NATIONAL CENTER FOR EDUCATION STATISTICS

The NAEP 1996 Technical Report

THE NATION'S REPORT CARD

U.S. Department of Education
Office of Educational Research and Improvement

NCES 1999-452

What is The Nation's Report Card?

THE NATION'S REPORT CARD, the National Assessment of Educational Progress (NAEP), is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history, geography, and other fields. By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement is collected under this program. NAEP guarantees the privacy of individual students and their families.

NAEP is a congressionally mandated project of the National Center for Education Statistics, the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations. NAEP reports directly to the Commissioner, who is also responsible for providing continuing reviews, including validation studies and solicitation of public comment, on NAEP's conduct and usefulness.

In 1988, Congress established the National Assessment Governing Board (NAGB) to formulate policy guidelines for NAEP. The Board is responsible for selecting the subject areas to be assessed from among those included in the National Education Goals; for setting appropriate student performance levels; for developing assessment objectives and test specifications through a national consensus approach; for designing the assessment methodology; for developing guidelines for reporting and disseminating NAEP results; for developing standards and procedures for interstate, regional, and national comparisons; for determining the appropriateness of test items and ensuring they are free from bias; and for taking actions to improve the form and use of the National Assessment.

The National Assessment Governing Board

NATIONAL CENTER FOR EDUCATION STATISTICS

The NAEP 1996 Technical Report

Nancy L. Allen
James E. Carlson
Christine A. Zelenak

in collaboration with
Luz Bay, Patrick B. Bourgeacq, Mary Lyn Bourque, Nancy W. Caldwell,
Hua-Hua Chang, Wen-Hung Chen, John R. Donoghue, John J. Ferris,
David S. Freund, Lucy M. Gray, Steven P. Isham, Frank Jenkins, Eugene G. Johnson,
Bruce A. Kaplan, Edward Kulick, Stephen Lazer, Jo-Lin Liang, Susan C. Loomis,
Eiji Muraki, Norma A. Norris, Ingeborg U. Novatkoski, Jiahe Qian,
Katharine E. Pashley, Timothy Robinson, Alfred M. Rogers, Keith F. Rust,
Terry L. Schoeps, Spencer S. Swinton, Bradley Thayer, Neal Thomas,
Mark M. Waksberg, Leslie Wallace, Lois H. Worthington, Jinming Zhang

July 1999

U.S. Department of Education
Office of Educational Research and Improvement **NCES 1999-452**

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
C. Kent McGuire
Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

Assessment Division
Peggy G. Carr
Associate Commissioner

July 1999

This Internet edition of *The NAEP 1996 Technical Report* includes the changes to the printed report that are being distributed as an errata sheet.

SUGGESTED CITATION
U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. *The NAEP 1996 Technical Report,* NCES 1999–452, by Allen, N.L., Carlson, J.E., & Zelenak, C.A. (1999). Washington, DC: National Center for Education Statistics.

FOR MORE INFORMATION
Content contact:
Sheida White
202–219–1675

To obtain single copies of this report, while supplies last, or ordering information on other U.S. Department of Education products, call toll free 1–877– 4ED PUBS (877–433–7827), or write:

Education Publications Center (ED Pubs)
U.S. Department of Education
P.O. Box 1398
Jessup, MD 20794–1398

TTY/TDD 1–877–576–7734
FAX 301–470–1244

Online ordering via the Internet: *http://www.ed.gov/pubs/edpubs.html*
Copies also are available in alternate formats upon request.
This report also is available on the World Wide Web: *http://nces.ed.gov/naep*

The work upon which this publication is based was performed for the National Center for Education Statistics, Office of Educational Research and Improvement, by Educational Testing Service.

THE NAEP 1996 TECHNICAL REPORT

◆ Table of Contents ◆

Chapter 13 (continued)

Chapter 17 (continued)

THE NAEP 1996 TECHNICAL REPORT

◆ List of Tables ◆

THE NAEP 1996 TECHNICAL REPORT

◆ List of Figures ◆

ACKNOWLEDGMENTS

The design, development, administration, analysis, and reporting of the 1996 National Assessment of Educational Progress (NAEP) program was a collaborative effort among staff from the National Center for Education Statistics (NCES), the National Assessment Governing Board (NAGB), the Council of Chief State School Officers (CCSSO), Educational Testing Service (ETS), Westat, Inc., and National Computer Systems (NCS). This report documents NAEP design, administration, and data analysis procedures, indicating what technical decisions were made and the rationale behind those decisions. The development of this report and of the national assessment program is the result of the considerable knowledge, experience, creativity, and dedication of many individuals. I would like to acknowledge these individuals for their contribution to NAEP.

The 1996 NAEP assessment was funded through NCES, in the Office of Educational Research and Improvement of the U.S. Department of Education. The NCES staff, played a crucial role in all aspects of the program. We are grateful for the reviews of this report contributed by Janis Brown, Arnold Goldstein Andrew Kolstad, Ralph Lee, Larry Ogle, Andrew Malizio, Marilyn McMillen, Holly Spurlock, Michael Ross, Sheida White, and Shi-Chang Wu.

I would also like to gratefully acknowledge Huynh Huynh of the University of South Carolina, Edward Kifer of the University of Kentucky, and Gary Marco of ETS for conducting a technical review of this report.

Special thanks also go to the members of the National Assessment Governing Board (NAGB) and the NAGB staff who provided advice and guidance in the preparation of this report, particularly Mary Lyn Bourque, who provided information for some of the report appendices.

ETS management and the NAEP program area within ETS's School and College Services Division have been very supportive of NAEP's technical work. Special thanks go to Nancy Cole as well as to Paul Ramsey, Henry Braun, Charles Davis, Juliet Shaffer (formerly of ETS), and James Carlson of ETS research management, and Paul Williams (formerly of ETS), Steve Lazer, John Mazzeo, Beth Durkin, and Kim Whittington of the NAEP program within ETS's School and College Services Division. Significant contributions to the project were also received from Nada Ballator, Jay Campbell, Patricia Donahue, Christine O'Sullivan, and Hilary Persky.

The guidance of the NAEP Design and Analysis Committee on technical aspects of NAEP has been outstanding. During the period of analysis of the 1996 data, the members were Sylvia Johnson (chair), Albert Beaton, Jeri Benson, Jeremy Finn, Paul Holland, Huynh Huynh, Gaea Leinhardt, David Lohman, Anthony Nitko, Ingram Olkin, Tej Pandey, Juliet Shaffer, and Hariharan Swaminathan.

The design and data analysis of the 1996 National Assessment was primarily the responsibility of the NAEP research and data analysis staff at ETS with significant contributions from NAEP management, Westat, and NCS staffs. In addition to managing day-to-day data analytic operations, these staff members have made many innovative statistical and psychometric contributions. Major contributions from the Division of Statistics and Psychometric Research were made by James Carlson, John Donoghue, Frank Jenkins, Hua-Hua Chang, Jiahe Qian, Spencer Swinton, and Jinming Zhang. Jo-lin Liang served as research associate. Eugene Johnson, Robert Mislevy, and Juliet Shaffer provided valuable statistical and psychometric advice.

The Division of Data Analysis and Technology Research at ETS, under the leadership of John Barone, was responsible for developing the operating systems and carrying out the data analyses. David Freund and Alfred Rogers developed and maintained the large and complex NAEP data management systems, and Katharine Pashley managed database activities. Alfred Rogers developed the production versions of key analysis and scaling systems. Special recognition goes to Phillip Leung and Patricia O'Reilly for their expertise in the design and development of the NAEP Web site. Thanks also go to David Freund, Steven Isham, Bruce Kaplan, Debbie Kline, and Edward Kulick for their continuing roles as leaders and developers of innovative solutions to NAEP data analysis challenges. Many other members of this division made important contributions of their time and talent to NAEP data analyses and analysis software and data products, including John Ferris, Shuyi Hua, Gerry Kokolis, Laura Jenkins, Mike Narcowich, Norma Norris, Ingeborg Novatkoski, Craig Pizzuti, Steve Szyszkiewicz, Xiaohui Wang, and Lois Worthington.

The staff at Westat contributed their talents and efforts in all areas of the sample design and data collection. Particular recognition is due to Leslie Wallace, Lucy Gray, Renee Slobasky, Nancy Caldwell, Keith Rust, Mark Waksberg, Debby Vivari, and Dianne Walsh for directing the sampling and data collection activities. Thanks are also due to John Burke, James Green, and Penny James.

Critical to the program was the contribution of NCS, which has been responsible for the printing, distribution, and processing of the assessment materials, as well as an increased role in professional scoring of constructed-response items. The leadership roles of Patrick Bourgeacq, Charles Brungardt, Bradley Thayer, and Timothy Robinson are especially acknowledged. Thanks also go to Brent Studer.

Joan Stoeckel, Sharon Stewart, and Donna Lembeck are acknowledged for their administrative assistance during the preparation of this report.

Carol Errickson also deserves special thanks for coordinating the design of the report's cover.

Special recognition and appreciation go to Christine Zelenak, Terry Schoeps, and Debbie Kline, editors of this report. They have been responsible for organizing, scheduling, editing, motivating, and ensuring the cohesiveness and correctness of the final report.

Finally, NAEP is grateful to the students and school staff members who participated in the assessment. Without their efforts, there would be no assessment.

Nancy L. Allen
Director, NAEP Analysis: Planning and National Analyses
NAEP Research, ETS
July 1999

THE NAEP 1996 TECHNICAL REPORT

Introduction[1]

James E. Carlson
Educational Testing Service

The 1996 National Assessment of Educational Progress (NAEP) monitored the performance of students in American schools in the subject areas of reading, mathematics, science, and writing. The national sample involved nearly 124,000 public and nonpublic-school students who were 9-, 13-, or 17 years old or in grades 4, 8, or 12.

The purpose of this technical report is to provide details on the instrument development, sample design, data collection, and data analysis procedures of the 1996 assessment. Detailed substantive results are not presented here but can be found in a series of NAEP reports on the status of and trends in student performance; several additional reports provide information on how the assessment was designed and implemented. The reader is directed to the following reports for 1996 results and supporting documentation:

- *NAEP 1996 Mathematics Report Card for the Nation and the States: Findings from the National Assessment of Educational Progress* (Reese, Miller, Mazzeo, & Dossey, 1997)
- *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 NAEP Guide: A Description of the Content and Methods of the 1994 and 1996 Assessments* (NAEP, 1996)
- *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)
- *NAEP 1996 Mathematics Cross-State Data Compendium for the Grade 4 and Grade 8 Assessment* (Shaughnessy, Nelson, & Norris, 1997)
- *NAEP 1996 Science Cross-State Data Compendium for the Grade 8 Assessment* (Keiser, Nelson, Norris, & Szyszkiewicz, 1998)
- *Mathematics Framework for the 1996 National Assessment of Educational Progress* (National Assessment Governing Board, 1994)
- *Science Framework for the 1996 National Assessment of Educational Progress* (National Assessment Governing Board, 1993)

[1] James E. Carlson is responsible for psychometric and statistical analyses of NAEP.

- *Technical Report of the NAEP 1996 State Assessment Program in Mathematics* (Allen, Jenkins, Kulick, & Zelenak, 1997)

- *Technical Report of the NAEP 1996 State Assessment Program in Science* (Allen, Swinton, Isham, & Zelenak, 1998)

- *NAEP 1996 National Assessment Secondary-Use Data Files User Guide* (Rogers, Kline, & Schoeps, 1999)

- *NAEP 1996 State Assessment Program in Mathematics Secondary-Use Data Files User Guide* (O'Reilly, Zelenak, Rogers, & Kline, 1999)

- *NAEP 1996 State Assessment Program in Science Secondary-Use Data Files User Guide* (O'Reilly, Zelenak, Rogers, & Kline, 1999)

- *NAEP 1996 Science Performance Standards: Achievement Results for the Nation and the States* (Bourque, Champagne, & Crissman, 1997)

- *School Policies Affecting Instruction in Mathematics: Findings from the National Assessment of Educational Progress* (Hawkins, Stancavage, & Dossey, 1998)

- *Student Work and Teacher Practices in Mathematics* (Mitchell, Hawkins, Jakwerth, Stancavage, & Dossey, 1999)

- *Estimation Skills, Mathematics-in-context, and Advanced Skills in Mathematics* (Hawkins, Mitchell, Stancavage, & Dossey, 1999)

- *Students Learning Science: A Report on Policies and Practices in U.S. Schools* (O'Sullivan, Weiss, & Askew, 1998)

- *Student Work and Teacher Practices in Science: A Report on What Students Know and Can Do* (O'Sullivan & Weiss, 1999)

- *The 1996 NAEP Sampling and Weighting Report* (Wallace & Rust, 1999)

- *Report on Data Collection Activities for the 1996 National Assessment of Educational Progress* (Westat, Inc., 1996)

- *Report of Processing and Professional Activities* (National Computer Systems, 1996)

The *Report Card* publications highlight results for the nation, states, and selected subgroups. Reports on student work and teacher practices focus on instructional variables related to mathematics and science education and are designed to meet the information needs of teachers and curriculum specialists. The aim of the reports on school policies, which focus on instruction-relevant variables from the school or community level, is to meet the information needs of principals, school boards, and interested citizens. Technical and other reports listed above provide more detailed information on the NAEP data and analysis procedures. Many of the NAEP reports, including the almanacs (summary data tables), are also available on the Internet at *http://nces.ed.gov/naep.* For ordering information on printed copies of these

reports, go to the Department of Education web page *http://www.ed.gov/pubs/edpubs.html*, call toll free 1–877– 4ED PUBS (877–433–7827), or write to:

Education Publications Center (ED Pubs)
U.S. Department of Education
P.O. Box 1398
Jessup, MD 20794–1398

The *Frameworks* are designed to assess the outcomes of students' education in mathematics and science in grade 4, 8, and 12 as part of NAEP. For ordering information on these reports, write:

National Assessment Governing Board
800 North Capitol Street NW
Suite 825
Washington, DC 20002

The *Frameworks* and other NAGB documents are also available through the web at *http://www.nagb.org*.

