Chapter 13

DATA ANALYSIS FOR THE SCIENCE ASSESSMENT[1]

John R. Donoghue, Jinming Zhang, Steven P. Isham,
Lois H. Worthington, and Ingeborg U. Novatkoski
Educational Testing Service

13.1 OVERVIEW

This chapter describes the analyses performed on the responses to the cognitive and background items in the 1996 assessment of science. These analyses led to the results presented in the *NAEP 1996 Science Report Card for the Nation and the States: Findings from the National Assessment of Educational Progress* (O'Sullivan, Reese, & Mazzeo, 1997). The emphasis of this chapter is on the methods and results of procedures used to develop the IRT-based scale scores that formed the basis of these reports. However, some attention is given to the analysis of constructed-response items as reported in the *NAEP 1996 Science Report Card for the Nation and the States*. The theoretical underpinnings of the IRT and plausible values methodology described in this chapter are given in Chapter 11, and several of the statistics are described in Chapter 9.

For 1996, the NAEP science 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 science. The 1996 assessment included the use of hands-on science tasks and theme blocks as well as considerably more constructed-response items than previous NAEP assessments.

The student samples that were administered science items in the 1996 assessment are shown in Table 13-1. Chapters 1 and 3 contain descriptions of the target populations and the sample design used for the assessment.

Table 13-1
NAEP 1996 Science Student Samples

Sample	Booklet IDs Number	Mode	Cohort Assessed	Time of Testing[1]	Number Assessed
4 [Science Main]	201-237	Print	Grade 4	1/3/96 - 3/29/96 (Winter)	7,305
8 [Science Main]	201-237	Print	Grade 8	1/3/96 - 3/29/96 (Winter)	7,774
12 [Science Main]	201-237	Print	Grade 12	1/3/96 - 3/29/96 (Winter)	7,537
12 [Sci-Advanced]	238-240	Print	Grade 12	1/3/96 - 3/29/96 (Winter)	2,431

[1]Final makeup sessions were held April 1-5, 1996.

LEGEND:

Print	Printed administration
Main	Main assessment
Advanced	Assessment with advanced booklets

[1] John Donoghue was the primary person responsible for the planning, specification, and coordination of the science analyses. He was assisted by Jinming Zhang. Computer activities for all science scaling and data analyses were directed by Steve Isham and completed by Lois Worthington and Ingeborg Novatkoski. Others contributing to the analysis of science data were David S. Freund, Katharine Pashley, and Norma A. Norris.

The objectives of the science analyses were to

- prepare scale values and estimate subgroup proficiency distributions for national samples of students who were administered science items from the main assessment, and

- prepare the analysis of the advanced science assessment. The advanced science sample 12[Sci-Advanced] is a separate sample from the 12[Science Main] sample. Analyses of the advanced science assessment will be described in a subsequent NAEP report. The 12[Sci-Advanced] sample is discussed further in section 13-3.

The 1996 science samples were analyzed to provide comparisons of science achievement for various subgroups of the 1996 target populations. The target populations were grade 4, grade 8, and grade 12 students in the United States. Unlike previous NAEP assessments, only grade-defined cohorts were assessed in the 1996 NAEP. The age of students was based on a calendar year, with birthdates in 1986, 1982, and 1978, respectively, for ages 9, 13, and 17. The sampled students in each of these three cohorts were assessed in the winter (January to March with final makeup sessions held during the first week of April). As described in Chapter 9, the reporting sample for the national science assessment consisted of students in the S2 sample (also see Chapter 19 for tables describing the students assessed and the reporting sample for each component of the science assessment).

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

1. conventional item and test analyses, and DIF analyses (Section 13.4);
2. item response theory (IRT) scaling (Section 13.5);
3. estimation of subgroup proficiency distributions based on the "plausible values" methodology (Section 13.6);
4. transforming the 1996 assessment scales to the final reporting metric for each of the fields of science, and (Section 13.7.1); and
5. creation of the science composite scale (Section 13.7.2).

Section 13.8 describes the results of partitioning the error variance, 13.9 discusses the matching of student responses to those of their teachers, and 13.10 provides a brief explanation of sampling weights.

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.

13.2 DESCRIPTION OF ITEMS AND ASSESSMENT BOOKLETS

The 1996 NAEP main science assessment differed from the long-term trend assessment in regard to the sample age definition, time of testing, the objectives that define the emphasis of the assessment, and most of the items used. It also differed from the 1990 main NAEP science assessment in the same regards. Because of these differences, equating or linking the main and the long-term trend assessments was not appropriate. Neither is a direct comparison to the results of the 1990 main science assessment. The 1996 main science assessment can be used to start a new baseline for measuring trends in the nation.

The pool of items used in the 1996 science assessment contained a range of constructed-response and multiple-choice questions measuring performance on sets of objectives. The items in the assessment

were based on the curriculum framework described in *Science Framework for the 1996 National Assessment of Educational Progress* (National Assessment Governing Board, 1993). The total number of scaled items was 136, 190, and 186, respectively, for grades 4, 8, and 12. Note that some items overlap across grade. Each of the items was classified into one of three fields of science: *earth science*; *physical science*; and *life science*. These three fields of science constituted the scales used in 1996 reporting. Table 13-2 shows the numbers of items within content area scales for each grade. The numbers presented in Table 13-2 show item counts both for the original item pool, and after the necessary adjustments were made during scaling (see Section 13.5.2, below).

Table 13-2
Number of Items in Scales in the Science Main Assessment by Field of Science

Grade		Physical Science	Earth Science	Life Science	Total
4	Pre-Scaling	45	53	47	145
	Post-Scaling	43	49	44	136
8	Pre-Scaling	63	65	66	194
	Post-Scaling	62	63	65	190
12	Pre-Scaling	60	64	66	190
	Post-Scaling	59	62	65	186

For each grade, the items were divided into 15 mutually exclusive, separately timed blocks. At grade 4, students were given 20 minutes to complete each block; at grades 8 and 12 each block required 30 minutes. As described in Chapter 2, the blocks were combined into booklets according to a complex spiraling design. (See Chapter 4 for more information about the blocks and booklets.) Each student's booklet contained three blocks of cognitive items. Four of the 15 blocks were hands-on tasks in which students were given a set of equipment and asked to conduct an investigation and answer questions (mostly constructed-response) relating to the investigation. These hands-on tasks were always presented in the last position, after two paper-and-pencil blocks. Three of the remaining 11 blocks were theme blocks. Theme blocks were placed randomly in student booklets, but not in every booklet. No student received more than one theme block. Each theme block was paired with each non-theme paper-and-pencil block just once. Each paper-and-pencil block appeared in the first or second position the same number (3 or 4) of times. For each of the grades, the composition of each block of items, in terms of content and format, is given in Tables 13-3 through 13-5.[2] Common labeling of these blocks across grade levels does not denote common items.

[2] The numbers in Tables 13-2 through 13-8 differ slightly from those given in Chapter 2. The numbers in Chapter 2 do not reflect the grouping of certain sets of items into cluster items for the purposes of scaling.

Table 13-3

1996 NAEP Science Block Composition by Item Type, Grade 4, As Defined Before Scaling

		Constructed-Response Items Scored Polytomously					
Block	**Multiple-Choice Items**	**2-category**	**3-category**	**4-category**	**5-category**	**Cluster Items**	**Total Items**
S3	0	1	6	0	0	0	7
S4	1	4	1	1	0	0	7
S5	2	0	4	1	1	0	8
S6	0	0	3	2	0	0	5
S7	2	0	7	1	0	0	10
S8	1	0	7	0	1	0	9
S9	2	0	6	1	0	0	9
S10	6	0	6	0	0	0	12
S11	6	1	4	0	0	0	11
S12	6	0	8	1	0	0	15
S13	6	0	4	1	0	0	11
S14	5	0	5	0	0	0	10
S15	3	0	5	1	0	0	9
S20	6	0	2	3	0	0	11
S21	5	0	4	1	1	0	11
Total	**51**	**6**	**72**	**13**	**3**	**0**	**145**

Table 13-4

1996 NAEP Science Block Composition by Item Type, Grade 8, As Defined Before Scaling

		Constructed-Response Items Scored Polytomously					
Block	**Multiple-Choice Items**	**2-category**	**3-category**	**4-category**	**5-category**	**Cluster Items**	**Total Items**
S3	0	0	4	1	0	1	6
S4	3	0	4	3	0	0	10
S5	0	0	8	0	0	0	8
S6	0	0	5	2	0	0	7
S7	2	2	8	0	0	0	12
S8	5	0	5	0	0	0	10
S9	3	0	9	1	0	0	13
S10	8	1	6	1	0	0	16
S11	8	0	7	1	0	0	16
S12	8	1	5	2	0	0	16
S13	8	0	7	1	0	0	16
S14	7	0	8	0	0	1	16
S15	7	1	6	1	1	0	16
S20	8	0	6	2	0	0	16
S21	7	0	7	2	0	0	16
Total	**74**	**5**	**95**	**17**	**1**	**2**	**194**

Table 13-5

1996 NAEP Science Block Composition by Item Type, Grade 12, As Defined Before Scaling

		Constructed-Response Items Scored Polytomously					
Block	Multiple-Choice Items	2-category	3-category	4-category	5-category	Cluster Items	Total Items
S3	0	0	4	1	1	0	6
S4	0	0	1	1	2	0	4
S5	0	1	7	0	0	0	8
S6	0	1	5	1	1	0	8
S7	5	0	7	3	0	0	15
S8	6	0	7	0	0	0	13
S9	4	0	8	1	1	0	14
S10	8	0	9	1	0	0	18
S11	8	1	4	3	0	0	16
S12	8	1	5	2	0	0	16
S13	8	0	7	1	0	0	16
S14	8	0	6	1	1	0	16
S15	0	1	5	2	0	0	8
S20	7	0	6	3	0	0	16
S21	8	0	4	4	0	0	16
Total	**70**	**5**	**85**	**24**	**6**	**0**	**190**

Some items (fewer than 10%) received special treatment during scaling. For each of the grades, Tables 13-6 through 13-8 show the composition of each block after deletions of items and collapsing of categories for polytomously-scored constructed-response items as a result of scaling. If data had poor fit with the response model for an item, the item was deleted. If a constructed-response item was scored in multiple categories but one category had no (or very few) responses, or one of the categories had responses that had poor fit to the response model, that category was combined with other categories ("collapsed"). All item deletions and all but one category collapse were performed in the course of scaling the national science assessment data; the remaining collapse was performed based on data in State Assessment, with the same collapse performed for the national scaling. In addition, categories of a small number of items were combined These changes were made so that the scaling model used for these items fit the data more closely, and are described more fully in Section 13.5.

For grade 4, each of the 11 paper-and-pencil blocks contained from five to nine constructed-response items. Seven of these blocks contained one or more constructed-response items scored on a 0-3 scale. Two items were scored on a 0-4 scale. The four hands-on task blocks contained from five to seven constructed-response items and up to two multiple-choice items.

For grade 8, each of the 11 paper-and-pencil blocks contained from five to ten constructed-response items. Eight of these blocks contained one or more constructed-response items scored on a 0-3 scale. One item was scored on a 0-4 scale. The four hands-on task blocks contained from six to eight constructed-response items. One of these blocks also contained three multiple-choice items.

For grade 12, each of the 11 paper-and-pencil blocks contained from seven to ten constructed-response items. Ten of these blocks contained one or more constructed-response items scored on a 0-3 scale. Two items were scored on a 0-4 scale. The four hands-on task blocks contained from four to eight constructed-response items. None of these blocks contained multiple-choice items.

Table 13-6

*1996 NAEP Science Block Composition by Item Type, Grade 4, As Defined After Scaling**

		Constructed-Response Items Scored Polytomously					
Block	Multiple-Choice Items	2-category	3-category	4-category	5-category	Cluster Items	Total Items
S3	0	1	4	0	0	1	6
S4	1	5	1	0	0	0	7
S5	2	0	4	0	1	0	7
S6	0	0	3	1	0	0	4
S7	2	3	4	1	0	0	10
S8	1	0	7	0	1	0	9
S9	1	0	7	0	0	0	8
S10	6	0	4	0	0	1	11
S11	6	2	3	0	0	0	11
S12	6	0	5	0	0	0	11
S13	6	1	3	1	0	0	11
S14	5	0	5	0	0	0	10
S15	3	0	5	1	0	0	9
S20	6	0	2	3	0	0	11
S21	5	1	3	1	1	0	11
Total	**50**	**13**	**60**	**8**	**3**	**2**	**136**

*Counts reflect items that were dropped and collapsed.

Table 13-7

*1996 NAEP Science Block Composition by Item Type, Grade 8, As Defined After Scaling**

		Constructed-Response Items Scored Polytomously					
Block	Multiple-Choice Items	2-category	3-category	4-category	5-category	Cluster Items	Total Items
S3	0	1	3	1	0	1	6
S4	2	1	4	1	0	1	9
S5	0	1	7	0	0	0	8
S6	0	1	4	1	0	0	6
S7	2	4	6	0	0	0	12
S8	5	1	4	0	0	0	10
S9	3	2	7	1	0	0	13
S10	8	1	7	0	0	0	16
S11	8	1	6	1	0	0	16
S12	8	1	5	2	0	0	16
S13	8	1	4	1	0	1	15
S14	7	0	8	0	0	1	16
S15	6	1	6	1	1	0	15
S20	8	2	4	2	0	0	16
S21	7	0	7	2	0	0	16
Total	**72**	**18**	**82**	**13**	**1**	**4**	**190**

* Counts reflect items that were dropped and collapsed.

Table 13-8

*1996 NAEP Science Block Composition by Item Type, Grade 12, As Defined After Scaling**

		Constructed-Response Items Scored Polytomously					
Block	Multiple-Choice Items	2-category	3-category	4-category	5-category	Cluster Items	Total Items
S3	0	1	4	1	0	0	6
S4	0	0	2	2	0	0	4
S5	0	2	6	0	0	0	8
S6	0	4	2	1	1	0	8
S7	5	3	5	2	0	0	15
S8	6	2	4	0	0	0	12
S9	4	3	5	1	1	0	14
S10	8	0	9	1	0	0	18
S11	8	1	4	3	0	0	16
S12	8	1	5	2	0	0	16
S13	8	0	5	1	0	1	15
S14	8	0	6	2	0	0	16
S15	0	2	4	2	0	0	8
S20	6	0	6	2	0	0	14
S21	8	0	4	4	0	0	16
Total	**69**	**19**	**71**	**24**	**2**	**1**	**186**

* Counts reflect items that were dropped and collapsed.

All constructed-response items were scored by specially trained readers, as described in Chapter 5. In addition, a small number of "cluster items" were formed. A cluster item is an aggregation of a group of items (in the case of NAEP science, typically two to four items) that are related to a single content strand, topic, or stimulus, and are developed and scored as a single unit (see Wainer & Kiely, 1987, for further details and examples of different types of cluster items). Some items were initially scored as cluster items, and the additional clusters were formed in scaling due to data dependencies.

In the main samples, each student was administered a booklet containing two paper-and-pencil blocks and one block consisting of a hands-on task. In addition, the booklet contained a block of background questions common to all booklets for a particular grade level, a block of questions concerning the student's motivation and his or her perception of the difficulty of the cognitive items, and a block of science-related background questions common to all science booklets for a particular grade level.

The design of the 1996 science assessment required that each student be administered one of the 37 booklets in the design. Within each administration site, all booklets were "spiraled" together in a random sequence and distributed to students sequentially, in the order of the students' names on the Student Listing Form (see Chapter 4). As a result of the design and the spiraling of booklets, a considerable degree of balance was achieved in the data collection process. Each block of items (and, therefore, each item) was administered to randomly equivalent samples of students.

13.3 SPECIAL SCIENCE ASSESSMENT

As stated previously, there was a special study in the 1996 national NAEP assessment in addition to the main and long-term trend assessments. This was the advanced study, denoted by the 12[Sci-Advanced] sample in Table 13-1. This study examined the performance of twelfth-

grade students who were taking advanced science courses. Students were assessed for approximately two hours, and each student received four cognitive blocks, consisting of a common block (SS, composed of 18 items) and three special blocks (each composed of 16 items) designed to assess advanced material. Each block assessed specific science content: one block for physics, one for chemistry, and one for life sciences. The common block was composed of items from the main assessment, although these items were drawn from several different blocks from the main assessment. The block structure of the special study booklets is provided in Table 13-9. In addition to the cognitive blocks, the special study booklets had three blocks in common with the main assessment:

1. a general student background block (CS)
2. a science background block (SB), and
3. a motivation block (SX).

Table 13-9
Block Structure of the 12[Sci-Advanced] Special Study Booklets

Booklet IDs	Cognitive Blocks			
	First	Second	Third	Fourth
238	SS	SR	SV	SP
239	SS	SV	SP	SR
240	SS	SP	SR	SV

The advanced study was not part of the main assessment and analyses for these booklets will be described in an NCES publication by Christine O'Sullivan to be published in the third quarter of 1999. Therefore, the special 12[Sci-Advanced] sample will not be discussed further in this chapter.

13.4 ITEM ANALYSES

13.4.1 Conventional Item and Test Analyses

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

Tables 13-10, 13-11, and 13-12 show the number of items, average weighted item score, average weighted item-to-total score correlation (biserial or polyserial), and weighted alpha reliability for each block administered at each grade level for the main assessment. The table also gives the number of students who were administered the block and the weighted percent reaching the last item in the block. Student sampling weights were used to compute all statistics except for the sample sizes. Preliminary item analyses for all items within a block were completed before scaling; however, the results shown here reflect the characteristics of the items that contributed to the final science scales.

Table 13-10

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

Statistic	Position	S7	S8	S9	S10	S11	S12	S13	S14	S15	S20	S21
Number of scaled items		10	9	8	11	11	11	11	10	9	11	11
Unweighted sample size												
	1	812	804	747	630	608	609	607	782	577	561	557
	2	745	822	749	578	595	593	616	798	599	608	583
	ALL	1557	1626	1496	1208	1203	1202	1223	1580	1176	1169	1140
Weighted average item score												
	1	.47	.47	.38	.46	.52	.48	.55	.32	.31	.49	.49
	2	.46	.47	.37	.47	.52	.47	.52	.30	.29	.47	.45
	ALL	.47	.47	.38	.47	.52	.48	.53	.31	.30	.48	.47
Weighted alpha reliability												
	1	.69	.57	.58	.67	.64	.73	.62	.50	.53	.61	.64
	2	.69	.52	.61	.67	.69	.76	.66	.49	.53	.65	.67
	ALL	.69	.55	.59	.67	.67	.74	.64	.49	.53	.63	.66
Weighted average r-polyserial												
	1	.59	.57	.61	.57	.61	.62	.53	.53	.59	.54	.59
	2	.59	.57	.64	.57	.64	.66	.57	.54	.60	.58	.60
	ALL	.59	.57	.63	.57	.62	.64	.55	.54	.60	.56	.60
Weighted proportion of students attempting last item												
	1	.87	.63	.79	.85	.88	.87	.87	.79	.73	.80	.56
	2	.91	.74	.85	.87	.87	.93	.92	.84	.80	.88	.71
	ALL	.88	.68	.82	.86	.88	.90	.90	.82	.77	.84	.63

Table 13-11

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

Statistic	Position	S7	S8	S9	S10	S11	S12	S13	S14	S15	S20	S21
Number of scaled items		12	10	13	16	16	16	15	16	15	16	16
Unweighted sample size												
	1	833	838	830	631	649	638	629	836	625	627	622
	2	828	849	843	609	621	634	640	846	630	619	621
	ALL	1661	1687	1673	1240	1270	1272	1269	1682	1255	1246	1243
Weighted average item score												
	1	.45	.57	.49	.35	.43	.33	.41	.42	.37	.42	.44
	2	.43	.54	.47	.33	.42	.33	.38	.42	.36	.42	.42
	ALL	.44	.55	.48	.34	.43	.33	.40	.42	.37	.42	.43
Weighted average r-polyserial												
	1	.76	.65	.68	.61	.71	.70	.59	.72	.70	.70	.72
	2	.75	.64	.68	.64	.71	.73	.61	.74	.69	.70	.71
	ALL	.75	.65	.68	.63	.71	.71	.61	.73	.70	.70	.72
Weighted alpha reliability												
	1	.65	.68	.58	.47	.52	.52	.45	.54	.55	.51	.52
	2	.65	.66	.57	.50	.52	.54	.46	.56	.54	.52	.51
	ALL	.65	.67	.58	.49	.52	.53	.46	.55	.55	.51	.51
Weighted proportion of students attempting last item												
	1	.89	.97	.96	.92	.84	.82	.92	.96	.89	.86	.88
	2	.89	.97	.93	.90	.83	.82	.89	.96	.85	.84	.88
	ALL	.89	.97	.95	.91	.84	.82	.90	.96	.87	.85	.88

Table 13-12

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

Statistic	Position	S7	S8	S9	S10	S11	S12	S13	S14	S15	S20	S21
Number of scaled items		15	12	14	18	16	16	15	16	8	14	16
Unweighted sample size												
	1	813	835	750	613	651	603	626	830	583	601	611
	2	783	842	801	602	571	609	591	838	597	645	619
	ALL	1596	1677	1551	1215	1222	1212	1217	1668	1180	1246	1230
Weighted average item score												
	1	.49	.56	.44	.37	.47	.48	.50	.40	.17	.41	.42
	2	.49	.56	.41	.35	.45	.46	.51	.40	.17	.41	.42
	ALL	.49	.56	.43	.36	.46	.47	.50	.40	.17	.41	.42
Weighted alpha reliability												
	1	.76	.56	.79	.76	.67	.77	.64	.75	.62	.70	.72
	2	.78	.63	.80	.77	.66	.78	.69	.76	.62	.69	.73
	ALL	.77	.59	.80	.76	.67	.77	.67	.75	.62	.70	.72
Weighted average r-polyserial												
	1	.63	.61	.62	.57	.46	.58	.54	.58	.69	.54	.54
	2	.65	.64	.65	.58	.47	.58	.57	.58	.69	.53	.55
	ALL	.64	.63	.64	.58	.47	.58	.55	.58	.69	.53	.54
Weighted proportion of students attempting last item												
	1	.80	.93	.82	.87	.92	.81	.91	.85	.79	.79	.91
	2	.79	.92	.88	.85	.95	.80	.93	.87	.77	.80	.86
	ALL	.80	.92	.85	.86	.93	.81	.92	.86	.78	.80	.88

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 observed response (omitted). Standard practice at ETS is to treat all nonrespondents to the last item as if they had not reached the item. Items that were not reached were treated as if they had not been presented to the examinee, while omitted items were regarded as incorrect. The proportion of students attempting the last item of a block (or, equivalently, one minus the proportion not reaching the last item) can be used as an index of the degree of speededness of the block of items.

As is evident from Tables 13-10 through 13-12, the difficulty and the average item-to-total correlations of the blocks varied somewhat. Such variability was expected since these blocks were not created to be parallel in either difficulty or content. Based on the proportion of students attempting the last item, no block seemed to be speeded, by the criterion of a proportion less than .8 attempting the last item. The most speeded block showed 84 percent of the students reaching the last item in the block.

For the 11 paper-and-pencil blocks, small but consistent differences were noted based upon whether a block occurred in the first or second position within a booklet. When the block appeared first in the booklet, the average item score tended to be higher and the average polyserial correlation tended to be lower. The largest differences were noted in the proportion of students not attempting the last item in the block; more students attempted the last item when the block appeared in the second position. It appears that the students learned to pace themselves through the second block, based on their experience with the first block. Similar effects (slightly larger) were noted in the 1992 NAEP reading assessment (Donoghue, 1994). At that time, a study was completed to examine the effect of the serial position differences. Due to the balance of the complex design of the booklets, the serial position differences were found to have minimal effects on scaling.

13.4.2 Scoring the Constructed-Response Items

As indicated in Table 13-4 through 13-6, about two-thirds of the science items were constructed-response. Two-category constructed-response items were given a right/wrong scoring. The categories of responses for the items and the number of responses that were rescored for each item are indicated in Appendix I. The percent agreement for the raters and Cohen's (1968) Kappa, a reliability estimate appropriate for items that are dichotomized, are also given in the tables. For grades 4 and 12, a 20 percent sample was used in calculating the reliability. At grade 8, the national and State Assessment data were combined. A 6 percent sample of these combined data was used to calculate the reliability results.

In general, the rater reliability of the scoring for dichotomized responses was quite high. Cohen's Kappa reliabilities ranged over items from .86 to .93 for grade 4, from .75 to .96 for grade 8, and from .71 to .95 for grade 12.

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 regular, short and extended constructed-response items as well as the range of interrater reliabilities that were obtained. Extended constructed-response items were scored on a scale from 1 to 5 to reflect degrees of knowledge. In scaling, this scale is shifted to 0 to 4. Rating information on extended constructed-response items can be found in Appendix I, which lists the sample sizes, percent agreement, and Cohen's Kappa reliability index.

13.4.3 Differential Item Functioning

A differential item functioning (DIF) analysis of the science items was done to identify potentially biased items. Sample sizes were large enough to compare male and female students, White and Black students, and White and Hispanic students. 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.

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 DIF analyses, weights were rescaled separately for each comparison, as described in Chapter 9. DIF analyses were conducted separately by grade. For dichotomous items, the DIF index generated by the Mantel-Haenszel procedure was used to place items into one of three categories: "A," "B," or "C." "A" items exhibit no DIF, "B" items exhibit a weak indication of DIF, and "C" items exhibit a strong indication of DIF. "C" items were examined by a DIF committee for presence of bias.

Table 13-13 summarizes the results of the DIF analyses for dichotomous items. Focal groups are female, Black, and Hispanic groups.

Table 13-13
DIF Category by Grade for Dichotomous Items

Grade	DIF Category[1]	Analysis Male/Female	White/Black	White/Hispanic
4	C-	3	0	0
	B-	13	4	1
	A-	21	26	33
	A+	19	26	23
	B+	2	2	1
	C+	0	0	0
8	C-	3	0	0
	B-	10	6	2
	A-	41	33	34
	A+	23	40	43
	B+	4	2	2
	C+	0	0	0
12	C-	1	0	0
	B-	13	4	2
	A-	38	31	33
	A+	22	35	41
	B+	5	9	3
	C+	0	0	0

[1] 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.

Positive values indicate items that were differentially easier for the focal group. Table 13-14 summarizes the results of the DIF analyses for polytomous items. Again, focal groups are female, Black, and Hispanic groups, and positive values indicate that the item was differentially easier for the focal group. The Mantel statistic provides a statistical test of the hypothesis of no DIF. To aid in interpreting the results for polytomous items, the standardized mean difference between focal and reference groups was produced. This statistic was rescaled

by dividing the standardized mean differences by the standard deviation of the respective item. The description of this procedure can be found in Chapter 9. For polytomous items, a standardized mean difference ratio of .25 or greater (coupled with a significant Mantel statistic) was considered a strong indication of DIF. It can be shown that standardized mean difference ratios of .25 are at least as extreme as Mantel-Haenszel statistics corresponding to "C" items (see Chapter 9 for details).

Table 13-14
DIF Category by Grade for Polytomous Items

Grade	DIF Category[1]	Analysis Male/Female	White/Black	White/Hispanic
4	CC-	5	2	0
	BB-	4	2	1
	AA-	39	39	40
	AA+	29	36	41
	BB+	4	2	1
	CC+	2	2	0
8	CC-	7	0	2
	BB-	7	5	3
	AA-	43	48	54
	AA+	57	57	54
	BB+	3	4	3
	CC+	1	4	2
12	CC-	3	3	2
	BB-	7	2	2
	AA-	40	54	54
	AA+	49	45	45
	BB+	9	5	4
	CC+	3	2	4

[1] 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.

Following standard practice at ETS, all items identified as showing DIF were reviewed by a committee of trained test developers and subject-matter specialists. As described in Chapter 9, such committees are charged with making judgments about whether the differential difficulty of an item is *unfairly* related to group membership; that is, whether the item is biased. The committee assembled to review NAEP items that were identified as "C" or "CC" items. The committee included both ETS staff and outside members with expertise in the field. It was the committee's judgment that one of the items for the national data was functioning differentially due to factors irrelevant to test objectives. The item asked the student to list two ways that cold temperatures could cause problems. Although the item appeared to be disadvantaging Hispanic students, the committee concluded that this was probably because a large proportion of Hispanic Americans live in warmer parts of the country, and that anyone without experience of cold weather would be disadvantaged in answering this question. The item was removed from scaling due to this differential item functioning.

13.5 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 *Science Framework for the 1996 National Assessment of Educational Progress* (National Assessment Governing Board, 1993) specifies that the 1996 science assessment be developed within-grade. Likewise, all IRT scaling was performed within-grade.

Within each grade, items were grouped into three distinct sets corresponding to the three fields of science: earth, physical, and life. IRT-based scales corresponding to each of the fields of science above were developed using the scaling models described in Chapter 11. The scales summarize student performance across all three item types in the assessment (multiple-choice, short constructed-response, and extended constructed-response).

13.5.1 Item Parameter Estimation

For each fields of science scale, item parameter estimates were obtained by 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 11). The calibration was performed using all the available examinees in the reporting group. Student sampling weights were used for the analysis.

As described in Chapter 11, multiple-choice items were dichotomously scored (scored 0,1) and were scaled using the three-parameter logistic model. Omitted responses to multiple-choice items were treated as fractionally correct, with the fraction being set to the reciprocal of the number of response options for an item. All constructed-response items with two categories were dichotomously scored and were scaled using the two-parameter logistic model with the lower asymptote parameter set at 0. Omitted responses to these items were treated as incorrect. For calibration, all items that were not reached were treated as if they were not presented to the examinees. Responses to extended constructed-response items that were off-task were also treated as if they had not been presented.

A key assumption associated with IRT scales is that of conditional independence. Conditional on proficiency, examinee's item responses are assumed to be independent. When sets of items are logically dependent on each other, or are based on a single stimulus, this assumption can be violated to a degree that results in aberrant scaling results. In order to avoid possible problems with inter-item dependencies, a small number of additional cluster items was created by combining examinee responses to sets of related items into a single score for each set. The cluster items, rather than their original constituent items, were used in scaling the 1996 science assessment. Examinees omitting all constituents of the cluster item were placed in the "zero correct" category of the cluster item. Examinees classified as "not reaching" all constituent parts were treated as having not been presented the cluster item. All cluster items were scaled using the generalized partial credit model.

