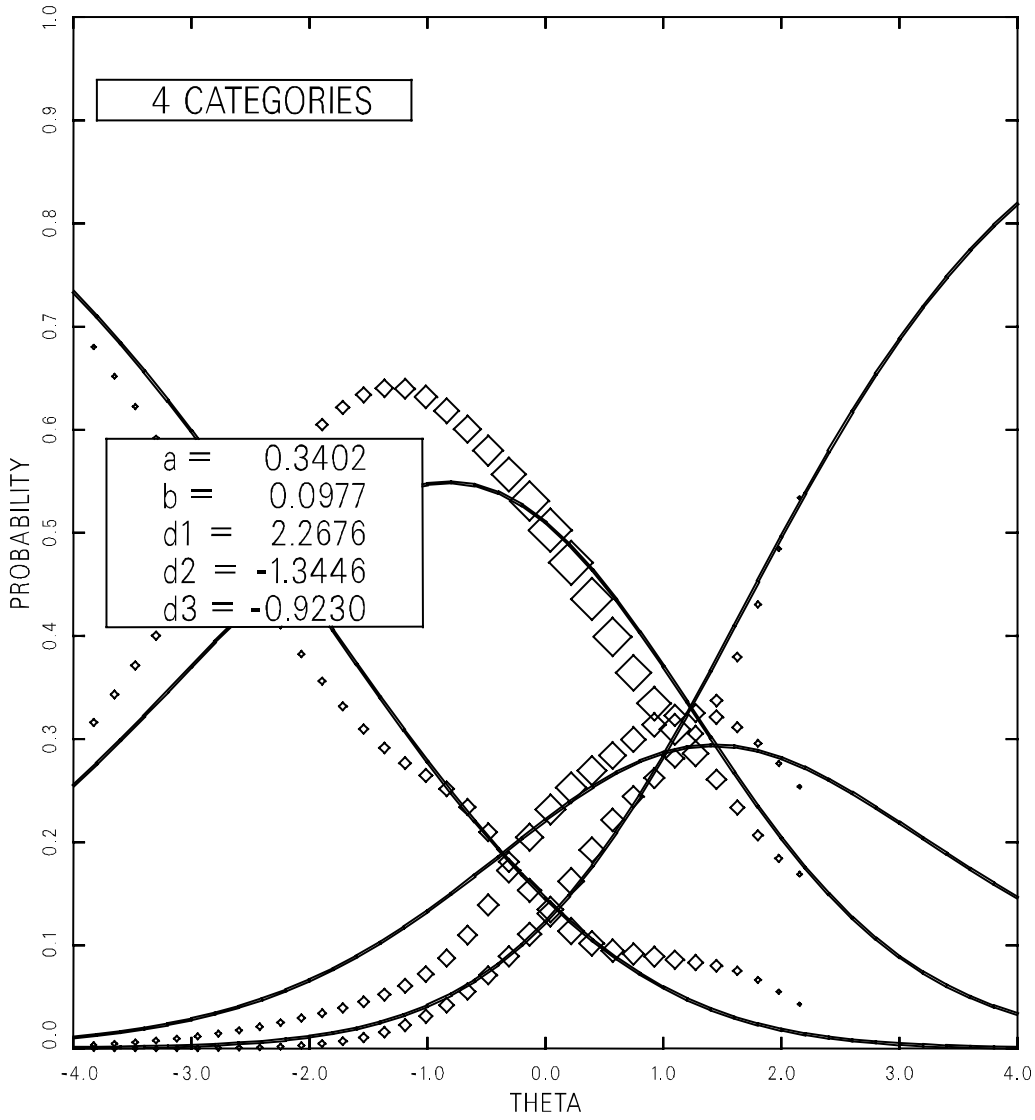
Figure 24-5
*Dichotomous Item (P040506) Exhibiting Poor Model Fit**
(Deleted from the Assessment)