Additional samples of approximately 125,000 fourth- and 125,000 eighth-graders in 48 jurisdictions were assessed in the 1996 state assessment in mathematics. Also a sample of approximately 125,000 fourth-graders in 47 states and jurisdictions was assessed as part of the 1996 state assessment in science. A representative sample of about 2,500 students was selected in each jurisdiction for each subject at each grade level. The state-level sampling plan allowed for cross-state comparisons and comparisons with the nation in fourth-grade science and fourth- and eighth-grade mathematics achievement. Technical details of the state assessments are not presented in this technical report but can be found in the state technical reports.

AN OVERVIEW OF NAEP IN 1996

For the 1996 assessment, NAEP researchers continued to build on the original design technology outlined in *A New Design for a New Era* (Messick, Beaton, & Lord, 1983). In order to maintain its links to the past and still implement innovations in measurement technology, NAEP continued its multistage sampling approach. Long-term trend and short-term trend samples use the same methodology and population definitions as in previous assessments. Main assessment samples use innovations associated with new NAEP technology and address current educational issues. Long-term trend data are used to estimate changes in performance from previous assessments; main assessment sample data are used primarily for analyses involving the current student population, but also to estimate short-term trends for a small number of recent assessments. In continuing to use this two-tiered approach, NAEP reaffirms its commitment to maintaining long-term trends while at the same time implementing the latest in measurement technology.

A major new design feature was introduced for 1996 to permit the introduction of new inclusion rules for students with disabilities (SD) and limited English proficient (LEP) students, and the introduction of testing accommodations for those students. The 1996 national NAEP incorporated a multiple sampling plan that allowed for the study of changes in NAEP inclusion and accommodation procedures. In order to provide for studies of the effects of these changes, students from different samples were administered the NAEP instruments using different sets of inclusion rules and accommodation procedures. Testing accommodations were provided for SD and LEP students in certain samples who could be assessed, but not with standard instruments or administration procedures.

In the 1996 assessment, many of the innovations that were implemented for the first time in 1988 were continued and enhanced. For example, a variant of the focused balanced incomplete block (focused-BIB) booklet design that was used in 1988 and has continued to be used in other assessment years, was used in the 1996 main assessment samples in mathematics and science. In the focused-BIB design, an individual receives blocks of cognitive items in the same subject area. The focused-BIB design allows for improved estimation within a particular subject area, and estimation continues to be optimized for groups rather than individuals.

In 1996, NAEP continued to apply the plausible values approach to estimating means for demographic as well as curriculum-related subgroups. Proficiency estimates were based on draws from a posterior distribution that was based on an optimum weighting of two sets of information: the student's responses to cognitive items, and his or her demographic and associated educational process variables. This Bayesian procedure was developed by Mislevy (see Chapter 11 or Mislevy, 1991). The 1996 procedures continued to use an improvement that was implemented first in 1988 and refined for the 1994 assessment. This is a multivariate procedure that uses information from all scales within a given subject area in the estimation of the proficiency distribution on any one scale in that subject area.

A major improvement used in the 1992 and 1994 assessments, and continued in 1996, was the use of the generalized partial credit model for item response theory (IRT) scaling. This allowed the incorporation of constructed-response questions that are scored on a multipoint rating scale into the NAEP scale in a way that utilizes the information available in each response category.

One important innovation in reporting the 1990 assessment data that was continued through 1996 was the use of simultaneous comparison procedures in carrying out significance tests for the differences across assessment years. Methods such as the Bonferroni allow one to control for the type I error rate for a fixed number of comparisons. In 1996, a new procedure that provided more powerful procedures that control for the false discovery rate were implemented for some comparisons. Tests for linear and quadratic trends were also applied to the national trend data in reading, mathematics, science, and writing.

ORGANIZATION OF THE TECHNICAL REPORT

Part I of this report presents the details of the design of the 1996 National Assessment, summarized in Chapter 1. Chapters 2 through 8 describe the development of the objectives and the items used in the assessment, the sample selection procedures, the assessment booklets and questionnaires, the administration of the assessment in the field, the processing of the data from the assessment instruments into computer-readable form, the professional scoring of constructed-response items, and the methods used to create a complete NAEP database.

The 1996 NAEP data analysis procedures are described in Part II of the report. Chapter 9 provides a summary of the analysis steps. Subsequent chapters provide a general discussion of the weighting and variance estimation procedures used in NAEP, an overview of NAEP scaling methodology, and details of the trend and main assessment analyses performed for each subject area in the 1996 assessment.

Chapter 19 presents basic data from the 1996 assessment, including the properties of the measuring instruments and characteristics of the sample.

Chapter 1

OVERVIEW OF PART I: THE DESIGN AND IMPLEMENTATION OF THE 1996 NAEP[1]

Nancy L. Allen and Eugene G. Johnson
Educational Testing Service

1.1 INTRODUCTION

The 1996 National Assessment collected information on the knowledge, skills, and attitudes of young Americans in mathematics, science, reading, and writing. The three components of the National Assessment were the main assessments of mathematics and science, the long-term trend assessments of mathematics, science, reading, and writing, and special assessments of aspects of mathematics and science. The basis for the information collected for the National Assessment was a complex sample survey involving nearly 124,000 students, consisting of national samples of public- and nonpublic-school students who were in grades 4, 8, and 12 or were 9-, 13-, or 17-year olds. Additional NAEP data came from the State Assessment program, which in 1996 assessed mathematics at grades 4 and 8 in representative samples of public- and nonpublic-school students in 44 states, the District of Columbia, Department of Defense Domestic Dependent Elementary and Secondary Schools (DDESS), Department of Defense Dependents Schools (DoDDS), and Guam. The 1996 State Assessment program also assessed science at grade 8 in representative samples of public- and nonpublic-school students in 43 states, the District of Columbia, DDESS, DoDDS, and Guam; DDESS and DoDDS fourth-grade students were assessed as part of a separate special science assessment.

This chapter describes the design for the 1996 assessment and gives an overview of the steps involved in its implementation, from the planning stage through the creation of edited data files. The major components of the implementation are presented here with references to other chapters in Part I that provide greater detail on each aspect of the assessment. The procedures used for the analysis of the data are summarized in the overview to Part II and discussed in detail in the remaining chapters in that part of the report. Excluded from this technical report are the details of the design and analysis of the 1996 State Assessments, which instead appear in the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics* (Allen, Jenkins, Kulick, & Zelenak, 1997), and in the *Technical Report of the NAEP 1996 State Assessment Program in Science* (Allen, Swinton, Isham, & Zelenak, 1998). Also excluded are the details of the analyses of special studies of advanced mathematics and science students and of students receiving special theme-related and estimation mathematics items. The analyses will be described in the appendices of the reports containing the results of these special studies.

[1] Nancy L. Allen is responsible for the psychometric and statistical analysis of national and state NAEP data. Eugene G. Johnson is a senior psychometrician, contributing to the design of NAEP and to discussions of sampling issues. Previously, he was responsible for the psychometric and statistical analysis of NAEP data. The authors are indebted to the authors of Chapters 2 through 8 for portions of this chapter.

The organization of this chapter, and of Part I, is as follows:

- Section 1.2 provides an overview of the NAEP design for 1996 and includes a description of the constituent samples. To provide background information, the section also includes the assessment schedule from the inception of NAEP in 1969 through the 1996 assessment.

- Section 1.3 provides a summary of the development of the objectives for each subject area in the assessment and a description of the development and review of the items written to fit those objectives. Details of the objective and item development processes appear in Chapter 2.

- Section 1.4 provides a summary of the sampling design used for the 1996 assessment with a fuller description provided in Chapter 3.

- Section 1.5 includes a discussion of the assignment of the cognitive and background questions to assessment booklets and a description of the complex block designs that were the basis for assigning cognitive items to assessment booklets and assessment booklets to individuals. Chapter 4 provides a detailed description of the assessment booklets.

- Section 1.6 provides a summary of the field administration procedures, including the processes of training field administrators, attaining school cooperation, administering the assessment, and conducting quality control. Further details appear in Chapter 5.

- Section 1.7 includes a description of the flow of data from the receipt of the assessment materials through data entry, validation, and resolution to the creation of edited data files. Chapter 6 provides a detailed description of the process.

- Section 1.8 contains a discussion of the professional scoring of students' responses to the constructed-response items in the assessment. Details of the process are given in Chapter 7.

- Section 1.9 provides a summary of the creation of the database, the quality control of data entry, and lists the 1996 database products. This section also includes a description of the use of the World Wide Web for dissemination of NAEP information. Further details appear in Chapter 8.

1.2 THE 1996 NAEP DESIGN

A major purpose of NAEP is the reliable measurement of trends in educational achievement over time. To do this well, confounding effects due to changes from one assessment to the next in assessment instrumentation or in assessment procedures must be minimized. This implies a stability in the measurement process over time. At the same time, the assessment must remain current by allowing the introduction of new curriculum concepts and changes in educational priorities and by permitting the use of new measurement technology. The objectives for an assessment are determined through a consensus process in which committees of subject matter experts, scholars, and citizens representing many diverse

constituencies and points of view are assembled to determine the educational goals that students should achieve. Satisfying these objectives often requires changes in assessment instrumentation and methodology.

In order to meet the goals of measuring trends reliably and responding to changes in the current thinking about subject areas, NAEP has instituted a multicomponent assessment system where each component is itself a set of assessments designed to accomplish a specific goal. There are three components in the 1996 National Assessment design: (1) main assessments; (2) assessments for long-term trend; and (3) special assessments. In particular, the main assessments respond to changes in curriculum on a regular basis, while the long-term trend assessments are not changed and measure longer-term trends in a valid way. These are discussed in detail in this chapter.

Several improvements were made in the design of NAEP in the 1984 and succeeding assessments. Until the 1984 assessment, NAEP was administered using matrix sampling and tape recorders; that is, by administering booklets of exercises using an aurally presented stimulus that paced groups of students through the individual assessment exercises in a common booklet. In the 1984 assessment, BIB spiraling, which does not include aural pacing, was introduced in place of taped matrix sampling. The NAEP design now includes sampling grade populations as well as the age populations that NAEP originally assessed. The definitions of student age and the time of year in which the assessment takes place have been made uniform so that students in the fourth-, eighth-, and twelfth-grades are assessed. To shorten the timetable for reporting results, the period for national data collection was decreased in 1992, 1994, and 1996 from the five-month period used in 1990 to a three-month period in the winter (corresponding to the period used for the winter half-sample of the 1990 National Assessment). To enhance the coverage of the subject areas assessed, the number of items measuring knowledge and skills was increased for the 1992, 1994, and 1996 assessments.

A special feature of the 1996 main and state assessments was the introduction of new rules for inclusion of students with disabilities (SD) and limited English proficient (LEP) students in NAEP assessments (presented in Chapter 5). A subsample of the schools selected for participation in the 1996 assessments used the old inclusion rules (sample type 1; S1) to determine whether students should be included in the assessment, and another subsample used the new inclusion rules (sample type 2; S2). In addition to the two groups of schools using the old and new inclusion rules without offering students special testing accommodations, the 1996 main assessments included a third group of schools that used the new inclusion rules and offered students within those schools accommodations to the standard NAEP administration procedures (sample type 3; S3). Figure 1-1 contains the layout of the pieces of the sample collected for each grade of the main assessment of mathematics.

The accommodations provided by NAEP in the main assessments were meant to match those specified in the student's individualized education program (IEP) or those ordinarily provided in the classroom for testing situations. The most common accommodation was extended time. The samples of students from the third group of schools that used the new inclusion rules and offered students accommodations were not included for most analysis and reporting purposes, although the results for these samples were studied in follow-up analyses. In the State Assessments, no special accommodations were offered. The information in Chapters 3 and 5 applies to schools and students in all of the sample types, while the data analysis chapters reflect schools and students in reporting samples only.