Each of the multi-category constructed response items was also scaled using the generalized partial credit model. These items had two, three, four, or five categories of partial credit. One cluster item had six categories. Scoring levels were labeled as shown in Table 13-15.

Table 13-15
Labels for Score Levels of Polytomous Items

Score	3-Category	4-Category	5-Category	6-Category
5				Complete
4			Complete	Essential
3		Complete	Essential	Adequate
2	Complete	Partially correct	Partially correct	Partially correct
1	Partially correct	Unsatisfactory	Unsatisfactory	Unsatisfactory
0	Wrong, off-task, or omitted	Wrong, off-task, or omitted	Wrong, off-task, or omitted	Wrong, off-task, or omitted

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 parameter estimation proceeded in two phases. First, the subject ability distribution was assumed fixed (normal [0,1]) and a stable solution was obtained. The parameter estimates from this solution were then used as starting values for a subsequent set of runs in which the subject ability distribution was freed and estimated concurrently with item parameter estimates. After each estimation cycle, the subject ability distribution was restandardized to have a mean of zero and standard deviation of one. Correspondingly, parameter estimates for that cycle were also linearly restandardized.

13.5.2 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 the final composition of the item pool making up the scales by identifying misfitting items that should not be included. Evaluations of model fit were based primarily on graphical analyses. For dichotomously-scored multiple-choice and two-category response items, model fit was evaluated by examining plots of estimates of the expected conditional (on θ) probability of a correct response that do not assume a two-parameter or three-parameter logistic model versus the probability predicted by the estimated item characteristic curve (see Mislevy & Sheehan, 1987, p. 302). For the cluster items and multiple-category constructed-response items, similar plots were produced for each item category characteristic curve (see Chapter 9).

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 (1987) point out that the usefulness of these indices appears particularly limited in situations like NAEP where examinees have been administered relatively short tests. A study by Stone, Mislevy, and Mazzeo (1994) using simulated data suggests that the correct reference chi-square distributions for these indices have considerably fewer degrees of freedom than the value indicated by the BILOG/PARSCALE program and require additional adjustments of scale. However, it is not yet clear how to estimate the correct number of degrees of freedom and necessary scale factor adjustment factors. Consequently, pseudo chi-square goodness-of-fit indices are used only as rough guides in interpreting the severity of model departures.

Second, as discussed in Chapter 9, it is almost certainly the case that, for most items, item-response models hold only to a certain degree of approximation. Given the large samples sizes used in NAEP, there will be sets of items for which one is almost certain to reject the hypothesis that the model fits the data (since the likelihood of rejecting the null increases with sample size) even though departures are minimal in nature or involve kinds of misfit unlikely to impact on important model-based inferences about student achievement. In practice, it is always wise to temper statistical decisions with judgments about the severity of model misfit and the potential impact of such misfit on final results.

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 too lenient, hence including items with model fit poor enough to invalidate the types of inferences made from NAEP results. Items that clearly did not fit the model were not included; however, a certain degree of misfit was tolerated for a number of items included in the final scales.

For the large majority of the items, the fit of the model was extremely good. Figure 13-1 provides a typical example of what the plots look like for a dichotomously-scored item in this class of items. The plot that is shown is for an item from the physical science scale. In the plot, the y-axis indicates the probability of a correct response and the x-axis indicates scale score level (θ). The crosses show estimates of the conditional (on θ) probability of a correct response that do not assume a logistic form (referred to subsequently as nonlogistic-based estimates). The sizes of the crosses are proportional to the estimated density of the theta distribution at the indicated value. The solid curve shows the estimated theoretical item response function. The item response function provides estimates of the conditional probability of a correct response based on an assumed logistic form. The vertical dashed line indicates the estimated location parameter (b) for the item and the horizontal dashed line indicates the estimated lower asymptote (c). Also shown in the plot are the actual values of the item parameter estimates (lower right-hand corner). As is evident from the plot, the 'empirical' or non-logistic-based item trace is in extremely close agreement with the model-based item response function logistic curve.

Figure 13-2 provides an example of a plot for a 4-category extended constructed-response item exhibiting good model fit. Like the plots for the dichotomously-scored multiple-choice items, this plot shows two estimates of each item category characteristic curve, one set that does not assume the partial credit model (the empirical trace shown as asterisks) and one that does (the theoretical trace shown as solid curves). The estimates for all parameters for the item in question are also indicated on the plot. As with Figure 13-1, the two sets of estimates agree quite well, although there is a slight tendency for the nonlogistic-based estimates for category two to be somewhat higher than the model-based estimates for theta values less than 1. An aspect of Figure 13-2 worth noting is the large proportion of examinees that responded in the two lowest response categories for this item[3]. Such results were typical for the extended constructed-response items. Substantial proportions of examinees were either unable or unwilling to provide even minimally adequate answers to such items.

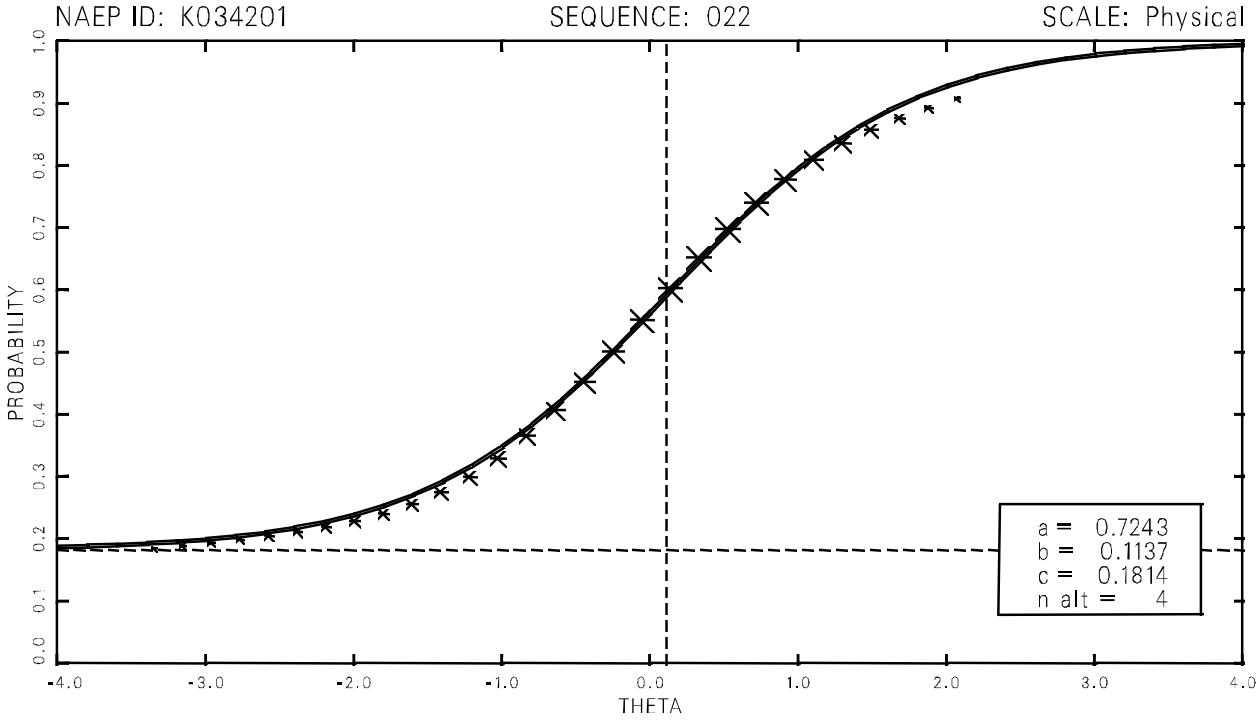
As discussed above, some of the items retained for the final scales display some degree of model misfit. Figures 13-3 (a dichotomously-scored multiple-choice item) and 13-4 (an extended constructed-response item) provide typical examples of such items. Note that in Figure 13-4, the empirical curve lies above the theoretical curve in the lower part of the ability scale for the lowest category, but below the theoretical curve for the next higher category. Combining these two categories would have improved the model fit, but it was judged that the misfit was not sufficiently pronounced in this case to warrant such collapsing. In general, good agreement between empirical and theoretical item traces were found for the regions of the theta scale that includes most of the examinees. Misfit was confined to conditional probabilities associated with theta values in the tails of the subject ability distributions.

[3] This is evidenced by the relatively large size of the asterisks indicating estimated conditional probabilities for these two categories.

Item K049907 (grade 12, earth science scale) did not fit in a run with and unconstrained (to normal) prior and with all the adjustments that had been made in the national scaling (Figure 13-5). Categories 1 and 2 were combined to yield a 0-1 dichotomous item, and the fit improved substantially (Figure 13-6).

Figure 13-1

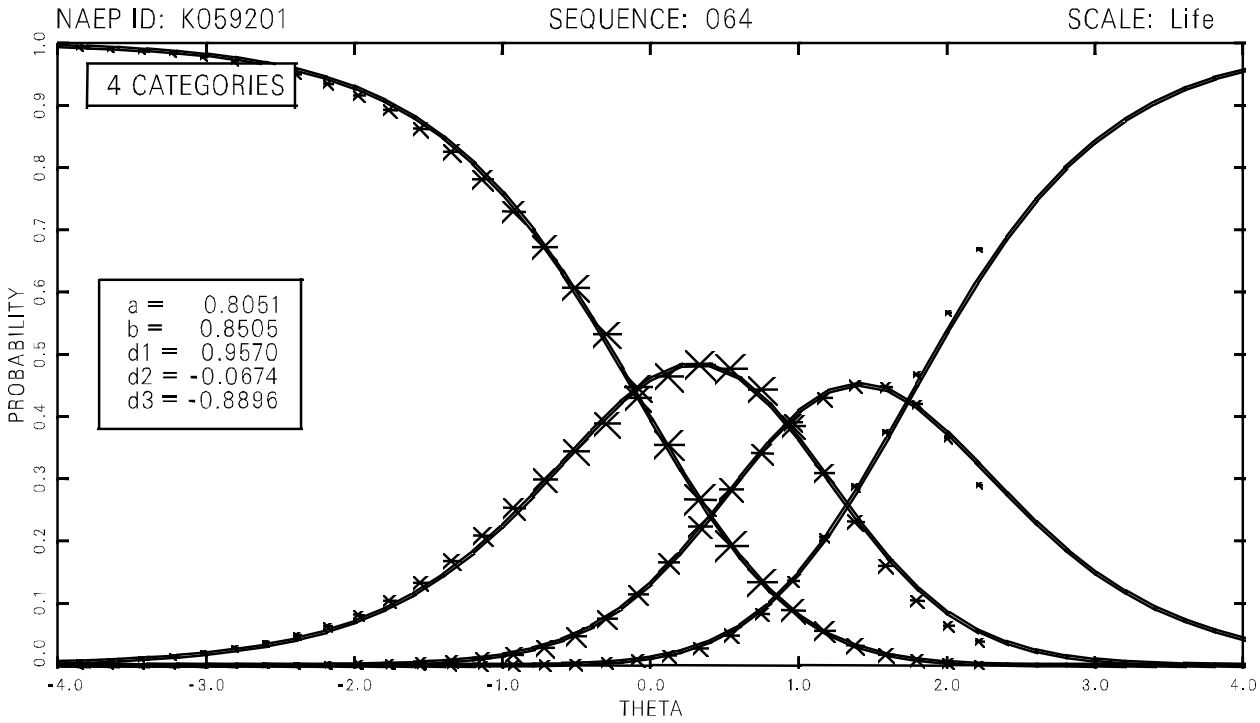
Plot Comparing Empirical and Model-Based Estimates of Item Response Functions for Binary Scored (Multiple-Choice) Items Exhibiting Good Model Fit*



*Asterisks indicate estimated conditional probabilities obtained without assuming a logistic form; the solid curve indicates estimated item response function assuming a logistic form.

Figure 13-2

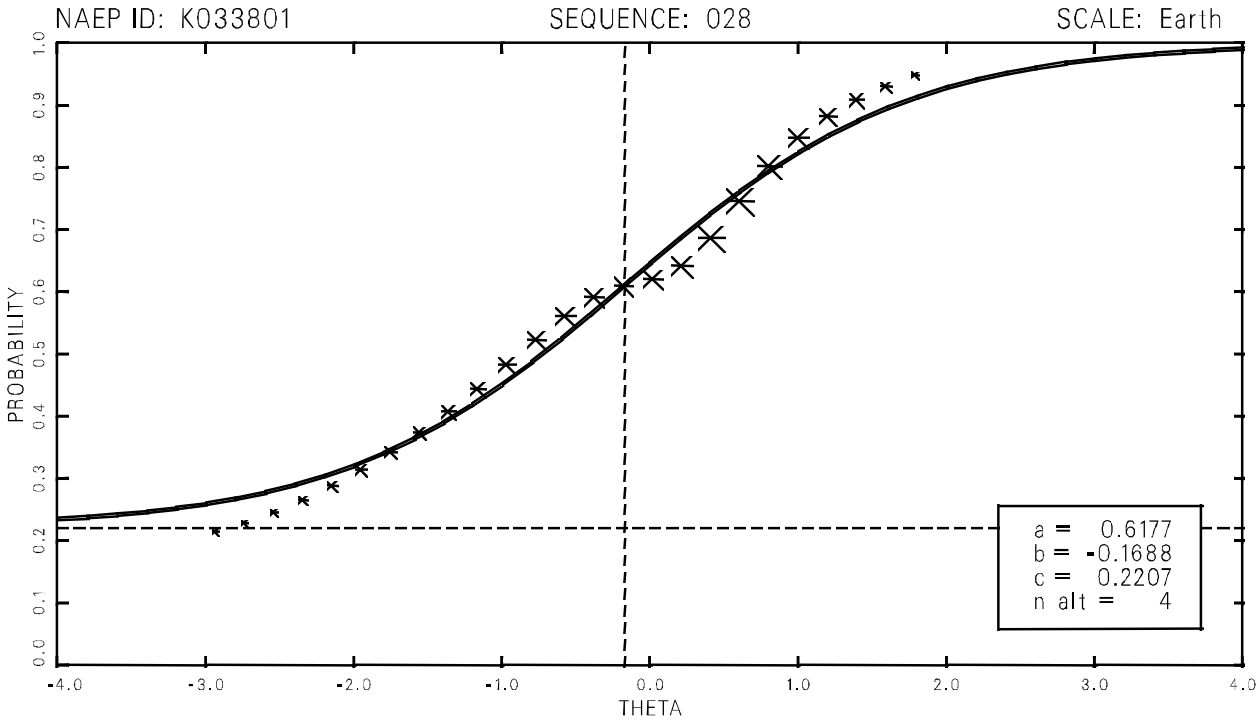
Plot Comparing Empirical and Model-Based Estimates of Item Category Characteristic Curves for a Polytomously Scored Item Exhibiting Good Model Fit*



*Asterisks indicate estimated conditional probabilities obtained without assuming a model-based form; the solid curve indicates estimated item response function assuming a model-based form.

Figure 13-3

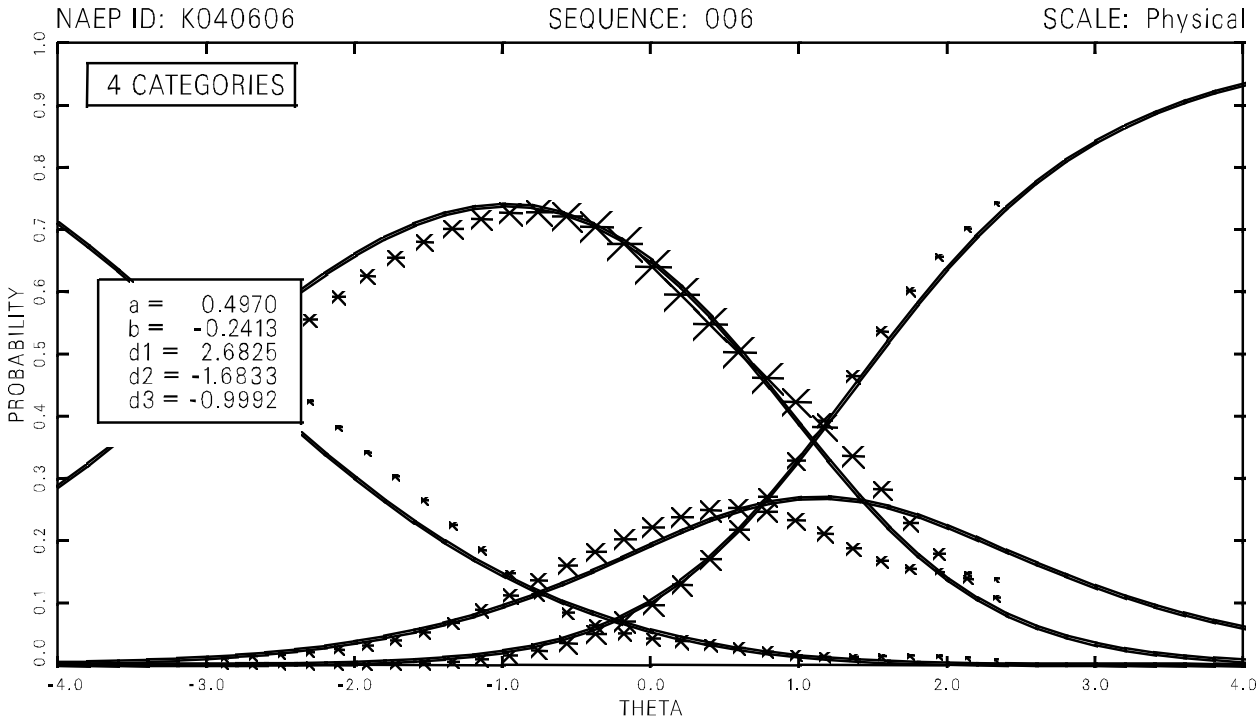
Plot Comparing Empirical and Model-Based Estimates of Item Response Functions for Binary-Scored (Multiple-Choice) Item Exhibiting Some Model Misfit*



*Asterisks indicate estimated conditional probabilities obtained without assuming a logistic form; the solid curve indicates estimated item response function assuming a logistic form.

Figure 13-4

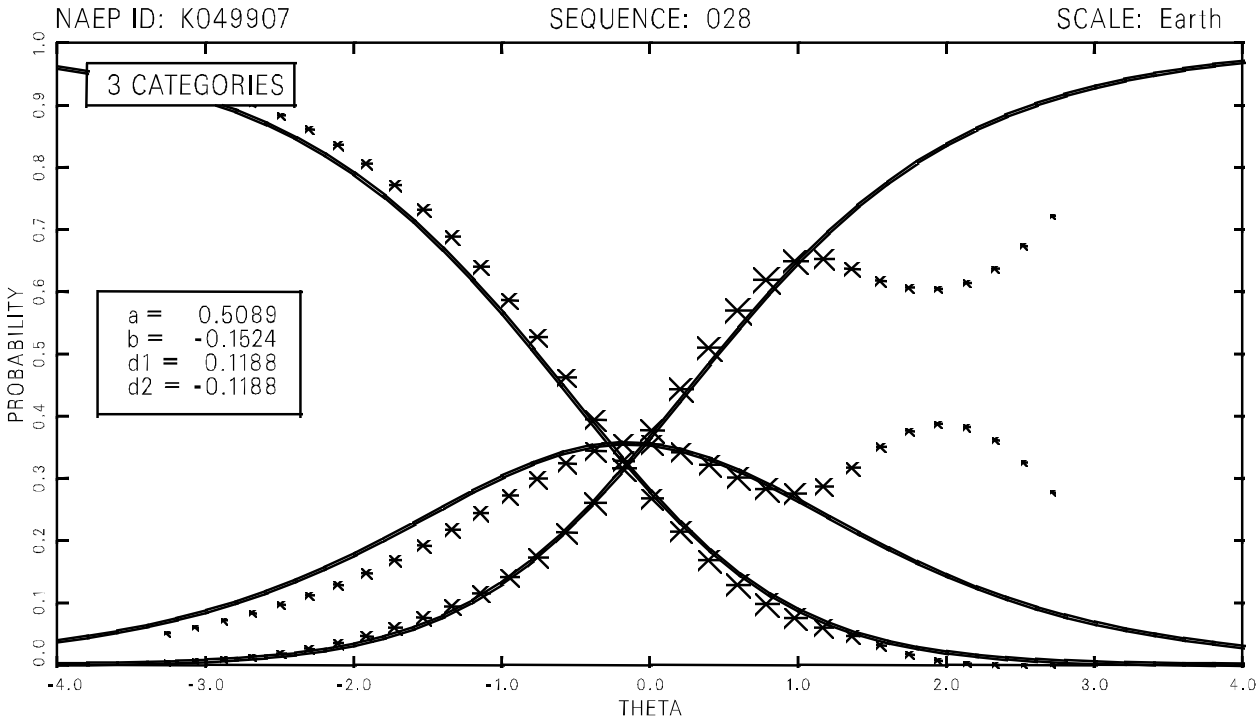
Plot Comparing Empirical and Model-Based Estimates of Item Category Characteristic Curves for a Polytomously Scored Item Exhibiting Some Model Misfit*



*Asterisks indicate estimated conditional probabilities obtained without assuming a model-based form; the solid curve indicates estimated item response function assuming a model-based form.

Figure 13-5

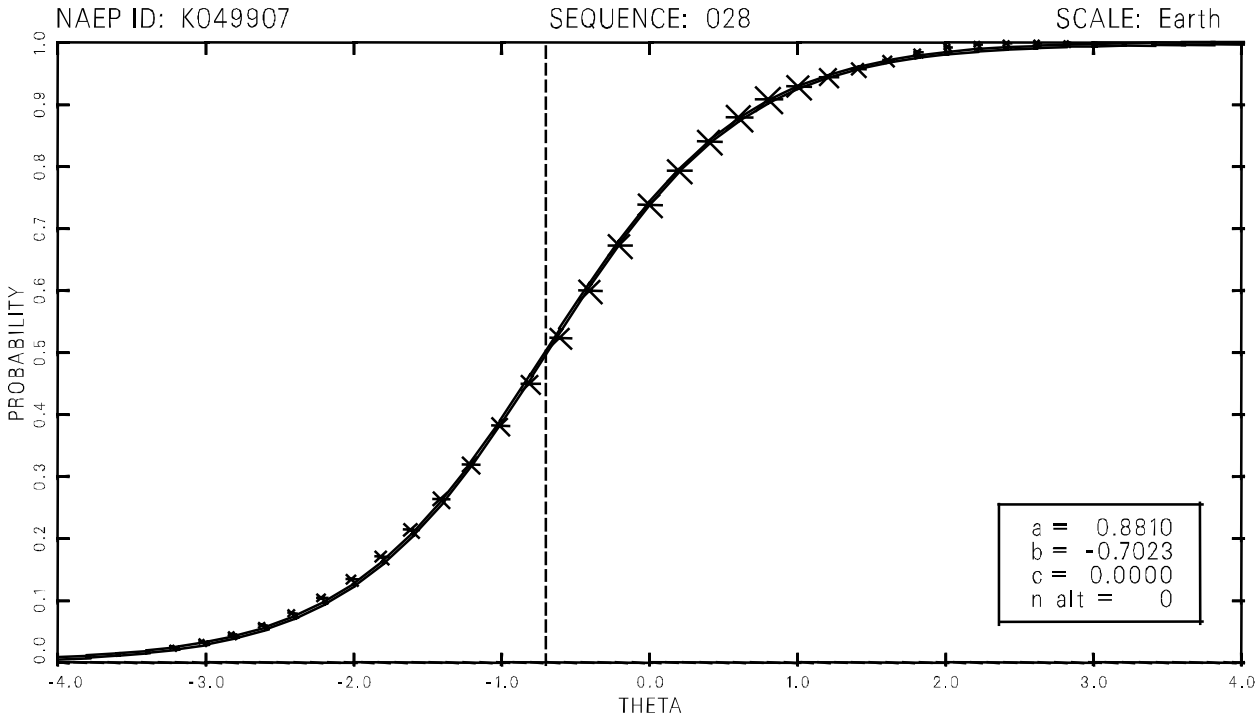
Plot Comparing Empirical and Model-Based Estimates of Item Category Characteristic Curves for a Polytomously Scored Item (K044101) Exhibiting Poor Model Fit*



*Asterisks indicate estimated conditional probabilities obtained without assuming a model-based form; the solid curve indicates estimated item response function assuming a model-based form.

Figure 13-6

Plot Comparing Empirical and Model-Based Estimates of Item Category Characteristic Curves for Polytomously Scored Item (K044101) After Collapsing Categories 2 and 3*



*Asterisks indicate estimated conditional probabilities obtained without assuming a logistic form; the solid curve indicates estimated item response function assuming a logistic form.

Note: When the number of alternatives of a constructed-response item equaled zero, the item was scored in only two categories.

On the following pages, Table 13-16 lists the items that received special treatment during the scaling process. Included in the table are the block locations and item numbers for the items that were combined into cluster items as well as for those that were excluded from the final scales. At grade 8, all items received identical special treatment in the development of the 1996 State Assessment scales. No other items in either assessment received special treatment. The IRT parameters for the items included in the science assessment are listed in Appendix D.

Table 13-16

Items from the 1996 Science Assessment Receiving Special Attention

Grade	NAEP ID	Block/Item Number	Field of Science	Disposition	Reason
4	K031105	S4 - 5	Physical science	Collapsed categories: 0,1,2 become 0,0,1	Lack of Fit
4	K031107	S4 - 7	Physical science	Collapsed categories: 0,1,2,3 becomes 0,0,1,2	Lack of Fit
4	K031203	S5 - 3	Physical science	Dropped	Dependency
4, 8	K031301	S6 - 6	Physical science	Dropped	Dependency
4	K034802	S13 - 2	Physical science	Collapsed categories 0,1,2 become 0,0,1	Lack of Fit
4	K031402	S7 - 2	Earth science	Collapsed categories : 0,1,2 become 0,0,1	Lack of Fit
4	K031404	S7 - 4	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
4	K031407	S7 - 7	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
4	K034501	S12 - 10	Earth science	Collapsed categories: 0,1,2,3 become 0,1,2,2	Zero Frequency
4	K040501	S21 - 11	Earth science	Collapsed categories: 0,1,2 become 0,0,1	Lack of Fit
4	K031001	S3 - 1	Life science	Dropped to form cluster item	Dependency
4	K031002	S3 - 2	Life science	Dropped to form cluster item	Dependency
4	K031605	S9 - 5	Life science	Dropped	Lack of Fit
4	K031607	S9 - 9	Life science	Collapsed categories 0,1,2,3 become 0,1,2,2	Lack of Fit
4	K032501	S10 - 9A	Life science	Dropped to form cluster item	Dependency
4	K032502	S10 - 9G	Life science	Dropped to form cluster item	Dependency
4	K033501	S11 - 9	Life science	Collapsed categories 0,1,2 become 0,0,1	Lack of Fit

(continued)

Table 13-16 (continued)
Items from the 1996 Science Assessment Receiving Special Attention

Grade	NAEP ID	Block/Item Number	Field of Science	Disposition	Reason
4	KZ34101	S12 - 6	Earth Science	Dropped	— [1]
4	KZ34401	S12 - 9	Earth Science	Dropped	— [1]
4	KZ34501	S12 - 10	Earth Science	Dropped	— [1]
4	KZ34502	S12 - 11	Earth Science	Dropped	— [1]
8	K040601	S3 - 1	Physical science	Collapsed categories 0,1,2 become 0,0,1	Lack of Fit
8	K040702	S4 - 3	Physical science	Collapsed categories: 0,1,2,3 become 0,1,1,2	Lack of Fit
8	K040705	S4 - 4	Physical science	Collapsed categories: 0,1,2 become 0,1,1	Zero Frequency
8	K031306	S6 - 9	Physical science	Collapsed categories: 0,1,2 become 0,1,1	Zero Frequency
8	K043603	S11 - 16	Physical science	Collapsed categories: 0,1,2 become 0,1,1	Zero Frequency
8	K040711	S4 - 12	Earth science	Dropped to form cluster item	Dependency
8	K040712	S4 - 13	Earth science	Dropped to form cluster item	Dependency
8	K040803	S5 - 2	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
8	K040901	S7 - 1	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
8	K040905	S7 - 5	Earth science	Collapsed categories: 0,1,2 become 0,0,1	Lack of Fit
8, 12	K049403	S13 - 15	Earth science	Dropped to form cluster item	Dependency
8, 12	K049404	S13 - 16	Earth science	Dropped to form cluster item	Dependency

[1] This item was deleted due to an error discovered after scaling was completed.

(continued)

Table 13-16 (continued)
Items from the 1996 Science Assessment Receiving Special Attention

Grade	NAEP ID	Block/Item Number	Field of Science	Disposition	Reason
8	K044101	S20 - 5	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
8	K044401	S20 - 8	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
8	K041306	S8 - 6	Life science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
8	K031603	S9 - 3	Life science	Collapsed categories: 0,1,2 become 0,0,1	Lack of Fit
8	K031611	S9 - 11	Life science	Collapsed categories: 0,1,2 become 0,1,1	Zero Frequency
8	K042602	S10 - 15	Life science	Collapsed categories: 0,1,2,3 become 0,1,1,2	Lack of Fit
8	K049301	S13 - 12	Life science	Collapsed categories: 0,1,2 become 0,1,1	Zero Frequency
8	K037001	S15 - 1	Life science	Dropped	Lack of Fit
12	K041306	S8 - 6	Life science	Dropped	Lack of Fit
12	K049502	S3 - 4	Physical science	Collapsed categories: 0,1,2,3,4 become 0,0,1,2,3	Lack of Fit
12	K049503	S3 - 5	Physical science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
12	K049602	S4 - 2	Physical science	Collapsed categories: 0,1,2,3,4 become 0,0,1,2,2	Lack of Fit
12	K049603	S4 - 3	Physical science	Collapsed categories: 0,1,2,3,4 become 0,1,2,3,3	Lack of Fit
12	K049702	S6 - 2	Physical science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
12	K049703	S6 - 4	Physical science	Collapsed categories: 0,1,2 become 0,0,1	Lack of Fit
12	K049706	S6 - 7	Physical science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
12	K058201	S20 - 1	Physical science	Dropped	DIF

(continued)

Table 13-16 (continued)
Items from the 1996 Science Assessment Receiving Special Attention

Grade	NAEP ID	Block/Item Number	Field of Science	Disposition	Reason
12	K040801	S5 - 0	Earth science	Collapsed categories: 0,1,2 become 0,0,1	Lack of Fit
12	K049804	S7 - 4	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
12	K049810	S7 - 10	Earth science	Collapsed categories: 0,1,2,3 become 0,1,2,2	Lack of Fit
12	K089811	S7 - 11	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
12	K049812	S7 - 12	Earth science	Collapsed categories: 0,1,2 become 0,0,1	Lack of Fit
12	K049903	S9 - 3	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
12	K049904	S9 - 4	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
12	K049907	S9 - 7	Earth science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
12	K057201	S20 - 2	Earth science	Dropped	Lack of Fit
12	K049506	S3 - 6	Life science	Collapsed categories: 0,1,2,3 become 0,1,2,2	Zero Frequency
12	K041401	S8 - 8	Life science	Collapsed categories: 0,1,2 become 0,0,1	Lack of Fit
12	K041404	S8 - 11	Life science	Collapsed categories: 0,1,2 become 0,1,1	Lack of Fit
12	K053601	S14 - 13	Life science	Collapsed categories: 0,1,2,3,4 become 0,1,2,3,3	Lack of Fit
12	K054006	S15 - 6	Life science	Collapsed categories: 0,1,2 become 0,0,1	Lack of Fit

13.6 GENERATION OF PLAUSIBLE VALUES

13.6.1 Principal Components (NSWEEP Program)

Multivariate plausible values were generated for the entire age/grade sample using the multivariate conditioning program CGROUP as revised by Thomas (1994). This procedure employed student weights. Prior to 1990, 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. 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.