** Diamonds represent 1998 grade 4 civics 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 24-6

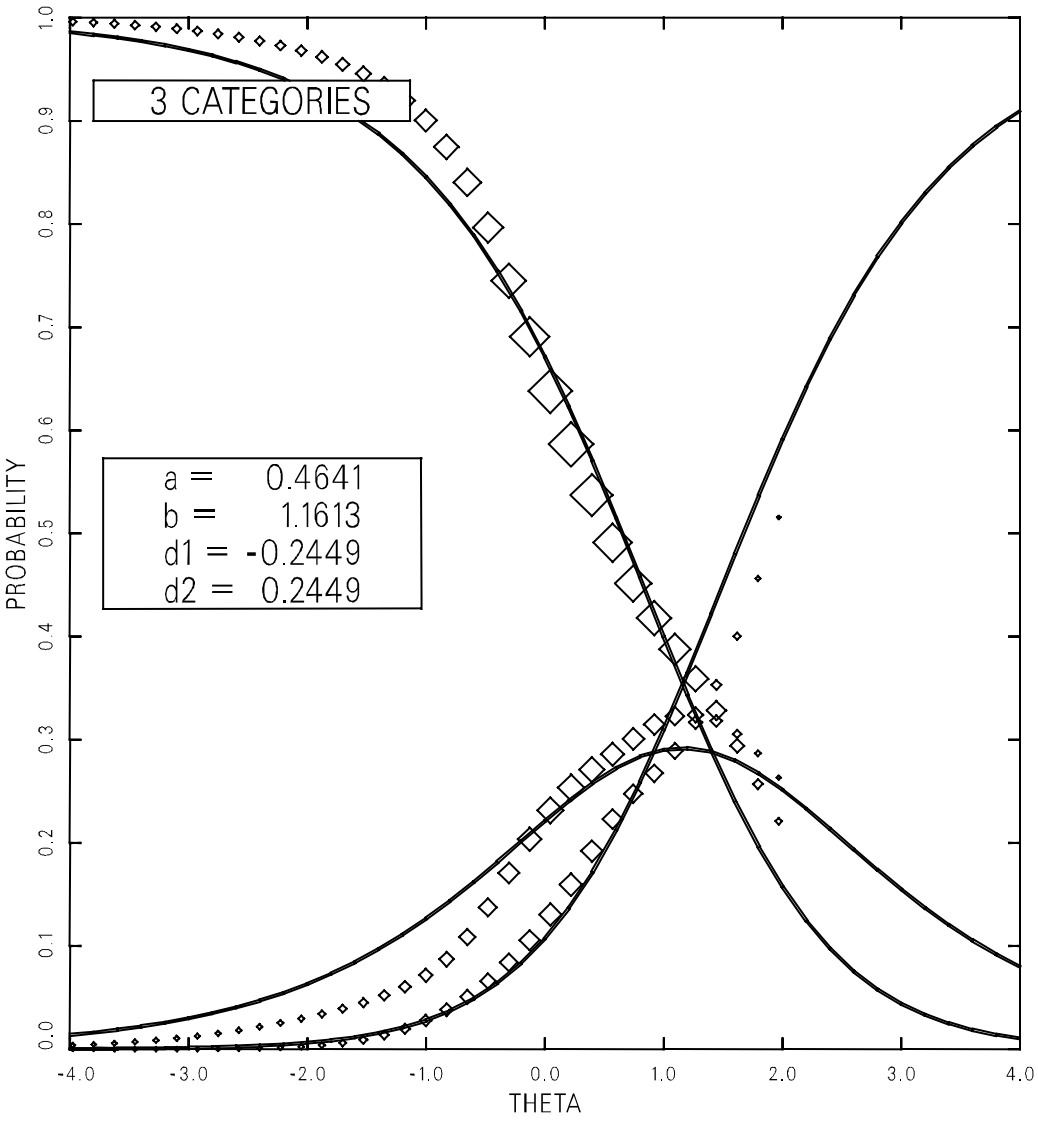
*Polytomous Item (P040402) Exhibiting Poor Model Fit in the Lower Two Categories**



** Diamonds represent 1998 grade 4 civics 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 24-7

*Same Polytomous Item (P040402) with the Lower Two Categories Collapsed, Now Exhibiting Improved Model Fit**



** Diamonds represent 1998 grade 4 civics 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.*

24.4 GENERATION OF PLAUSIBLE VALUES

For the grade sample, univariate plausible values for a single overall civics score scale were generated using the univariate conditioning program BGROUP. As with the scaling, student weights were used at this stage of the analysis. 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 869 for grade 4, 866 for grade 8, and 699 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. 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). The principal component procedure reduced the number of variables to 318 in grade 4, 320 in grade 8, and 263 in grade 12.

Research based on data from the 1990 trial state assessment suggests that results obtained using the 90 percent subset of components will differ only slightly from those obtained using the full set (Mazzeo, Johnson, Bowker, & Fong, 1992). Table 24-5 contains a list of the number of principal components included in conditioning, as well as the proportion of scale score variance accounted for by the conditioning model (as described in Chapter 12) for each grade.