The new inclusion rules are applied only when a student has been categorized in his or her IEP as a student with disabilities (SD) or as a limited English proficient (LEP) student; all other students were asked to participate in the assessment. For this reason, the sample of students that was selected for most analysis and reporting purposes for the main mathematics assessment consisted of students from two

Figure 1-1
Subsamples of the Mathematics Assessment:1996

GROUPS OF STUDENTS	GROUPS OF SCHOOLS		
	Sample Type 1 OLD INCLUSION RULES - NO ACCOMMODATIONS -	Sample Type 2 NEW INCLUSION RULES - NO ACCOMMODATIONS -	Sample Type 3 NEW INCLUSION RULES - ACCOMMODATIONS -
NOT SD/LEP[1]	A_1	A_2	A_3[2]
INCLUDED SD/LEP[1]	B_1	B_2[2]	B_3[2]
EXCLUDED SD/LEP[1]	C_1[3]	C_2[3]	C_3[3]

[1] Students with Disabilities/Limited English Proficient
[2] Results for students in subsamples B_2, A_3, and B_3 were not reported in *NAEP 1996 Mathematics: Report Card for the Nation and the States*.
[3] Students in subsamples C_1, C_2, and C_3 were not included in the assessment.

groups: those who were not categorized as SD or LEP students and who were from schools providing no accommodations and used either set of inclusion rules (A_1 and A_2 in Figure 1-1); and those who were categorized as SD or LEP students and who were from schools providing no accommodations and using only the old inclusion rules (B_1 in Figure 1-1). The advantage of this reporting sample is that it preserves trend with previous assessments and it makes use of most of the data from the assessment. The main science assessment sample did not include students from schools using the old inclusion rules. The sample of students that was selected for most analysis and reporting purposes for the main science assessment consisted of students from the schools using new inclusion rules, but not providing any accommodations (A_2 and B_2 in Figure 1-1). The advantage of this reporting sample is that it makes use of the most up-to-date inclusion rules and begins a science trend line.

Special analyses that used the national science and mathematics assessment data to compare the old and new inclusion rules and examine the effect of offering testing accommodations indicated little difference in proportions of students included in the assessment. More students were included in the assessment when they were offered accommodations; however, a portion of students who would have participated in the assessment under standard conditions was assessed with accommodations when they were offered. A result of this is that fewer students were assessed under standard conditions when accommodations were offered. The students from the schools offering accommodations were not included in the analyses or results contributing to the Mathematics or Science Report Cards, so they did not affect the measurement of the 1990, 1992, and 1996 trend for mathematics. The results from the science assessment were not compared to those from previous assessments.

NAEP's design for 1996 required collecting 24 different samples in order to conduct the assessments. The various samples collected and reported for the 1996 assessment are summarized in Table 1-1.

Table 1-1
NAEP 1996 Student Samples

Sample	Booklet IDs	Mode	Cohort Assessed	Time of Testing[1]	Age Definition	Modal Grade	Reporting Sample Size[2]
9[Math-MainP]	101-129, 921	Print	Grade 4	1/3/96 - 3/29/96			6,627
13[Math-MainP]	101-130, 921	Print	Grade 8	1/3/96 - 3/29/96			7,146
17[Math-MainP]	101-130	Print	Grade 12	1/3/96 - 3/29/96			6,904
4 [Math-Estimation]	127	Tape	Grade 4	1/3/96 - 3/29/96			2,023
8 [Math-Estimation]	127	Tape	Grade 8	1/3/96 - 3/29/96			2,183
12 [Math-Estimation]	127	Tape	Grade 12	1/3/96 - 3/29/96			1,849
4 [Math-Theme]	128, 129	Print	Grade 4	1/3/96 - 3/29/96			3,790
8 [Math-Theme]	128, 129	Print	Grade 8	1/3/96 - 3/29/96			4,027
12 [Math-Theme]	128, 129	Print	Grade 12	1/3/96 - 3/29/96			3,735
8 [Math-Advanced]	130	Print	Grade 8	1/3/96 - 3/29/96			2,337
12 [Math-Advanced]	130	Print	Grade 12	1/3/96 - 3/29/96			2,965
9[Sci-MainP]	201-237	Print	Grade 4	1/3/96 - 3/29/96			7,305
13[Sci-MainP]	201-237	Print	Grade 8	1/3/96 - 3/29/96			7,774
17[Sci-MainP]	201-240	Print	Grade 12	1/3/96 - 3/29/96			7,537
12 [Sci-Advanced]	238-240	Print	Grade 12	1/3/96 - 3/29/96			2,431
9[Math-State]	101-126	Print	Grade 4	1/29/96 - 3/4/96			*[3]
13[Math-State]	101-126	Print	Grade 8	1/29/96 - 3/4/96			*[3]
13[Sci-State]	201-237	Print	Grade 8	1/29/96 - 3/4/96			*[3]
9[RW-LTTrend]	51 - 56	Print	Age 9/Grade 4	1/3/96 - 3/8/96	CY	4	5,019
13[RW-LTTrend]	51 - 56	Print	Age 13/Grade 8	10/9/95 - 12/22/95	CY	8	5,493
17[RW-LTTrend]	51 - 56	Print	Age 17/Grade 11	3/11/96 - 5/10/96	Not CY	11	4,669
9[MS-LTTrend]	91 - 93	Tape	Age 9	1/3/96 - 3/8/96	CY	4	5,414
13[MS-LTTrend]	91 – 93	Tape	Age 13	10/9/95 - 12/22/95	CY	8	5,658
17[MS-LTTrend]	84 – 85	Tape	Age 17	3/11/96 - 5/10/96	Not CY	11	3,539

[1] Final makeup sessions for the winter session (January 3-March 29, 1996) were held April 1-5, 1996.
[2] The total number of students assessed in the reporting sample of the national assessment was 43,293 for the main assessment, 29,792 for the long-term trend assessment, and 25,340 for the special studies.
[3] Note: consists of distinct samples in 48 jurisdictions

LEGEND:	Math	Mathematics	Main	Main assessment, print administration
	Sci	Science	LTTrend	Long-term trend assessment
	RW	Reading and writing	CY	Calendar year: birthdates in 1986 and 1982 respectively for ages 9 and 13
	MS	Mathematics and science		
	Print	Print administration	Not CY	Age 17 only: birthdates between October 1, 1978 and September 30, 1979
	Tape	Audiotape administration		

Each row of Table 1-1 corresponds to a particular sample and each column of the table indicates the following major features of that sample:

1. *Sample* is the sample identifier. The first part of the sample code is a number (the age class) representing the student cohort included in the sample (note that this part of the code does not indicate whether an age or grade sample was selected); the second part, in brackets, denotes the specific sample type. For example, 9[Math Main] is a main

assessment mathematics sample for grade 4, assessed in print mode. A full description of the purposes for the various sample types is given below.

2. *Booklets* gives the identifier numbers for the booklets used for the assessment of the particular sample.

3. *Mode* indicates the mode of assessment, which may be print or tape. NAEP originally assessed students using a tape recorder in addition to booklets, thus pacing the students through exercises at a fixed rate. In 1996, NAEP used a paced audiotape for its mathematics and science long-term trend assessments. However, most other assessments in 1996 used printed instructions with the student expected to read the exercises. The only other exception was the 1996 assessment of mathematics estimation skills.

4. The *cohort assessed* denotes the age, grade, or age/grade of the population being sampled. For example, *grade 4* represents students who are in the fourth grade; an *age 17* cohort consists of students (in any grade) who are 17 years old. Samples for the 1996 main assessments were selected on the basis of grade only. The traditional NAEP samples used in long-term trend estimation were defined by age only. However, the 1996 reading and writing long-term trend assessments were defined by being either of a particular age *or* of the modal grade for students of that age. For reading and writing, results are reported for grade and age samples, respectively. The definitions of age, and thus the corresponding grade, have changed in ways that are described in Section 1.2.1.

5. *Time of testing* indicates the time of year in which the assessment is performed. NAEP traditionally assessed 9-year-olds in the winter, 13-year-olds in the fall, and 17-year-olds in the spring; like the 1994 main assessment, in 1996, all grades were assessed in the winter (between January 3-March 29, 1996; final makeup sessions were held April 1-5, 1996).

6. *Age definition* is denoted as calendar year (CY) or not calendar year (Not CY). NAEP originally defined age by birth within a calendar year at ages 9 and 13 but defined age 17 as being born between October 1 of one year and September 30 of the next. In the 1996 main assessments, no students were selected on the basis of their age.

7. The *modal grade* is the grade attended by most of the students of the sampled age. For example, if an age 17 sample is listed as having a modal grade of 11, then most of the 17-year-old students, as defined, are in the eleventh grade. The definition of age affects the modal grade of the sample. All students sampled for the 1996 main assessments were in the grade defined by the cohort assessed.
8. The *reporting sample size* is the number of students in the sample who were actually administered the assessment and whose results were used in the NAEP subject-area reports.

1.2.1 The 1996 NAEP Samples

The NAEP samples in 1996 consisted of four types: the samples from the National Assessment—the main NAEP samples, the long-term trend samples, and the special studies samples—and the State Assessment samples.

***The Main NAEP Samples*.** The main NAEP samples are labeled in Table 1-1 as [Math Main] and [Science Main]. The samples used complex spiraling procedures (defined in Section 1.5), and were intended to form the basis for future assessments. Each sample was assessed in the winter period (January 3 through March 29, 1996). In these samples, only grade populations were sampled, although age/grade populations were assessed in previous assessment years. The main NAEP samples, and their purposes, are as follows:

> [Math Main] are grades 4, 8, and 12 mathematics assessment samples used for measuring mathematics achievement in 1996. The fourth- and eighth-grade samples also provided the comparison groups for the 1996 State Assessment of mathematics in grades 4 and 8. These samples used print administration.
>
> [Science Main] are grades 4, 8, and 12 science assessment samples used for measuring science achievement in 1996. The eighth-grade samples also provided the comparison groups for the 1996 State Assessment in science in grade 8. These samples used print administration.

The Long-Term Trend Samples. The long-term trend samples are labeled as [RW-LTTrend] and [MS-LTTrend] in Table 1-1. Each sample was defined in the same way as equivalent samples in previous assessments and used the same assessment technology as was used in those assessments. Therefore, the long-term trend samples are directly comparable to those from previous assessments and so can be used for continuing the NAEP long-term trend lines. Because these samples were designed to link the 1996 data with data from previous assessments, they are also referred to as bridge samples. The long-term trend samples and their purposes are as follows:

> [RW-LTTrend] are age/grade samples used for estimating long-term trends in reading and writing. These samples used assessment booklets identical to those initially used in 1984 and subsequently used in 1988, 1990, 1992, and 1994 (many of the items were also used in pre-1984 assessments). As in 1984, 1988, 1990, 1992, and 1994, print administration was used. These samples used the age definitions and time of testing originally used by NAEP in the 1970s and the early 1980s. The estimates of reading achievement from these samples link to eight previous reading assessments (1971, 1975, 1980, 1984, 1988, 1990, 1992, and 1994); the estimates of writing achievement link to five previous writing assessments (1984, 1988, 1990, 1992, and 1994). *[Please note that a review of the 1996 long-term trend writing assessment data analyses has been undertaken by NCES. Additional analyses may be required before revised results are released.]*
>
> [MS-LTTrend] are age-only samples used for estimating long-term trends in mathematics and science achievement. These samples used the same age definitions and time of testing as were used since 1969 and used the same assessment instruments as were used in the 1986, 1990, 1992, and 1994 long-term trend assessments of mathematics and science. As in previous assessments, the administration of the mathematics and science questions was paced with an audiotape. The estimates of science achievement from these samples link to eight previous science assessments (1970, 1973, 1977, 1982, 1986, 1990, 1992, and 1994); the estimates of mathematics achievement link to seven previous assessments (1973, 1978, 1982, 1986, 1990, 1992, and 1994).

The Special Studies Samples. Three sets of samples were collected as part of special NAEP studies. The samples used special innovative procedures to allow the study of specific aspects of mathematics and science. Each sample was assessed in the winter period (January 3 through March 29, 1996). In these samples, only grade populations were sampled. The special studies samples, and their purposes, are as follows:

> [Math-Estimation] are samples of specially selected students in grades 4, 8, and 12 who were administered mathematics estimation booklets in separate paced-tape sessions. The students are representative of the fourth-, eighth-, and twelfth-grade students in the nation.
>
> [Math-Theme] are samples of specially selected students in grades 4, 8, and 12 who were administered mathematics theme booklets. These samples were assessed in print administrations. The students were selected to represent the national populations of fourth-, eighth-, and twelfth-grade students. The students in these samples were assessed in separate sessions.
>
> [Math-Advanced] and [Sci-Advanced] are samples of specially selected students in grade 8 (for mathematics only) and grade 12 (for mathematics and science) who received advanced mathematics and science booklets. They were assessed in separate sessions. The students were selected from students who were taking advanced courses in mathematics or science.