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, Johnson, Bowker, & Fong, 1994). Table 13-17 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 13-17
Proportion of Proficiency Variance Accounted for by the Conditioning Model for the Science Main Assessment

Grade	Number of Conditioning Contrasts[1]	Number of Principal Components[1]	Proportion of Proficiency Variance		
			Physical Science	Earth Science	Life Science
4	948	317	.64	.57	.59
8	1,041	326	.63	.63	.64
12	808	290	.71	.71	.70

[1] Excluding the constant term

13.6.2 Conditioning (CGROUP Program)

The codings of the original science-specific conditioning variables, before principal components were calculated, are presented in Appendix C. NAEP-CGROUP (described in Chapter 11) creates posterior distributions of proficiencies by combining information from item responses of individuals and information from linear regression of proficiency on conditioning variables. For each individual, five plausible values are randomly drawn from their posterior proficiency distribution.

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 B.

13.6.3 Analysis of Dimensionality

As mentioned earlier, the main assessment is multivariate with three content area scales. Tables 13-18 and 13-19 give conditional and marginal correlations for the three scales for the three grades. The conditional correlations can be thought of as correlations from information pooled within the demographic subgroups corresponding to grouping variables used to condition the data with CGROUP. The conditional correlations correspond to the error correlations of a CGROUP analysis. The conditional correlations are high, averaging .79 for grade 4, .90 for grade 8, and .83 for grade 12. The marginal correlations are the average of the scale correlations over five plausible values generated by CGROUP. Since these correlations are not pooled within

background groupings, marginal correlations tend to be larger than conditional correlations, averaging .84 for grade 4, .91 for grade 8, and .90 for grade 12. Although it is of substantive interest to analyze the scales separately, the correlations indicate that they are highly redundant with each other.

Table 13-18
Conditional Correlations from Conditioning (CGROUP)

Grade	Field of Science Scale	Physical Science	Earth Science	Life Science
4	Physical Science	1.00	—	—
	Earth Science	0.79	1.00	—
	Life Science	0.79	0.78	1.00
8	Physical Science	1.00	—	—
	Earth Science	0.92	1.00	—
	Life Science	0.89	0.88	1.00
12	Physical Science	1.00	—	—
	Earth Science	0.87	1.00	—
	Life Science	0.82	0.80	1.00

Table 13-19
Marginal Correlations of Science Scales[1]

Grade	Field of Science Scale	Physical Science	Earth Science	Life Science
4	Physical Science	1.00	—	—
	Earth Science	0.84	1.00	—
	Life Science	0.84	0.84	1.00
8	Physical Science	1.00	—	—
	Earth Science	0.93	1.00	—
	Life Science	0.91	0.90	1.00
12	Physical Science	1.00	—	—
	Earth Science	0.92	1.00	—
	Life Science	0.90	0.89	1.00

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

13.7 THE FINAL PROFICIENCY SCALES

13.7.1 Field of Science Scales

Like all IRT scales, the field of science scales have a linear indeterminacy that may be resolved by an arbitrary choice of origin and unit-size for each scale. The 1996 science assessment was developed using a new framework. Because it was not appropriate to compare results from the 1996 assessment with those of previous NAEP science 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. NAEP assessments developed earlier (such as the 1992 reading assessment) were developed with a cross-grade framework, in which the trait being measured is conceptualized as cumulative across the grades of the assessment. Accordingly, a single 0-to-500 scale was established for all three grades in each of these assessments.

In 1993, the National Assessment Governing Board (NAGB) determined that future NAEP assessments should be developed using within-grade frameworks. This removes the constraint that the trait being measured is cumulative. 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. The NAEP 1994 U.S. history and geography assessments were developed and scaled within-grade. After scaling, the scales were aligned so that grade 8 had a higher mean than grade 4, and grade 12 had a higher mean than grade 8. The results were reported on a final 0-to-500 scale that looked similar to those used in reading, in spite of the differences in development and scaling. This choice of the reporting scale was the source of potential confusion and misinterpretation.

Therefore, for the NAEP 1996 science assessment—which was also developed and scaled using within-grade procedures—a new reporting metric was adopted. The results are reported on 0-to-300 scales and the means for each of the grades are identical. For each grade, the mean for each field of science was set at 150 and the standard deviation was 35. Constraining the mean and standard deviation of the scales to 150 and 35 also constrained, to some degree, the percentiles for the total group of students at each grade. However, within-grade comparisons of percentiles across subgroups can still provide valuable comparative information. The reporting metric was developed using data from the national assessment program, and the results for the state assessment program were linked to these scales.

For each field of science, the scale mean and standard deviation were set to 150.0 and 35.0 using the transformation:

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

where θ_{target} denotes values on the final transformed scale and $\theta_{calibrated}$ denotes values on the original calibration scale from BILOG/PARSCALE. The calculation of the value of "A" and "B" is described in Chapter 9, Section 9.3.5. The constants for the linear transformation for each scale are given in Table 13-20.

Table 13-20
Coefficients of Linear Transformations of the Fields of Science Scales from the Calibrating Scale Units to the Units of Reporting Proficiency

Grade	Field of Science Scale	A	B
4	Physical Science	34.91	151.17
	Earth Science	34.09	150.67
	Life Science	35.09	150.51
8	Physical Science	35.85	150.23
	Earth Science	34.56	150.58
	Life Science	35.64	150.25
12	Physical Science	37.76	149.65
	Earth Science	36.59	149.77
	Life Science	35.91	150.19

13.7.2 The Composite Science Proficiency Scale

In addition to the plausible values for each scale, a composite of the three fields of science scales was created as a measure of overall science performance. The composite was a weighted average of plausible values on the fields of science scales (earth, physical, and life). The weights for the scales were proportional to the importance assigned to each field of science contained in the assessment specifications given in the *Science Framework*. The weights are given in Table 13-21. As indicated in Table 2-4 of Chapter 2, the weights for each of the fields of science are similar to the actual proportion of assessment time devoted to that field. In

developing the composite scale, the weights were applied to the plausible values for each fields of science as expressed in terms of the final scale (i.e., after transformation from the provisional BILOG/PARSCALE scales).

Table 13-21
Weights Used for Each Field of Science Scale to Form the Science Composite

Field of Science Scale	Grade 4	Grade 8	Grade 12
Physical science	.33	.30	.33
Earth science	.33	.30	.33
Life science	.33	.40	.33

Finally, it is necessary to caution that, although the science composite is expressed in units that seem somewhat similar to the long-term trend science scale, it is not appropriate to compare scores. The transformation chosen to resolve the linear indeterminacy in the science composite is a convenient transformation, but it is only one of a conceptually infinite number of such transformations that could have been chosen. Any one of these transformations would have provided equivalent information about the relative standings of subgroups in the population. *Because there is no link, real or implied, in the construction of the science composite and the field of science scales to either the mathematics assessment or to the previous science assessments, the comparison of students' science proficiencies to students' proficiencies in other subject areas is devoid of meaning.*

13.8 PARTITIONING OF THE ESTIMATION ERROR VARIANCE

For each grade, the error variance of the final, transformed scale mean was partitioned as described in Chapter 11. This analysis yielded estimates of the proportion of error variance due to sampling students and the proportion due to the latent nature of θ. These estimates are given in Table 13-22 for each field of science scale and the composite scale (for stability, the estimates of the between-imputation variance, *B*, in Equation 11.9). Additional results, including those by gender and race/ethnicity, are presented in Appendix E.

Table 13-22
Estimation Error Variance and Related Coefficients for the Science Main Assessment

Grade	Field of Science Scale	Total Estimation Error Variance	Proportion of Variance Due to . . . Student Sampling	Latency of θ
4	Physical Science	1.16	0.82	0.18
	Earth Science	0.72	0.85	0.15
	Life Science	0.86	0.78	0.22
	Composite	0.64	0.89	0.11
8	Physical Science	0.91	0.92	0.08
	Earth Science	0.89	0.91	0.09
	Life Science	1.07	0.86	0.14
	Composite	0.78	0.94	0.06
12	Physical Science	1.03	0.92	0.08
	Earth Science	0.91	0.94	0.06
	Life Science	0.80	0.89	0.11
	Composite	0.76	0.96	0.04

13.9 SCIENCE TEACHER QUESTIONNAIRES

Teachers of fourth- and eighth-grade students were surveyed about their educational background and teaching practices. The students in a particular classroom had their records matched with their 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 non-matching students. Of the 7,305 fourth-grade students in the sample, 89.0 percent were matched with both parts of the teacher questionnaire and 2.2 percent were matched with only the first, teacher background, part of the questionnaire. For the eighth-grade students sample, 82.4 percent were matched with both parts of the teacher questionnaire and 1.4 percent were matched with only the first part 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 teacher, resulting in higher match rates. Thus, 91.4 percent of the fourth graders and 83.8 percent of the eighth graders were matched with at least the background information about their science teachers.

13.10 THE WEIGHT FILES

Westat produced the final student and school weights and the corresponding replicate weights for the 1996 science assessment. Information for the creation of the weight files was supplied by NCS under the direction of ETS.

As was described in the *Technical Report of the 1996 State Assessment Program in Science* (Allen, Swinton, Isham, & Zelenak, 1998), the State Assessment sample was split into two subsamples, one using the 1992 inclusion rules (S1) and one using the 1996 inclusion rules (S2) the weighting process was more complex than in previous assessments (see Allen, Swinton, Isham, & Zelenak, 1998 for more details).

In the national science assessment, only the 1996 inclusion rules (S2) were used. Also, there were no accommodations offered for students with disabilities or for students with limited English proficiency in the national science assessment. Thus, a single sample was used for both analysis and reporting, and only a single set of student sampling weights. The student sampling weights have replicate weights associated with them. Replicate weights are used to estimate jackknife standard errors for each statistic estimated for the national science assessment.

Chapter 14

DATA ANALYSIS FOR THE LONG-TERM TREND READING ASSESSMENT[1]

Jo-Lin Liang and Lois H. Worthington
Educational Testing Service

14.1 INTRODUCTION

This chapter describes the analyses performed on the responses to the cognitive and background items in the 1996 long-term trend reading assessment. These analyses led to the results presented in the *NAEP 1996 Trends in Academic Progress: Achievement of U.S. Students in Science, 1969 to 1996; Mathematics, 1973 to 1996; Reading, 1971 to 1996; and Writing, 1984 to 1996* (Campbell, Voelkl, & Donahue, 1997). The emphasis of this chapter is on the methods and results of procedures used to develop the IRT-based scale scores that formed the basis of this report. The theoretical underpinnings of the IRT and plausible values methodology described in this chapter are given in Chapter 11, and several of the statistics are described in Chapter 9.

The objectives of the reading long-term trend analysis were to prepare scale values and perform all analyses necessary to produce a long-term trend report in reading. The reading long-term trend results include the years 1971, 1975, 1980, 1984, 1988, 1990, 1992, 1994, and 1996. The major analysis components are discussed in turn. Some aspects of the analysis, such as procedures for item analysis, scoring of constructed-response items, and methods of scaling, are described in previous chapters and are therefore not detailed here.

The student samples that were administered reading items in the 1996 long-term trend reading assessment are shown in Table 14-1. See Chapters 1 and 3 for descriptions of the target populations and the sample design used for the assessment.

The long-term trend results reported in *Trends in Academic Progress* are based on print administrations and occur at all of the age levels. The samples involved in the analysis are shown in Table 14-1 as 9[RW-LTTrend], 13[RW-LTTrend], and 17[RW-LTTrend]. The long-term trend booklets for these samples contained blocks of reading and writing items administered in print form. All students received a block of common background questions, distinct for each age, and subject-area background questions that were presented in the cognitive blocks. The booklets are identical to those used for long-term trend assessments in 1984, 1988, 1990, 1992, and 1994. The booklets and the blocks within those booklets are listed in Chapter 4. Additional information about all of the items in these blocks also appears in that chapter. This chapter includes specific information about the long-term trend items that were scaled. Both age- and grade-selected students contributed to the long-term trend scaling. However, only students in the "age-only" portion of the reading long-term trend samples contributed to the results presented in *Trends in Academic Progress*.

---

[1] Jo-Lin Liang was the primary person responsible for the planning, specification, and coordination of the reading long-term trend analyses, advised by Eiji Muraki, and Nancy L. Allen. Data analyses and scaling were performed by Lois H. Worthington, advised by David S. Freund. Others contributing to the analysis of data were Bruce A. Kaplan, and Norma A. Norris. John R. Donoghue provided consultation.

Table 14-1

NAEP 1996 Long-Term Trend Reading Student Samples

Sample	Booklet IDs	Mode	Cohort Assessed	Time of Testing	Age Definition	Modal Grade	Number Assessed
9 [RW-LTTrend]	51-56	Print	Age 9/Grade 4	1/3/96 - 3/8/96 (Winter)	CY	4	5,019
13 [RW-LTTrend]	51-56	Print	Age 13/Grade 8	10/9/95 - 12/22/95 (Fall)	CY	8	5,493
17 [RW-LTTrend]	51-56	Print	Age 17/Grade 11	3/11/96 - 5/10/96 (Spring)	Not CY	11	4,669

LEGEND

RW	Reading and writing
LTTrend	Long-term trend assessment
Print	Print administration
CY	Calendar year: birthdates in 1986, and 1982 for ages 9, and 13
Not CY	Age 17 only: birthdates between October 1, 1978, and September 30, 1979

Table 14-2 clarifies the relationships between the 1996 long-term trend samples and samples from previous years. For ages 9, 13, and 17, the [RW-LTTrend] samples allow direct comparisons with 1994, 1992, 1990, 1988, and 1984 samples. The long-term trend scale, established in 1984, was linked to the 1971, 1975, and 1980 assessments using a complex equating strategy described in *Implementing the New Design: The NAEP 1983-84 Technical Report* (Beaton, 1987). At each age, several intact booklets were retained from the 1984 assessment, forming the basis of the reading long-term trend assessment in 1988, 1990, 1992, 1994, and 1996. Information about the 1988 assessment is available in *Focusing the New Design: The NAEP 1988 Technical Report* (Johnson & Zwick, 1990); information about the 1990 assessment is given in *The NAEP 1990 Technical Report* (Johnson & Allen, 1992); information about the 1992 assessment is given in *The NAEP 1992 Technical Report* (Johnson & Carlson, 1994); and information about the 1994 assessment is given in *The NAEP 1994 Technical Report* (Allen, Kline, & Zelenak, 1996).

The 1996 long-term trend included, at each age level, six of the assessment booklets administered in 1984. These booklets (51-56) contained both reading and writing blocks, as well as background items. Although these long-term trend booklets represented only about one-tenth of the reading booklets administered in the complex 1984 BIB design,[2] they contained 10 of the 12 reading blocks that were scaled at each age/grade level in 1984. The samples of students who received these long-term trend booklets are described in Table 14-1 and in Chapter 3. The purpose of the reading long-term trend analysis was to add to the reading trend results that extended from 1971 to 1994 for ages 9, 13, and 17. The numbers of scaled items for each age are presented in Table 14-3. Each age was scaled separately. The numbers of items scaled in 1996 that were common across assessment years are given in Table 14-4. As was the case for previous long-term trend analyses, the long-term trend scale is univariate. Dimensionality analyses conducted following the 1984 assessment showed that the reading items were well summarized by a unidimensional scale (Zwick, 1987).

[2] The long-term trend assessment included 1984 Booklets 16, 17, 27, 34, 55, and 60 at age 9 and Booklets 13, 16, 17, 21, 34, and 57 at ages 13 and 17 (see J. R. Johnson, 1987, pp. 120-121). The 1984 main assessment focused-BIB design included 57 booklets that contained at least one scaled reading block at age 9 and 56 such booklets at ages 13 and 17.

Table 14-2

NAEP Reading Samples Contributing to 1996 Long-Term Trend Results, 1971-1996

Cohort	Year	Sample	Subjects	Time of Testing	Mode of Administration	Age Definition	Modal Grade
Age 9	1971	Main	RL	Winter	Tape	CY	4
	1975	Main	RA	Winter	Tape	CY	4
	1980	Main	RA	Winter	Tape	CY	4
	1984	Main	RW	Winter, Spring	Print	CY	4
	1984	T-84	RW	Winter	Tape	CY	4
	1988	LTTrend[1]	RW	Winter	Print	CY	4
	1990	LTTrend[1]	RW	Winter	Print	CY	4
	1992	LTTrend[1]	RW	Winter	Print	CY	4
	1994	LTTrend[1]	RW	Winter	Print	CY	4
	1996	LTTrend[1]	RW	Winter	Print	CY	4
Age 13	1971	Main	RL	Fall	Tape	CY	8
	1975	Main	RA	Fall	Tape	CY	8
	1980	Main	RA	Fall	Tape	CY	8
	1984	Main	RW	Winter, Spring	Print	CY	8
	1984	T-84	RW	Fall	Tape	CY	8
	1988	LTTrend[1]	RW	Fall	Print	CY	8
	1990	LTTrend[1]	RW	Fall	Print	CY	8
	1992	LTTrend[1]	RW	Fall	Print	CY	8
	1994	LTTrend[1]	RW	Fall	Print	CY	8
	1996	LTTrend[1]	RW	Fall	Print	CY	8
Age 17	1971	Main	RL	Spring	Tape	Not CY	11
	1975	Main	RABS	Spring	Tape	Not CY	11
	1980	Main	RA	Spring	Tape	Not CY	11
	1984	Main	RW	Winter, Spring	Print	Not CY	11
	1984	T-84	RW	Spring	Tape	Not CY	11
	1988	LTTrend[1]	RW	Spring	Print	Not CY	11
	1990	LTTrend[1]	RW	Spring	Print	Not CY	11
	1992	LTTrend[1]	RW	Spring	Print	Not CY	11
	1994	LTTrend[1]	RW	Spring	Print	Not CY	11
	1996	LTTrend[1]	RW	Spring	Print	Not CY	11

[1]**Note:** Within a cohort, these samples received common booklets.

LEGEND

RL	Reading and literature	LTTrend	Long-term trend (these samples received common booklets within an age group)
RA	Reading and art		
RABS	Reading, art, index of basic skills	Print	Print administration
RW	Reading and writing	Tape	Audiotape administration
Main	Main assessment	CY	Calendar year: birthdates (1996 sample) in 1986 and 1982 for ages 9 and 13
T-84	Special sample in the 1984 assessment that was used to establish links to previous assessments (1971-1980) for the purposes of long-term trend	Not CY	Age 17 only (1996 sample): birthdates between October 1 and September 30 of the appropriate years

Table 14-3
Numbers of Scaled Reading Long-Term Trend Items Common Across Ages

Age	Number of Items
9 only	61[1]
13 only	22[1]
17 only	23
9 and 13 only	13
9 and 17 only	2
13 and 17 only	42
9, 13, and 17	26[1]
Total	189[1]

[1] These figures have been updated since their publication in the 1992 and 1994 NAEP technical reports (Table 12-3 and Table 15-3, respectively).

Table 14-4
Numbers of Scaled Reading Long-Term Trend Items Common Across Assessments

	Number of Items		
Assessment Year	Age 9	Age 13	Age 17
1984, 1992, 1994, 1996	102	103	93[1]
1984, 1990, 1992, 1994, 1996	101	101	92[1]
1984, 1988, 1990, 1992, 1994, 1996	98	98	87
1980, 1984, 1988, 1990, 1992, 1994, 1996	67	71	52
1971, 1975, 1980, 1984, 1988, 1990, 1992, 1994, 1996	36	45	37

[1] These figures have been updated since their publication in the NAEP 1992, and 1994 Technical Reports (Table 12-4, and Table 15-4, respectively).

The steps in the reading long-term trend analysis are documented in the following sections. As is usual in NAEP analyses, the first step was to gather item and block information. The trend items were then calibrated according to the IRT model. Plausible values were generated after conditioning on available background variables. Finally, the scale values were placed on the final reading long-term trend scale used in previous trend assessments.

14.2 ITEM ANALYSIS FOR THE READING LONG-TERM TREND ASSESSMENT

Conventional item analyses did not identify any difficulties with the long-term trend data. The results displayed in Table 14-5 contain the number of items, size of the unweighted sample administered the block, average weighted proportion correct, average weighted r-biserial, and average weighted alpha as a measure of reliability for each block. Because the blocks were presented in self-paced, print-administered form, the weighted proportion of students attempting the last item is included in the table to give an indication of the speededness of each block. Common labeling of these blocks across ages does not denote common items. Booklet information is detailed in Chapter 4. Student weights were used for

all statistics except for the sample sizes. The average values reflect only the items in the block that were scaled. Overall, the 1996 item-level statistics were not very different from those for the 1994 assessment.

Table 14-5
Descriptive Statistics for Item Blocks in the Reading Long-Term Trend Samples

	Blocks										
Statistics	**B8**	**B10**	**B11**	**B12**	**B13**	**B14**	**B15**	**B16**	**B17**	**B18**	**B22**
Age 9											
Number of scaled items	10	8	11	7	11	12	11	—[1]	11	12	9
Number of scaled constructed-response items	1	0	0	1	1	1	0	—[1]	0	0	3
Unweighted sample size	630	610	610	603	624	624	595	—[1]	603	1222	592
Average weighted proportion correct	.62	.43	.43	.50	.41	.60	.51	—[1]	.56	.47	.60
Average weighted r-biserial	.75	.62	.62	.82	.61	.75	.60	—[1]	.74	.64	.73
Weighted alpha reliability	.75	.67	.71	.76	.71	.84	.64	—[1]	.81	.75	.74
Weighted proportion of students attempting last item	.92	.82	.77	.72	.66	.66	.86	—[1]	.86	.82	.97
Age 13											
Number of scaled items	12	9	8	5	11	12	10	9	16	11	—[2]
Number of scaled constructed-response items	1	0	0	0	1	1	1	1	0	0	—[2]
Unweighted sample size	629	642	630	673	630	615	658	642	673	629	—[2]
Average weighted proportion correct	.64	.59	.63	.64	.58	.66	.66	.73	.59	.67	—[2]
Average weighted r-biserial	.71	.59	.73	.80	.65	.68	.64	.77	.56	.74	—[2]
Weighted alpha reliability	.71	.60	.68	.57	.66	.78	.55	.67	.70	.77	—[2]
Weighted proportion of students attempting last item	.94	.83	.98	.95	.91	.81	.84	.92	.78	.97	—[2]
Age 17											
Number of scaled items	12	4	8	6	11	12	13	10	10	7	—[2]
Number of scaled constructed-response items	1	1	0	1	1	1	1	1	0	0	—[2]
Unweighted sample size	605	638	585	643	585	588	622	638	643	605	—[2]
Average weighted proportion correct	.72	.72	.79	.72	.68	.84	.65	.74	.53	.68	—[2]
Average weighted r-biserial	.75	.78	.83	.81	.76	.78	.62	.71	.63	.79	—[2]
Weighted alpha reliability	.72	.43	.67	.41	.68	.80	.73	.74	.70	.70	—[2]
Weighted proportion of students attempting last item	.95	.97	1.00	.96	.96	.89	.69	.82	.72	.98	—[2]

[1] Block B16 was not administered at age class 9.

[2] Block B22 was not administered at age class 13 or 17.

14.3 TREATMENT OF CONSTRUCTED-RESPONSE ITEMS

Data for constructed-response items in the long-term trend analysis were used for the 1984, 1990, 1992, 1994, and 1996 assessments only. Constructed-response items were not included in the original scoring of the 1988 reading assessment because a previous study (Zwick, 1988) had shown that scoring inconsistencies (drops in interrater reliability and/or scorer drift—that is, scorers showed evidence of rating items more strictly or more leniently than did the original 1984 scorers) had affected these items.

Rater reliability within year was computed for the 1996 constructed-response items at each age. Between year reliability was also studied with the 1994, and the 1984 responses. In general, the 1996 scoring did not show irregularities. All of the 1996 constructed-response items were used in the trend analysis except the items that were excluded from calibration in the previous assessment. The deleted items are listed in Table 14-6. The remaining constructed-response items were dichotomized according to criteria developed by subject-area experts. The dichotomized versions of the constructed-response items were included in the calibration.

Table 14-6
Items Deleted from the Reading Long-Term Trend Analysis

Age	Block	Item	Reason for Exclusion
9	B10	N001801	Excluded in previous assessment
	B13	N003003	Excluded in previous assessment
	B10	N008905	Excluded in previous assessment (constructed-response item)
13	B10	N001801	Excluded in previous assessment
	B10	N001904	Excluded in previous assessment (constructed-response item)
	B11	N002302	Excluded in previous assessment
	B12	N002804	Excluded in previous assessment (constructed-response item)
	B17	N005001	Excluded in previous assessment
17	B10	N001702	Excluded in previous assessment
	B11	N002302	Excluded in previous assessment
	B17	N015905	Excluded in previous assessment (constructed-response item)

14.4 IRT SCALING FOR THE READING LONG-TERM TREND ASSESSMENT

14.4.1 Item Parameter Estimation

The first step in the scaling process was the estimation of item parameters for the long-term trend items. This item calibration was performed using the BILOG/PARSCALE program described in Chapter 11. Items were calibrated separately for each of the three age/grade groups. Item parameters were estimated using combined data from the assessment years 1994 and 1996, treating each assessment as a sample from a separate subpopulation. Student weights were used for the calibration. To ensure that each assessment year had a similar influence on the calibration, student weights for 1994 examinees were multiplied by a constant, to adjust them to have the same sum as the sum of the weights for the 1996 examinees. Approximately 600-700 examinee responses for each item were present in each assessment year.

Starting values for item parameters were based on the final item parameter values from the analysis of the 1994 long-term trend assessment. As described in Chapter 9, BILOG/PARSCALE calibrations were completed in two stages. At stage one, the proficiency distribution of each assessment year was constrained to be normal, although the means and variances differed across assessment years. The values of the item parameters from this normal solution were then used as starting values for a second-stage estimation run in which the proficiency distribution (modeled as a separate multinomial distribution for each assessment year) was estimated concurrently with item parameters. Calibration was concluded when changes in item parameters became negligibly small (i.e., less than .005).

14.4.2 Evaluation of Model Fit

During and subsequent to item parameter estimation, evaluations of the fit of the IRT models were carried out for each of the items. These evaluations were based primarily on graphical analysis. First, model fit was evaluated by examining plots of nonmodel-based estimates of the expected proportion correct (conditional on proficiency) versus the proportion correct predicted by the estimated item response function (see Chapter 9, and Mislevy & Sheehan, 1987, p. 302). In making decisions about excluding items from the final scales, a balance was sought between being too stringent, hence deleting too many items and possibly damaging the content representativeness of the pool of scaled items, and being too lenient, hence including items with model fit poor enough to endanger the types of model-based inferences made from NAEP results. A certain degree of misfit was tolerated for a number of items included in the final scales.

Most of the items fit the model well. Items excluded from the analysis of the 1996 assessment were the same items that were deleted from the 1994 reading long-term trend analysis. Table 14-6 lists items that were excluded from the analysis of the 1996 long-term trend assessment.

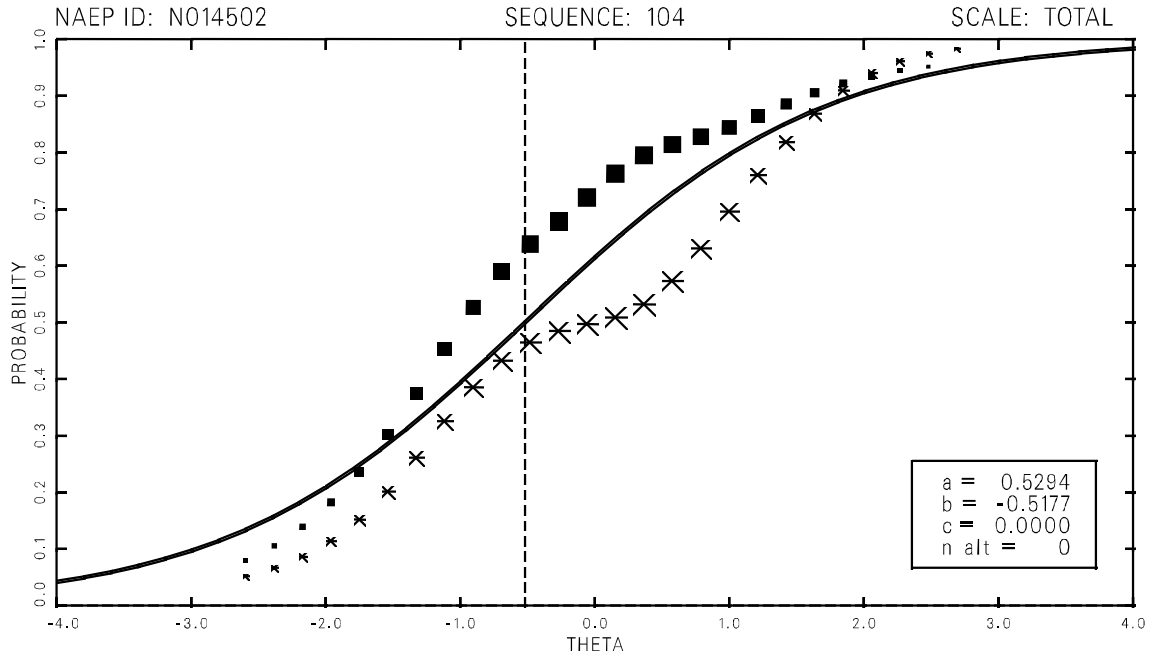
The adequacy of the assumption of a common item response function across assessment years was also evaluated by comparing the nonmodel-based expected proportions for each assessment year to the single, model-based item response function fit by BILOG/PARSCALE. Items that showed clear evidence of functioning differently across assessments were treated as separate items for each assessment year--that is, separate item response functions were estimated for each assessment. As was the case with deleting items, in making decisions about scaling items separately by assessment year, a balance was sought between being too stringent, hence splitting too many items and possibly damaging the common item link between the assessment years, and being too lenient, hence including items with model fit poor enough to endanger the model-based trend inferences. These separately scaled items will be reexamined in future long-term trend assessments.