The codings of the original civics-specific conditioning variables, before principal components were calculated, are presented in Appendix F. The BGROUP program estimates distributions of scale scores by combining information from item responses of individuals and information from linear regression of scale score on conditioning variables. For each individual, five plausible values are randomly drawn from their estimated scale score distribution.

Table 24-5
Proportion of Scale Score Variance Accounted for by the Conditioning Model for the National Main Civics Assessment

Grade	Number of Conditioning Contrasts	Number of Principal Components	Proportion of Scale Score Variance Accounted for
4	869	319	.64
8	866	320	.58
12	699	262	.55

The conditioning model reduces redundancy by extracting principal components from a large number of conditioning variables and basing conditioning on the components that account for 90 percent of the variance of the components (see Sections 17.4 and 20.4).

The proportion of variance of each original conditioning variable accounted for by the principal components included in the conditioning model is listed in Appendix C. The estimated conditioning effects for the principal components of the samples defined by the three grade groups are also given in Appendix C. The values of the conditioning effects are expressed in the metrics of the original calibration scale. Definitions of derived conditioning variables are given in Appendix G.

24.5 TRANSFORMATION OF THE CIVICS CALIBRATION SCALE FOR REPORTING

Since the 1998 civics assessment was developed and scaled using within-grade procedures, and since there was no prior civics assessment with a comparable framework to which it was being linked, a new reporting metric was adopted. The results are reported on 0–300 scales with identical means at each grade. As is shown in Table 24-6, the mean of the civics scale was set at 150 for each grade, and the standard deviation at 35.

Table 24-6
Means and Standard Deviations for the Civics Scale

	All Five Plausible Values	
Grade	**Mean**	**S. D.**
4	150.0	35.0
8	150.0	35.0
12	150.0	35.0

If the achievement distribution were normal, we would expect this range to cover about 99.998 percent of the distribution. Note that any transformed scale scores below 0 were censored to values of 0. A total of three scores in grade 4, six scores in grade 8, and five scores in grade 12 were censored to values of 0. Had any transformed scale scores been greater than 300, they would have been censored to values of 300; however, no such cases were encountered.

Constraining the mean and standard deviation of the scales in this way also constrained, to some degree, the percentile distributions for the total group. However, within-grade comparisons of percentiles across subgroups continue to provide valuable comparative information, although cross-grade comparisons, with each grade set to the same mean and standard deviation, do not have meaning.

For each grade, the target mean and standard transformation resulted from applying the linear transformation:

$$\theta_{target} = A \bullet \theta_{calibrated} + B,$$

where A and B are linear transformation constants. The values of A and B for each grade are given in Table 24-7. These numbers are documented for researchers who wish to reproduce these analyses, and equally, for archival purposes for those who carried out these analyses.

Table 24-7
Transformation Constants for the National Main Civics Assessment

Grade	A	B
4	39.98	149.36
8	38.49	149.68
12	37.87	149.46

24.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE

Within each grade, the error variance of the reporting scale mean was partitioned according to the procedure described in Chapter 12. The variance is partitioned into 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. Table 24-8 contains 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 (for stability, the estimates of the between-imputation variance, *B*, in Equation 12.12 are based on 100 imputations for each student). Table 24-8 shows that the preponderance of error variance is attributable to student sampling. More detailed information by gender and race/ethnicity is presented in Appendix H.

Table 24-8
Estimation Error Variance and Related Coefficients for the National Main Civics Assessment

		Proportion of Variance Due to...	
Grade	Total Estimation Error Variance	Student Sampling	Latency of θ
4	.54	.90	.10
8	.32	.91	.09
12	.62	.95	.05

* Since θ is unobserved, or "latent," a proportion of the estimation error is due to the fact that θ is known imperfectly.

24.7 CIVICS TEACHER QUESTIONNAIRE

Teachers of fourth- and eighth-grade students assessed in civics were surveyed. Along with a variable that indicated whether a student record had been matched with a teacher record, variables derived from the questionnaire were used in the conditioning models for the grade 4 and the grade 8 samples. These variables were included, so that means for subgroups defined by these variables could be compared with no bias. Of the 5,948 fourth-grade students in the main sample, 5,110 (86%) were matched with both parts of the teacher questionnaire and 277 (5%) were matched with only the first part of the questionnaire. Of the 8,212 eighth-grade students in the main sample, 6,053 (74%) were matched with both parts of the teacher questionnaire and 649 (8%) were matched with only the first part of the questionnaire. Thus, 91 percent of the fourth-graders and 82 percent of the eighth-graders were matched with at least the background information about their civics teachers.