The State Assessment Samples. In Table 1-1, 9[Math-State], 13[Math-State], and 13[Sci-State] are samples of fourth- and eighth-grade public- and nonpublic-school students from each of the states and jurisdictions participating in the 1996 State Assessment of mathematics, and eighth-grade public- and nonpublic-school students from each of the states and jurisdictions participating in the 1996 State Assessments of science.[2] The assessment booklets were the same print-administered booklets as those used for the matching samples 9[Math Main], 13[Math Main], and 13[Science Main] but the administrative procedures varied from that of the national assessment in that state personnel collected the data.

1.2.2 NAEP Assessments Since 1969

Table 1-2 shows the subject areas, grades, and ages assessed since the NAEP project began in 1969. As can be seen, in addition to the 1996 subject areas of mathematics, science, reading, and writing, several other subject areas have been assessed over the years—social studies, U.S. history, civics, citizenship, geography, literature, music, career development, art, and computer competence. Many subject areas are reassessed periodically to measure trends over time.

Assessments were conducted annually through 1980, but budget restrictions since then have reduced data collection to a biennial basis. Since its inception, NAEP has assessed 9-year-olds, 13-year-olds, and in-school 17-year-olds, although the age definitions changed in 1986 and again in 1988. Because of budget restrictions, NAEP no longer routinely assesses out-of-school 17-year-olds or young adults. (A separate assessment of young adults of ages 21 to 25 was conducted in 1985 under a separate grant.)

[2] Fourth-grade students in DDESS and DoDDS schools were assessed as a separate special science assessment.

Table 1-2

National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969-1996

		Grades/Ages Assessed										
Assessment Year	**Subject Area(s)**	**Grade 3**	**Grade 4**	**Age 9**	**Grade 7**	**Grade 8**	**Age 13**	**Grade 11**	**Grade 12**	**Age 17**	**Age 17OS[1]**	**Adult**
1969-70	Science			X			X			X	X	X
	Writing			X			X			X	X	X
	Citizenship			X			X			X	X	X
1970-71	Reading			X			X			X	X	X
	Literature			X			X			X	X	X
1971-72	Music			X			X			X	X	X
	Social Studies			X			X			X	X	X
1972-73	Science			X			X			X	X	X
	Mathematics			X			X			X	X	X
1973-74	Career and Occupational Development			X			X			X	X	X
	Writing			X			X			X	X	
1974-75	Reading			X			X			X	X	
	Art			X			X			X	X	
	Index of Basic Skills									X	X	
1975-76	Citizenship/Social Studies			X			X			X	X	
1976-77	Science			X			X			X		
	Basic Life Skills[2]									X		
	Sci/ RD / Energy / Health[2]											X
1977-78	Mathematics			X			X			X		
	Consumer Skills[2]									X		
1978-79	Writing, Art, and Music			X			X			X		
1979-80	Reading/Literature			X			X			X	X	
	Art						X					
1981-82	Science[2]			X			X			X		
	Mathematics and Citizenship/Social Studies			X			X			X		
1984	Reading		X	X		X	X	X		X		
	Writing		X	X		X	X	X		X		
	Writing (long-term trend)		X	X		X	X	X		X		
1985	Adult Literacy[2]											X
1986	Reading	X		X	X		X	X		X		
	Mathematics	X		X	X		X	X		X		
	Science	X		X	X		X	X		X		
	Computer Competence	X		X	X		X	X		X		
	U.S. History[2]							X		X		
	Literature[2]							X		X		
	Reading (long-term trend)		X	X		X	X	X		X		
	Mathematics (long-term trend)		X	X		X	X	X		X		
	Science (long-term trend)		X	X		X	X	X		X		

[1] *Age 17 students who had dropped out of school or had graduated prior to assessment.*
[2] *Small, special-interest assessment conducted on limited samples at specific grades or ages.*

(continued)

Table 1-2 (continued)

National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969-1996

		Grades/Ages Assessed										
Assessment Year[3]	**Subject Area(s)**	**Grade 3**	**Grade 4**	**Age 9**	**Grade 7**	**Grade 8**	**Age 13**	**Grade 11**	**Grade 12**	**Age 17**	**Age 17OS[1]**	**Adult**
1988	Reading		X	X		X	X		X	X		
	Writing		X	X		X	X		X	X		
	Civics		X	X		X	X		X	X		
	U.S. History		X	X		X	X		X	X		
	Document Literacy[2]		X	X		X	X		X	X		
	Geography[2]		X	X		X	X		X	X		
	Reading (long-term trend)		X	X		X	X	X		X		
	Writing (long-term trend)		X	X		X	X	X		X		
	Mathematics (long-term trend)			X			X	X		X		
	Science (long-term trend)			X			X	X		X		
	Civics (long-term trend)						X			X		
1990	Reading		X	X		X	X		X	X		
	Mathematics		X	X		X	X		X	X		
	Science		X	X		X	X		X	X		
	Reading (long-term trend)		X	X		X	X	X		X		
	Writing (long-term trend)		X	X		X	X	X		X		
	Mathematics (long-term trend)			X			X			X		
	Science (long-term trend)			X			X			X		
	Trial State Mathematics					X						
1992	Reading		X	X		X	X		X	X		
	Writing		X	X		X	X		X	X		
	Mathematics		X	X		X	X		X	X		
	Reading (long-term trend)		X	X		X	X	X		X		
	Writing (long-term trend)		X	X		X	X	X		X		
	Mathematics (long-term trend)			X			X			X		
	Science (long-term trend)			X			X			X		
	Trial State Mathematics		X			X						
	Trial State Reading		X									
1994	Reading		X	X		X	X		X	X		
	U.S. History		X	X		X	X		X	X		
	Geography		X	X		X	X		X	X		
	Reading (long-term trend)		X	X		X	X	X		X		
	Writing (long-term trend)		X	X		X	X	X		X		
	Mathematics (long-term trend)			X			X			X		
	Science (long-term trend)			X			X			X		
	Trial State Reading		X									

[1] Age 17 students who had dropped out of school or had graduated prior to assessment.
[2] Small, special-interest assessment conducted on limited samples at specific grades or ages.
[3] It should be noted that somewhat different age definitions were used in the 1984, 1986, and 1988 assessments. In the 1984 assessment, the two younger ages were defined on a calendar-year basis, while the 17-year-olds were defined on an October 1 to September 30 basis. This resulted in modal grades of 4, 8, and 11. To allow for age cohorts that were exactly four years apart, in the 1986 main assessment all ages were defined on an October 1 to September 30 basis, resulting in modal grades of 3, 7, and 11. Special studies (Kaplan, Beaton, Johnson, & Johnson, 1988) were conducted to measure the effect of the changes in age definition. Because of problems encountered in assessing third graders, in 1988 the ages were redefined on a calendar-year basis, with the modal grades being 4, 8, and 12. These were the age definitions used in the 1990, 1992, 1994, and 1996 main assessments.

(continued)

Table 1-2 (continued)
National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969-1996

		Grades/Ages Assessed										
Assessment Year[3]	**Subject Area(s)**	**Grade 3**	**Grade 4**	**Age 9**	**Grade 7**	**Grade 8**	**Age 13**	**Grade 11**	**Grade 12**	**Age 17**	**Age 17OS[1]**	**Adult**
1996	Mathematics		X			X			X			
	Science		X			X			X			
	Reading (long-term trend)		X	X		X	X	X		X		
	Writing (long-term trend)		X	X		X	X	X		X		
	Mathematics (long-term trend)			X			X			X		
	Science (long-term trend)			X			X			X		
	State Mathematics		X			X						
	State Science		X			X						

[1] Age 17 students who had dropped out of school or had graduated prior to assessment.
[3] It should be noted that somewhat different age definitions were used in the 1984, 1986, and 1988 assessments. In the 1984 assessment, the two younger ages were defined on a calendar-year basis, while the 17-year-olds were defined on an October 1 to September 30 basis. This resulted in modal grades of 4, 8, and 11. To allow for age cohorts that were exactly four years apart, in the 1986 main assessment all ages were defined on an October 1 to September 30 basis, resulting in modal grades of 3, 7, and 11. Special studies (Kaplan, Beaton, Johnson, & Johnson, 1988) were conducted to measure the effect of the changes in age definition. Because of problems encountered in assessing third graders, in 1988 the ages were redefined on a calendar-year basis, with the modal grades being 4, 8, and 12. These were the age definitions used in the 1990, 1992, 1994, and 1996 main assessments.

The table also indicates that in 1984, NAEP began gathering data by grade as well as by age, a practice that had been continued in assessments up to 1994; the 1996 national assessment included gathered data by grade only. It should be noted that somewhat different age definitions were used in the 1984, 1986, and 1988 assessments. In the 1984 assessment, the two younger ages were defined on a calendar-year basis, while the 17-year-olds were defined on an October 1 to September 30 basis. This resulted in modal grades of 4, 8, and 11. To allow for age cohorts that were exactly four years apart, in the 1986 main assessment all ages were defined on an October 1 to September 30 basis, resulting in modal grades of 3, 7, and 11. Special studies (Kaplan, Beaton, Johnson, & Johnson, 1988) were conducted to measure the effect of the changes in age definition. Because of problems encountered in assessing third graders, in 1988 the ages were redefined on a calendar-year basis, with the modal grades being 4, 8, and 12. These were the age definitions used in the 1990, 1992, and 1994 main assessments.

1.3 DEVELOPMENT OF ASSESSMENT OBJECTIVES, ITEMS, AND BACKGROUND QUESTIONS

In 1996, NAEP conducted main assessments of students at all three grade levels in mathematics and science. These assessments entailed the generation of a large number of cognitive items—items measuring knowledge and skills. In addition, a large number of background and attitude questions were asked of students, and school, teacher and instructional questions were asked of principals and teachers. Details on the item development procedures for the 1996 main assessment are given in Chapter 2; this section provides an overview. (In addition to the main assessments and the special assessments, long-term trend studies were conducted in reading, mathematics, science, and writing. Since the instruments used for these studies consisted entirely of items used in previous assessments, no developmental tasks were required for their use in the 1996 assessment.)

In addition to the cognitive items, several questionnaires were developed: a common student Background Questionnaire given to all assessed students of a given grade, a School Characteristics and Policies Questionnaire, Teacher Questionnaires for teachers of fourth- and eighth-grade students in mathematics and science and for teachers of twelfth-grade students who were assessed in advanced mathematics, and an SD/LEP Student Questionnaire. Each of these questionnaires was developed through a broad-based consensus process.

All items in the assessment underwent extensive reviews by subject area and measurement specialists, as well as careful scrutiny to eliminate any potential bias or lack of sensitivity to any group. Further, the items were field tested on a representative group of students. Based on the results of the field test, items were revised or modified as necessary and then again reviewed for bias. With the help of staff and outside reviewers, the Instrument Development Panels selected the items to include in the assessment. After the items were selected and formed into the final groupings or blocks of items, they were carefully reviewed by the National Center for Education Statistics (NCES), the Office of Management and Budget (OMB), and the National Assessment Governing Board (NAGB).

The assessment instruments included multiple-choice items, constructed-response items scored dichotomously, constructed-response items scored polytomously, and cluster items in both mathematics and science. The constructed-response items were professionally scored as described in Chapter 7.

1.4 THE 1996 SAMPLE DESIGN

The sample for the 1996 NAEP assessment was selected using a complex three-stage sample design. The three-stage sample design includes (1) the sampling of students from (2) selected schools within (3) 94 selected geographic areas, called primary sampling units (PSUs), across the United States. The 1996 sample design differed from previous years due to oversampling of SD/LEP students. With the inclusion of the different inclusion rules and the availability of accommodations, the sample design was similar to that used in 1986, 1988, 1990, 1992, and 1994 and is described in detail by Westat, Inc., the firm contracted by NCES to select the sample, in *The 1996 NAEP Sampling and Weighting Report* (Wallace & Rust, 1999). The following sections provide an overview of the five steps used to draw NAEP samples using the three-stage sample design. Further details are given in Chapter 3. Steps 3 and 4 describe the assignment of sample types and assessment sessions to the second sampling unit schools. Steps 5a through 5c contain procedures for the collection of data for SD/LEP students, teachers, and schools.

➪ Step 1: Primary Sampling Units

In the first stage of sampling, the United States (the 50 states and the District of Columbia) was divided into geographic primary sampling units (PSUs). Each PSU met a minimum size requirement and generally comprised either a consolidated metropolitan statistical area (CMSA), a metropolitan statistical area (MSA), a single county, or a group of contiguous counties. The PSUs were classified into four regions (Northeast, Southeast, Central, West), each containing about one-fourth of the U.S. population. In each region, PSUs were additionally classified as MSA or nonMSA. This resulted in eight subuniverses of PSUs.