At age 9, one constructed-response item was calibrated separately for each assessment year. Examination of residual plots identified the item as functioning differently across assessments. Figure 14-1 shows item N014502 from the analysis for grade 4/age 9. Data are presented for 1994 (squares), and for 1996 (asterisks)[3]. For middle proficiency values, the two sets of symbols diverge, and the discrepancy of the item characteristic curves of the two years is substantial. The top (1994 data), and the bottom (1996 data) of Figure 14-2 show the plots for the item treated separately by assessment year. The 1996 data showed poorer fit. In order to maintain the link for the trend, this item was kept in the analysis but with the 1994 data calibrated separately and the 1996 data excluded from the final calibration. The remaining misfit is relatively small. Overall, one long-term trend reading item was calibrated separately by assessment year. Table 14-7 lists the item that was calibrated separately across assessment years.

[3] The size of the symbols are proportional to the estimated number of students at a particular scale score level. The symbols are ordinarily larger in the middle of the theta scale, where most students' scale scores fall.

Figure 14-1

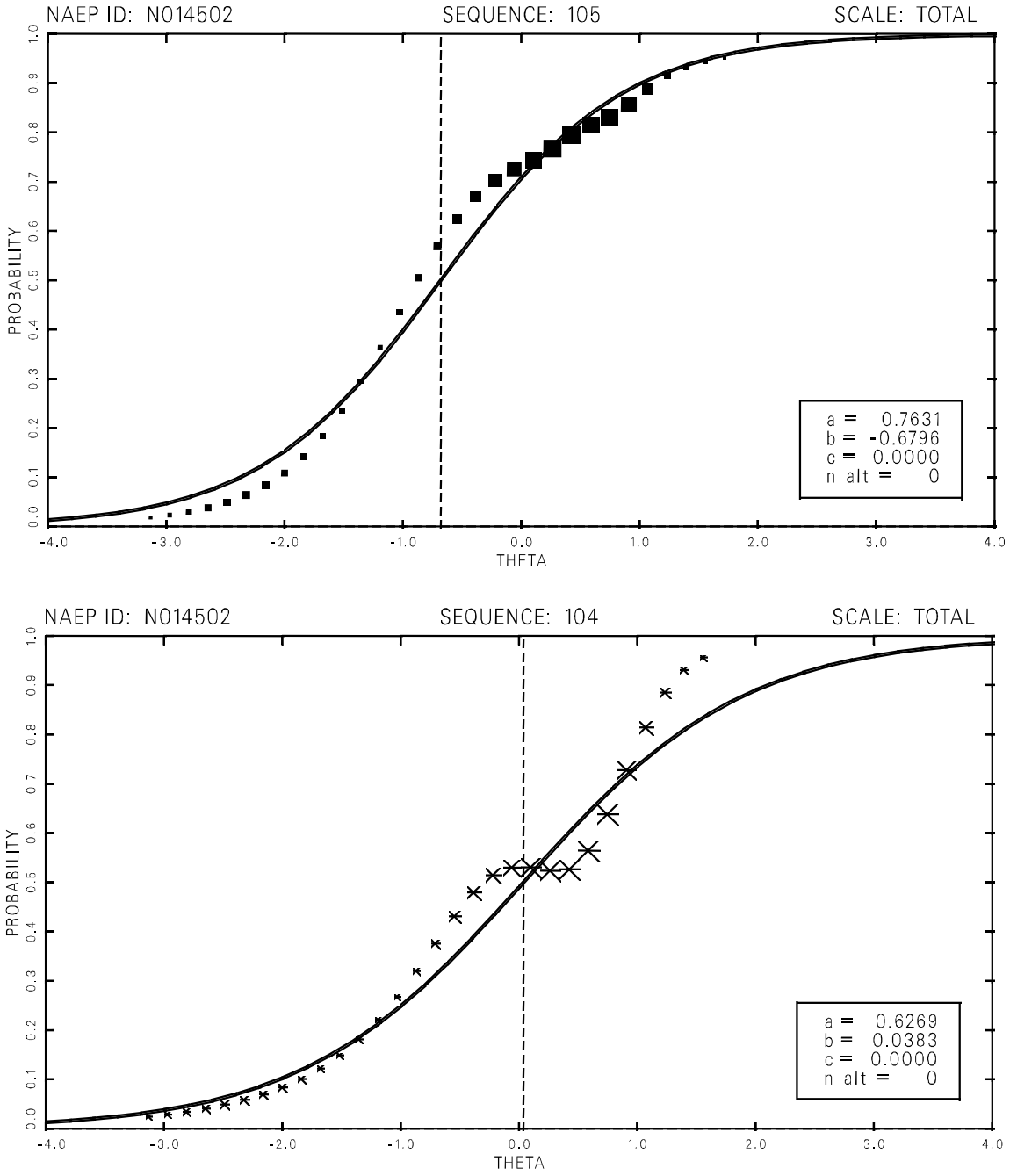
Example Long-Term Trend Item (N014502, Age 9)
Demonstrating Differential Item Functioning Across Assessment Years 1994 and 1996[1]



[1]This plot compares empirical and model-based estimates of the item response function (IRF). The smooth curve represents the model-based estimate at each provisional proficiency level. The squares represent 1994 data; asterisks represent 1996 data.

Note: When the number of alternatives of a constructed-response item equaled zero, the item was scored in only two categories.

Figure 14-2
Example Long-Term Trend Item (N014502, Age 9)
Fitting Separate Item Response Functions for Each Assessment Year[1]



[1]The plot compares empirical and model-based estimates of the item response function (IRF). The smooth curve represents the model-based estimate at each provisional proficiency level. The squares represent 1994 data; asterisks represent 1996 data.

Note: When the number of alternatives of a constructed-response item equaled zero, the item was scored in only two categories.

Table 14-7
Item Calibrated Separately by Assessment Year in the Reading Long-Term Trend Analysis

Age	Block	Item	Reason for Separate Calibration
9	B22	N014502	Poor fit across assessments to common item response function

At age 17, two items (N002201 and N002202) caused difficulty in scaling. In preliminary calibrations, both items did not fit the model well and had large slope-parameter values (3.9 and 5.0, respectively). The item response function of N002202 also demonstrated an elevated tail. Further examination of the items indicated that this might be due to local dependence of the items, although neither item had been problematic at this age group in the 1994 assessment. The approach of fixing the slope-parameter was taken to obtain stable item parameter estimates. At calibration stage-two, after the estimation of the proficiency distribution was constrained to be normally distributed and calibrated to convergence, the slope-parameter of N002201 was fixed at the value, and all items were calibrated to convergence. Parameters estimates from this run served as the final estimates for age 17.

A list of the items scaled for each of the ages, along with their item parameter estimates, appears in Appendix D.

14.5 GENERATION OF PLAUSIBLE VALUES

The generation of plausible values was conducted independently for each age/grade level for each of the assessment years. The item parameters from BILOG/PARSCALE, final student weights, item responses, and selected background variables were used with the computer program BGROUP (described in Chapter 11) to generate the values for each age. The background variables included student demographic characteristics (i.e., race/ethnicity of the student, highest level of education attained by parents), students' perceptions about reading, and student behavior both in and out of school (e.g., amount of television watched daily, amount of homework done each day). Appendix C gives the codings for the conditioning variables for the three age groups. Table 14-8 contains a list of the number of background contrasts included in conditioning, as well as the proportion of variance accounted for by the conditioning model for each age/grade.

Table 14-8
Proportion of Proficiency Variance Accounted for by the Conditioning Model for the Reading Long-Term Trend Assessment

Age/Grade	Number of Conditioning Contrasts[1]	Proportion of Proficiency Variance
9/4	49	.32
13/8	49	.35
17/11	47	.34

[1] Excluding the constant term.

14.6 THE FINAL READING LONG-TERM TREND SCALE

The linear indeterminacy of the long-term trend scale was resolved by linking the 1996 long-term trend scales to previous long-term trend scales. For each age, the item parameters from the joint calibration based on data from both 1994 and 1996 were used with the 1994 data to reestimate plausible values for the 1994 data. The mean and standard deviation of the new 1994 estimates were calculated and matched to the mean and standard deviation of the old 1994 plausible values that were reported previously. The linear constants of this transformation were then applied to transform the 1996 scales to the 1994 proficiency metric. The transformation equations (described in Chapter 9) that resulted from this matching of the first two moments for the 1994 data are

$$\text{Age } 9: \theta_{target} = 52.46 \bullet \theta_{calibrated} + 206.38,$$

$$\text{Age } 13: \theta_{target} = 39.73 \bullet \theta_{calibrated} + 256.65, \text{ and}$$

$$\text{Age } 17: \theta_{target} = 44.13 \bullet \theta_{calibrated} + 283.89,$$

where θ_{target} denotes values on the final transformed scale, and $\theta_{calibrated}$ denotes values on the calibration scale. Overall summary statistics for the long-term trend samples are given in Table 14-9.

As in the past, interpretation of the long-term trend results was facilitated through the provision of scale anchoring information. In 1984, five NAEP reading scale levels were selected as anchor points. These points (described in *Trends in Academic Progress*) are:

150 = simple, discrete reading tasks;
200 = partially developed skills and understanding;
250 = interrelation of ideas and generalizations;
300 = understanding complicated information; and
350 = learning from specialized reading materials.

Detailed descriptions of the skills required to read at each level were derived and benchmark exercises were selected to exemplify each level. These same anchor points were used in the 1988, 1990, 1992, 1994, and 1996 reading long-term trend reports. The estimated proportion of students in each reporting category who are at or above each anchor point was examined in *Trends in Academic Progress*.

14.7 PARTITIONING OF THE ESTIMATION ERROR VARIANCE

For each age, the error variance of the final, transformed proficiency mean was partitioned into two parts as described in Chapter 11. This analysis yielded estimates of the proportion of error variance due to sampling students, and the proportion of error variance due to the latent nature of θ. These estimates are given in Table 14-10 (for stability, the estimates of the between-imputation variance, *B*, in Equation 11.9 are based on 100 imputations). More detailed information is available for gender and race/ethnicity subgroups in Appendix E.

Table 14-9

Means and Standard Deviations on the Reading Long-Term Trend Scale

Age	Assessment Year	All Five Plausible Values	
		Mean	S.D.
9	1984	211.0	41.1
	1988	211.8	41.2
	1990	209.2	44.7
	1992	210.5	40.4
	1994	211.0	40.5
	1996	212.4	39.0
13	1984	257.1	35.5
	1988	257.5	34.7
	1990	256.8	36.0
	1992	259.8	39.4
	1994	257.9	39.8
	1996	257.9	39.1
17	1984	288.8	40.3
	1988	290.1	37.1
	1990	290.2	41.3
	1992	289.7	43.0
	1994	288.1	44.4
	1996	287.6	42.2

Table 14-10

Estimation Error Variance and Related Coefficients for the Reading Long-Term Trend Assessment

Age	Total Estimation of Error Variance	Proportion of Variance Due to ...	
		Student Sampling	Latency of θ
9	1.01	0.86	0.14
13	0.93	0.85	0.15
17	1.08	0.82	0.18

Chapter 15

DATA ANALYSIS FOR THE LONG-TERM TREND MATHEMATICS ASSESSMENT[1]

Jiahe Qian and Norma A. Norris
Educational Testing Service

15.1 INTRODUCTION

This chapter describes the analyses performed on the responses to the cognitive and background items in the 1996 long-term trend assessment of mathematics. The emphasis of this chapter is on the methods and results of procedures used to develop the IRT-based scale scores; however, some attention is given to the analysis of constructed-response items. The theoretical underpinnings of the IRT and plausible values methodology described in this chapter are given in Chapter 11.

The objectives of the mathematics analyses were to prepare scale values and perform all analyses necessary to produce a long-term trend report in mathematics. The mathematics long-term trend results include the years 1973, 1978, 1982, 1986, 1990, 1992, 1994, and 1996. The results of 1996 long-term trend assessment of mathematics are presented in the *NAEP 1996 Trends in Academic Progress: Achievement of U.S. Students in Science, 1969 to 1996; Mathematics, 1973 to 1996; Reading, 1971 to 1996; and Writing, 1984 to 1996* (Campbell, Voelkl, & Donahue, 1997).

The student samples that were administered mathematics items in the 1996 long-term trend assessment are shown in Table 15-1. See Chapters 1 and 3 for descriptions of the target populations and the sample design used for the assessment.

Table 15-1
NAEP 1996 Mathematics Long-Term Trend Student Samples

Sample	Booklet IDs	Mode	Cohort Assessed	Time of Testing	Age Definition	Modal Grade	Number Assessed
9 [MS-LTTrend]	91-93	Tape	Age 9	1/3/96 - 3/8/96 (Winter)	CY	4	5,414
13 [MS-LTTrend]	91-93	Tape	Age 13	10/9/95 - 12/22/95 (Fall)	CY	8	5,658
17 [MS-LTTrend]	84-85	Tape	Age 17	3/11/96 - 5/10/96 (Spring)	Not CY	11	3,539

LEGEND

MS	Mathematics and science
LTTrend	Long-term trend assessment: booklets are identical to 1986 long-term trend assessments
Tape	Audiotape administration
CY	Calendar year: birthdates (1996 sample) in 1986 and 1982 for ages 9 and 13
Not CY	Age 17 only (1996 sample): birthdates between October 1 and September 30 of the appropriate years

[1] Jiahe Qian was the primary person responsible for the planning, specification, and coordination of the mathematics long-term trend analyses. Computer activities for all long-term trend mathematics scaling and data analyses were performed by Norma Norris. Nancy Allen and Eiji Muraki provided consultation.

Data from the 1996 long-term trend samples that contributed to the trends in mathematics achievement were scaled separately from the 1996 mathematics main samples. Accordingly, the long-term trend analysis and main analysis are presented in separate chapters. This chapter pertains to the scaling of the long-term trend data; information about the scaling of the data from the mathematics main assessment samples is presented in Chapter 12.

The long-term trend results reported in the *NAEP 1996 Trends in Academic Progress* are based on paced-tape administrations and occur at all age levels. The samples involved in the analysis are shown as 9[MS-LTTrend], 13[MS-LTTrend], and 17[MS-LTTrend] in Table 15-1. For ages 9 and 13, the long-term trend booklets contained blocks of reading, mathematics and science items. In the assessments, the mathematics and science blocks were administered by audiotape to pace the students through blocks (the reading blocks were only presented in print form). The age 17 long-term trend booklets contained only mathematics and science blocks, both administered by paced-tape recordings as well. All students received a block of common background questions, yet distinct for each age. Subject-area background questions were presented in the cognitive blocks. The booklets for the age 9 and age 13 samples (Booklets 91-93) are the same as those used for long-term trend assessments in 1986, 1988, 1990, 1992, and 1994. The booklets for the age 17 sample (Booklets 84-85) are the same as those used for the 1986, 1990, 1992, and 1994 long-term trend assessments. The booklets and the blocks within those booklets are listed in Tables 4-20 through 4-22 in chapter 4.

Table 15-2 clarifies the relationships among the 1996 long-term trend samples and samples from previous years. For ages 9, 13, and 17, the paced-tape bridge to the 1986 samples allows direct comparisons between the samples from the long-term assessments after 1990 and the 1986 long-term trend samples. There was also a paced-tape administration in 1988, at ages 9 and 13, that was comparable to the other years. However, a paced-tape administration was not conducted at age 17 in 1988. Instead, a noncomparable paper-based assessment was administered. Hence, 1988 is not included as a point in the long-term trend reporting. In 1986, the mathematics long-term trend items were scaled with common items from the 1978 and 1982 assessments. Because the 1973 assessment had few items in common with the current assessment, data from that assessment was not scaled using the IRT model but was linked to the trend line by a linear transformation involving the mean proportion correct for common items (See *Expanding the New Design: The NAEP 1985-86 Technical Report* (Beaton, 1988)). When comparisons were made including the 1970 and 1973 assessment results, z-tests rather than t-tests were used to test statistical significance (See Section 18.5.1). The 1996 long-term trend assessment was linked to the 1973, 1978, and 1982 assessments through the 1986 assessment. Information about previous assessment years is available in *Expanding the New Design: The NAEP 1985-86 Technical Report* (Beaton, 1988), *The NAEP 1992 Technical Report* (Johnson & Carlson, 1994), and *The NAEP 1994 Technical Report* (Allen, Kline, & Zelenak, 1996).

Table 15-3 indicates the number of items in common across different age combinations. Table 15-4 shows the number of items scaled in 1996 that were common across assessment years. The 1986, 1990, 1992, 1994, and 1996 assessments had all items in common. For age 9, the number of items common across assessment years 1978 to 1996 was only 35; for age 13, it was 56; and for age 17, it was 54. Item parameters were estimated assuming a univariate scale, since the number of items presented to each student was small and there were too few items to estimate several content area scales separately.

Table 15-2

NAEP Mathematics Samples Contributing to 1996 Long-Term Trend Results, 1973-1996

Cohort Assessed	Year	Sample	Subjects	Time of Testing	Mode of Administration	Age Definition	Modal Grade
Age 9	1973	Main	MS	Winter	Tape	CY	4
	1978	Main	M	Winter	Tape	CY	4
	1982	Main	MCS	Winter	Tape	CY	4
	1986	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1990	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1992	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1994	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1996	LTTrend[1]	MS	Winter	Tape[2]	CY	4
Age 13	1973	Main	MS	Fall	Tape	CY	8
	1978	Main	M	Fall	Tape	CY	8
	1982	Main	MCS	Fall	Tape	CY	8
	1986	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1990	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1992	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1994	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1996	LTTrend[1]	MS	Fall	Tape[2]	CY	8
Age 17	1973	Main	MS	Spring	Tape	Not CY	11
	1978	Main	M	Spring	Tape	Not CY	11
	1982	Main	MCS	Spring	Tape	Not CY	11
	1986	LTTrend[1]	MS	Spring	Tape	Not CY	11
	1990	LTTrend[1]	MS	Spring	Tape	Not CY	11
	1992	LTTrend[1]	MS	Spring	Tape	Not CY	11
	1994	LTTrend[1]	MS	Spring	Tape	Not CY	11
	1996	LTTrend[1]	MS	Spring	Tape	Not CY	11

[1] Within an age group, these samples received common booklets.
[2] Mathematics and science administered by audiotape, reading administered by print.

LEGEND

M	Mathematics	Main	Main assessment
MS	Mathematics and science	Tape	Audiotape administration
MCS	Mathematics, civics, and science	CY	Calendar year: birthdates (1996 sample) in 1986 and 1982 for ages 9 and 13
LTTrend	Long-term trend: booklets are identical to the long-term trend assessment of 1986.	Not CY	Age 17 only (1996 sample): birthdates between October 1 and September 30 of the appropriate years

The steps in the mathematics long-term trend analysis are documented in the following sections. Consistent with the procedures in earlier NAEP analyses, the first step was to calculate standard item statistics. The results served as a check for data entry errors and as a reasonableness check against results from previous assessments.

The second step was to fit an IRT model to the data from the 1996 and 1994 assessments for each age separately. This procedure puts item parameters and ability estimates on the same scale across years. The same item may have different item parameters for different age groups.

Table 15-3
Numbers of Scaled Mathematics Long-Term Trend Items Common Across Ages

Age	Booklet Numbers	Number of Items
Total		153
9 only	91-93	32
13 only	91-93	30
17 only	84-85	41
9 and 13 only	91-93, 91-93	20
9 and 17 only	91-93, 84-85	0
13 and 17 only	91-93, 84-85	27
9, 13, and 17	91-93, 91-93, 84-85	3

Table 15-4
Numbers of Scaled Mathematics Long-Term Trend Items Common Across Assessments

	Number of Items		
Assessment Year	Age 9	Age 13	Age 17
1986, 1990, 1992, 1994, 1996	55	80	71
1982, 1986, 1990, 1992, 1994, 1996	53	79	65
1978, 1986, 1990, 1992, 1994, 1996	35	56	54
1978, 1982, 1986, 1990, 1992, 1994, 1996	35	56	54

Next, the analysis for an age group was completed by the creation of plausible values through a multiple imputation estimation procedure in which item parameter estimates, student responses, and student background information were combined to produce the most precise possible estimates of student subgroup ability. Plausible values were used to calculate proficiency means for the entire sample and for the selected subgroups.

Finally, the scales of the 1996 trend assessment were transformed to proficiency scale used in previous mathematics trend assessments. These proficiency means constitute the last point in the mathematics long-term trend from 1973 to 1996. The only available estimates of the proficiency means for 1973 were linked via extrapolation to the IRT scale, but the data from that year was never scaled using an IRT model.

15.2 ITEM ANALYSIS FOR THE MATHEMATICS LONG-TERM TREND ASSESSMENT

No problems in coding, formats, or data were detected. The conventional item analysis, with results displayed in Table 15-5, was performed at the block level on the paced-tape long-term trend data.

Table 15-5 contains the number of items, size of the sample administered to the block, mean weighted proportion correct, mean weighted r-biserial, and mean weighted alpha as a measure of

reliability for each block. The average values were calculated using examinee weights and the items in the block that were scaled. The 1996 item-level statistics were not very different from those for the 1994 assessment. Similar statistics for the 1994 assessment were reported in Table 16-5 of *The NAEP 1994 Technical Report*.

Table 15-5
Descriptive Statistics for Item Blocks in the Mathematics Long-Term Trend Samples (1996)

Statistic	Block M1	M2	M3[1]
Age 9			
Number of scaled items	24	26	5
Number of scaled constructed response items	9	9	0
Unweighted sample size	1,852	1,841	1,721
Average weighted proportion correct	.62	.64	.68
Average weighted r-biserial	.61	.65	.83
Weighted alpha reliability	.82	.86	.50
Age 13			
Number of scaled items	36	36	8
Number of scaled constructed response items	9	8	0
Unweighted sample size	1,928	1,864	1,866
Average weighted proportion correct	.68	.62	.66
Average weighted r-biserial	.59	.55	.68
Weighted alpha reliability	.87	.85	.61
Age 17			
Number of scaled items	33	33	5
Number of scaled constructed response items	10	5	1
Unweighted sample size	1,848	1,848	1,691
Average weighted proportion correct	.66	.67	.56
Average weighted r-biserial	.69	.63	.76
Weighted alpha reliability	.90	.87	.54

[1] This block is mostly calculator items, which were not analyzed. For the item analysis, students who did not respond to any items in the block were omitted; however, such students were assigned proficiencies in the final database.

In the 1996 mathematics long-term trend assessment, 20 percent of the samples of the constructed-response items were used to check the interrater reliability—the score agreement between first and second raters. The percent of exact agreement ranged from 96.3 to 100 percent; and the intraclass correlation ranged from .902 to 1.00—except .886 for item N269201 in the age 13 sample. In general, the interrater reliability was very high in the 1996 mathematics long-term trend assessment.

The correspondence between blocks, booklets, and samples is given for the mathematics long-term trend assessment in Tables 4-20 through 4-22 in Chapter 4. Common labeling of these blocks across ages does not denote common items.

15.3 IRT SCALING FOR THE MATHEMATICS LONG-TERM TREND ASSESSMENT

15.3.1 Item Parameter Estimation

The scaling process began with the estimation of item parameters. IRT parameters were estimated using the NAEP version of the BILOG/PARSCALE program (Mislevy & Bock, 1982; Muraki & Bock, 1991) described in Chapter 11. Item calibration was performed separately for each of the three age groups, using the total combined data from the 1994 and 1996 assessments. Including the 1994 assessment data assures that item parameters will be similar for adjacent assessments so that year-to-year trends will not be distorted by abrupt changes in calibration. The calibration was performed on the entire sample of students, resulting in a range of about 1,700 to 1,900 examinee responses to each item in each assessment year. The calibration was based on student weights that were rescaled for the 1996 data so that the sum of the weights equaled the unweighted sample size. Also, weights for the 1996 data were restandardized to give equal weight to the two assessment years included in the scaling. As with the previous assessment, calculator items were excluded from the analysis. Because calculators have changed greatly since the start of the long-term trend assessment, it was judged that calculator questions are no longer comparable across time. These items were kept in the assessment, since excluding them would have changed the testing context.

Since parameters for items in blocks M1, M2, and M3 were estimated separately for ages 9, 13, and 17, items administered at more than one age have multiple sets of item parameter estimates. Items were examined for lack of fit with the data. Those that exhibited extreme violation of IRT assumptions (i.e., did not have monotonically increasing item characteristic curves) were deleted from the analysis, as they were in previous assessments. Other items were deleted because they were calculator items, which were not considered part of the regular assessment. These excluded items appear in Tables 15-6, 15-7, and 15-8. As a result of these deletions, 55 items were scaled for age 9, 80 items were scaled for age 13, and 71 items were scaled for age 17. Of the 153 noncalculator items that were part of the assessment, seven items (5%) were excluded due to poor fit with the data. A list of the items scaled for each of the ages, along with their item parameter estimates, appears in Appendix D.

15.4 DERIVED BACKGROUND VARIABLES

In the long-term trend analysis, all derived background variables were used to define subgroups of students for reporting. For this reason, these variables were also used in conditioning. Derived reporting variables are described in Appendix B.

Table 15-6

Items Deleted from the Age 9 Mathematics Long-Term Trend Analysis

Booklet IDs	Block	Item	Reason for Exclusion
91	M1	N252601	Was deleted in prior assessment
		N262502	Was deleted in prior assessment
92	M3	N268221	Calculator item[1]
		N276021	Calculator item
		N276022	Calculator item
		N276821	Calculator item
		N276822	Calculator item
		N276823	Calculator item
		N277621	Calculator item
		N277622	Calculator item
		N277623	Calculator item
		N284021	Calculator item
		N284022	Calculator item

[1] All calculator items were deleted from the analysis.

Table 15-7

Items Deleted from the Age 13 Mathematics Long-Term Trend Analysis

Booklet IDs	Block	Item	Reason for Exclusion
91	M1	N262502	Was deleted in prior assessment
93	M2	N261601	Was deleted in prior assessment
92	M3	N264521	Calculator item[1]
		N259921	Calculator item
		N276821	Calculator item
		N276822	Calculator item
		N276823	Calculator item
		N278921	Calculator item
		N278922	Calculator item
		N278923	Calculator item
		N278924	Calculator item
		N278925	Calculator item
		N280621	Calculator item
		N280622	Calculator item
		N280623	Calculator item
		N280624	Calculator item
		N280625	Calculator item
		N280626	Calculator item

[1] All calculator items were deleted from the analysis.

Table 15-8
Items Deleted from the Age 17 Mathematics Long-Term Trend Analysis

Booklet IDs	Block	Item	Reason for Exclusion
84	M1	N282801	Was deleted in prior assessment
		N285701	Was deleted in prior assessment
84	M2	N266801	Was deleted in prior assessment
		N255301	Was deleted in prior assessment
85	M3	N259921	Calculator item[1]
		N264321	Calculator item
		N264521	Calculator item
		N267921	Calculator item
		N276821	Calculator item
		N276822	Calculator item
		N276823	Calculator item
		N278921	Calculator item
		N278922	Calculator item
		N278923	Calculator item
		N278924	Calculator item
		N278925	Calculator item
		N280621	Calculator item
		N280622	Calculator item
		N280623	Calculator item
		N280624	Calculator item
		N280625	Calculator item
		N280626	Calculator item
		N285321	Calculator item

[1]All calculator items were deleted from the analysis.

15.5 GENERATION OF PLAUSIBLE VALUES

Plausible values were calculated separately for each age group. In this phase of analysis, student background information was used to condition item responses in order to more accurately estimate average subgroup abilities. The conditioning program BGROUP was used to combine NAEP BILOG/PARSCALE item parameters with weighted item responses and background variables to produce posterior ability estimates called plausible values. As defined in Chapter 11, BGROUP is an enhanced version of the original conditioning program, MGROUP. *Plausible values are not test scores* in the usual sense, but can be used to provide consistent estimates of population characteristics. There were 53 contrasts in the conditioning model at age 9, 56 at age 13, and 63 at age 17. Appendix C gives the codings for the conditioning variables for the three age groups. A check was made on the distributions of the plausible values for each age, including inspection of the whole group and subgroup means and standard deviations. Table 15-9 contains a list of the number of background contrasts included in conditioning, as well as the proportion of variance accounted for by the conditioning model for each age/grade.

Table 15-9
Proportion of Proficiency Variance Accounted for by the Conditioning Model for the Mathematics Long-Term Trend Assessment

Age/Grade	Number of Conditioning Contrasts[1]	Proportion of Proficiency Variance
9/4	53	.37
13/8	56	.35
17/12	63	.57

[1] Including the constant term.

15.6 THE FINAL MATHEMATICS LONG-TERM TREND SCALE

Since the plausible value (theta) scales have a linear indeterminacy, comparisons with previous assessments will be sensible only if the scale is linearly transformed to a meaningful metric. This indeterminacy was resolved by linking the 1996 scales to previous long-term trend scales. The 1996 data had to be transformed to compensate for linear changes in the scale due to employing newly estimated item parameters and new BGROUP conditioning parameters in 1996. The transformation was accomplished by first reestimating the 1994 student abilities using 1996 item parameters and 1996 BGROUP parameters. (For score metric transformation, see Chapter 9.) The new 1994 ability estimates were then equated to the old 1994 ability estimates by matching the first two moments (i.e., the mean and standard deviation). The constants for this transformation were then applied to the 1996 data. The transformation equations that resulted are

$$\text{Age 9: } \theta_{target} = 34.04 \bullet \theta_{calibrated} + 230.46$$

$$\text{Age 13: } \theta_{target} = 33.08 \bullet \theta_{calibrated} + 273.91$$

$$\text{Age 17: } \theta_{target} = 30.46 \bullet \theta_{calibrated} + 306.57,$$

where θ_{target} denotes an individual's value on the final transformed scale of the 1996 data and $\theta_{calibrated}$ denotes an individual's value on the original 1996 theta scale. Overall summary statistics for the long-term trend samples are given in Table 15-10. For the descriptions of the results of the mathematics long-term trend study, see *NAEP 1996 Trends in Academic Progress* (Campbell, Voelkl, & Donahue, 1997).