Ninety-four of the PSUs were selected for the 1996 main assessment. Twenty-two PSUs were designated as certainty units because of their size, and were included in the sample with certainty. The remaining smaller PSUs were not guaranteed to be selected and were accordingly designated as noncertainty PSUs. Within each major stratum (subuniverse), further stratification was achieved by ordering the noncertainty PSUs according to several additional socioeconomic characteristics. Seventy-

two PSUs were selected, one per stratum from each of the noncertainty strata, with probability proportional to size (total population from the 1990 census). To enlarge the samples of Black and Hispanic students, thereby enhancing the reliability of estimates for these groups, PSUs from the high-minority subuniverses were sampled at twice the rate of those from the other subuniverses. This was achieved by creating smaller strata within the high-minority subuniverses.

For the long-term trend samples, 52 PSUs were selected: 10 PSUs were selected with certainty; six additional PSUs were selected from the 12 remaining main sample certainty PSUs; and 36 PSUs were selected from the 72 noncertainty strata independently of PSU selections for the main samples.

➪ Step 2: Selection of Schools

In the second stage of sampling for the main assessments, the public schools (including Bureau of Indian Affairs (BIA) schools and Department of Defense Education Activity (DoDEA) schools) and nonpublic schools (including Catholic schools) within each of the selected PSUs were listed according to the grade ranges associated with the three age classes. An independent sample of schools was selected separately for each of the grades so that some schools were selected for assessment of two grades, and a few were selected for all three. Schools within each PSU were selected (without replacement) with probabilities proportional to assigned measures of size with oversampling of nonpublic schools and of schools with high minority enrollment. Overall probabilities of selection for high-minority schools were twice those for other schools while the probabilities of selection of nonpublic schools were triple those of low-minority public schools of the same size. The increased probabilities of selection enlarged the samples of Black and Hispanic students and the samples of students from nonpublic schools, thereby enhancing the reliability of estimates for these groups. Details of the probabilities used for school selection appear in Chapter 3.

The samples of schools for the long-term trend assessments were drawn in a manner very similar to that used for the main assessments. The chief difference in the two samples was that nonpublic schools and schools with high minority enrollment were not oversampled for the long-term trend assessments. Schools were not selected for both main and long-term trend assessments at the same age/grade.

For the main samples, the overall school cooperation rate was 86 percent for grade 4, 83 percent for grade 8, and 79 percent for grade 12. For the long-term trend samples, the overall school cooperation rate was 85 percent for age class 9, 84 percent for age class 13, and 81 percent for age class 17. In certain instances, refusing schools were replaced by substitutes according to the rules indicated in Chapter 3.

➪ Step 3: Assigning Sample Type to Schools

In order to determine the effect of using different criteria for excluding students from the assessment, three different sample types were assigned to the schools selected for the main assessment. In sample type 1 schools, the inclusion criteria for the main samples were identical to those used in 1990 and 1992. In sample type 2 schools, new 1996 inclusion criteria were used. In sample type 3 schools, the new 1996 inclusion criteria were used and accommodations were offered to SD/LEP students. More detailed information on assigning sample type to schools is provided in Chapter 3.

➪ Step 4: Assigning Assessment Sessions to Schools

Sessions were assigned to the selected schools found to be in-scope at the time of session assignment, as described in Chapter 3. Sessions were assigned to schools with three aims in mind. The first was to distribute students to the different session types across the whole sample for each age class so that the target numbers of assessed students would be achieved (in each sample type separately in the

main assessments). The second was to maximize the number of different session types that were administered within a given selected school, without creating unduly small sessions. The third was to give each student an equal chance of being selected for a given session type regardless of the number of sessions conducted in the school.

➪ Step 5: Sampling Students

In the third stage of sampling, a consolidated list was prepared for each school of all grade- and age-eligible students (for long-term trend) or all grade-eligible students (for the main assessments) for the age class for which the school was selected. To provide the target sample size, a systematic selection of eligible students was made from this list, if necessary. In small- and medium-sized schools all eligible students were in the sample. For schools assigned to more than a single session type, students were assigned by Westat district supervisors to one of the various session types (audiotape or print administration) using specified procedures. No student was assigned to more than one session.

Step 5a: Excluded Students

Despite NAEP's goal to assess all selected students, certain selected students were judged by school authorities as being incapable of participating meaningfully in the assessment. For each student who was excluded, school staff who had knowledge of the student's capabilities completed an SD/LEP student questionnaire, listing the reason for exclusion and providing some background information. For each SD/LEP student who was included in the assessment, school staff also completed an SD/LEP student questionnaire.

Specific guidelines for exclusion were provided for all samples in the 1996 assessment. However, somewhat different criteria were used for the long-term trend samples than for the main assessment samples. In addition, the inclusion criteria for the main samples differed by sample type.

The exclusion guidelines for the long-term trend samples were the same as those used in previous assessments. Three types of students could be excluded under these guidelines—non-English speaking students, students with mental retardation who are educable but who were judged incapable of meaningfully responding to exercises appropriate to their age level, and students so functionally disabled that they could not perform in the NAEP assessment situation.

As stated previously, for the main samples, the procedures for assessing students with disabilities (SD) and students of limited English proficiency (LEP) varied by sample type. The exclusion procedure used in sample type 1 differed somewhat from that used in sample types 2 and 3. In sample type 1 schools, the inclusion criteria for the main samples were identical to those used in 1990 and 1992. These criteria were intended to be somewhat more rigorously defined than those used in the long-term trend samples. In sample type 2 schools, new 1996 inclusion criteria were used. In sample type 3 schools, the new 1996 inclusion criteria were used and accommodations were offered to SD/LEP students. The new inclusion criteria was developed to more closely match the procedures used by many states and school districts in testing situations. Both sets of the inclusion rules are presented in Chapter 5.

Step 5b: Sampling Teachers

Teachers of fourth- and eighth-grade students assessed in mathematics and science and twelfth-grade students assessed in advanced mathematics were identified and asked to complete a questionnaire (described in Chapter 2) about their background and experiences and about instructional practices, by class, for any classes containing assessed students.

Step 5c: The School Characteristics and Policies Questionnaires

A School Characteristics and Policies Questionnaire was mailed to every sampled school by Westat before the assessment for completion by the principal or school administrator. The Westat supervisor then collected the questionnaires and returned them to ETS. The School Characteristics and Policies Questionnaire is described in Chapter 2.

1.5 ASSESSMENT INSTRUMENTS

Four types of instruments were used in the 1996 assessment: student assessment booklets (which included the student common **Background Questionnaire** as well as cognitive items), SD/LEP Student Questionnaires, Teacher Questionnaires, and a School Characteristics and Policies Questionnaire. This section provides an overview of these instruments; more detailed information can be found in Chapter 4.

The student common Background Questionnaires were completed by the students participating in the 1996 assessment. These questionnaires included questions about the students' race/ethnicity, parental education levels, and other background variables specified by NCES and a committee of survey, content, and education experts. These questionnaires appeared at the beginning of some student assessment booklets and at the end of others. The student assessment booklets also included subject-related **background questions** about instructional opportunity, and interest in and attitudes towards the subject area.

1.5.1 Student Assessment Booklets—Main Assessment

1.5.1.1 Student Assessment Booklets—Main Assessment—Mathematics

Each student assessed in mathematics received a booklet containing a set of general background questions, content questions, subject-specific background questions, and questions about his or her motivation and familiarity with the assessment materials. The content questions were assembled into sections or blocks. Students in the main assessment were given three 15-minute blocks. Those sampled for the theme assessment completed one 15-minute block and one 30-minute block. Those sampled for the advanced study at grade 8 completed three 20-minute blocks; at grade 12 advanced sample students completed three 30-minute blocks. Students in the estimation sample completed one 15-minute block from the main assessment and two paced-tape sections. The overall assessment time for each student was approximately 63 minutes.

The assembly of blocks into booklets for the main assessment and their subsequent assignment to sampled students was determined by a balanced incomplete block (BIB) design with spiraled

administration. The student booklets contained two five-minute background sections, a one-minute background section, and three 15-minute blocks of items according to a BIB design.

The BIB design for the 1996 national mathematics assessment was focused by subject area, so that students received booklets containing only blocks of mathematics questions (not science). The BIB design also balances the order of presentation of the 15-minute blocks of items—every 15-minute block appears as the first cognitive block in two booklets, as the second cognitive block in two other booklets, and as the third cognitive block in another two booklets.

The design used in 1996 required that 13 blocks of mathematics items at each grade be assembled into 26 booklets. Theme blocks were placed in two other booklets, and estimation blocks in one other booklet. At grades 8 and 12, the advanced study was placed in one additional booklet. Once assembled, the main assessment booklets were then spiraled and bundled. Spiraling involves interweaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in each position in a bundle.

The final step in the BIB-spiraling procedure was the assigning of the booklets to the assessed students. The students within an assessment session were assigned booklets in the order in which the booklets were bundled. Thus, most students in an assessment session received different booklets. In the assessment design, representative and randomly equivalent samples of students responded to each item at a given grade level.

Chapter 4 provides more detail on the contents of the mathematics instruments.

1.5.1.2 Student Assessment Booklets—Main Assessment—Science

Each student assessed in science received a booklet containing general background questions, content questions, subject-specific background questions, and questions about his or her motivation and familiarity with the assessment materials. The content questions were assembled into sections or blocks. Students in the main assessment were given three 20-minute blocks at grade 4, and three 30-minute blocks at grades 8 and 12. The last block in every book was a hands-on block. Those sampled for the advanced study at grade 12 completed four 30-minute blocks. The overall assessment time for each student was, on average, 120 minutes.

The assembly of blocks into booklets for the main assessment and their subsequent assignment to sampled students was determined by a complex design with spiraled administration. The student booklets contained two five-minute background sections, a one-minute background section, and three blocks of items.

The design for the 1996 national assessment was focused by subject area, so that students received booklets containing only blocks of science questions (not mathematics). The design also balances the order of presentation of the blocks of items, except for the hands-on blocks, which always appear in position three of a booklet. All other blocks appear an equal number of times in position one and position two. Further, the design was set up to ensure that no student answered more than one theme-based block (though some students did not receive any). This design allows for some balancing of the impact of context and fatigue effects to be measured and reported, but makes allowance for the difficulties and disruption of administering hands-on blocks. It also takes into account the limited breadth of content coverage included in the theme blocks.

The design used in 1996 required that fifteen blocks of mathematics items at each grade be assembled into 37 booklets. At grade 12, the advanced study was composed of three additional booklets. Once assembled, the main assessment booklets were then spiraled and bundled. Spiraling involves interweaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in each position in a bundle.

The final step in the spiraling procedure was the assigning of the booklets to the assessed students. The students within an assessment session were assigned booklets in the order in which the booklets were bundled. Thus, most students in an assessment session received different booklets. In the assessment design, representative and randomly equivalent samples of about 2,000 students responded to each item at a given grade level.

Chapter 4 provides more detail on the contents of the science instruments.

1.5.2 Student Assessment Booklets—Long-Term Trend Samples

There were two distinct long-term trend samples in the 1996 assessment, each of which required reprinting booklets used in previous assessments:

Reading-Writing Long-Term Trend: Six booklets were used at each of the three age/grades for the purposes of measuring long-term trends in reading and writing. These booklets were identical to booklets used in the 1984 main assessments of reading and writing and in the 1988, 1990, 1992, and 1994 long-term trend assessments of those subjects. Each booklet consisted of a common background block in the beginning of each booklet and three cognitive blocks, either two reading and one writing or one reading and two writing. All cognitive blocks also contained subject-related background questions. The booklets were administered without audiotape and were spiraled together for administration.

***Mathematics-Science Long-Term Trend*:** These instruments were used for the measurement of mathematics and science and were identical to booklets administered in 1990, 1992 and 1994. These booklets contained 15-minute blocks of mathematics and science items; each mathematics block and each science block was administered using audiotape pacing. (At the younger two ages, the booklets also contain a block of reading items, which was print-administered.) There were three booklets each at age 9 and age 13 and two booklets at age 17. The common background questions appeared at the beginning of each booklet. Combined, the booklets at an age contain three blocks of mathematics items and three blocks of science items. Because of the audiotape pacing, each booklet was administered in a separate session.

1.5.3 Other Instruments

Besides the student assessment booklets, other instruments provided data relating to the assessment:

The ***SD/LEP Student Questionnaires*** were completed by the teachers of those students who were selected to participate in the assessment sample who had disabilities (SD) or were classified as Limited English Proficient (LEP). The questionnaires were completed for all SD or LEP students, whether or not they actually participated in the assessment. The questionnaires asked about the nature of the student's disability and the special programs in which the student participated. The response rates for

this questionnaire ranged from 92 to 95 percent for the different student samples. The criteria used for excluding students are described in Chapter 5.