To provide a context for interpreting the overall mathematics long-term trend results, the NAEP mathematics results were "anchored" at five NAEP mathematic scale levels. These points (described in the *NAEP 1996 Trends in Academic Progress*) are:

150 = simple arithmetic facts;
200 = beginning skills and understanding;
250 = numerical operations and beginning problem solving;
300 = moderately complex procedures and reasoning; and
350 = multi-step problem solving and algebra.

These same anchor points were used in the 1978, 1982, 1986, 1990, 1992, and 1994 mathematics long-term trend reports.

Table 15-10
Means and Standard Deviations on the Mathematics Long-Term Trend Proficiency Scale

Age	Assessment	All Five Plausible Values Mean	S. D.
9	1978	218.6	36.0
	1982	219.0	34.8
	1986	221.7	34.0
	1990	229.6	32.9
	1992	229.6	33.1
	1994	231.1	33.2
	1996	231.0	33.8
13	1978	264.1	39.0
	1982	268.6	33.4
	1986	269.0	30.8
	1990	270.4	31.3
	1992	273.1	30.9
	1994	274.3	32.4
	1996	274.3	31.6
17	1978	300.4	34.9
	1982	298.5	32.4
	1986	302.0	31.0
	1990	304.6	31.3
	1992	306.7	30.1
	1994	306.2	30.2
	1996	307.2	30.2

15.7 PARTITIONING OF THE ESTIMATION ERROR VARIANCE

For each age's scale, the error variance of the final transformed proficiency mean was partitioned as described in Chapter 11. 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 proficiency, θ, is a latent variable that is estimated rather than observed. Table 15-11 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 11.9 are based on 100 imputations.).

Table 15-11
Estimation Error Variance and Related Coefficients for the Mathematics Long-Term Trend Assessment

Age	Total Estimation Error Variance	Proportion of Variance Due to . . . Student Sampling	Latency of θ
9	0.66	0.85	0.15
13	0.66	0.89	0.11
17	1.24	0.93	0.07

More detailed information is available for gender and race/ethnicity subgroups in Appendix E.

Chapter 16

DATA ANALYSIS FOR THE LONG-TERM TREND SCIENCE ASSESSMENT[1]

Jinming Zhang and Norma A. Norris
Educational Testing Service

16.1 INTRODUCTION

This chapter describes the analyses performed on the responses to the cognitive and background items in the 1996 long-term trend assessment of science. The objectives of the science analyses are to prepare scale values and perform all analyses necessary to produce a long-term trend report in science. The results obtained from these analyses includes the years 1969-1970, 1973, 1977, 1982, 1986, 1990, 1992, 1994, and 1996, and are presented in the *NAEP 1996 Trends in Academic Progress: Achievement of U.S. Students in Science, 1969 to 1996; Mathematics, 1973 to 1996; Reading, 1971 to 1996; and Writing, 1984 to 1996* (Campbell, Voelkl, & Donahue, 1997). The theoretical underpinnings of the IRT and the plausible values methodology used in this chapter are described in Chapter 9 and Chapter 11, and are therefore not detailed here.

The student samples that were administered science items in the 1996 long-term trend assessment are shown in Table 16-1 as 9[MS-LTTrend], 13[MS-LTTrend], and 17[MS-LTTrend]. (See Chapters 1 and 3 for descriptions of the target populations and the sample design used for the assessment.) Data from the long-term trend samples that contributed to the trends in science achievement were scaled separately from the 1996 science main focused-BIB samples. Accordingly, the long-term trend and main analyses are presented in separate chapters. Information about the scaling of the data from the science main focused-BIB samples is presented in Chapter 13.

The science long-term trend results reported in the *NAEP 1996 Trends in Academic Progress* are based on paced-tape administrations at all three age levels. For ages 9 and 13, the long-term trend booklets contain one mathematics block, one reading block, and one science block. The science and mathematics blocks were paced by tape-recordings (i.e., tape-recordings were used to be sure that the items were read in a consistent manner in every session and pace students through the blocks) and the reading block was presented in print form only and were not paced by tape-recordings. The age 17 long-term trend booklets contain only mathematics and science blocks, both paced by tape-recordings. All students received a block of common background questions, distinct for each age. Subject-area background questions were presented in the cognitive blocks. The booklets for the age 9 and age 13 samples (Booklets 91-93) and the booklets for the age 17 samples (Booklets 84-85) are the same as those used for long-term trend assessments in 1986, 1990, 1992, and 1994. The booklets and the blocks within those booklets are listed in Chapter 4. Additional information about all of the items in these blocks is also found in that chapter. This chapter includes specific information about the long-term trend items that were scaled.

[1] Jinming Zhang was the primary person responsible for the planning, specification, and coordination of the science long-term trend analyses. Computer activities for all long-term trend science scaling and data analyses were performed by Norma Norris. Nancy Allen, Eiji Muraki, and John Donoghue provided consultation.

Table 16-1
NAEP 1996 Long-Term Trend Science Student Samples

Sample	Booklet IDs	Mode	Cohort Assessed	Time of Testing	Age Definition	Modal Grade	Number Assessed
9 [MS-LTTrend]	91-93	Tape	Age 9	1/3/96 – 3/8/96 (Winter)	CY	4	5,414
13 [MS-LTTrend]	91-93	Tape	Age 13	10/9/95 – 12/22/95 (Fall)	CY	8	5,658
17 [MS-LTTrend]	84-85	Tape	Age 17	3/11/96 – 5/10/96 (Spring)	Not CY	11	3,539

LEGEND

MS	Mathematics and science
LTTrend	Long-term trend assessment: booklets are identical to 1986 long-term trend assessments
Tape	Audiotape administration
CY	Calendar year: birthdates in 1986 and 1982 for ages 9 and 13, respectively
Not CY	Age 17 only: birthdates between October 1, 1978, and September 30, 1979

Table 16-2 clarifies the relationships among the 1996 long-term trend samples and samples from previous years. For all ages, the 1996 science long-term trend samples allow direct comparisons with 1986, 1990, 1992, and 1994 long-term trend samples because the same booklets were used in these assessments. There was also a tape administration in 1988 at ages 9 and 13 that was comparable to the other years. However, a tape administration was not conducted at age 17 in 1988. Instead, a noncomparable paper-based assessment was conducted. Hence, 1988 is not included as a point in the long-term trend reporting. In 1986, the science long-term trend items were scaled with common items from the 1977 and 1982 assessments. Because of the small number of items in common with those in the 1969-70 and 1973 assessments, data from the 1969-70 and 1973 assessments were not scaled using the IRT model, but were linked to the long-term trend line by a linear transformation involving the logit of mean proportion correct for common items. When comparisons were made including the 1969-70 and 1973 assessment results, z-tests rather than t-tests were used to test statistical significance (See Section 18.5.1). From 1990, each new long-term trend assessment was linked to the previous assessments through the last assessment. For instance, the 1996 long-term trend assessment was linked to the previous assessments through the 1994 long-term trend assessment. Information about previous assessment years, including 1969-70 and 1973, is available in Chapter 11 of *Expanding the New Design: The NAEP 1985-86 Technical Report* (Yamamoto, 1988), Chapter 14 of *The NAEP 1990 Technical Report* (Allen, 1992), Chapter 14 of *The NAEP 1992 Technical Report* (Allen & Isham, 1994), and Chapter 17 of *The NAEP 1994 Technical Report* (Swinton, Allen, Isham & Chen, 1996).

The numbers of scaled items in common across different ages are presented in Table 16-3. As was done with previous long-term trend analyses, each age was scaled separately and the long-term trend scales are univariate. Derivation of scales for specific content areas was not feasible given the limited number of items presented to students in the long-term trend samples. The number of items scaled in 1996 that were common across assessment years is presented in Table 16-4.

The steps in the science long-term trend analysis are documented in the following sections. As is usual in NAEP analyses, the first step was to gather item-level and block-level information. Then, the cognitive items were calibrated according to the IRT model. Next, derived background variables were calculated, and plausible values were generated after conditioning on available background variables and selected two-way interactions. Finally, the scale values were placed on the final science long-term trend scale used in previous trend assessments.

Table 16-2

NAEP Science Samples Contributing to 1996 Long-Term Trend Results, 1970-1996

Cohort Assessed	Year	Sample	Subjects	Time of Testing	Mode of Administration	Age Definition	Modal Grade
Age 9	1970	Main	SWC	Winter	Tape	CY	4
	1973	Main	MS	Winter	Tape	CY	4
	1977	Main	SCI	Winter	Tape	CY	4
	1982	Main	MSC	Winter	Tape	CY	4
	1986	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1990	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1992	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1994	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1996	LTTrend[1]	MS	Winter	Tape[2]	CY	4
Age 13	1970	Main	SWC	Fall	Tape	CY	8
	1973	Main	MS	Fall	Tape	CY	8
	1977	Main	SCI	Fall	Tape	CY	8
	1982	Main	MSC	Fall	Tape	CY	8
	1986	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1990	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1992	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1994	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1996	LTTrend[1]	MS	Fall	Tape[2]	CY	8
Age 17	1969	Main	SWC	Spring	Tape	Not CY	11
	1973	Main	MS	Spring	Tape	Not CY	11
	1977	Main	SCI	Spring	Tape	Not CY	11
	1982	Main	MSC	Spring	Tape	Not CY	11
	1986	LTTrend[1]	MS	Spring	Tape	Not CY	11
	1990	LTTrend[1]	MS	Spring	Tape	Not CY	11
	1992	LTTrend[1]	MS	Spring	Tape	Not CY	11
	1994	LTTrend[1]	MS	Spring	Tape	Not CY	11
	1996	LTTrend[1]	MS	Spring	Tape	Not CY	11

[1] Within an age group, these samples received common booklets.

[2] Mathematics and science administered by audiotape, reading administered by print.

LEGEND

SCI	Science	LTTrend	Long-term trend: booklets are identical to the long-term trend assessment of 1986
MS	Mathematics and science		
MSC	Mathematics, science, and citizenship	Tape	Audiotape administration
SWC	Science, writing, and citizenship	CY	Calendar year: birthdates in 1986 and 1992 for ages 9 and 13 in the 1996 assessment
Main	Main assessment	Not CY	Age 17 only: birthdates between October 1 and September 30 of the appropriate years

Table 16-3
Numbers of Scaled Science Long-Term Trend Items Common Across Ages

Age	Booklet Numbers	Number of Items
9 only	91-93	55
13 only	91-93	30
17 only	84-85	32
9 and 13 only	91-93, 91-93	0
9 and 17 only	91-93, 84-85	0
13 and 17 only	91-93, 84-85	45[1]
9, 13, and 17	91-93, 91-93, 84-85	1
Total		163

[1] One of these items (N406303) was treated as a different item from 1990 in the scaling of the 1992 assessment, but only for age 13. It was treated as an item common to 1992, 1994 and 1996 for all ages in the 1994 and 1996 assessments.

Table 16-4
Numbers of Scaled Science Long-Term Trend Items Common Across Assessments

	Number of Items		
Assessment Years	Age 9	Age 13	Age 17
1986, 1990, 1992, 1994,1996	56	76	78
1982, 1986, 1990, 1992, 1994,1996	10[1]	58	47
1977, 1986, 1990, 1992, 1994,1996	56	76	76
1977, 1982, 1986, 1990, 1992, 1994, 1996	10[1]	58[2]	45

[1] Twenty-four items common to years 1977 and 1982, but not later years, were included in the 1986 scaling of these items to stabilize the estimation of the item parameters. See *Expanding the New Design: The NAEP 1985-86 Technical Report* for more information.
[2] One of these items (N406303) was treated as a different item from 1990 in the scaling of the 1992 assessment, but only for age 13. It was treated as an item common to 1992, 1994 and 1996 in the 1994 and 1996 assessments for all ages.

16.2 ITEM ANALYSIS FOR THE SCIENCE LONG-TERM TREND ASSESSMENT

Conventional item analyses did not identify any difficulties with the 1996 long-term trend data for the 1996 samples that bridge to 1986. Table 16-5 contains information about the science long-term trend blocks. These blocks were presented to samples 9[MS-LTTrend], 13[MS-LTTrend], and 17[MS-LTTrend]. At all ages, the blocks labeled S1, S2, and S3 were presented intact to students in the 1986, 1990, 1992, 1994, and 1996 long-term trend samples. The age 9 and age 13 blocks appeared in Booklets 91 through 93. For age 17, Block S3 was in Booklet 84, and Blocks S1 and S2 were in Booklet 85. The correspondence between blocks, booklets, and samples is given for the long-term trend assessment in Tables 4-14 through 4-16 in Chapter 4. Common labeling of these blocks across ages does not denote common items.

Table 16-5 contains the number of scaled items, size of the sample administered to the block, mean weighted proportion correct, mean weighted r-biserial, and mean weighted alpha as a measure of reliability for each block. The average values were calculated using examinee sampling weights and the

responses to the items in the block that were scaled. On average, the 1996 item-level statistics were not very different from those for the 1994 assessments. The percent of examinees not reaching items in the science long-term trend blocks was almost always zero because the items were administered with a tape-recording to pace response time.

Table 16-5
Descriptive Statistics for Item Blocks in the Science Long-Term Trend Samples (1996)

Statistic	Blocks S1	S2	S3
Age 9			
Number of scaled items	17	20	19
Number of scaled constructed-response items	0	0	0
Unweighted sample size	1,852	1,721	1,841
Average weighted proportion correct	0.62	0.58	0.71
Average weighted r-biserial	0.57	0.48	0.58
Weighted alpha reliability	0.71	0.64	0.73
Age 13			
Number of scaled items	23	30	23
Number of scaled constructed-response items	0	0	0
Unweighted sample size	1,928	1,866	1,864
Average weighted proportion correct	0.53	0.56	0.61
Average weighted r-biserial	0.53	0.50	0.52
Weighted alpha reliability	0.74	0.79	0.72
Age 17			
Number of scaled items	24	31	23
Number of scaled constructed-response items	0	0	0
Unweighted sample size	1,691	1,691	1,848
Average weighted proportion correct	0.65	0.65	0.61
Average weighted r-biserial	0.49	0.54	0.64
Weighted alpha reliability	0.68	0.79	0.82

16.3 IRT SCALING FOR THE SCIENCE LONG-TERM TREND ASSESSMENT

16.3.1 Item Parameter Estimation

The first step in the scaling process was the estimation of item parameters for the long-term trend items. This item calibration was performed using the NAEP version of the BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs. Items were calibrated separately for each of the three age groups, using combined data from the 1994 and 1996 assessment years and treating each assessment sample as a sample from a separate subpopulation. In several previous long-term trend analyses, combined data from the last assessment and the current assessment were used for item parameter estimation. The purposes for including the last long-term trend assessment data are to assure that item parameter estimates will be similar for adjacent assessments so that year-to-year trends will not be distorted by abrupt changes in calibration, and to make it possible to link the current long-term trend assessment to the previous assessments through the last assessment. Student weights were used for the calibration. To ensure that

each assessment year had a similar influence on the calibration, student weights for each 1994 age group were multiplied by a constant, to adjust them to have the same sum as the sum of the student weights for the corresponding 1996 age group.

Although other items were examined for irregularities, only items that were deleted from the previous scaling of the paced-tape long-term trend data were excluded in the 1996 analysis. Eight percent of the items (18 items) administered to the long-term trend sample were excluded from analyses of previous assessments. The deleted items appear in Table 16-6. As a result of these deletions, 56 items were scaled for age 9, 76 items were scaled for age 13, and 78 items were scaled for age 17. A list of the items scaled for each of the ages, along with their item parameter estimates, appears in Appendix D.

Table 16-6
Items Deleted from the Paced-Tape Science Long-Term Trend Analysis

Age	Booklet IDs	Block	Item	Reason for Exclusion
9	91	S1	N400201	Excluded in previous assessment
	92	S2	N401701	Excluded in previous assessment
	92	S2	N402003	Excluded in previous assessment
	92	S2	N402004	Excluded in previous assessment
	92	S2	N402601	Excluded in previous assessment
	92	S2	N402603	Excluded in previous assessment
	93	S3	N403802	Excluded in previous assessment
13	91	S1	N404902	Excluded in previous assessment
	91	S1	N404903	Excluded in previous assessment
	92	S2	N407501	Excluded in previous assessment
	93	S3	N409401	Excluded in previous assessment
	93	S3	N409402	Excluded in previous assessment
	93	S3	N409403	Excluded in previous assessment
	93	S3	N409801	Excluded in previous assessment
17	85	S1	N410001	Excluded in previous assessment
	85	S1	N410002	Excluded in previous assessment
	85	S1	N410301	Excluded in previous assessment
	85	S2	N407402	Excluded in previous assessment

16.3.2 Derived Background Variables

In the long-term trend analysis, all variables derived from background questions were used both in generating plausible values and in reporting (to define subgroups). Derived conditioning and reporting variables are described in Appendix B.

16.4 GENERATION OF PLAUSIBLE VALUES

The generation of plausible values was conducted independently for each age group. The item parameters from NAEP-BILOG/PARSCALE, final student weights, item responses and selected background variables (conditioning variables) were used with the computer program BGROUP (described in Chapter 11) in order to generate the plausible values for each student. There were 49 contrasts in the conditioning model (11.8) at age 9, excluding an overall constant, 52 at age 13, and 58 at age 17. Appendix C gives the codings for the conditioning variables for the three age groups. A check on the distributions of the plausible values for each age was made. The generation of plausible values is described in more detail in Chapters 9 and 11. Table 16-7 contains a list of the number of background contrasts included in conditioning, as well as the proportion of variance accounted for by the conditioning model for each age. This proportion is the ratio of the difference between the total variance and the BGROUP residual variance, divided by the total variance. The total variance is the mean of the five theta-scale variances obtained by their respective plausible values.

Table 16-7
Proportion of Proficiency Variance Accounted for by the Conditioning Model for the Science Long-Term Trend Assessment

Age	Number of Conditioning Contrasts[1]	Proportion of Proficiency Variance
9	49	0.33
13	52	0.37
17	58	0.46

[1] Excluding the constant term.

16.5 THE FINAL SCIENCE LONG-TERM TREND SCALE

The linear indeterminacy of the long-term trend scale was resolved by linking the 1996 long-term trend scales to the previous long-term trend scales using the following procedure. For each age, the item parameters based on combined data from 1994 and 1996 were used with the 1994 data to find plausible values for the 1994 data. The mean and standard deviation of all of the plausible values (theta scale) were calculated and matched to the mean and standard deviation of all of the science long-term trend scale scores (final reporting scale) based on the 1994 item parameters and 1994 data as reported in the *NAEP 1994 Technical Report*. The transformations that resulted from this matching of the first two moments for the 1994 data are

$$\text{Age 9: } \theta_{target} = 38.57 \bullet \theta_{calibrated} + 232.56,$$

$$\text{Age 13: } \theta_{target} = 40.11 \bullet \theta_{calibrated} + 255.54, \text{ and}$$

$$\text{Age 17: } \theta_{target} = 48.28 \bullet \theta_{calibrated} + 293.82,$$

where θ_{target} denotes values on the final reporting scale of the 1996 data and $\theta_{calibrated}$ denotes values on the original 1996 calibration (theta) scale. Overall summary statistics for the long-term trend scales are given

in Table 16-8. The detailed science long-term trend results from the analyses described in this chapter are reported in *NAEP 1996 Trends in Academic Progress*.

Table 16-8
Means and Standard Deviations on the Science Long-Term Trend Scale

		All Five Plausible Values	
Age	**Assessment**	**Mean**	**S. D.**
9	1977	219.9	44.9
	1982	220.8	40.9
	1986	224.3	41.6
	1990	228.7	40.2
	1992	230.6	39.9
	1994	231.0	40.9
	1996	229.7	42.2
13	1977	247.4	43.5
	1982	250.1	38.6
	1986	251.4	36.6
	1990	255.2	37.6
	1992	258.0	36.9
	1994	256.8	37.2
	1996	256.0	38.4
17	1977	289.5	45.0
	1982	283.3	46.7
	1986	288.5	44.4
	1990	290.4	46.2
	1992	294.1	44.7
	1994	294.0	45.6
	1996	295.7	45.1

As in the past, interpretation of the science long-term trend results was facilitated through the provision of scale anchoring information. In 1986, five science scale levels were selected as anchor points, using the process described in *Expanding the New Design: The 1985-86 Technical Report* (Beaton, 1988). Because the 1996 science long-term trend scale was tied to the 1986 long-term trend scale through the 1990, 1992, and 1994 data, the distribution of proficiency scores derived from the long-term trend samples can be described in terms of scale anchors. The five levels of science proficiency are

150 = Knows everyday science facts;
200 = Understands simple scientific principles;
250 = Applies basic scientific information;
300 = Analyzes scientific procedures and data; and
350 = Integrates specialized scientific information.

16.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE

For each age, the error variance of the final reporting scale mean was partitioned into two parts as described in Chapter 11. One part of the error variance is due to the sampling of students (sampling variance) and the other is due to the fact that proficiency θ is a latent variable that is estimated rather than observed. These estimates are given in Table 16-9 (for stability, the estimates of the between-imputation variance, *B*, in Equation 11.9 are calculated based on 100 imputations). More detailed information for gender and race/ethnicity subgroups is available in Appendix E.

Table 16-9
Estimation Error Variance and Related Coefficients for the Science Long-Term Trend Assessment

		Proportion of Variance Due To . . .	
Age	**Total Estimation Error Variance**	**Student Sampling**	**Latency of θ**
9	1.13	0.81	0.19
13	1.04	0.87	0.13
17	1.40	0.86	0.14

Chapter 17

DATA ANALYSIS FOR THE LONG-TERM TREND WRITING ASSESSMENT

[This chapter is intended to provide information about the 1996 long-term trend assessment in writing; however, the data from this assessment are currently under review. After additional examination and analyses, this chapter will be included in a revised web version of the complete report.]

Chapter 18

CONVENTIONS USED IN HYPOTHESIS TESTING AND REPORTING NAEP RESULTS[1]

Spencer S. Swinton, David S. Freund, and Nancy L. Allen
Educational Testing Service

18.1 OVERVIEW

Results for the 1996 NAEP Assessments were disseminated in several different reports: the *NAEP 1996 Mathematics Report Card for the Nation and the States*, the *NAEP 1996 Science Report Card for the Nation and the States*, *NAEP 1996 Trends in Academic Progress*, *Cross-State Data Compendium from the NAEP 1996 Mathematics Assessment, Cross-State Compendium from the NAEP 1996 Science Assessment*, and, distributed only in electronic form, six sections of summary data tables for each report. These reports are published on the NCES/NAEP website *http://nces.ed.gov/naep*. Several other reports based on 1996 NAEP data will be forthcoming.

The *NAEP 1996 Mathematics Report Card for the Nation and the States* and the *NAEP 1996 Science Report Card for the Nation and the States* highlight key assessment results for the nation and summarize results across the jurisdictions participating in the assessments. These reports contain composite scale score results (scale score means, etc.) for the nation, for each of the four regions of the country, and for public-school students within each jurisdiction participating in the State Assessment[2], both overall and by primary reporting variables. The seven key reporting variables (referred to here as primary reporting variables) are gender, race/ethnicity, level of parents' education, Title I participation, eligibility for free or reduced cost school lunch, type of location, and type of school (public, Catholic schools, other religious schools, and other private schools). For public-school students, scale score means were reported for a variety of other subpopulations defined by responses to items from the student, teacher, and school questionnaires and by school and location demographic variables provided by Westat, Inc.[3] Upcoming reports will include estimates of scale score means and selected percentiles for specific subgroups of students of interest in each report.

The report *NAEP 1996 Trends in Academic Progress* provides a look at NAEP results for Science, Mathematics, Reading, and Writing since the first NAEP assessments of those subjects in 1969-70. This report includes scale score results for the nation overall and by gender, race/ethnicity, gender and race/ethnicity, region, level of parents' education and type of school (public and nonpublic). It also provides percentages of students in categories defined by subject specific background variables (such as students who reported having experimented with living plants), along with their average scale scores. The report contains trends in average scale scores by quartile and percentages of students performing at or above selected performance levels. An additional report gives data for the mechanics of writing long-term trend.

---

[1] Spencer S. Swinton played a role in making decisions about hypothesis testing methods and procedures and worked with David S. Freund who implemented many of the methods and procedures in computer programs. Nancy L. Allen contributed to the current version of this chapter.

[2] Further technical documentation for the State Assessments appears in the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics* and the *Technical Report of the NAEP 1996 State Assessment Program in Science.*

[3] Some of these variables were used by Westat, Inc., in developing the sampling frame for the assessment and in drawing the sample of participating schools.

The third type of report consists of a number of data compendia. Two of these are entitled the *Cross-State Data Compendium from the NAEP 1996 Mathematics Assessment* and the *Cross-State Data Compendium from the NAEP 1996 Science Assessment*. Like the *Report Cards*, the *Compendia* report results for the nation and for all of the jurisdictions participating in the State Assessment. The *Compendia* contain most of the tables included in the *Report Cards* plus tables that provide composite scale results for a large number of secondary reporting variables (e.g., amount of homework, teacher preparation).

The fourth type of summary report is an electronically-delivered collection of summary data tables that contain detailed breakdowns of the science scale score data for each sample according to the responses to the student, teacher, and school questionnaires for the public-school, nonpublic-school, and combined populations as a whole and for important subgroups of the public-school population, as defined by the primary reporting variables. There are six sections in each collection of summary data tables:

> *The Distribution Data Section* provides selected composite-scale and science subscale percentiles for the public-school, nonpublic-school, and total populations and for the major demographic subgroups of the national school population.
>
> *The Student Questionnaire Section* breaks down the composite scale score data according to the students' responses to questions in the three student questionnaires (common core, subject-specific background, and motivational section) included in the assessment booklets.
>
> *The Teacher Questionnaire Section* breaks down the composite scale score data according to the teachers' responses to questions in teacher questionnaires, where they are available.
>
> *The School Questionnaire Section* breaks down the composite scale score data according to the principals' (or other administrators') responses to questions in the school characteristics and policies questionnaire.
>
> *The Scale Section* breaks down the scale score data for the mathematics content strands or the fields of science according to selected items (such as the amount of science homework done per day) from the questionnaires.
>
> *The Item Section* provides the response data (percent of students choosing each option) for each cognitive item in the assessment.

The production of these reports required many decisions about a variety of data analysis and statistical issues. For example, certain categories of the reporting variables contained limited numbers of examinees. A decision was needed as to what constituted a sufficient sample size to permit the reliable reporting of subgroup results, and which, if any, estimates were sufficiently unreliable to need to be "flagged" as a caution to readers. As a second example, the performance for subgroups of students were compared. A number of inferential rules, based on logical and statistical considerations, had to be developed to ensure that conclusions are adequately supported by the data from the assessment. Practical comparison procedures were required to control for Type I errors without paying too large a penalty with respect to the statistical power for detecting real and substantively interesting differences. For most tests, the number of related tests was not so large that the Bonferroni test (Hochberg, 1988) exacted too large a penalty in power in exchange for protection from Type I error. For sets of comparisons with very large numbers of related tests, such as comparing a state to all other states, a new multiple comparison

criterion, False Discovery Rate or FDR (Benjamini & Hochberg, 1994), was implemented. FDR controls the *rate* of false rejections (e.g., five false rejections per 100 rejections), rather than controlling the probability of one such error (Familywise Error Rate, or FWE), as the Bonferroni procedure does.

The purpose of this chapter is to document the major conventions and statistical procedures used in generating the *Report Cards, NAEP 1996 Trends*, the *Data Compendia*, and the summary data tables. Additional details about procedures relevant to the *Report Card* and *Cross-State Data Compendia* can be found in the text and technical appendices of those reports.

18.2 MINIMUM SCHOOL AND STUDENT SAMPLE SIZES FOR REPORTING SUBGROUP RESULTS

In all of the reports, estimates of quantities such as composite and scale score means and percentages of students indicating particular levels of background variables (as measured in the student, teacher, and school questionnaires) are reported for the population of students in each grade. These estimates are also reported for certain key subgroups of interest as defined by primary NAEP reporting variables. Where possible, NAEP reports results for gender, for five racial/ethnic subgroups (White, Black, Hispanic, Asian American/Pacific Islander, and American Indian/Alaskan Native), three types of locations (central cities, urban fringes/large towns, rural/small town areas), four levels of parents' education (did not finish high school, high school graduate, some college, college graduate), Title 1 participation, eligibility for the free or reduced-cost school lunch component of the National School Lunch Program, and type of school. However, for some regions of the country and sometimes for the nation as a whole, school and/or student sample sizes were too small for one or more of the categories of these variables to permit accurate reporting.

A consideration in deciding whether to report an estimated quantity is whether the sampling error is too large to permit effective use of the estimates. A second, and equally important, consideration is whether the standard error estimate that accompanies a statistic is itself sufficiently accurate to inform potential readers about the reliability of the statistic. The precision of a sample estimate (be it sample mean or standard error estimate) for a population subgroup from a three-stage sample design (the one used to select samples for the national assessments) is a function of the sample size of the subgroup and of the distribution of that sample across first-stage sampling units (i.e., PSUs in the case of the national assessments). Hence, both of these factors were used in establishing minimum sample sizes for reporting.

Here a decision was reached to report subgroup results only if the student sample size exceeded 61.[5] A design effect of two was assumed for this decision, implying a sample design-based variance twice that of simple random sampling. This assumption is consistent with previous NAEP experience (Johnson & Rust, 1992). In carrying out the statistical power calculations when comparing a subgroup to the total group, it was assumed that the total population sample size is large enough to contribute negligibly to standard errors. Furthermore, it was required that the students within a subgroup be adequately distributed across PSUs to allow for reasonably accurate estimation of standard errors. In consultation with Westat, a decision was reached to publish only those statistics that had standard error estimates based on five or more degrees of freedom. The same minimum student and PSU sample size restrictions were applied to proportions and to comparisons of percentages or proportions as well as average scale scores and comparisons of average scale scores.