Teacher Questionnaires were administered to the teachers of fourth- and eighth-grade students assessed in mathematics and science and teachers of twelfth-grade students who were assessed in advanced mathematics. The Teacher Questionnaire included a general section that contained questions about the teacher's background and experience. The rest of the questionnaire contained questions about instructional practices, by class, for any classes containing assessed students. The response rates ranged from 95 to 100 percent for the different student samples.

School Characteristics and Policies Questionnaires were completed by school principals or their representatives, who provided information about school administration, staffing patterns, special programs, subject requirements, and school resources. The response rates for the different student samples ranged from 92 to 95 percent.

1.6 FIELD OPERATIONS AND DATA COLLECTION

Field operations and data collection for the 1996 assessment were the responsibility of Westat, Inc., and are documented in Chapter 5. The field operation was conducted by a staff at Westat's home office and a larger staff in the field. The Westat home-office staff coordinated all activities related to field operations and managed materials distribution and home-office receipt of assessment reporting forms. The field staff consisted of area supervisors, assessment supervisors, and exercise administrators. The assessment supervisors, who were trained by Westat, were each responsible for the assessment activities in one or more PSUs. Although ETS made initial contact with participating school districts, each assessment supervisor was primarily responsible for making follow-up contacts with these districts, recruiting and training exercise administrators to work with them in administering the assessment sessions, arranging the assessment sessions, and selecting the sample of students to be assessed within each school. The assessment supervisors and the exercise administrators administered the assessments, filled out the necessary forms, performed process control, and shipped the assessment booklets and forms to National Computer Systems (NCS), the subcontractor responsible for processing NAEP materials and data.

Gaining school cooperation was the joint responsibility of Westat and ETS. ETS made the preliminary contacts preparatory to obtaining school cooperation by first contacting the Chief State School Officers, informing them that schools within their states had been selected for the assessment and, in a later letter, listing the selected schools and districts. Later mailings were sent to superintendents of public schools and parochial schools and principals of other nonpublic schools for all schools selected in the assessment. These materials provided an explanation of NAEP, a list of the selected schools in the official's jurisdiction, and a cover letter explaining that a Westat district supervisor would contact them to set up an introductory meeting. Westat district supervisors then scheduled and conducted introductory meetings (both by telephone and in person), worked with the schools to schedule the assessments, and, with the exercise administrators, conducted the assessments. The overall participation rate of schools originally selected in the 1996 assessments was 83 percent for the main samples and 84 percent for the long-term trend samples. Further detail on school participation rates is given in Chapter 3.

The main assessment sessions were conducted between January 3 and March 29, 1996, at all three grade levels. The age 9/grade 4 long-term trend assessments were carried out between January 3 and March 8, 1996; the age 17/grade 11 long-term trend samples were conducted between March 11 and May 10, 1996. The age 13/grade 8 long-term trend assessments were carried out between October 9 and December 22, 1995. When the main assessments of the long-term trend subjects were first collected in

1987 and 1986, studies were completed to take into account the difference in assessment time across the samples (Kaplan, Beaton, Johnson, & Johnson, 1988).

Two special studies that required additional steps in the sampling process were included in the 1996 main assessment. One of these special studies involved students who were eligible for advanced mathematics or science sessions. Advanced sessions were only available to designated students at grade 8 in mathematics and at grade 12 in mathematics and science. Further details on the advanced sessions is provided in Chapters 3 and 5.

The other special study involved applying two versions of the SD/LEP "inclusion" criteria for NAEP assessments and, in some schools, offering accommodations for testing students designated as SD/LEP. In the study, the school sample was divided into three subsamples: S1 (sample type 1), S2 (sample type 2), and S3 (sample type 3). The purpose of these subsamples was to collect data under the same conditions as previous assessments in order to maintain trend in mathematics within NAEP; evaluate the impact of a revised, more specific set of inclusion criteria; and evaluate the combined effect of the new criteria and the use of accommodations for testing students. Further details on this special study are provided in Chapters 3 and 5. Results for the study appear in the SD/LEP report along with a technical procedural appendix describing the special analyses completed for the study.

An automated management system tracked and recorded the progress of field work throughout the 1996 assessment period. In addition, progress was constantly monitored through telephone reports held between the area supervisors and the assessment supervisors and between the area supervisors and the home office staff.

Both Westat and ETS participated in the quality control of the field administration, which involved on-site visits by Westat and ETS staff to verify the sampling of the students and to observe the conduct of the assessment by the supervisors and the exercise administrators.

1.7 MATERIALS AND DATA PROCESSING

After completing an assessment session, Westat field supervisors and exercise administrators shipped the assessment booklets and forms from the field to National Computer Systems for entry into computer files, professional scoring, and creating the data files for transmittal to ETS. Careful checking assured that all data from the field were received. More than 134,000 booklets and questionnaires were received and processed for the national portion of the 1996 assessment. The extensive processing of these data is detailed in Chapter 6.

The student data were transcribed into machine-readable form by scanning the student instruments with an optical scanning machine. An intelligent data entry system was used for resolution of the scanned data, the entry of documents rejected by the scanning machine, and the entry of information from the questionnaires. Additionally, each piece of input data was checked to verify that it was of an acceptable type, that it was within a specified range or ranges of values, and that it was consistent with other data values. The entry and editing of materials is discussed in Chapter 6.

1.8 PROFESSIONAL SCORING

Items requiring a written response from the student (constructed-response items) were included in the main and state assessments in mathematics and science and in the long-term trend assessments in

reading, mathematics, and writing. More than nine million constructed responses were read and marked by the professional scoring staff for the national and state portions of the 1996 assessment.

Image processing and scoring were again used in 1996. Images of students' responses to the constructed-response items were scanned into computerized form, then scored online by professional raters.

Chapter 7 describes the professional scoring operation, including an overview of the scoring guides, the training procedures, and the scoring process for each subject area.

1.9 CREATION OF THE DATABASE

Before analyses could begin, the student response data, school, teacher, and SD/LEP student questionnaire data, and all sampling weights had to be integrated into a coherent and comprehensive database. This database, which was used for all analyses, was also the source for the creation of two NAEP database products—the item information database and the secondary-use data files. Secondary-use data files include sample control statement files for SAS and SPSS statistical systems and the NAEP Data on Disk product suite. The Data on Disk products, including a complete set of secondary-use data files on CD-ROM, PC-based NAEP data extraction software, and NAEP analysis modules, make secondary use of NAEP data much easier than it has been in the past. The quality of the data resulting from the complete data entry system, from the actual instruments collected in the field to the final machine-readable database used in analysis, was verified by selecting field instruments at random and performing a character-by-character comparison of these instruments with their representations in the final database. Chapter 8 provides details on the database, quality control activities, and database products.

Chapter 2

DEVELOPING THE NAEP OBJECTIVES, ITEMS, AND BACKGROUND QUESTIONS FOR THE 1996 ASSESSMENTS OF MATHEMATICS AND SCIENCE[1]

Stephen Lazer
Educational Testing Service

2.1 INTRODUCTION

In 1996, the main NAEP assessments were conducted in mathematics and science.[2] Long-term trend assessments were also conducted in reading, writing, mathematics, and science; these assessments are composed of instruments identical to those used in previous years. Additional data were gathered under the auspices of the State Assessment program, which in 1996 assessed mathematics at grades 4 and 8 using representative samples of public- and nonpublic-school students in 44 states[3], the District of Columbia, Department of Defense Domestic Dependent Elementary and Secondary Schools (DDESS), Department of Defense Dependents Schools (DoDDS), and Guam. The 1996 State Assessment program also assessed science at grade 8 using representative samples of public- and nonpublic-school students in 43 states, the District of Columbia, DDESS, DoDDS, and Guam; DDESS and DoDDS students were also assessed in science at grade 4. A summary of the main assessment subject areas follows:

Mathematics: A mathematics assessment was administered to national samples at grades 4, 8, and 12. This assessment was designed around the measurement of five mathematics content areas, and continued a trend line begun in 1990. In other words, many of the assessment questions were used in the 1990 and 1992 NAEP assessments; others were newly developed for 1996. A mixture of multiple-choice, short constructed-response, and extended constructed-response questions made up the assessment; in aggregate, well over half of the student assessment time was spent answering constructed-response questions. On some portion of the assessment, students were required to use calculators and other hands-on materials. In addition to the instruments used to generate the main reporting scales, three supplemental mathematics surveys were conducted at the national level. At each grade, special instruments were administered to representative national samples that were designed to measure the estimation skills of students. Second, separate samples at all grades were given "thematic" instruments constructed to measure the ability of students to solve in-depth mathematics problems. Finally, at grades eight and twelve a special study was conducted. In this study, students with advanced mathematics training were administered special assessment booklets whose contents were more advanced than those of the main assessment.

Science: A science assessment was administered to national samples at grades 4, 8, and 12. The assessment measured a broad range of science-education outcomes. Because the 1996 science assessment was based on a new framework, it represents the beginning of a

[1] Stephen Lazer manages assessment development activities for the NAEP program at ETS.
[2] Copies of the frameworks for these assessments are available from the National Assessment Governing Board (NAGB) *http://www.nagb.org*.
[3] Not all participating jurisdictions gathered both public- and nonpublic-school samples.

new trend line. The assessment involved three different types of testing. Some portions covered general scientific knowledge and skills. Others sections of the survey tested students abilities to answer questions in an in-depth thematic or topical area. Finally, each sampled student completed a component that involved conducting a hands-on science experiment. A combination of multiple-choice, short constructed-response, and extended constructed-response questions made up the assessment; in aggregate, well over half of the student assessment time was spent answering constructed-response questions. In addition to the main assessment, a special study of students with advanced scientific training was conducted at grade 12.

From its inception, NAEP has developed assessments through a consensus process and the 1996 instruments were no exception. Under the direction of the National Assessment Governing Board (NAGB), educators, scholars, and citizens representative of many diverse constituencies and points of view designed assessment frameworks for both subject areas. Staff at Educational Testing Service (ETS) who were subject-area experts in their respective fields worked with subject-area consultants well versed in assessment methodology to develop assessment questions appropriate to the objectives. All questions underwent extensive reviews by subject-matter specialists, measurement specialists, and ETS employees. Questions were assembled and printed into booklets suitable for matrix sampling and then administered either by a trained field staff (for the national program) or by state or local school district staff (for the State Assessment program) to stratified, multistage probability samples of students.

All 1996 assessment development efforts were governed by four major considerations:

1. The primary goal of the development process was to craft instruments that matched the content definitions included in the assessment frameworks, which was developed through consensus processes conducted under the auspices of the National Assessment Governing Board (NAGB).

2. As outlined in the ETS proposal for the administration of the NAEP cooperative agreement (ETS, 1992), the development of the items was guided by an Instrument Development Committee and further reviewed by state representatives and classroom teachers from across the country. In addition, the items had to be carefully reviewed for potential bias.

3. As described in the *ETS Standards of Quality and Fairness* (ETS, 1987), all materials developed at ETS were in compliance with specified procedures. In particular, all questions were carefully reviewed for content accuracy, testworthiness, and potential bias.

4. As per federal regulations, all NAEP cognitive and background items were submitted to a federal clearance process. This process involved review of all cognitive items by NCES and NAGB, and review of all background questions by the Office of Management and Budget (OMB), the Information Management Team (IMT) of the Department of Education, and NCES.

The development effort for the 1996 assessment included questionnaires[4] for students, teachers, and school administrators, in addition to a substantial number of cognitive items for both subject areas.

[4] These questionnaires can be obtained from the National Center for Education Statistics (NCES).

The following sections include general overviews about setting objectives and developing items and specific details about developing subject-specific objectives and assessments. A list of the consultants who participated in the 1996 development process is included in Appendix A.

2.2 GENERAL OVERVIEW OF THE 1996 ASSESSMENT FRAMEWORKS

The subject-area objectives for each NAEP assessment are determined through a legislatively mandated consensus process. These objectives typically take the form of frameworks, or matrices, delineating the important content and process areas to be assessed. In addition to these broad frameworks, the Council of Chief State School Officers (CCSSO), and therefore NAGB, provided detailed descriptions of item types and the numbers of items to be selected for each category. The various frameworks for the 1996 assessments are described below.

The frameworks for the main 1996 NAEP assessments were developed through consensus processes conducted by the CCSSO working under contract to NAGB. The projects involved participation and review by many groups, including teachers, content-area scholars, educational policymakers, and members of the general public. In addition to people directly involved in the framework development processes, the documents were reviewed by state education and testing officials, by representatives of professional associations, and by researchers. In addition, the frameworks were the subject of testimony at public hearings arranged to allow the widest possible participation in the consensus process. The objectives resulting from these processes reflect neither a narrowly defined theoretical framework nor the view of every participant. They do, however, represent the thinking of a broad cross-section of individuals who are expert in the various content areas and who are deeply committed to the improvement of American education.