[5] This number was obtained by determining the sample size necessary to detect an effect size of 0.5 with a probability of 0.8 or greater.

18.3 IDENTIFYING ESTIMATES OF STANDARD ERRORS WITH LARGE MEAN SQUARED ERRORS

As noted above, standard errors of average scale scores, proportions, and percentiles play an important role in interpreting subgroup results and in comparing the performances of two or more subgroups. The jackknife standard errors reported by NAEP are statistics whose quality depends on certain features of the sample from which the estimate is obtained. In certain cases, the mean squared error[6] associated with the estimated standard errors may be quite large. This result typically occurred when the number of students upon which the standard error is based is small or when this group of students comes from a small number of participating PSUs. The minimum PSU and student sample sizes that were imposed in most instances suppressed statistics where such problems existed. However, the possibility remained that some statistics based on sample sizes that exceed the minimum requirements had standard errors that were not well estimated. Therefore, in the reports, estimated standard errors for published statistics that are themselves subject to large mean squared errors are followed by the symbol "!".

The magnitude of the mean squared error associated with an estimated standard error for the mean or proportion of a group depends on the coefficient of variation (*CV*) of the estimated size of the population group, denoted as $\hat{N}$ (Cochran, 1977, Section 6.3). The coefficient of variation is estimated by:

$$CV(\hat{N}) = \frac{SE(\hat{N})}{\hat{N}}$$

where $\hat{N}$ is a point estimate of *N* and $SE(\hat{N})$ is the jackknife standard error (described in Chapter 10 of this report) of $\hat{N}$.

Experience with previous NAEP assessments suggests that when this coefficient exceeds 0.2, the mean squared error of the estimated standard errors of means and proportions based on samples of this size may be quite large. (Further discussion of this issue can be found in Johnson & Rust, 1992.) Therefore, the standard errors of means and proportions for all subgroups for which the coefficient of variation of the population size exceeds 0.2 are marked as described above. In the *Report Cards, NAEP Trends*, the *Data Compendia*, and the summary data tables, statistical tests involving one or more quantities that have standard errors, confidence intervals, or significance tests so flagged should be interpreted with caution.

18.4 TREATMENT OF MISSING DATA FROM THE STUDENT, TEACHER, AND SCHOOL QUESTIONNAIRES

As previously described, responses to the student, teacher, and school questionnaires played a prominent role in all reports. Although the return rate on all three types of questionnaire was high,[7] there were missing data for each type of questionnaire.

---

[6] The mean squared error of the estimated standard error is defined as $\mathscr{E}[\hat{S} - \sigma]^2$, where $\hat{S}$ is the estimated standard error, σ is the "true" standard error, and $\mathscr{E}$ is the expectation, or expected value operator.

[7] Information about survey participation rates (both school and student), as well as proportions of students excluded by each jurisdiction from the assessment, is given in Appendix B. Sampling adjustments intended to account for school and student nonresponse are described in Chapter 7.

The reported estimated percentages of students in the various categories of background variables, and the estimates of the average scale score of such groups, were based on only those students for whom data on the background variable were available. In the terminology of Little and Rubin (1987), the analyses pertaining to a particular background variable presented in the reports are contingent on the assumption that the data are missing completely at random.[8]

The estimates of proportions and proficiencies based on "missing-completely-at-random" assumptions are subject to potential nonresponse bias if, as may be the case, the assumptions are not correct. The amount of missing data was small (usually, less than 2%) for most of the variables obtained from the student, school, and teacher questionnaires. For analyses based on these variables, reported results are subject to little, if any, nonresponse bias. However, for particular background items from the student, school, and teacher questionnaires, the level of nonresponse was somewhat higher. As a result, the potential for nonresponse bias in the results of analyses based on this latter set of background items is also somewhat greater. Background items for which more than 10 percent of the returned questionnaires were missing are identified in the questionnaire sections (as specified at the beginning of this chapter) of the summary data tables. Again, results for analyses involving these background variables should be interpreted with caution.

To analyze the relationships among teachers' questionnaire responses and their students' achievement, each teacher's questionnaire had to be matched to the students who were taught by that teacher. If a student could not be matched to a teacher, all teacher questionnaire responses are missing for that student. The percentages of students that were matched to teacher questionnaires in each sample for which a teacher questionnaire was administered are reported in the subject area Chapters 12 and 13. Lower percentages of students with teacher questionnaire data indicate that there is less certainty about results for variables from the teacher questionnaire. Note that these match rates do not reflect the additional missing data due to item-level nonresponse. The amount of additional item-level nonresponse in the returned teacher questionnaires can be found in the summary data tables.

18.5 HYPOTHESIS TESTING CONVENTIONS

18.5.1 Comparing Means and Proportions for Different Groups of Students

Many of the group comparisons explicitly commented on in the reports involved mutually exclusive sets of students. Examples include comparisons of the average scale score for male and female students, White and Hispanic students, students attending schools in central city and urban fringe/large town locations, students who reported watching six or more hours of television each night and students who report watching less than one hour each night.

The text in the reports indicate that means or proportions from two groups were different only when the difference in the point estimates for the groups being compared was statistically significant at an approximate simultaneous α level of 0.05. An approximate procedure was used for determining statistical significance NAEP staff judged to be statistically defensible, as well as being computationally tractable. Although all pairs of levels within a variable were tested and reported in the summary data tables, some text within the reports was developed for only a subset of these comparisons although the family size was maintained at that of the original tests. For example, text was included in the reports to compare the majority ethnic group and each minority group, but text for all possible comparisons of groups may not have been included.

[8] The mechanism generating the missing data is independent of both the response to the particular background items and the scale score.

The procedure used to make statistical tests is described in the following paragraphs. This procedure was used in all cases except when comparisons were made with students assessed in assessment years for which average scale scores were extrapolated as part of the long-term trend analyses. In those cases, z-tests comparing the test statistics to the appropriate value from the standard normal distribution was used.

Let A_i be the statistic in question (e.g., a mean for group *i*) and let S_{A_i} be the jackknife standard error of the statistic. The text in the reports identified the means or proportions for groups *i* and *j* as being different if:

$$\frac{|A_i - A_j|}{\sqrt{S^2_{A_i}(A_i) + S^2_{A_j}(A_j)}} \geq T_{\frac{.05}{2c}}$$

where T_α is the (1 - α) percentile of the *t* distribution with degrees of freedom, *df*, as estimated below, and *c* is the number of related comparisons being tested. See the following section (Section 18.5.2) for a more specific description of multiple comparisons. In cases where group comparisons were treated as individual units, the value of *c* was taken as 1, and the test statistic was approximately equivalent to a standard two-tailed t-test for the difference between group means or proportions from large independent samples with the α level set at 0.05. When *c* is not 1, this test is based on the Bonferroni procedure described in Hochberg (1988). The degrees of freedom of this t-test is defined by a Satterthwaite (Johnson & Rust, 1992) approximation as follows:

$$df = \frac{(\sum_{k=1}^{N} S^2_{A_k})^2}{\sum_{k=1}^{N} \frac{S^4_{A_k}}{df_{A_k}}}$$

where N is the number of subgroups involved, and df_{A_k} is as follows:

$$df_{A_k} = \left(3.16 - \frac{2.77}{\sqrt{m}}\right) \left[\frac{\left(\sum_{j=1}^{m} (t_{j_k} - t_k)^2\right)^2}{\sum_{j=1}^{m} (t_{j_k} - t_k)^4}\right]$$

where *m* is the number of replicates, t_j is the j^{th} replicated estimate for the mean of a subgroup, and *t* is the estimate of the subgroup mean using the overall weights and the first plausible value.

The procedures in this section assume that the data being compared are from independent samples. Because of the sampling design in which PSUs, schools, and students within school are randomly sampled, the data from mutually exclusive sets of students may not be strictly independent. Therefore, the significance tests employed are, in many cases, only approximate. As described in Section

10.4, another procedure, one that does not assume independence, could have been conducted. However, that procedure is computationally burdensome. A comparison of the standard errors using the independence assumption and the correlated group assumption was made using NAEP data. The estimated standard error of the difference based on independence assumptions was approximately ten percent larger than the more complicated estimate based on correlated groups. In almost every case, the correlation of NAEP data across groups was positive. Because, in NAEP, significance tests based on assumptions of independent samples are only somewhat conservative, the approximate procedure was used for most comparisons.

The procedures described above were used for testing differences of both means *and* nonextreme percentages. The approximation for the test for percentages works best when sample sizes are large, and the percentages being tested have magnitude relatively close to 50 percent. Statements about group differences should be interpreted with caution if at least one of the groups being compared is small in size and/or if "extreme" percentages are being compared. Differences in percentages were treated as involving "extreme" percentages if for either percentage, *P*:

$P < P_{lim} = \frac{200}{N_{EFF} + 2}$, where the effective sample size is $N_{EFF} = \frac{P(100-P)}{(SE_{JK})^2}$, and SE_{JK} is the jackknife standard error of *P*. Similarly, at the other end of the 0 - 100 scale, a percentage is deemed extreme if 100 - $P < P_{lim}$. In either extreme case, the normal approximation to the distribution is a poor approximation, and the value of *P* was reported, but no standard error was estimated and hence no significance tests were conducted.

18.5.2 Multiple Comparison Procedures

Frequently, groups (or families) of comparisons were made and were presented as a single set. The appropriate text, usually a set of sentences or a paragraph, was selected for inclusion in a report based on the results for the entire set of comparisons. For example, some reports contain a section that compared average scale scores for a predetermined group, generally the majority group (in the case of race/ethnicity, for example, White students) to those obtained by other minority groups. The entire set of tests was presented in the summary data tables. For families of comparisons like these, a Bonferroni procedure (Miller, 1966), controlling the Familywise Error Rate (FWE), was used. This procedure defines the value of T_{α}, as in the previous section, where *c* is the number of contrasts in the set. In the race/ethnicity example, *c* was taken to be the number of minority groups meeting minimum sample size requirements, and each statistical test was consequently carried out at an α level of 0.05/*c*.

However, in an attempt to gain greater power, two separate definitions of family size were employed for comparisons in two-way tables. For *n* levels of a control variable (e.g., ethnicity) and *m* levels of a comparison variable (e.g., number of hours of homework), the standard Bonferroni family size of $n \times m \times (m\text{-}1)/2$ was used. In addition, when the $m \times (m\text{-}1)/2$ marginal tests yielded a significant difference for a pair of categories of the comparison variable, the *n* levels of the control variable corresponding to that pair of categories were tested with a family size of *n*. Significance was reported if either definition of family size met the criterion.

Further, in the *Report Card* and summary data tables, two-way interactions were tested directly for some variables. The tests for an $m \times n$ table were t-tests using a family size $n \times (n\text{-}1) \times m \times (m\text{-}1)/4$. In these cases, a modification due to Hochberg of the standard Bonferroni procedure was employed, in which probabilities associated with outcomes are ordered, and α is divided by an integer which increases

from 1 to the family size as successively smaller probabilities are tested. More formally, the Hochberg Stagewise Procedure (Hochberg, 1988) is defined as follows:

Let q be the number of significance tests made (the family size) and let $P_1 \leq P_2 \leq ... \leq P_m$ be the ordered significance levels for the q tests. Let α be the combined significance level. The Hochberg procedure compares P_q with α, P_{q-1} with $\alpha/2$, ..., P_j with $\alpha/(q-j+1)$, stopping comparisons with the first j such that $P_j < \alpha/(q-j+1)$. All tests associated with P_1, ..., P_j are declared significant; all tests associated with P_{j+1}, ..., P_q are declared nonsignificant.

To compare a jurisdiction in a State Assessment with the nation and all other participating jurisdictions, as many as 46 different comparisons need to be computed. This is done in the comparisons of overall scale score maps in the State Assessment reports and in the comparisons of short-term trends in mathematics achievement in the *Mathematics Report Card*. A potentially more powerful multiple comparison procedure was used to judge significance in this case. The procedure, described by Benjamini and Hochberg (1994), was the procedure chosen. Unlike the Bonferroni procedure that controls the FWE, the procedure described by Benjamini and Hochberg (1994) controls the expected proportion of falsely rejected hypotheses among all rejections (FDR). For example, at the 0.05 level, for every 100 rejections of the null hypothesis, the procedure ensures that no more than five will be expected to be false. Note that control of the FDR is a less conservative type of error control than that of the FWE. Simulations have shown that "the FDR is controlled at level α for the dependent tests involved in pairwise comparisons as well as for independent tests" (Shaffer, 1994).

The Benjamini and Hochberg application of the False Discovery Rate (FDR) criterion can be described as follows. Let q be the number of significance tests made and let $P_1 \leq P_2 \leq . . . \leq P_q$ be the ordered significance levels of the q tests, from lowest to highest probability. Let α be the combined significance level desired, usually 0.05. The procedure will compare P_q with α, P_{q-1} with $\alpha(q\text{-}1)/q$, . . ., P_j with $\alpha j/q$, stopping the comparisons with the first j such that $P_j \leq \alpha j/q$. All tests associated with P_1, . . ., P_j are declared significant; all tests associated with P_{j+1}, . . ., P_q are declared nonsignificant.

18.5.3 Linear and Quadratic Tests of Trends

Tests of significance designed to identify consistent patterns of trend data are available and, although they are more complex, they provide more power to identify those specific patterns than a series of t- or z-tests would provide.

One such set of tests of significance is the test of linear and test of quadratic trends applied to the long-term trend data for the nation and selected subpopulations. The purpose of this first set of general tests was to determine whether the results of the series of assessments in a given subject could be generally characterized as increasing or decreasing, and whether the results have steadily increased (or decreased) over the time period of interest. Simple curvilinear (i.e., quadratic) relationships capture more complex patterns. For example, one possible pattern is to have initial score declines over part of the time period followed by score increases in more recent assessments. Another possible pattern is to have a sequence of several assessments in which scores increased followed by a period of relative stable performance. These examples are two, but not all, of the simple curvilinear relationships that were tested.

The linear and quadratic components of the trend in average scale scores for a given subject area and age group were estimated by applying two sets of contrasts to the set of average scale scores by year. The linear component of the trend was estimated by the sum $b_1 = \Sigma c_j x_j$, where the x_j are the average scale

scores by year and the c_j are defined such that b_1 corresponds to the slope of an unweighted regression of the average scale scores on the assessment year. In other words,

$$c_j = \frac{y_j - \frac{1}{N}\sum_i y_i}{\sum_k (y_k - \frac{1}{N}\sum_i y_i)^2}$$

where y_j represents an assessment year. The quadratic component was estimated by the sum $b_2 = \Sigma d_j x_j$ in which the d_j are formally orthogonal to the c_j and are defined such that b_2 is the quadratic term in the unweighted regression of the average scale scores on the assessment year and the square of the assessment year. In other words,

$$d_j = \frac{h_j}{\sum_i h_i^2}$$

where

$$h_j = (y_j{}^2 - \frac{1}{N}\sum_i y_i{}^2) - \left[\sum_k c_k (y_k{}^2 - \frac{1}{N}\sum_i y_i{}^2)\right] \bullet (y_j - \frac{1}{N}\sum_i y_i)$$

Both c_j and d_j match expected linear quadratic contrasts in common texts *when the years are equally spaced through time* (Winer, 1962/1971). The statistical significance of b_1 and b_2 was evaluated by comparing each estimate to its estimated standard error. The standard error of b_1 was estimated as the square root of the sum $\sum c_j^2 SE_j^2$, in which SE_j is the estimated standard error of x_j. The estimated standard error of the b_2 was analogously defined.

The linear and quadratic trend tests allow statements to be made about results across assessment years in a more powerful way than is possible if results for each year had been compared to those of every other year, using a multiple-comparison procedure such as the Bonferroni method. These tests do not control the overall Type I error rate when they are applied to several related subgroups, such as the students in each region of the country. For this reason, the Bonferroni method for controlling Type I error was used when the trends for related subgroups were tested. For example, when tests were conducted for linear trend for the separate race/ethnicity groups (i.e., White, Black, and Hispanic), these tests were treated as a single family of comparisons of size 3. The significance level for each of the separate tests was adjusted by the Bonferroni procedure to yield a family-wise error rate of .05.

18.5.4 Comparing Proportions Within a Group

Certain analyses involved the comparison of proportions. One example was the comparison of the proportion of students who reported that a parent graduated from college to the proportion of students who indicated that their parents did not finish high school to determine which proportion was larger. There are other such proportions of interest in this example, such as the proportion of students with at least one parent graduating from high school but neither parent graduating from college. For these types of analyses, NAEP staff determined that the dependencies in the data could not be ignored.

Unlike the case for analyses of the type described in Section 18.5.1, the correlation between the proportion of students reporting a parent graduated from college and the proportion reporting that their parents did not finish high school is likely to be negative and large. For a particular sample of students, it is likely that the higher the proportion of students reporting “at least one parent graduated from college” is, the lower the proportion of students reporting “neither parent graduated from high school” will be. A negative dependence will result in underestimates of the standard error if the estimation is based on independence assumptions (as is the case for the procedures described in Section 18.5.1). Such underestimation can result in an unacceptably large number of “nonsignificant” differences being identified as significant.

The procedures of Section 18.5.1 were modified for analyses that involved comparisons of proportions within a group. The modification involved using a jackknife method for obtaining the standard error of the difference in dependent proportions. The standard error of the difference in proportions was obtained by first obtaining a separate estimate of the difference in question for each jackknife replicate, using the first plausible value only, then taking the standard deviation of the set of replicate estimates as the estimate. The procedures used for proportions within a group differed from the procedures of Section 18.5.1 only with respect to estimating the standard error of the difference; all other aspects of the procedures were identical.

Chapter 19

STATISTICAL SUMMARY OF THE 1996 NAEP SAMPLES[1]

Bruce A. Kaplan
Educational Testing Service

19.1 INTRODUCTION

The analysis of the 1996 NAEP data has resulted in the production of thousands of tables presenting estimates of the proficiency of students, and various subgroups of students, in American schools. This chapter provides a statistical summary of the 1996 NAEP national samples. The chapter assumes a general familiarity with the structure of NAEP as summarized in the Introduction and in the overviews presented in Chapters 1 and 9. Similar results for the state samples appear in the data compendia for the state mathematics and science assessments.

Two of the many types of NAEP results are presented here:

1. the results of the instrument development process, including the sizes of the item pools and numbers of booklets; and

2. the results of the sampling process, including the numbers of students in each sample by selected subgroups.

19.2 MEASUREMENT INSTRUMENTS

For the 1996 assessment, 79 different assessment booklets and questionnaires were printed for age class 9, 80 for age class 13, and 81 for age class 17. These instruments are shown by age level and type in Table 19-1.

The item pool contributing to all main and long-term trend booklets is described in Table 19-2. In general, there are two types of items, cognitive and noncognitive. The cognitive items are developed to measure proficiency in particular subject areas, such as reading and mathematics. Cognitive items may be constructed-response or multiple-choice. The noncognitive items are usually questions about the student's or teacher's backgrounds and perceptions but may also probe other areas, such as school policies or teaching methods. Because many items were used at more than one age class, the total number of items in an item pool is not the sum of the item pools used for the three age classes. However, results for cognitive items that were common across two or three age classes were not compared, due to a NAEP policy of within grade scaling.

The SD/LEP Student Questionnaires, Teacher Questionnaires, and School Characteristics and Policies Questionnaires contained only noncognitive questions. The number of items in the noncognitive

[1] Bruce A. Kaplan was responsible for the text, specifying the tables, and coordinating table production. Shuyi Hua produced most of the tables in this chapter. David Freund's advice was invaluable in the production of this chapter.

pools is the same as the number of items in the questionnaires. More information about the instruments that were developed is provided in Chapters 2 and 4.

19.3 SAMPLE CHARACTERISTICS

In this section, the characteristics of the final reporting NAEP samples are described. The process by which the samples were selected is discussed in Chapter 3.

In the 1996 main assessment, NAEP contacted 2,267 schools (2,263 original and 4 replacements), of which 1,791 contributed data to the assessment. The disposition of these schools is shown in Table 19-3. Some of the schools were unwilling to cooperate; others were believed to be eligible from the sampling frame, but were not. The cooperation rate is calculated as the sum of cooperating schools and the schools that were found to have no eligible students divided by the same sum plus the schools that refused or were from districts that refused to cooperate.

Table 19-3 also shows the number of schools in several categories: region of the country (Northeast, Southeast, Central, West), school type (public, nonpublic, Catholic, Bureau of Indian Affairs, Department of Defense Education Activity), type of location, number of teachers, and number of students.

For the 1996 long-term trend studies, NAEP contacted 856 schools (844 original and 12 replacements), of which 681 contributed data to the various trend assessments. Table 19-4 supplies the same information for the schools assessed for the long-term trend studies that Table 19-3 supplies for the main assessment schools.

The numbers of respondents to the teacher questionnaires are summarized in Table 19-5. The first column in this table includes the number of teachers who responded, by grade and subject area. The second column is the number of students who were not linked to teachers. The third column is the number of students linked to teachers, but not specific classes of these teachers (for eighth grade) or teachers who did not answer classroom information (for fourth grade). The last column is the number of students linked to their teachers and their specific classes.

NAEP is administered in units called assessment sessions. If the number of students attending an assessment session is fewer than a predetermined number, the students missing from the session are assigned to a makeup session and then assessed. Table 19-6 shows the number of regular and makeup sessions in 1996 NAEP by age class for the main and long-term trend samples. Altogether, 103,814 assessed and excluded students were involved in the 1996 NAEP.

Tables 19-7 through 19-9 display the distribution of the students assessed in the cross-sectional NAEP assessment in several basic categories: gender, racial/ethnic grouping, region of the country, parental education, type of location, school type, and modal age. These data are presented for assessed students in the mathematics main and estimation samples in Table 19-7, the mathematics theme and advanced samples in Table 19-8, and the science main and advanced samples in Table 19-9. Tables 19-10, 19-11, and 19-12 provide equivalent information, respectively, for excluded students.

Tables 19-13, 19-14, and 19-15 display the distribution of students assessed in the long-term trend reading and writing assessment for several basic categories: gender, racial/ethnic grouping, region of the country, parental education, type of location, and school type.

There is one table for each age/grade. The tables have four columns:

- eligible by age, which means that the students were in an appropriate age group;
- eligible by grade, which means that the students were in an appropriate grade;
- eligible by age and by grade, which means that the students were of both an appropriate age and appropriate grade; and
- eligible by age or by grade, which is the total number of students for whom data were collected.

Tables 19-16, 19-17, and 19-18 provide similar information for the long-term trend science and mathematics assessment. Note that since these are age-only samples, the number of students who are age-eligible only will be the same as the number of students who are age- *or* grade-eligible. Likewise, the number of students who are grade-eligible only will be the same as the number of students who are both age- *and* grade-eligible. Tables 19-19 through 19-24 enumerate the excluded students across the various long-term trend samples.

19.4 POPULATION ESTIMATES

The 1996 NAEP samples were designed for estimating the size and attributes of a number of different populations of students. The estimation procedures use sampling weights, developed by Westat, Inc., that are associated with the members of the sample (see Chapter 3). In this chapter, all estimates of population parameters are calculated using these sampling weights. Note these estimates are for the reporting samples (see Chapter 3 for an explanation of the reporting and modular samples).

The sum of the initial weights for a given sample is an estimate of the number of students who are in the population represented by the sample. In other words, the sum of the initial weights is taken as the estimated population size. In analyses, however, this sum of weights was rescaled to sum to the sample size. For example, in Table 19-25, the estimated number of fourth graders in the nation is 3,711,786, as estimated from the main mathematics sample, as opposed to the 6,627 students in the sample given in Table 19-7.

Due to design considerations the main assessment was divided into subsamples, and were administered, and therefore weighted, independently, so that the sum of the initial weights for each subsample estimates the population size. The subsamples for mathematics were main, estimation, theme and advanced; for science, the subsamples were main and advanced.

Note that the samples for the main (cross-sectional) assessment are grade-only samples, while reading and writing long-term trend are grade and age samples. The samples for the mathematics and science long-term trend are age-only samples. The sum of the initial weights of the excluded students estimates the number of ineligible students at the respective age/grade levels.

In most cases, the number of students in an age/grade combination is not of interest; a researcher will be interested in estimating the number of students at either a grade or an age level. For the samples that contain both grade- and age-eligible students, an estimate of the total number of students at an age level can be made by summing the initial weights of only the age-eligible students and adding the

corresponding sample of age-eligible excluded students' initial weights. An estimate of the total number of students in a grade sample can be made by summing the initial weights of grade-eligible students plus the initial weights of grade-eligible students from the appropriate excluded student sample.

Tables 19-25 to 19-42 show the sizes of the estimated populations of assessable students and the weighted percentages for the NAEP reporting categories of gender, race/ethnicity, region of the country, parents' education level, type of location, school type and modal age. The estimated subpopulation percentages for the cross-sectional samples are shown in Tables 19-25 through 19-30. Tables 19-31 to 19-36 show the same information for the long-term trend samples. In a similar manner, Tables 19-37 to 19-42 show the estimated total population of excluded students and the weighted percentages by demographic subgroups (data about parents' education level is not collected for excluded students and therefore not reported; data about reasons for exclusion are included instead).

In previous years, this chapter also provided several tables showing selected proficiency results for assessed students, as an aid to readers who are interested in the estimates of proficiency that led to the interpretive results provided in the NAEP subject area reports. These tables are no longer included in this report. Instead, readers are encouraged to take advantage of the electronic version of these results, in the form of thousands of summary data tables computed to analyze the 1996 data. The summary data tables are available both on CD-ROM and via the World-Wide Web at http://nces.ed.gov/naep.

Table 19-1
Measurement Instruments Developed for 1996 NAEP

Student Assessment Booklets	Age Class 9	13	17
Total Number of Cross-Sectional (MAIN)	66[1]	67[1]	70
Mathematics	29[1]	30[1]	30
Main	26[1]	26[1]	26
Estimation	1	1	1
Theme	2	2	2
Advanced	—[2]	1	1
Science	37	37	40
Main	37	37	37
Advanced	—[3]	—[3]	3
Total Number of Long-Term Trend	9	9	8
Reading and Writing	6	6	6
Mathematics and Science	3	3	2
Total Number of Questionnaires	4	4	3
Excluded Students (Long-Term Trend only)	1	1	1
SD/LEP (Cross-Sectional (main) only)	1	1	1
Teacher	1	1	0
School	1	1	1
Total Number of Assessment Instruments	79[1]	80[1]	81

[1] A bilingual book was also used, but not counted as a separate book for this table.
[2] No advanced mathematics booklets were administered to age/class 9.
[3] No advanced science booklets were administered to age/class 9 or 13.

Table 19-2

Number of Items Administered, by Sample and Age Class

	Age Class			
	9	**13**	**17**	**Distinct Items**
COMMON BACKGROUND				
Cross-Sectional (Main Math)	24	26	36	42
Cross-Sectional (Main Science)	24	26	36	45
Reading and Writing Long-Term Trend	37	34	48	48
Math and Science Long-Term Trend	28	30	48	58
MATH MAIN				
Background	25	31	44	56
Cognitive — Main	144	178	183	358
Cognitive — Estimation	31	32	38	76
Cognitive — Theme	14	22	18	45
Cognitive — Advanced	0	22	22	44
Motivation	5	5	5	5
SCIENCE MAIN				
Background	39	42	53	68
Cognitive — Main	141	194	190	439
Cognitive — Advanced	0	0	66	66
Motivation	5	5	5	5
LONG-TERM TREND				
Reading Background	40	42	78	81
Reading Cognitive	105	108	96	193
Writing Background	53	65	65	77
Writing Cognitive	6	6	6	12
Mathematics Background	3	29	39	49
Mathematics Cognitive	68	96	94	184
Science Background	16	29	29	45
Science Cognitive	63	83	82	180
SD/LEP STUDENT QUESTIONNAIRE	114	114	114	58
MATH TEACHER QUESTIONNAIRE				
Teacher Background	77	0	59	79
Math Background	17	17	22	36
Math Classroom	40	49	57	96
SCIENCE TEACHER QUESTIONNAIRE				
Teacher Background	77	59	0	79
Science Background	14	13	0	14
Science Classroom	59	59	0	59
SCHOOL QUESTIONNAIRE	100	105	127	196

Table 19-3

School Characteristics in Main Samples (All Samples)

	Grade 4	Grade 8	Grade 12	Total
TOTAL ORIGINAL SAMPLE	723	761	779	2,263
Cooperating	604	592	591	1,787
No Eligibles Enrolled	20	42	28	90
School Refused	99	127	160	386
COOPERATION RATE	86	83	79	83
COOPERATING REPLACEMENTS FOR REFUSALS	1	1	2	4
TOTALS				
Cooperating Schools	605	593[1]	593[1]	1,791
Completed Questionnaires	605	594[1]	595[1]	1,794
REGION				
Northeast	130	126	123	379
Southeast	134	133	151	418
Central	165	163	145	473
West	176	170	174	520

[1] Occasionally schools with a completed questionnaire had no eligible students, so they were not included as participating cooperating schools.

Table 19-3 (continued)
School Characteristics in Main Samples (All Samples)

	Grade 4	Grade 8	Grade 12	Total
SCHOOL TYPE				
Public	387	335	428	1,150
Nonpublic	200	243	151	594
Private	81	91	90	262
Catholic	119	152	61	332
BIA	0	0	1	1
DoDea	0	0	0	
NUMBER OF TEACHERS				
Unclassified	0	0	0	0
1-4	15	9	4	28
5-9	84	61	18	163
10-19	141	149	52	342
20-49	311	244	202	757
50-74	30	89	122	241
75-99	5	20	69	94
100+	1	8	113	122
Missing	18	14	15	47
NUMBER OF STUDENTS				
Unclassified	0	0	0	0
1-99	34	18	22	74
100-299	176	184	103	463
300-499	173	113	69	355
500-749	133	105	89	327
750-999	53	86	55	194
1000-1499	12	51	89	152
1500+	6	23	153	182
Missing	18	14	15	47

Table 19-4
School Characteristics in Long-Term Trend Samples

	Age Class			
	9	13	17	TOTAL
TOTAL ORIGINAL SAMPLE	291	316	237	844
Cooperating	240	238	191	669
No Eligibles Enrolled	8	27	2	37
School Refused	43	51	44	138
COOPERATION RATE	85	84	81	81
COOPERATING REPLACEMENTS FOR REFUSALS	8	4	0	12
TOTALS				
Cooperating Schools	248	242	191	681
Completed Questionnaires	248	242	191	681
REGION				
Northeast	51	54	36	141
Southeast	62	64	56	182
Central	59	54	42	155
West	76	70	57	203

Table 19-5

Numbers of Responses to Teacher Questionnaires and Students Matched with Teacher Data

	Number of Teachers Responding	Number of Students with No Match	Number of Students with Partial Match	Number of Students with Complete Match
MATH				
GRADE 4				
Main	752	408	99	6,105
Estimation	320	154	11	1,841
Theme	608	351	34	3,405
Advanced	0	0	0	0
GRADE 8				
Main	607	953	49	6,144
Estimation	242	274	8	1,901
Theme	437	603	51	3,373
Advanced	330	343	9	1,985
GRADE 12				
Main	0	0	0	0
Estimation	0	0	0	0
Theme	0	0	0	0
Advanced	404	393	241	2,331
SCIENCE				
GRADE 4				
Main	535	646	159	6,500
Advanced	0	0	0	0
GRADE 8				
Main	371	1,258	112	6,404
Advanced	0	0	0	0
GRADE 12				
Main	0	0	0	0
Advanced	0	0	0	0

Table 19-6

Number of Students Assessed and Excluded by Sample and Age Class

	Age Class		
	9	13	17
ASSESSED STUDENTS	30,178	34,618	33,629
Cross-Sectional	19,745	23,467	25,421
Math	12440	15,693	15,453
Main	6,627	7,146	6,904
Estimation	2,023	2,183	1,849
Theme	3,790	4,027	3,735
Advanced	0	2,337	2,965
Science	7305	7,774	9,968
Main	7305	7,774	7,537
Advanced	0	0	2,431
Long-Term Trend	10,433	11,151	8,208
Reading and Writing	5,019	5,493	4,669
Math and Science	5,414	5,658	3,539
EXCLUDED STUDENTS	2,256	1,698	1,435
Cross-Sectional	1,139	765	722
Math	383	339	297
Main	204	166	116
Estimation	43	56	75
Theme	136	113	99
Advanced	0	4	7
Science	756	426	425
Main	756	426	425
Advanced	0	0	0
Long-Term Trend	1,117	933	713
Reading and Writing	532	481	412
Math and Science	585	452	301

Table 19-7

Number of Students in the Mathematics Main and Estimation Samples by Subgroup Classification, Grades 4, 8, and 12

	MAIN			ESTIMATION		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4**	**Grade 8**	**Grade 12**
TOTAL	6,627	7,146	6,904	2,023	2,183	1,849
GENDER						
Male	3,290	3,597	3,244	994	1,052	898
Female	3,337	3,549	3,660	1,029	1,131	951
RACE/ETHNICITY						
White	4,125	4,501	4,596	1,193	1,407	1,258
Black	1,106	1,193	1,106	348	370	273
Hispanic	974	911	732	328	247	229
Asian American	250	408	339	98	124	72
American Indian	149	110	115	53	26	10
Unclassified	23	23	16	3	9	7
REGION						
Northeast	1,414	1,312	1,414	471	489	297
Southeast	1,669	1,883	1,924	540	520	509
Central	1,606	1,726	1,675	396	549	470
West	1,938	2,225	1,891	616	625	573
PARENT'S EDUCATION						
Less Than High School	219	466	462	77	113	125
High School	837	1,503	1,300	227	438	305
Greater Than High School	462	1,310	1,741	146	361	390
Graduated College	2,804	3,112	3,177	852	994	985
Unknown	2,232	736	200	681	247	39
TYPE OF LOCATION						
Central City	2,380	3,218	2,555	859	988	823
Urban Fringe/Large Town	2,794	2,186	2,428	721	698	618
Rural/Small Town	1,453	1,742	1,921	443	497	408
SCHOOL TYPE						
Public	5,215	5,590	5,398	1,528	1,707	1,340
Nonpublic	1,412	1,556	1,455	495	476	509
Private	458	576	521	164	117	200
Catholic	954	980	934	331	359	309
BIA	0	0	51	0	0	0
DoDEA	0	0	0	0	0	0
MODAL AGE						
Younger	35	48	92	12	4	21
At Modal Age	4,197	4,380	4,441	1,335	1,333	1,194
Older	2,395	2,718	2,371	676	846	634

Table 19-8

Number of Students in the Mathematics Theme and Advanced Samples by Subgroup Classification, Grades 4, 8 and 12

	THEME			ADVANCED		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4***	**Grade 8**	**Grade 12**
TOTAL	3,790	4,027	3,735	0	2,337	2,971
GENDER						
Male	1,905	2,030	1,797	0	1,130	1,532
Female	1,885	1,997	1,938	0	1,207	1,439
RACE/ETHNICITY						
White	2,206	2,440	2,279	0	1,650	2,001
Black	655	731	695	0	280	319
Hispanic	672	641	497	0	216	327
Asian American	169	140	228	0	149	306
American Indian	84	65	26	0	33	12
Unclassified	4	10	10	0	9	6
REGION						
Northeast	723	608	851	0	413	638
Southeast	1,037	1,125	1,025	0	552	800
Central	887	937	742	0	626	666
West	1,143	1,357	1,117	0	746	867
PARENT'S EDUCATION						
Less Than High School	162	320	296	0	84	129
High School	498	922	746	0	308	398
Greater Than High School	277	732	980	0	445	664
Graduated College	1,494	1,648	1,578	0	1,352	1,709
Unknown	1,354	396	125	0	121	54
TYPE OF LOCATION						
Central City	1,721	1,495	1,452	0	968	1,106
Urban Fringe/Large Town	1,289	1,798	1,339	0	902	1,106
Rural/Small Town	780	734	944	0	467	759
SCHOOL TYPE						
Public	3,034	3,438	3,075	0	1,661	2,130
Nonpublic	756	589	660	0	676	841
Private	299	219	235	0	234	346
Catholic	457	370	425	0	442	495
BIA	0	0	0	0	0	0
DoDEA	0	0	0	0	0	0
MODAL AGE						
Younger	16	21	63	0	18	51
At Modal Age	2,467	2,417	2,427	0	1,594	2,113
Older	1,307	1,589	1,245	0	725	807

*Advanced students not sampled for Grade 4.

Table 19-9

Number of Students in the Science Main and Advanced Samples by Subgroup Classification, Grades 4, 8, and 12

	MAIN			ADVANCED		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4***	**Grade 8***	**Grade 12**
TOTAL	7,305	7,774	7,537	0	0	2,431
GENDER						
Male	3,651	3,872	3,547	0	0	1,167
Female	3,654	3,902	3,990	0	0	1,264
RACE/ETHNICITY						
White	4,106	4,292	4,748	0	0	1,714
Black	1,251	1,492	1,225	0	0	293
Hispanic	1,352	1,426	1,015	0	0	197
Asian American	356	382	458	0	0	209
American Indian	223	149	70	0	0	12
Unclassified	17	33	21	0	0	6
REGION						
Northeast	1,503	1,068	1,562	0	0	541
Southeast	1,843	2,246	2,148	0	0	695
Central	1,699	1,595	1,589	0	0	634
West	2,260	2,865	2,238	0	0	561
PARENT'S EDUCATION						
Less Than High School	271	553	606	0	0	87
High School	938	1,471	1,414	0	0	272
Greater Than High School	544	1,428	1,879	0	0	526
Graduated College	2,994	3,400	3,308	0	0	1,476
Unknown	2,433	774	211	0	0	0
TYPE OF LOCATION						
Central City	3,228	3,055	3,080	0	0	949
Urban Fringe/Large Town	2,769	2,963	2,488	0	0	895
Rural/Small Town	1,308	1,756	1,969	0	0	587
SCHOOL TYPE						
Public	5,814	6,376	6,112	0	0	1,739
Nonpublic	1,491	1,398	1,425	0	0	692
Private	499	597	499	0	0	185
Catholic	992	801	926	0	0	507
BIA	0	0	0	0	0	0
DODEA	0	0	0	0	0	0
MODAL AGE						
Younger	46	46	93	0	0	38
At Modal Age	4,739	4,553	4,802	0	0	1,720
Older	2,520	3,175	2,642	0	0	673

*Advanced students not sampled for Grade 4 and Grade 8.

Table 19-10

Number of Excluded Students in the Mathematics Main and Estimation Samples by Subgroup Classification, Grades 4, 8, and 12

	MAIN			ESTIMATION		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4**	**Grade 8**	**Grade 12**
TOTAL	204	166	116	43	56	75
GENDER						
Male	122	104	72	33	34	44
Female	82	62	44	10	22	31
RACE/ETHNICITY						
White	92	100	65	24	33	32
Black	34	30	22	7	8	16
Hispanic	66	18	25	9	12	23
Asian American	8	10	3	2	3	4
American Indian	3	3	1	1	0	0
Unclassified	1	5	0	0	0	0
REGION						
Northeast	21	45	22	5	19	8
Southeast	49	34	27	9	11	9
Central	29	36	18	14	11	9
West	105	51	49	15	15	49
PARENT'S EDUCATION						
Less Than High School	0	0	0	0	0	0
High School	0	0	0	0	0	0
Greater Than High School	0	0	0	0	0	0
Graduated College	0	0	0	0	0	0
Unknown	0	0	0	0	0	0
TYPE OF LOCATION						
Central City	82	64	40	10	29	36
Urban Fringe/Large Town	61	50	44	21	15	13
Rural/Small Town	61	52	32	12	12	26
SCHOOL TYPE						
Public	197	162	115	43	56	75
Nonpublic	7	4	1	0	0	0
Private	1	0	0	0	0	0
Catholic	6	4	1	0	0	0
BIA	0	0	0	0	0	0
DoDEA	0	0	0	0	0	0
MODAL AGE						
Younger	0	3	1	0	0	0
At Modal Age	106	48	31	22	21	17
Older	98	115	84	21	35	58

Table 19-11

Number of Excluded Students in the Mathematics Theme and Advanced Sample by Subgroup Classification, Grades 4, 8, and 12

	THEME			ADVANCED		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4***	**Grade 8**	**Grade 12**
TOTAL	136	113	99	0	4	7
GENDER						
Male	71	59	55	0	4	6
Female	65	54	44	0	0	1
RACE/ETHNICITY						
White	37	34	55	0	2	2
Black	26	30	17	0	0	0
Hispanic	60	44	19	0	0	0
Asian American	9	3	7	0	2	5
American Indian	3	2	1	0	0	0
Unclassified	1	0	0	0	0	0
REGION						
Northeast	18	15	30	0	0	2
Southeast	23	37	25	0	2	0
Central	29	15	8	0	0	2
West	66	46	36	0	2	3
PARENT'S EDUCATION						
Less Than High School	0	0	0	0	0	0
High School	0	0	0	0	0	0
Greater Than High School	0	0	0	0	0	0
Graduated College	0	0	0	0	0	0
Unknown	0	0	0	0	0	0
TYPE OF LOCATION						
Central City	91	45	52	0	2	4
Urban Fringe/Large Town	35	47	31	0	0	1
Rural/Small Town	10	21	16	0	2	2
SCHOOL TYPE						
Public	133	112	91	0	4	7
Nonpublic	3	1	8	0	0	0
Private	1	0	8	0	0	0
Catholic	2	1	0	0	0	0
BIA	0	0	0	0	0	0
DoDEA	0	0	0	0	0	0
MODAL AGE						
Younger	2	3	1	0	0	0
At Modal Age	77	46	21	0	2	3
Older	57	64	77	0	2	4

*Advanced students not sampled for Grade 4.

Table 19-12

Number of Excluded Students in the Science Main and Advanced Samples by Subgroup Classification, Grades 4 ,8, and 12

	MAIN			ADVANCED		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4**	**Grade 8**	**Grade 12**
TOTAL	756	426	425	0	0	0
SEX						
Male	457	265	259	0	0	0
Female	299	161	166	0	0	0
RACE/ETHNICITY						
White	239	145	185	0	0	0
Black	124	98	103	0	0	0
Hispanic	317	159	99	0	0	0
Asian American	65	15	33	0	0	0
American Indian	5	7	2	0	0	0
Unclassified	6	2	3	0	0	0
REGION						
Northeast	91	38	84	0	0	0
Southeast	170	119	112	0	0	0
Central	132	40	57	0	0	0
West	363	229	172	0	0	0
PARENT'S EDUCATION						
Less Than High School	0	0	0	0	0	0
High School	0	1	0	0	0	0
Greater Than High School	0	0	0	0	0	0
Graduated College	0	0	0	0	0	0
Unknown	0	0	0	0	0	0
TYPE OF LOCATION						
Central City	482	187	212	0	0	0
Urban Fringe/Large Town	178	136	127	0	0	0
Rural/Small Town	96	103	86	0	0	0
SCHOOL TYPE						
Public	752	424	419	0	0	0
Nonpublic	4	2	6	0	0	0
Private	2	0	5	0	0	0
Catholic	2	2	1	0	0	0
BIA	0	0	0	0	0	0
DoDEA	0	0	0	0	0	0
MODAL AGE						
<Modal Age	8	5	6	0	0	0
=Modal Age	363	144	113	0	0	0
>Modal Age	385	277	306	0	0	0

Table 19-13

Number of Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 9/Grade 4

	Age	Grade	Age and Grade	Age or Grade
TOTAL	3,654	3,789	2,424	5,019
GENDER				
Male	1,808	1,838	1,128	2,518
Female	1,846	1,951	1,296	2,501
RACE/ETHNICITY				
White	2,067	2,183	1,356	2,894
Black	598	634	421	811
Hispanic	741	727	462	1,006
Asian American	139	151	126	164
American Indian	99	83	50	132
Unclassified	10	11	9	12
REGION				
Northeast	795	861	630	1,026
Southeast	1,027	1,019	622	1,424
Central	707	747	413	1,041
West	1,125	1,162	759	1,528
PARENT'S EDUCATION				
Less Than High School	158	176	99	235
High School	574	615	367	822
Greater Than High School	187	181	117	251
Graduated College	1,430	1,562	1,016	1,976
Unknown	1,274	1,212	806	1,680
TYPE OF LOCATION				
Central City	1,513	1,635	1,019	2,129
Urban Fringe/Large Town	1,271	1,297	893	1,675
Rural/Small Town	870	857	512	1,215
SCHOOL TYPE				
Public	3,237	3,342	2,116	4,463
Nonpublic	417	447	308	556
Private	174	169	116	227
Catholic	243	278	192	329
BIA	0	0	0	0
DoDEA	0	0	0	0

Table 19-14

Number of Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 13/Grade 8

	Age	Grade	Age and Grade	Age or Grade
TOTAL	3,847	4,150	2,504	5,493
GENDER				
Male	1,870	2,060	1,124	2,806
Female	1,977	2,090	1,380	2,687
RACE/ETHNICITY				
White	2,389	2,550	1,518	3,421
Black	540	593	348	785
Hispanic	565	635	373	827
Asian American	222	226	178	270
American Indian	125	141	84	182
Unclassified	6	5	3	8
REGION				
Northeast	811	894	582	1,123
Southeast	1,088	1,159	662	1,585
Central	771	854	478	1,147
West	1,177	1,243	782	1,638
PARENT'S EDUCATION				
Less Than High School	217	301	143	375
High School	1,089	1,172	690	1,571
Greater Than High School	406	441	281	566
Graduated College	1,725	1,846	1,164	2,407
Unknown	394	375	218	551
TYPE OF LOCATION				
Central City	1,441	1,560	937	2,064
Urban Fringe/Large Town	1,412	1,455	952	1,915
Rural/Small Town	994	1,135	615	1,514
SCHOOL TYPE				
Public	3,421	3,720	2,217	4,924
Nonpublic	410	396	274	532
Private	190	190	138	242
Catholic	220	206	136	290
BIA	16	34	13	37
DODEA	0	0	0	0

Table 19-15

Number of Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 17/Grade 11

	Age	Grade	Age and Grade	Age or Grade
TOTAL	3,681	3,737	2,749	4,669
GENDER				
Male	1,874	1,943	1,356	2,461
Female	1,807	1,794	1,393	2,208
RACE/ETHNICITY				
White	2,528	2,573	1,986	3,115
Black	449	440	279	610
Hispanic	465	468	302	631
Asian American	163	178	123	218
American Indian	69	67	52	84
Unclassified	7	11	7	11
REGION				
Northeast	682	721	523	880
Southeast	1,063	1,038	748	1,353
Central	871	900	686	1,085
West	1,065	1,078	792	1,351
PARENT'S EDUCATION				
Less Than High School	251	254	151	354
High School	945	947	675	1,217
Greater Than High School	672	692	528	836
Graduated College	1,673	1,717	1,316	2074
Unknown	103	96	61	138
TYPE OF LOCATION				
Central City	1,111	1,101	780	1,432
Urban Fringe/Large Town	1,537	1,577	1,196	1,918
Rural/Small Town	1,033	1,059	773	1,319
SCHOOL TYPE				
Public	3,384	3,411	2,511	4,284
Nonpublic	289	318	230	377
Private	140	145	105	180
Catholic	149	173	125	197
BIA	8	8	8	8
DODEA	0	0	0	0

Table 19-16

Number of Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 9[1]

	Age	Grade	Age and Grade	Age or Grade
TOTAL	5414	3,665	3,665	5,414
GENDER				
Male	2,709	1,766	1,766	2,709
Female	2,705	1,899	1,899	2,705
RACE/ETHNICITY				
White	3,204	2,146	2,146	3,204
Black	801	578	578	801
Hispanic	1,075	687	687	1,075
Asian American	188	156	156	188
American Indian	134	88	88	134
Unclassified	12	10	10	12
REGION				
Northeast	1,142	918	918	1,142
Southeast	1,436	906	906	1,436
Central	1,188	698	698	1,188
West	1,648	1,143	1,143	1,648
PARENT'S EDUCATION				
Less Than High School	230	146	146	230
High School	713	466	466	713
Greater Than High School	386	289	289	386
Graduated College	2,274	1,604	1,604	2,274
Unknown	1,792	1,150	1,150	1,792
TYPE OF LOCATION				
Central City	2,485	1,721	1,721	2,485
Urban Fringe/Large Town	1,670	1,198	1,198	1,670
Rural/Small Town	1,259	746	746	1,259
SCHOOL TYPE				
Public	4,790	3,231	3,231	4,790
Nonpublic	609	422	422	609
Private	162	104	104	162
Catholic	447	318	318	447
BIA	15	12	12	15
DODEA	0	0	0	0

[1] Note: Since this is an age-only sample, the number of students who are age-eligible only will be the same as the number of students who are age- *or* grade-eligible. Likewise, the number of students who are grade-eligible only will be the same as the number of students who are both age- *and* grade-eligible.

Table 19-17

Number of Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 13[1]

	Age	Grade	Age and Grade	Age or Grade
TOTAL	5,658	3,662	3,662	5,658
GENDER				
Male	2,736	1,652	1,652	2,736
Female	2,922	2,010	2,010	2,922
RACE/ETHNICITY				
White	3,528	2,272	2,272	3,528
Black	776	509	509	776
Hispanic	943	565	565	943
Asian American	293	234	234	293
American Indian	112	76	76	112
Unclassified	6	6	6	6
REGION				
Northeast	1,221	900	900	1,221
Southeast	1,589	937	937	1,589
Central	1,129	693	693	1,129
West	1,719	1,132	1,132	1,719
PARENT'S EDUCATION				
Less Than High School	353	188	188	353
High School	1,295	815	815	1,295
Greater Than High School	943	672	672	943
Graduated College	2,458	1,655	1,655	2,458
Unknown	587	320	320	587
TYPE OF LOCATION				
Central City	2,063	1,357	1,357	2,063
Urban Fringe/Large Town	2,047	1,386	1,386	2,047
Rural/Small Town	1,548	919	919	1,548
SCHOOL TYPE				
Public	5,096	3,260	3,260	5,096
Nonpublic	562	402	402	562
Private	224	181	181	224
Catholic	338	221	221	338
BIA	0	0	0	0
DODEA	0	0	0	0

[1] Note: Since this is an age-only sample, the number of students who are age-eligible only will be the same as the number of students who are age- *or* grade-eligible. Likewise, the number of students who are grade-eligible only will be the same as the number of students who are both age- *and* grade-eligible.

Table 19-18

Number of Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 17[1]

	Age	Grade	Age and Grade	Age or Grade
TOTAL	3,539	2,532	2,532	3,539
GENDER				
Male	1,755	1,196	1,196	1,755
Female	1,784	1,336	1,336	1,784
RACE/ETHNICITY				
White	2,401	1,836	1,836	2,401
Black	531	329	329	531
Hispanic	401	244	244	401
Asian American	155	94	94	155
American Indian	43	23	23	43
Unclassified	8	6	6	8
REGION				
Northeast	712	519	519	712
Southeast	1,122	803	803	1,122
Central	733	529	529	733
West	972	681	681	972
PARENT'S EDUCATION				
Less Than High School	236	122	122	236
High School	757	506	506	757
Greater Than High School	835	616	616	835
Graduated College	1,619	1,238	1,238	1,619
Unknown	71	37	37	71
TYPE OF LOCATION				
Central City	1,311	896	896	1,311
Urban Fringe/Large Town	1,189	883	883	1,189
Rural/Small Town	1,039	753	753	1,039
SCHOOL TYPE				
Public	3,257	2,309	2,309	3,257
Nonpublic	282	223	223	282
Private	124	99	99	124
Catholic	158	124	124	158
BIA	0	0	0	0
DODEA	0	0	0	0

[1] Note: Since this is an age-only sample, the number of students who are age-eligible only will be the same as the number of students who are age- *or* grade-eligible. Likewise, the number of students who are grade-eligible only will be the same as the number of students who are both age- *and* grade-eligible.

Table 19-19

Number of Excluded Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 9/Grade 4

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
TOTAL	345	404	217	532
SEX				
Male	207	243	124	326
Female	138	161	93	206
RACE/ETHNICITY				
White	133	161	66	228
Black	54	66	30	90
Hispanic	122	134	90	166
Asian American	30	35	26	39
American Indian	3	6	3	6
Unclassified	3	2	2	3
REGION				
Northeast	30	46	19	57
Southeast	96	130	52	174
Central	49	47	18	78
West	170	181	128	223
TYPE OF LOCATION				
Central City	189	211	130	270
Urban Fringe/Large Town	106	117	69	154
Rural/Small Town	50	76	18	108
SCHOOL TYPE				
Public	340	402	215	527
Nonpublic	4	2	2	4
Private	0	0	0	0
Catholic	4	2	2	4
BIA	0	0	0	0
DoDea	0	0	0	0

Table 19-20

Number of Excluded Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 13/Grade 8

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
TOTAL	265	303	87	481
SEX				
Male	175	211	67	319
Female	90	92	20	162
RACE/ETHNICITY				
White	153	187	47	293
Black	33	35	8	60
Hispanic	58	58	23	93
Asian American	10	11		16
American Indian	12	4	18	
Unclassified	1	0	0	1
REGION				
Northeast	45	46	17	74
Southeast	93	104	21	176
Central	41	74	14	101
West	86	79	35	130
TYPE OF LOCATION				
Central City	118	121	35	204
Urban Fringe/Large Town	79	80	35	124
Rural/Small Town	68	102	17	153
SCHOOL TYPE				
Public	257	291	84	464
Nonpublic	3	1	0	4
Private	2	1	0	3
Catholic	1	0	0	1
BIA	5	11	3	13
DoDea	0	0	0	0

Table 19-21

Number of Excluded Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 17/Grade 11

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
TOTAL	277	227	92	412
SEX				
Male	171	139	43	267
Female	106	88	49	145
RACE/ETHNICITY				
White	161	151	63	249
Black	65	33	18	80
Hispanic	35	26	7	54
Asian American	10	10	1	19
American Indian	4	7	3	8
Unclassified	2	0	0	2
REGION				
Northeast	36	44	18	62
Southeast	117	75	32	160
Central	51	41	15	77
West	73	67	27	113
TYPE OF LOCATION				
Central City	80	68	25	123
Urban Fringe/Large Town	107	105	46	166
Rural/Small Town	90	54	21	123
SCHOOL TYPE				
Public	277	224	92	409
Nonpublic	0	1	0	1
Private	0	0	0	0
Catholic	0	1	0	1
BIA	0	2	0	2
DoDea	0	0	0	0

Table 19-22

Number of Excluded Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 9/Grade 4

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
TOTAL	585	316	316	585
SEX				
Male	360	192	192	360
Female	225	124	124	225
RACE/ETHNICITY				
White	220	103	103	220
Black	96	41	41	96
Hispanic	217	133	133	217
Asian American	45	35	35	45
American Indian	2	1	1	2
Unclassified	5	3	3	5
REGION				
Northeast	65	38	38	65
Southeast	202	72	72	202
Central	80	35	35	80
West	238	171	171	238
TYPE OF LOCATION				
Central City	334	200	200	334
Urban Fringe/Large Town	148	91	91	148
Rural/Small Town	103	25	25	103
SCHOOL TYPE				
Public	578	314	314	578
Nonpublic	7	2	2	7
Private	0	0	0	0
Catholic	7	2	2	7
BIA	0	0	0	0
DoDea	0	0	0	0

Table 19-23

Number of Excluded Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age13/Grade 8

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
TOTAL	452	150	150	452
SEX				
Male	286	86	86	286
Female	166	64	64	166
RACE/ETHNICITY				
White	239	75	75	239
Black	76	20	20	76
Hispanic	116	47	47	116
Asian American	16	7	7	16
American Indian	2	1	1	2
Unclassified	3	0	0	3
REGION				
Northeast	87	34	34	87
Southeast	156	41	41	156
Central	89	25	25	89
West	120	50	50	120
TYPE OF LOCATION				
Central City	187	57	57	187
Urban Fringe/Large Town	125	57	57	125
Rural/Small Town	140	36	36	140
SCHOOL TYPE				
Public	450	150	150	450
Nonpublic	2	0	0	2
Private	2	0	0	2
Catholic	0	0	0	0
BIA	0	0	0	0
DoDea	0	0	0	0

Table 19-24

Number of Excluded Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age17/Grade 11

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
TOTAL	301	110	110	301
SEX				
Male	202	67	67	202
Female	99	43	43	99
RACE/ETHNICITY				
White	170	75	75	170
Black	55	17	17	55
Hispanic	55	14	14	55
Asian American	16	3	3	16
American Indian	5	1	1	5
Unclassified	0	0	0	0
REGION				
Northeast	57	36	36	57
Southeast	103	24	24	103
Central	51	18	18	51
West	90	32	32	90
TYPE OF LOCATION				
Central City	109	36	36	109
Urban Fringe/Large Town	107	50	50	107
Rural/Small Town	85	24	24	85
SCHOOL TYPE				
Public	298	110	110	298
Nonpublic	3	0	0	3
Private	3	0	0	3
Catholic	0	0	0	0
BIA	0	0	0	0
DoDea	0	0	0	0

Table 19-25

Weighted Percentage of Students in the Mathematics Main and Estimation Samples by Subgroup Classification, Grades 4, 8, and 12

	MAIN			ESTIMATION		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4**	**Grade 8**	**Grade 12**
TOTAL	3,711,786	3,566,392	2,827,040	3,688,821	3,598,564	2,740,931
GENDER						
Male	50.8	52.3	47.6	49.3	50.4	47.4
Female	49.2	47.7	52.4	50.7	49.6	52.6
RACE/ETHNICITY						
White	67.8	69.1	69.6	68.0	69.6	70.3
Black	14.6	14.2	14.1	14.6	14.1	13.8
Hispanic	12.9	12.3	11.2	13.0	11.9	11.4
Asian American	2.7	3.3	3.6	2.5	3.3	3.9
American Indian	1.7	1.1	1.3	1.9	0.9	0.4
Unclassified	0.2	0.1	0.2	0.1	0.2	0.2
REGION						
Northeast	21.9	20.3	21.8	21.6	20.7	23.5
Southeast	21.0	23.3	21.6	22.1	21.3	20.6
Central	24.8	24.3	24.0	24.6	24.4	24.5
West	32.3	32.1	32.6	31.7	33.7	31.5
PARENT'S EDUCATION						
Less Than High School	3.8	6.9	6.4	3.8	6.0	6.6
High School	12.5	21.8	18.8	11.6	22.0	17.2
Greater Than High School	7.0	18.5	25.4	7.5	16.3	21.0
Graduated College	39.4	41.7	46.4	40.0	42.5	52.9
Unknown	35.6	10.7	2.7	34.9	11.5	2.1
TYPE OF LOCATION						
Central City	30.1	33.3	31.7	37.3	37.1	39.3
Urban Fringe/Large Town	46.2	36.2	39.8	36.5	36.8	34.1
Rural/Small Town	23.7	30.5	28.5	26.2	26.1	26.6
SCHOOL TYPE						
Public	89.1	89.3	87.3	85.4	88.6	83.1
Non Public	10.9	10.7	12.0	14.6	11.4	16.9
Private	3.7	4.5	4.2	5.7	4.0	6.9
Catholic	7.2	6.3	7.9	8.9	7.4	10.0
BIA	0.0	0.0	0.6	0.0	0.0	0.0
DoDEA	0.0	0.0	0.0	0.0	0.0	0.0
MODAL AGE						
Younger	0.6	0.6	1.3	0.6	0.1	1.1
At Modal Age	59.6	56.2	64.9	60.8	55.6	65.2
Older	39.8	43.1	33.8	38.6	44.3	33.7

Table 19-26

Weighted Percentage of Students in the Mathematics Theme and Advanced Mathematics Samples by Subgroup Classification, Grades 4, 8, and 12

	THEME			ADVANCED		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4***	**Grade 8**	**Grade 12**
TOTAL	3,690,245	3,566,103	2,845,023	0	809,085	696,805
GENDER						
Male	51.8	52.6	49.2	0.0	48.3	50.6
Female	48.2	47.4	50.8	0.0	51.7	49.4
RACE/ETHNICITY						
White	68.5	70.5	69.3	0.0	71.0	74.1
Black	14.5	13.7	14.2	0.0	14.3	7.2
Hispanic	12.5	11.6	11.6	0.0	6.5	7.9
Asian American	2.9	2.4	4.2	0.0	5.6	10.3
American Indian	1.5	1.7	0.5	0.0	2.2	0.3
Unclassified	0.1	0.1	0.2	0.0	0.4	0.2
REGION						
Northeast	21.0	20.7	23.2	0.0	26.9	24.8
Southeast	22.9	22.5	21.2	0.0	16.9	21.0
Central	25.7	23.0	23.4	0.0	29.9	28.6
West	30.4	33.8	32.1	0.0	26.3	25.6
PARENT'S EDUCATION						
Less Than High School	4.5	7.9	6.9	0.0	3.0	3.8
High School	12.6	23.8	19.1	0.0	13.7	13.3
Greater Than High School	7.4	18.1	26.5	0.0	20.2	22.6
Graduated College	40.1	40.5	44.2	0.0	57.4	58.5
Unknown	35.3	9.4	3.2	0.0	4.7	1.5
TYPE OF LOCATION						
Central City	39.9	28.1	31.6	0.0	36.1	32.0
Urban Fringe/Large Town	37.7	49.2	41.3	0.0	38.7	38.5
Rural/Small Town	22.4	22.7	27.1	0.0	25.2	29.4
SCHOOL TYPE						
Public	87.6	90.5	88.7	0.0	83.5	81.5
Nonpublic	12.4	9.5	11.3	0.0	16.5	18.5
Private	5.0	3.9	3.9	0.0	5.3	7.9
Catholic	7.4	5.5	7.4	0.0	11.3	10.6
BIA	0.0	0.0	0.0	0.0	0.0	0.0
DoDEA	0.0	0.0	0.0	0.0	0.0	0.0
MODAL AGE						
Younger	0.4	0.6	1.5	0.0	0.7	1.5
At Modal Age	59.0	55.5	64.4	0.0	68.7	72.2
Older	40.5	43.9	34.1	0.0	30.6	26.3

*Advanced students not sampled for Grade 4.