The framework that governs the 1996 NAEP **mathematics** assessment is an enhanced version of the framework that governed the development of the 1990 and 1992 assessments. This framework was originally developed under the auspices of the Council of Chief State School Officers (CCSSO); it was revised in 1991-1992 through a consensus project managed by the College Board. It was this enhanced framework that governed the instrument development activities related to the 1996 assessment. The revisions to the framework were minor, and allow for continued measurement of educational progress through comparisons with 1990 and 1992 results. The mathematics framework is organized according to a five-by-three matrix of content strands by mathematical abilities. The content strands, which make up the main reporting subscales, are:

- Number Sense, Properties, and Operations;
- Measurement;
- Geometry and Spatial Sense;
- Data Analysis, Statistics, and Probability; and
- Algebra and Functions.

In addition, the assessment was designed to measure the three mathematical abilities of:

- Conceptual Understanding,
- Procedural Knowledge, and
- Problem Solving.

All exercises in the assessment were classified as measuring one of the content strands and one of the abilities. Additional specifications in the framework are related to an assessment dimension referred to as *mathematical power*. Mathematical power is conceived as consisting of mathematical abilities within a broader context of *reasoning* and with *connections* across the broad scope of mathematical content and thinking. *Communication* is viewed as both a unifying thread and a way for students to provide meaningful responses to tasks.

The framework for the 1996 **science** assessment is structured according to a matrix organized according to two major dimensions: *fields of science* and *knowing and doing science*. The fields of science, which make up the subscales on which assessment results are analyzed and reported, are the *earth*, *physical*, and *life* sciences. The cognitive dimension, knowing and doing science, is organized into three categories, *conceptual understanding*, *scientific investigation*, and *practical reasoning*. In addition, the framework includes requirements for measurement of content that crosses other categorical boundaries. Specifically, the *nature of science* and *themes* are categories that should integrate the three fields of science, rather than represent separate content.

2.3 GENERAL OVERVIEW OF PROCEDURES FOR DEVELOPING THE ITEMS

A carefully developed and tested series of steps, similar to those used for past NAEP assessments, was utilized to create assessment items that reflected mathematics and science objectives and that measured achievement related to them (see Sections 2.4 through 2.6 for more detail). The steps were as follows:

1. NAGB provided item specifications and frameworks in each subject area.

2. The Instrument Development Committees in both subject areas provided guidance to NAEP staff about how the objectives could be measured given the realistic constraints of resources and the feasibility of measurement technology. The committees made recommendations about priorities for the assessment (within the context of the assessment framework) and the types of items to be developed.

3. Items were chosen for the assessment through an extensive selection process that involved the input of practitioners from across the country as well as from members of the Instrument Development Committees.

4. Specialists with subject-matter expertise, skills, and experience in creating items according to specifications were identified from inside and outside ETS to develop and review the assessment questions.

5. The items were reviewed and revised by NAEP/ETS staff and external test specialists.

6. Representatives from the State Education Agencies met and reviewed all items and background questionnaires that were scheduled to be part of the state assessment.

7. Language editing and sensitivity reviews were conducted as required by *the ETS Standards for Quality and Fairness*.

8. Field test materials were prepared, including those necessary to secure clearance by the Office of Management and Budget.

9. A field test was conducted in many states, the District of Columbia, and three territories.[5]

10. Representatives from State Education Agencies met and reviewed the field test results for all exercises selected for the state assessment.

11. Based on the field test analyses, new items for the 1996 assessment were revised, modified, and re-edited, where necessary. The items once again underwent the full range of ETS reviews.

12. The Instrument Development Committees approved the selection of items to include in the 1996 assessment.

13. After a final review and check to ensure that each assessment booklet and each block met the overall guidelines for the assessment, the booklets were typeset and printed.

The following sections describe the development of the mathematics and science assessments in more detail.

2.4 DEVELOPING THE MATHEMATICS ASSESSMENT

2.4.1 Overview

The framework that governs the 1996 NAEP **mathematics** assessment is an enhanced version of the framework used on the 1990 and 1992 assessment. Similar to other NAEP assessments, the 1990 mathematics framework was developed through a broad-based consensus process managed by the CCSSO. In 1991-1992, the National Assessment Governing Board (NAGB) contracted with The College Board to review and revise the framework in preparation for the assessment originally planned for 1994 and administered, in fact, in 1996. The development process involved a committee of mathematicians and mathematics education specialists. Educators, scholars, and citizens, representative of many diverse constituencies and points of view, participated in the national consensus process to review and revise objectives for the assessment.

The instrument used in the 1996 mathematics assessment was composed of a combination of new items developed for administration in 1996[6] and items from the 1992 and 1990 assessments. Those items that were carried over from the 1992 and 1990 instruments comprised approximately 60 percent of the 1996 instrument. The remaining portion was made up of new items developed according to the recommendations included in the enhanced assessment specifications. Maintaining approximately 60 percent of the instrument across the two assessment years (1992 and 1996) allowed for the reporting of

[5] In this case, two field tests were conducted. The mathematics and science assessments were originally scheduled for 1994; thus a 1993 field test was conducted. This field test was designed to provide replacement items for those released in mathematics and all exercises needed in the new science assessment. Because the assessments were delayed until 1996, a supplementary field test was conducted in 1995. In mathematics, this field test was used for the development of theme and advanced blocks, and in the development of assessment accommodations. In science, the 1995 field test was used to develop more general science exercises for the main assessment.

[6] As was noted above, many of the new items were originally scheduled for use in 1994; however, the mathematics assessment was deferred until 1996. Other items were developed and field tested in 1995. For purposes of this report, we will refer to all exercises that were used in 1996 but that were not part of earlier NAEP surveys as having been "newly developed for 1996."

trends in mathematics performance. At the same time, developing a new set of items made it possible to release approximately 40 percent of the 1992 assessment for public use.

In developing the new portion of the 1996 NAEP mathematics assessment, the same procedures used in 1992 were followed; however, new items were constructed to meet the demands of the revised framework. All items underwent extensive reviews by specialists in mathematics, measurement, and bias/sensitivity; items developed for grades four and eight were also reviewed by state representatives. The core goals of the ETS assessment development process and procedures used to realize these goals are outlined in the introduction to this chapter and in Section 2.3.

The following sections include a detailed description of the development of the framework, objectives, and items for the 1996 NAEP mathematics assessment. Section 2.4.8 describes the student background questionnaires and the reading teacher questionnaire. Additional information on the structure and content of assessment booklets can be found in Chapter 4. Various committees worked on the development of the framework, objectives, and items for the mathematics assessment. The list of committee members and consultants who participated in the 1996 development process is provided in Appendix A.

2.4.2 Development of the Assessment Framework

NAGB is responsible for developing assessment objectives and test specifications for NAEP surveys. Appointed by the Secretary of Education from lists of nominees proposed by the board itself in various statutory categories, the 26-member board is composed of state, local, and federal officials, as well as educators and members of the public.

Under contract with NAGB, The College Board convened a committee during 1991 and 1992 to develop an enhanced version of the framework that had been used during the development of the 1990 and 1992 assessments (the 1992 mathematics assessment had already been developed at the time the development of this enhanced framework was begun). The enhanced version was needed to better reflect the rapid evolution of mathematics instruction that was underway in the early 1990s as a result of the emergence of the National Council of Teachers of Mathematics (NCTM) *Curriculum and Evaluation Standards for School Mathematics*. The *Standards* were rapidly becoming one of the acknowledged barometers for measuring achievement. The development process for the enhanced framework was based on consensus building, and included the committee listed in Appendix A.

During this development process, input and reactions were continually sought from a wide range of members of the mathematics field, experts in assessment, school administrators, and state staff. In particular, the process was informed by recommendations of leading professional organizations in mathematics.

2.4.3 Framework and Assessment Design Principles

The planning committee involved in the review and revision of the NAEP mathematics framework was given a number of working guidelines and goals by the National Assessment Governing Board and by the steering committee that oversaw the process. These guidelines directed the planning committee to develop assessment specifications that called for performance-oriented exercises that focus on problem-solving and provide students with opportunities to communicate their understandings in mathematics. The framework should, according to these guidelines, embody a broad view of mathematics

that addressed the high levels of mathematical literacy needed for employability, personal development, and citizenship. Also, the framework should take into account findings of contemporary research on mathematics and mathematics education, and would expand the range of assessment tools to include formats that more closely resembled classroom activities.

The development was further guided by the consideration that the assessment should reflect many of the states' curricular emphases and objectives in addition to what various scholars, practitioners, and interested citizens believed should be included in the curriculum. Accordingly, the committee focused on several general principles in revising the NAEP mathematics assessment. These principles are:

- The purpose of the NAEP mathematics assessment is to provide information about the progress and achievement of students in general rather than to test individual students' ability. NAEP is designed to inform policymakers and the public about mathematics ability in the United States. Furthermore, NAEP state data can be used to inform states of their students' relative strengths and weaknesses.

- The term "mathematical literacy" encompasses such broad skills and abilities as being able to reason numerically, algebraically, geometrically, spatially, and with data; identify and apply problem-solving strategies appropriately in situations; and use the language of mathematics to construct clear and coherent responses to problems or tasks.

- The mathematics assessment should use authentic problems and tasks that address important mathematics concepts and skills so that the assessment tool will demonstrate a close link to desired classroom instruction and students' mathematics experiences.

- Every effort should be made to make the best use of available methodology and resources in driving assessment capabilities forward.

- Every effort must be made in developing the assessment to represent a variety of opinions, perspectives, and emphases among professionals in universities, as well as in state and local school districts.

2.4.4 Framework for the 1996 Assessment

The framework for the 1996 mathematics assessment is organized according to a five-by-three matrix of content strands by mathematical abilities. The content strands are:

- Number Sense, Properties, and Operations;
- Measurement;
- Geometry and Spatial Sense;
- Data Analysis, Statistics, and Probability; and
- Algebra and Functions.

These content strands were assessed across the three mathematical abilities of:

- Conceptual Understanding,
- Procedural Knowledge, and
- Problem Solving.

Figures 2-1 and 2-2 describe the five content strands and three mathematical abilities that guided the development of the 1996 mathematics assessment.

Figure 2-1
Descriptions of Content Strands in Mathematics

Number Sense, Properties, and Operations

This strand focuses on students' understanding of numbers (whole numbers, fractions, decimals, integers, real numbers, and complex numbers), operations, and estimation, and their application to real-world situations. Students will be expected to demonstrate an understanding of numerical relationships as expressed in ratios, proportions, and percents. Students also will be expected to understand properties of numbers and operations, generalize from number patterns, and verify results.

Measurement

The measurement strand focuses on understanding of the process of measurement and on the use of numbers and measures to describe and compare mathematical and real-world objects. Students will be asked to identify attributes, select appropriate units and tools, apply measurement concepts, and communicate measurement-related ideas.

Geometry and Spatial Sense

As described in the NCTM *Standards*, spatial sense must be an integral component of the study and assessment of geometry. Understanding spatial relationships allows students to use the dynamic nature of geometry to connect mathematics to their world.

This content strand is designed to extend well beyond low-level identification of geometric shapes into transformations and combinations of those shapes. Informal constructions and demonstrations (including drawing representations), along with their justifications, take precedence over more traditional types of compass-and-straightedge constructions and proofs. While reasoning is addressed throughout all of the content strands, this strand continues to lend itself to the demonstration of reasoning within both formal and informal settings. The extension of proportional thinking to similar figures and indirect measurement is an important connection here.

(continued)

Figure 2-1 (continued)
Descriptions of Content Strands in Mathematics

Data Analysis, Statistics, and Probability

The important skills of collecting, organizing, reading, representing, and interpreting data will be assessed in a variety of contexts to reflect the pervasive use of these skills in dealing with information. Statistics and statistical concepts extend these basic skills to include analyzing and communicating increasingly sophisticated interpretations of data. Dealing with uncertainty and making predictions about outcomes require an understanding not only of the meaning of basic probability concepts but also the application of those concepts in problem-solving and decision-making situations.

Questions will emphasize appropriate methods for gathering data, the visual exploration of data, a variety of ways of representing data, and the development and evaluation of arguments based on data analysis. Students will be expected to apply these ideas in increasingly sophisticated situations that require increasingly comprehensive analysis and decision making.

Algebra and Functions

This strand extends from work with simple patterns at grade 4, to basic algebra concepts at grade 8, to sophisticated analysis at grade 12, and involves not only algebra but also precalculus and some topics from discrete mathematics. As described in the NCTM *Standards*, these algebraic concepts are developed throughout the grades with informal modeling done at the elementary level and with increased emphasis on functions at the secondary level. The nature of the algebraic concepts and procedures included in the assessment at all levels reflects the NCTM *Standards*. Students will be expected to use algebraic notation and thinking in meaningful contexts to solve mathematical and real-world problems, specifically addressing an increasing understanding of the use of functions (including algebraic and geometric) as a representational tool.