Table 19-27

Weighted Percentage of Students in the Science Main and Advanced Samples by Subgroup Classification, Grades 4, 8, and 12

	MAIN			ADVANCED		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4[1]**	**Grade 8[1]**	**Grade 12**
TOTAL	3,618,494	3,564,079	2,903,402	0	0	585,798
GENDER						
Male	50.3	50.8	48.4	0.0	0.0	49.2
Female	49.7	49.2	51.6	0.0	0.0	50.8
RACE/ETHNICITY						
White	68.8	69.6	69.9	0.0	0.0	74.1
Black	14.6	14.1	14.3	0.0	0.0	9.1
Hispanic	12.2	11.8	11.2	0.0	0.0	7.0
Asian American	2.6	2.6	3.8	0.0	0.0	9.1
American Indian	1.8	1.7	0.7	0.0	0.0	0.5
Unclassified	0.1	0.2	0.1	0.0	0.0	0.2
REGION						
Northeast	21.8	22.1	21.6	0.0	0.0	24.4
Southeast	22.4	21.8	21.2	0.0	0.0	21.2
Central	25.7	23.8	24.2	0.0	0.0	30.4
West	30.0	32.3	33.1	0.0	0.0	24.0
PARENT'S EDUCATION						
Less Than High School	4.2	6.2	6.6	0.0	0.0	2.7
High School	13.6	19.5	18.2	0.0	0.0	11.2
Greater Than High School	7.4	19.5	25.2	0.0	0.0	20.8
Graduated College	39.7	44.0	45.8	0.0	0.0	62.8
Unknown	33.2	9.1	2.5	0.0	0.0	0.0
TYPE OF LOCATION						
Central City	36.9	26.6	35.0	0.0	0.0	31.2
Urban Fringe/Large Town	38.1	45.1	36.3	0.0	0.0	41.3
Rural/Small Town	25.0	28.3	28.7	0.0	0.0	27.4
SCHOOL TYPE						
Public	87.5	88.8	87.7	0.0	0.0	83.0
Nonpublic	12.5	11.2	12.3	0.0	0.0	17.0
Private	4.8	4.7	4.4	0.0	0.0	5.4
Catholic	7.7	6.6	7.9	0.0	0.0	11.7
BIA	0.0	0.0	0.0	0.0	0.0	0.0
DODEA	0.0	0.0	0.0	0.0	0.0	0.0
MODAL AGE						
Younger	0.7	0.6	1.1	0.0	0.0	1.4
At Modal Age	59.3	54.5	63.6	0.0	0.0	70.9
Older	40.0	44.9	35.3	0.0	0.0	27.7

[1]Advanced students not sampled for Grade 4 or Grade 8.

Table 19-28

Weighted Percentage of Excluded Students in the Mathematics Main and Estimation Samples by Subgroup Classification, Grades 4, 8, and 12

	MAIN			ESTIMATION		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4**	**Grade 8**	**Grade 12**
TOTAL	229,564	162,944	88,046	180,417	142,276	110,197
GENDER						
Male	62.0	62.1	67.9	77.2	56.8	56.3
Female	38.0	37.9	32.1	22.8	43.2	43.7
RACE/ETHNICITY						
White	56.1	65.5	65.2	69.4	64.3	42.7
Black	16.4	18.3	15.9	11.8	12.6	24.4
Hispanic	22.9	11.2	17.6	14.2	19.2	29.3
Asian American	2.6	2.8	1.0	2.5	3.9	3.6
American Indian	1.6	1.4	0.3	2.1	0.0	0.0
Unclassified	0.3	0.8	0.0	0.0	0.0	0.0
REGION						
Northeast	12.2	22.0	29.8	12.4	22.1	16.4
Southeast	25.0	20.5	19.7	19.0	20.8	13.3
Central	16.4	24.6	14.4	35.8	18.1	14.9
West	46.4	32.9	36.2	32.8	39.0	55.4
PARENT'S EDUCATION						
Less Than High School	0.0	0.0	0.0	0.0	0.0	0.0
High School	0.0	0.0	0.0	0.0	0.0	0.0
Greater Than High School	0.0	0.0	0.0	0.0	0.0	0.0
Graduated College	0.0	0.0	0.0	0.0	0.0	0.0
Unknown	0.0	0.0	0.0	0.0	0.0	0.0
TYPE OF LOCATION						
Central City	28.8	38.1	27.9	23.5	49.7	51.1
Urban Fringe/Large Town	35.6	35.6	51.4	46.9	24.5	18.3
Rural/Small Town	35.6	26.3	20.8	29.5	25.7	30.6
SCHOOL TYPE						
Public	97.8	99.4	99.5	100.0	100.0	100.0
Non Public	2.2	0.6	0.5	0.0	0.0	0.0
Private	1.0	0.0	0.0	0.0	0.0	0.0
Catholic	1.1	0.6	0.5	0.0	0.0	0.0
BIA	0.0	0.0	0.0	0.0	0.0	0.0
DODEA	0.0	0.0	0.0	0.0	0.0	0.0
MODAL AGE						
Younger	0.0	2.4	0.8	0.0	0.0	0.0
At Modal Age	41.0	23.1	36.2	48.8	33.8	20.3
Older	59.0	74.5	63.0	51.2	66.2	79.7

Table 19-29

Weighted Percentage of Excluded Students in the Mathematics Theme and Advanced Samples by Subgroup Classification, Grades 4, 8, and 12

	THEME			ADVANCED		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4[1]**	**Grade 8**	**Grade 12**
TOTAL	238,555	160,860	115,182	0	1,042	1,701
GENDER						
Male	58.3	55.1	57.1	0.0	100.0	89.2
Female	41.7	44.9	42.9	0.0	0.0	10.8
RACE/ETHNICITY						
White	42.0	46.0	67.7	0.0	0.0	36.1
Black	18.2	25.3	13.8	0.0	0.0	0.0
Hispanic	31.3	24.3	14.8	0.0	0.0	0.0
Asian American	4.9	1.7	3.1	0.0	100.0	63.9
American Indian	2.8	2.8	0.6	0.0	0.0	0.0
Unclassified	0.7	0.0	0.0	0.0	0.0	0.0
REGION						
Northeast	12.7	12.9	28.5	0.0	0.0	41.3
Southeast	18.9	31.0	21.0	0.0	0.0	0.0
Central	25.7	21.1	16.3	0.0	0.0	26.5
West	42.6	35.0	34.1	0.0	100.0	32.3
PARENT'S EDUCATION						
Less Than High School	0.0	0.0	0.0	0.0	0.0	0.0
High School	0.0	0.0	0.0	0.0	0.0	0.0
Greater Than High School	0.0	0.0	0.0	0.0	0.0	0.0
Graduated College	0.0	0.0	0.0	0.0	0.0	0.0
Unknown	0.0	0.0	0.0	0.0	0.0	0.0
TYPE OF LOCATION						
Central City	54.8	31.3	42.3	0.0	100.0	41.0
Urban Fringe/Large Town	36.5	40.2	40.1	0.0	0.0	22.9
Rural/Small Town	8.7	28.5	17.6	0.0	0.0	36.1
SCHOOL TYPE						
Public	98.6	99.2	95.2	0.0	100.0	100.0
Nonpublic	1.4	0.8	4.8	0.0	0.0	0.0
Private	0.3	0.0	4.8	0.0	0.0	0.0
Catholic	1.1	0.8	0.0	0.0	0.0	0.0
BIA	0.0	0.0	0.0	0.0	0.0	0.0
DODEA	0.0	0.0	0.0	0.0	0.0	0.0
MODAL AGE						
Younger	2.7	1.5	0.4	0.0	0.0	0.0
At Modal Age	46.8	30.7	22.7	0.0	100.0	59.0
Older	50.6	67.8	76.8	0.0	0.0	41.0

[1] Advanced students not sampled for Grade 4.

Figure 19-30

Weighted Percentage of Excluded Students in the Science Main and Advanced Samples by Subgroup Classification, Grades 4, 8, and 12

	MAIN			ADVANCED		
	Grade 4	**Grade 8**	**Grade 12**	**Grade 4**	**Grade 8**	**Grade 12**
TOTAL	322,613	164,891	119,759	0	0	0
SEX						
Male	63.2	63.6	61.9	0.0	0.0	0.0
Female	36.4	38.1	38.1	0.0	0.0	0.0
RACE/ETHNICITY						
White	49.1	54.6	54.1	0.0	0.0	0.0
Black	16.6	19.4	22.4	0.0	0.0	0.0
Hispanic	27.6	21.6	18.1	0.0	0.0	0.0
Asian American	5.3	2.2	4.2	0.0	0.0	0.0
American Indian	0.8	2.0	0.5	0.0	0.0	0.0
Unclassified	0.5	0.2	0.7	0.0	0.0	0.0
REGION						
Northeast	14.6	12.7	21.8	0.0	0.0	0.0
Southeast	23.6	23.3	22.6	0.0	0.0	0.0
Central	22.2	21.9	18.3	0.0	0.0	0.0
West	39.6	42.2	37.2	0.0	0.0	0.0
PARENT'S EDUCATION						
Less Than High School	0.0	0.0	0.0	0.0	0.0	0.0
High School	0.0	0.5	0.0	0.0	0.0	0.0
Greater Than High School	0.0	0.0	0.0	0.0	0.0	0.0
Graduated College	0.0	0.0	0.0	0.0	0.0	0.0
Unknown	0.0	0.0	0.0	0.0	0.0	0.0
TYPE OF LOCATION						
Central City	52.0	29.1	40.1	0.0	0.0	0.0
Urban Fringe/Large Town	24.8	33.6	32.8	0.0	0.0	0.0
Rural/Small Town	23.3	37.3	27.1	0.0	0.0	0.0
SCHOOL TYPE						
Public	99.5	99.4	99.0	0.0	0.0	0.0
Nonpublic	0.5	0.6	1.0	0.0	0.0	0.0
Private	0.2	0.0	0.8	0.0	0.0	0.0
Catholic	0.3	0.6	0.2	0.0	0.0	0.0
BIA	0.0	0.0	0.0	0.0	0.0	0.0
DoDEA	0.0	0.0	0.0	0.0	0.0	0.0
MODAL AGE						
<Modal Age	0.9	1.0	1.1	0.0	0.0	0.0
=Modal Age	37.7	25.7	25.2	0.0	0.0	0.0
>Modal Age	61.3	73.2	73.7	0.0	0.0	0.0

Table 19-31

Weighted Percentage of Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 9/Grade 4

	Age	Grade	Age and Grade	Age or Grade
TOTAL	3,170,010	3,579,694	2,119,331	4,630,373
GENDER				
Male	49.5	49.3	46.3	50.8
Female	50.5	50.7	53.7	49.2
RACE/ETHNICITY				
White	66.7	68.3	67.5	67.6
Black	15.4	14.6	15.1	15.0
Hispanic	12.9	12.7	12.4	13.0
Asian American	2.6	2.6	3.5	2.3
American Indian	2.1	1.5	1.3	2.0
Unclassified	0.2	0.2	0.3	0.2
REGION				
Northeast	23.5	22.1	26.0	21.3
Southeast	23.9	23.8	22.2	24.6
Central	24.6	25.5	22.1	26.4
West	28.0	28.6	29.7	27.7
PARENT'S EDUCATION				
Less Than High School	4.0	4.4	3.9	4.4
High School	15.3	16.0	14.5	16.2
Greater Than High School	5.0	4.7	4.6	4.9
Graduated College	39.5	41.6	42.1	39.9
Unknown	35.5	32.2	34.1	33.6
TYPE OF LOCATION				
Central City	40.8	42.5	41.6	41.8
Urban Fringe/Large Town	34.3	33.4	36.6	32.6
Rural/Small Town	24.8	24.0	21.8	25.6
SCHOOL TYPE				
Public	86.2	86.3	85.2	86.7
Nonpublic	13.8	13.7	14.8	13.3
Private	6.8	6.2	6.7	6.4
Catholic	7.1	7.4	8.1	6.9
BIA	0.0	0.0	0.0	0.0
DODEA	0.0	0.0	0.0	0.0

Table 19-32

Weighted Percentage of Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 13/Grade 8

	Age	Grade	Age and Grade	Age or Grade
TOTAL	3,173,938	3,465,078	1,943,322	4,695,694
GENDER				
Male	48.6	50.3	44.4	51.6
Female	51.4	49.7	55.6	48.4
RACE/ETHNICITY				
White	66.6	69.1	68.9	67.4
Black	15.1	14.4	13.6	15.2
Hispanic	12.7	12.0	12.2	12.4
Asian American	3.5	2.8	3.8	2.9
American Indian	2.0	1.7	1.5	2.0
Unclassified	0.1	0.1	0.1	0.1
REGION				
Northeast	23.1	22.3	24.7	21.8
Southeast	25.5	25.0	23.8	25.9
Central	21.8	23.4	21.1	23.3
West	29.6	29.3	30.4	29.0
PARENT'S EDUCATION				
Less Than High School	5.3	7.0	5.3	6.5
High School	28.8	28.5	27.8	28.9
Greater Than High School	10.4	10.6	11.0	10.3
Graduated College	45.0	45.0	47.3	44.1
Unknown	10.1	8.5	8.4	9.7
TYPE OF LOCATION				
Central City	36.2	36.0	35.4	36.4
Urban Fringe/Large Town	35.7	33.9	38.2	33.3
Rural/Small Town	28.1	30.1	26.4	30.3
SCHOOL TYPE				
Public	88.5	89.4	87.9	89.4
Nonpublic	11.2	9.9	11.7	10.0
Private	4.8	4.6	5.7	4.3
Catholic	6.3	5.2	6.0	5.7
BIA	0.4	0.7	0.4	0.6
DODEA	0.0	0.0	0.0	0.0

Table 19-33

Weighted Percentage of Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 17/Grade 11

	Age	Grade	Age and Grade	Age or Grade
TOTAL	3,224,505	3,160,512	1,977,398	4,407,619
GENDER				
Male	51.4	52.4	48.9	53.3
Female	48.6	47.6	51.1	46.7
RACE/ETHNICITY				
White	69.3	68.3	74.2	66.4
Black	14.4	15.1	11.7	16.1
Hispanic	11.7	12.4	9.9	13.0
Asian American	3.1	2.9	2.6	3.2
American Indian	1.4	1.2	1.3	1.3
Unclassified	0.1	0.2	0.2	0.1
REGION				
Northeast	22.3	24.9	23.9	23.5
Southeast	24.2	22.7	21.4	24.4
Central	24.5	24.5	26.0	23.8
West	29.0	27.9	28.8	28.3
PARENT'S EDUCATION				
Less Than High School	6.8	6.7	4.9	7.6
High School	26.7	26.0	24.7	27.1
Greater Than High School	18.0	18.7	19.8	17.7
Graduated College	44.7	45.2	47.9	43.6
Unknown	2.8	2.7	2.1	3.0
TYPE OF LOCATION				
Central City	34.7	34.2	32.3	35.4
Urban Fringe/Large Town	39.2	39.5	41.4	38.5
Rural/Small Town	26.1	26.3	26.4	26.1
SCHOOL TYPE				
Public	91.7	90.8	90.7	91.5
Nonpublic	8.1	9.0	9.0	8.3
Private	3.5	3.4	3.6	3.4
Catholic	4.6	5.6	5.4	4.9
BIA	0.2	0.2	0.4	0.2
DODEA	0.0	0.0	0.0	0.0

Table 19-34

Weighted Percentage of Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 9[1]

	Age	Grade	Age and Grade	Age or Grade
TOTAL	3,320,984	2,207,888	2,207,888	3,320,984
GENDER				
Male	49.6	48.0	48.0	49.6
Female	50.4	52.0	52.0	50.4
RACE/ETHNICITY				
White	69.0	68.1	68.1	69.0
Black	14.3	15.5	15.5	14.3
Hispanic	12.4	11.6	11.6	12.4
Asian American	2.3	2.9	2.9	2.3
American Indian	1.9	1.8	1.8	1.9
Unclassified	0.2	0.2	0.2	0.2
REGION				
Northeast	21.7	25.7	25.7	21.7
Southeast	23.4	22.5	22.5	23.4
Central	25.0	21.4	21.4	25.0
West	30.0	30.4	30.4	30.0
PARENT'S EDUCATION				
Less Than High School	3.9	3.8	3.8	3.9
High School	12.5	12.2	12.2	12.5
Greater Than High School	7.1	7.9	7.9	7.1
Graduated College	42.7	45.0	45.0	42.7
Unknown	33.1	30.6	30.6	33.1
Type Of Location				
Central City	42.3	43.0	43.0	42.3
Urban Fringe/Large Town	32.4	34.3	34.3	32.4
Rural/Small Town	25.4	22.7	22.7	25.4
SCHOOL TYPE				
Public	86.8	86.3	86.3	86.8
Nonpublic	12.9	13.4	13.4	12.9
Private	4.0	3.8	3.8	4.0
Catholic	8.9	9.5	9.5	8.9
BIA	0.3	0.3	0.3	0.3
DODEA	0.0	0.0	0.0	0.0

[1] Note: Since this is an age-only sample, the number of students who are age-eligible only will be the same as the number of students who are age- *or* grade-eligible. Likewise, the number of students who are grade-eligible only will be the same as the number of students who are both age- *and* grade-eligible.

Table 19-35

Weighted Percentage of Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 13[1]

	Age	Grade	Age and Grade	Age or Grade
TOTAL	3,360,572	2,128,872	2,128,872	3,360,572
GENDER				
Male	48.5	45.2	45.2	48.5
Female	51.5	54.8	54.8	51.5
RACE/ETHNICITY				
White	68.6	68.0	68.0	68.6
Black	14.3	14.7	14.7	14.3
Hispanic	12.2	11.3	11.3	12.2
Asian American	3.6	4.4	4.4	3.6
American Indian	1.3	1.4	1.4	1.3
Unclassified	0.1	0.1	0.1	0.1
REGION				
Northeast	22.0	25.0	25.0	22.0
Southeast	24.9	22.9	22.9	24.9
Central	23.1	21.9	21.9	23.1
West	30.0	30.2	30.2	30.0
PARENT'S EDUCATION				
Less Than High School	5.5	4.3	4.3	5.5
High School	22.7	22.1	22.1	22.7
Greater Than High School	16.7	18.7	18.7	16.7
Graduated College	45.0	46.7	46.7	45.0
Unknown	9.7	7.9	7.9	9.7
TYPE OF LOCATION				
Central City	35.7	36.2	36.2	35.7
Urban Fringe/Large Town	35.9	38.1	38.1	35.9
Rural/Small Town	28.3	25.6	25.6	28.3
SCHOOL TYPE				
Public	88.8	88.2	88.2	88.8
Nonpublic	11.2	11.8	11.8	11.2
Private	4.5	5.1	5.1	4.5
Catholic	6.7	6.7	6.7	6.7
BIA	0.0	0.0	0.0	0.0
DODEA	0.0	0.0	0.0	0.0

[1] Note: Since this is an age-only sample, the number of students who are age-eligible only will be the same as the number of students who are age- *or* grade-eligible. Likewise, the number of students who are grade-eligible only will be the same as the number of students who are both age- *and* grade-eligible.

Table 19-36

Weighted Percentage of Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 17[1]

	Age	Grade	Age and Grade	Age or Grade
TOTAL	3,185,309	2,250,256	2,250,256	3,185,309
GENDER				
Male	49.5	47.0	47.0	49.5
Female	50.5	53.0	53.0	50.5
RACE/ETHNICITY				
White	69.3	73.8	73.8	69.3
Black	14.5	12.6	12.6	14.5
Hispanic	11.6	9.7	9.7	11.6
Asian American	3.5	3.0	3.0	3.5
American Indian	0.8	0.7	0.7	0.8
Unclassified	0.2	0.2	0.2	0.2
REGION				
Northeast	23.4	24.1	24.1	23.4
Southeast	22.4	22.1	22.1	22.4
Central	24.5	24.2	24.2	24.5
West	29.7	29.6	29.6	29.7
PARENT'S EDUCATION				
Less Than High School	6.4	4.4	4.4	6.4
High School	20.9	19.6	19.6	20.9
Greater Than High School	23.8	24.4	24.4	23.8
Graduated College	46.0	49.2	49.2	46.0
Unknown	2.0	1.5	1.5	2.0
TYPE OF LOCATION				
Central City	36.9	35.3	35.3	36.9
Urban Fringe/Large Town	37.9	39.4	39.4	37.9
Rural/Small Town	25.1	25.3	25.3	25.1
SCHOOL TYPE				
Public	91.4	90.7	90.7	91.4
Nonpublic	8.6	9.3	9.3	8.6
Private	3.5	3.7	3.7	3.5
Catholic	5.1	5.6	5.6	5.1
BIA	0.0	0.0	0.0	0.0
DODEA	0.0	0.0	0.0	0.0

[1] Note: Since this is an age-only sample, the number of students who are age-eligible only will be the same as the number of students who are age- *or* grade-eligible. Likewise, the number of students who are grade-eligible only will be the same as the number of students who are both age- *and* grade-eligible.

Table 19-37

Weighted Percentage of Excluded Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 9/Grade 4

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
SEX				
Male	61.2	63.6	57.7	64.0
Female	38.8	36.4	42.3	36.0
RACE/ETHNICITY				
White	48.3	57.6	42.7	57.5
Black	18.3	16.9	14.9	17.8
Hispanic	24.3	19.0	30.1	18.5
Asian American	7.6	4.7	10.4	4.5
American Indian	0.8	1.6	1.3	1.4
Unclassified	0.7	0.1	0.6	0.2
REGION				
Northeast	9.4	13.1	7.8	12.9
Southeast	26.9	36.4	23.5	35.8
Central	17.9	18.5	11.7	19.7
West	45.7	32.0	57.0	31.5
TYPE OF LOCATION				
Central City	50.2	42.8	55.0	42.8
Urban Fringe/Large Town	34.5	28.7	36.7	29.1
Rural/Small Town	15.3	28.5	8.3	28.1
SCHOOL TYPE				
Public	98.4	99.7	98.7	99.4
Nonpublic	1.3	0.3	1.3	0.4
Private	0.0	0.0	0.0	0.0
Catholic	1.3	0.3	1.3	0.4
BIA	0.0	0.0	0.0	0.0
DoDea	0.0	0.0	0.0	0.0
ESTIMATED TOTAL POPULATION	106,503	266,020	64,398	308,125

Table 19-38

Weighted Percentage of Excluded Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 13/Grade 8

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
SEX				
Male	68.4	67.3	77.6	66.8
Female	31.6	32.7	22.4	33.2
RACE/ETHNICITY				
White	57.5	69.5	60.5	66.6
Black	18.4	13.5	9.6	15.4
Hispanic	17.8	13.1	23.4	13.7
Asian American	3.4	1.2	3.7	1.6
American Indian	2.4	2.6	2.8	2.5
Unclassified	0.4	0.0	0.0	0.1
REGION				
Northeast	20.3	17.4	21.8	17.9
Southeast	34.1	32.9	23.1	34.1
Central	15.1	28.6	15.4	25.6
West	30.4	21.1	39.7	22.4
TYPE OF LOCATION				
Central City	45.4	40.0	41.3	41.5
Urban Fringe/Large Town	26.7	21.2	39.4	21.4
Rural/Small Town	27.9	38.8	19.4	37.1
SCHOOL TYPE				
Public	96.8	96.9	97.4	96.8
Nonpublic	1.3	0.5	0.0	0.8
Private	0.9	0.5	0.0	0.7
Catholic	0.4	0.0	0.0	0.1
BIA	1.9	2.6	2.6	2.4
DoDea	0.0	0.0	0.0	0.0
ESTIMATED TOTAL POPULATION	87,533	223,797	24,049	287,281

Table 19-39

Weighted Percentage of Excluded Students in the Reading and Writing Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 17/Grade 11

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
SEX				
Male	64.4	66.7	47.6	67.9
Female	35.6	33.3	52.4	32.1
RACE/ETHNICITY				
White	61.1	67.6	71.3	64.1
Black	23.1	14.0	19.4	17.6
Hispanic	11.7	12.3	6.4	12.7
Asian American	2.7	3.2	0.6	3.3
American Indian	0.9	2.9	2.3	2.0
Unclassified	0.5	0.0	0.0	0.2
REGION				
Northeast	15.0	25.0	23.7	20.5
Southeast	36.5	23.1	26.4	29.0
Central	19.9	19.5	18.1	19.9
West	28.6	32.3	31.8	30.6
TYPE OF LOCATION				
Central City	29.6	30.9	26.3	30.8
Urban Fringe/Large Town	38.3	46.3	52.2	41.8
Rural/Small Town	32.2	22.9	21.6	27.4
SCHOOL TYPE				
Public	100.0	97.8	100.0	98.6
Nonpublic	0.0	0.6	0.0	0.4
Private	0.0	0.0	0.0	0.0
Catholic	0.0	0.6	0.0	0.4
BIA	0.0	1.7	0.0	1.1
DoDea	0.0	0.0	0.0	0.0
ESTIMATED TOTAL POPULATION	121,771	167,734	29,838	259,667

Table 19-40

Weighted Percentage of Excluded Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 9

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
SEX				
Male	62.0	61.0	61.0	62.0
Female	38.0	39.0	39.0	38.0
RACE/ETHNICITY				
White	47.5	46.4	46.4	47.5
Black	19.5	13.2	13.2	19.5
Hispanic	25.4	30.7	30.7	25.4
Asian American	6.5	8.8	8.8	6.5
American Indian	0.3	0.3	0.3	0.3
Unclassified	0.8	0.6	0.6	0.8
REGION				
Northeast	12.0	11.7	11.7	12.0
Southeast	33.2	22.6	22.6	33.2
Central	18.0	15.6	15.6	18.0
West	36.8	50.0	50.0	36.8
TYPE OF LOCATION				
Central City	54.8	61.0	61.0	54.8
Urban Fringe/Large Town	27.2	30.0	30.0	27.2
Rural/Small Town	18.0	9.0	9.0	18.0
SCHOOL TYPE				
Public	98.5	99.2	99.2	98.5
Nonpublic	1.5	0.8	0.8	1.5
Private	0.0	0.0	0.0	0.0
Catholic	1.5	0.8	0.8	1.5
BIA	0.0	0.0	0.0	0.0
DoDea	0.0	0.0	0.0	0.0
ESTIMATED TOTAL POPULATION	173,491	91,294	91,294	173,491

Table 19-41

Weighted Percentage of Excluded Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 13

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
SEX				
Male	64.6	59.0	59.0	64.6
Female	35.4	41.0	41.0	35.4
RACE/ETHNICITY				
White	54.4	55.9	55.9	54.4
Black	20.1	14.1	14.1	20.1
Hispanic	21.2	26.9	26.9	21.2
Asian American	3.4	2.8	2.8	3.4
American Indian	0.3	0.3	0.3	0.3
Unclassified	0.5	0.0	0.0	0.5
REGION				
Northeast	19.9	23.1	23.1	19.9
Southeast	31.5	24.1	24.1	31.5
Central	21.0	19.0	19.0	21.0
West	27.6	33.8	33.8	27.6
TYPE OF LOCATION				
Central City	44.2	36.8	36.8	44.2
Urban Fringe/Large Town	25.9	38.3	38.3	25.9
Rural/Small Town	30.0	24.9	24.9	30.0
SCHOOL TYPE				
Public	99.6	100.0	100.0	99.6
Nonpublic	0.4	0.0	0.0	0.4
Private	0.4	0.0	0.0	0.4
Catholic	0.0	0.0	0.0	0.0
BIA	0.0	0.0	0.0	0.0
DoDea	0.0	0.0	0.0	0.0
ESTIMATED TOTAL POPULATION	146,608	43,419	43,419	146,608

Table 19-42

Weighted Percentage of Excluded Students in the Mathematics and Science Long-Term Trend Sample by Type of Eligibility and Subgroup Classification, Age 17

	Eligible by			
	Age	Grade	Age & Grade	Age or Grade
SEX				
Male	67.4	62.0	62.0	67.4
Female	32.6	38.0	38.0	32.6
RACE/ETHNICITY				
White	61.0	71.0	71.0	61.0
Black	18.6	16.1	16.1	18.6
Hispanic	15.4	10.8	10.8	15.4
Asian American	4.0	1.7	1.7	4.0
American Indian	1.0	0.4	0.4	1.0
Unclassified	0.0	0.0	0.0	0.0
REGION				
Northeast	22.1	36.4	36.4	22.1
Southeast	30.5	16.8	16.8	30.5
Central	17.9	17.6	17.6	17.9
West	29.6	29.2	29.2	29.6
TYPE OF LOCATION				
Central City	37.3	33.4	33.4	37.3
Urban Fringe/Large Town	36.3	47.7	47.7	36.3
Rural/Small Town	26.4	18.9	18.9	26.4
SCHOOL TYPE				
Public	98.6	100.0	100.0	98.6
Nonpublic	1.4	0.0	0.0	1.4
Private	1.4	0.0	0.0	1.4
Catholic	0.0	0.0	0.0	0.0
BIA	0.0	0.0	0.0	0.0
DoDea	0.0	0.0	0.0	0.0
ESTIMATED TOTAL POPULATION	131,897	38,085	38,085	131,897