Figure 2-2
Descriptions of Mathematical Abilities

Conceptual Understanding

Students demonstrate conceptual understanding in mathematics when they provide evidence that they can recognize, label, and generate examples and nonexamples of concepts; use and interrelate models, diagrams, manipulatives, and varied representations of concepts; identify and apply principles (i.e., valid statements generalizing relationships among concepts in conditional form); know and apply facts and definitions; compare, contrast, and integrate related concepts and principles to extend the nature of concepts and principles; recognize, interpret, and apply the signs, symbols, and terms used to represent concepts; or interpret the assumptions and relations involving concepts in mathematical settings.

Conceptual understanding reflects a student's ability to reason in settings involving the careful application of concept definitions, relations, or representations of either. Such an ability is reflected by student performance that indicates the production of examples, common or unique representations, or communications indicating the ability to manipulate central ideas about the understanding of a concept in a variety of ways.

Procedural Knowledge

Students demonstrate procedural knowledge in mathematics when they select and apply appropriate procedures correctly; verify or justify the correctness of a procedure using concrete models or symbolic methods; or extend or modify procedures to deal with factors inherent in problem settings.

Procedural knowledge includes the various numerical algorithms in mathematics that have been created as tools to meet specific needs efficiently. Procedural knowledge also encompasses the abilities to read and produce graphs and tables, execute geometric constructions, and perform noncomputational skills such as rounding and ordering. These latter activities can be differentiated from conceptual understanding by the task context or presumed student background—that is, an assumption that the student has the conceptual understanding of a representation and can apply it as a tool to create a product or to achieve a numerical result. In these settings, the assessment question is how well the student executed a procedure or how well the student selected the appropriate procedure to effect a given task.

Procedural knowledge is often reflected in a student's ability to connect an algorithmic process with a given problem situation, to employ that algorithm correctly, and to communicate the results of the algorithm in the context of the problem setting. Procedural understanding also encompasses a student's ability to reason through a situation, describing why a particular procedure will give the correct answer for a problem in the context described.

Problem Solving

In problem solving, students are required to use their accumulated knowledge of mathematics in new situations. Problem solving requires students to recognize and formulate problems; determine the sufficiency and consistency of data; use strategies, data, models, and relevant mathematics; generate, extend, and modify procedures; use reasoning (i.e., spatial, inductive, deductive, statistical, or proportional) in new settings; and judge the reasonableness and correctness of solutions. Problem solving situations require students to connect all of their mathematical knowledge of concepts, procedures, reasoning, and communication/ representational skills in confronting new situations. As such, these situations are, perhaps, the most accurate measures of students' proficiency in mathematics.

Tables 2-1 and 2-2 show the percentages of assessment time that the framework indicates should be devoted to each content strand and mathematical ability.

Table 2-1
Percentage Distribution of Items by Grade and Content Strand as Specified in the NAEP Mathematics Framework

Content Strand	Grade 4	Grade 8	Grade 12
Number Sense, Properties, and Operations[1]	40%-70%	25%-60%	20%-50%
Measurement	20%	15%	15%
Geometry and Spatial Sense	15%	20%	20%
Data Analysis, Statistics, and Probability	10%	15%	20%
Algebra and Functions	15%	25%	25%

[1] For this category, these percentages are the minimum and maximum that are acceptable, respectively.

Table 2-2
Percentage Distribution of Items by Grade and Mathematical Ability as Specified in the NAEP Mathematics Framework

Mathematical Ability	Grade 4	Grade 8	Grade 12
Conceptual Understanding	33%	33%	33%
Procedural Knowledge	33%	33%	33%
Problem Solving	33%	33%	33%

Note: Some items carry multiple classifications.

2.4.5 Developing the Cognitive Items

The 1996 assessment was designed to serve a dual purpose: to meet the content specifications elaborated in the revised NAEP mathematics framework and to allow for the measurement of changes in student mathematics achievement (that is, to permit linking to the 1990 and 1992 NAEP mathematics assessment). Because of these objectives, the following strategies were adopted in developing the 1996 mathematics assessment. First, at each grade, 8 of the 13 blocks used in the 1992 assessment were carried forward and used in 1996. Items were developed and field tested in 1993 to replace the five blocks from the 1992 assessment that had been released for public use. In addition, other portions of the 1993 field test were used to develop special components of the assessment. To this end, a new block of "estimation" exercises was developed, as were exercises for a special study of students at grades eight and twelve who had received advanced mathematics training. Given the fact that the mathematics assessment was delayed from 1994 until 1996, a supplemental field test was held in 1995 (exercises were developed for this field test in 1994). This round of exercise development was used to build extra exercises for the estimation and advanced components of the assessment. It was also used to build "theme blocks," which were intended to measure in-depth mathematics problem-solving ability.

The development of cognitive items involved careful field testing, both locally and nationally, of grade-appropriate questions and tasks for the assessment. Items were selected from a pool of questions that were written by teachers from across the country as well as by mathematics assessment specialists on

staff at ETS. The framework stated that the assessment should include some performance-based questions and tasks that require students to reason and make connections within and across different content strands of mathematics. Final selections of questions used in the 1996 assessment were approved by the Mathematics Instrument Development Committee.

The assessment included constructed-response (short and extended) and multiple-choice items. The decision to use a specific item type was based on a consideration of the most appropriate format for assessing the particular objective. Both types of constructed-response items were designed to provide an in-depth view of students' ability to communicate their understanding of important concepts in mathematics. Short constructed-response questions (scored with either a 2- or 3-level scoring rubric) were used when students needed to respond briefly in order to demonstrate full comprehension. Extended constructed-response questions (scored with a 4- or 5-level scoring rubric) were used when the task required more thoughtful consideration of the problem and engagement in more complex reasoning processes. Multiple-choice items were used when a straightforward, single correct answer was required.

A carefully developed and proven series of steps was used to create the assessment items. These steps are described earlier in the chapter under Section 2.3.

As was mentioned above, the assessment was designed to allow for measurement of trends. Therefore, eight 15-minute blocks at each grade were included in the 1996 assessment that had been a part of the 1992 assessment (and in some cases the 1990 assessment). These blocks were used in precisely the same form as they were in 1992. In addition, one of the paced-tape estimation blocks was carried forward from 1992. The remainder of the exercises used in 1996 were newly developed during either the 1992-1993 or 1994-1995 development cycles.

2.4.6 Development of the Operational Forms

The field tests of new items for the 1996 assessment were conducted in February and March, 1993 and February and March, 1995. The field test involved a convenience sample in which roughly 500 responses were obtained to each item.

The field test data were collected, scored, and analyzed in preparation for meetings with the Mathematics Instrument Development Panel. The objectives that guided the review of these items were:

- to determine which items were most suitable for assessing understanding in mathematics in accordance with the framework;
- to select items that displayed appropriate statistical attributes;
- to determine the need for revisions of items that lacked clarity, or had ineffective item formats;
- to prioritize items to be included in the assessment; and,
- to determine appropriate timing for assessment items.

Committee members, ETS assessment staff, and NAEP/ETS staff reviewed the materials. Item analyses (which provided the mean percentage of correct responses, the r-biserial correlations, and the difficulty level for each item) were used as a guide in identifying and flagging for further review those

test questions that were not measuring the intended objective well. In addition, another meeting of representatives from state education agencies was convened to review the field test results for exercises included in the grade four or eight assessments.

Once the committees had selected the items, all items were rechecked for content, measurement, and sensitivity concerns. The federal clearance process was initiated in June 1993 with the submission of draft materials to NCES. The package containing the set of cognitive items assembled into blocks and questionnaires was submitted in August 1993. A revised package with the new thematic blocks and the adjusted advanced and estimation blocks was submitted in July, 1995. Throughout the clearance process, revisions were made in accordance with changes required by the government. After approval, the blocks (assembled into booklets) and questionnaires were readied for printing in preparation for the assessment.

2.4.7 Distribution of Assessment Items

The mathematics assessment developed for use in 1996 was organized according to a series of blocks, each containing a set of questions. Some of the blocks were unique to a particular grade level. Other blocks were designed to be given to students in two grades (either 4 and 8 or 8 and 12) and a few blocks (a small percentage of the blocks used to measure trend) overlap all three grades.

At each grade, the main component of the assessment used for the creation of reporting scales included thirteen different 15-minute blocks of multiple-choice and constructed-response questions (some with both regular and extended constructed-response questions). Two additional books containing 30-minute theme blocks of multiple-choice and constructed-response questions were also used. Two estimation blocks at each grade were part of a special study and were presented to students by a paced-audiotape to assess students' estimation skills. Two 20-minute grade 8 special study blocks and two 30-minute grade 12 special study blocks targeted students in advanced mathematics courses. Of the 13 blocks at each grade that were used for the main assessment:

- three to five blocks at each grade level included items designed to be answered using a calculator. For the grade 4 calculator blocks, students were provided with a 4-function calculator, while at grades 8 and 12 the students were provided with a scientific calculator (students were also provided calculators for the theme blocks and allowed to use their own calculators on the advanced blocks)

- one block at all grade levels contained questions requiring the use of a protractor/ruler (ruler only for grade 4)

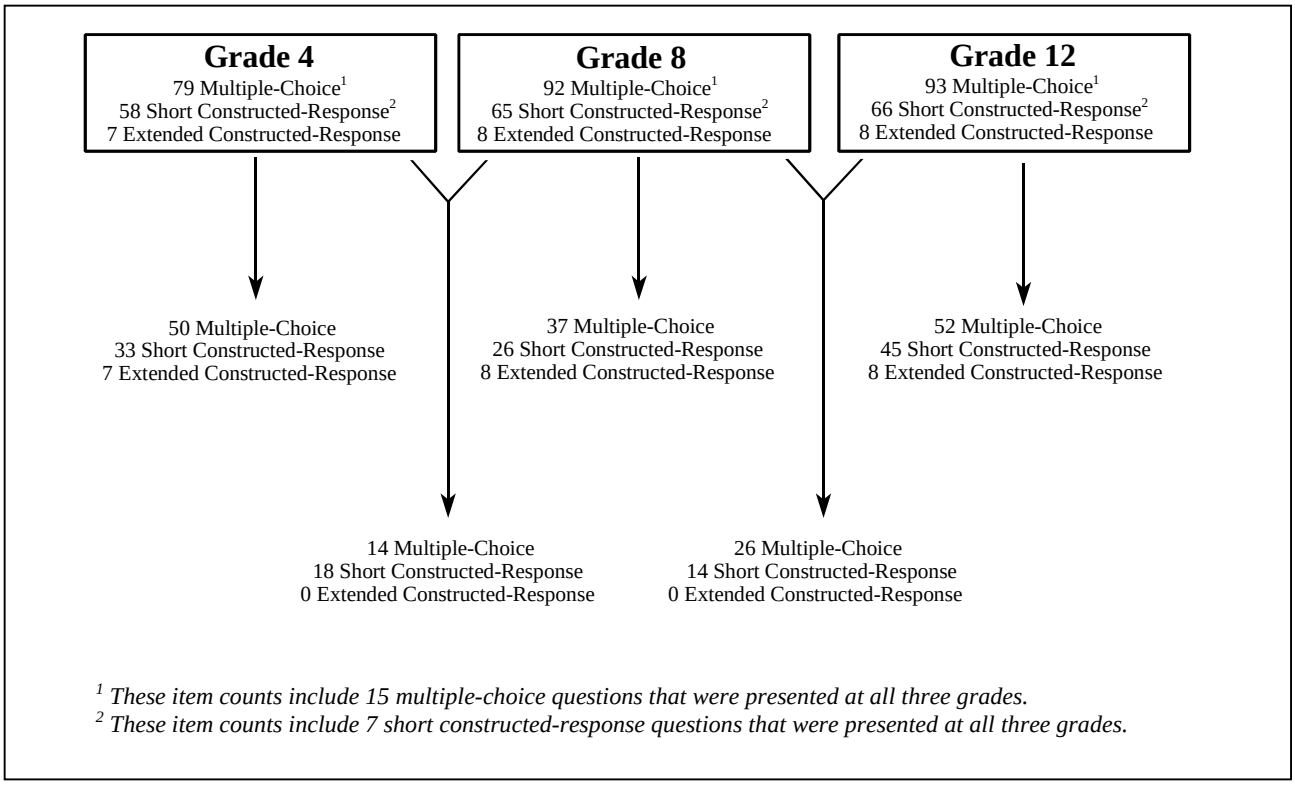
- two blocks at each grade involved the use of manipulatives; several of the theme blocks involved the use of manipulatives as well

- seven blocks at grade 4 and eight blocks at each of grades 8 and 12 included extended constructed-response items. The extended constructed-response mathematics items call for the student to work through a complex problem, require about five minutes to complete, and were scored on a 0-5 point scale. The theme and advanced blocks also included extended constructed-response items.

The sections that follow discuss the distribution of exercises on the mathematics assessment. For purposes of this discussion, we will limit the calculations to exercises that appeared as part of the main assessment, or BIB. Special study, theme, and estimation blocks will be discussed below.

Figure 2-3 lists the total number of items at each grade level in the main portion of the 1996 assessment. Of the total of 360 items, there were 194 unique multiple-choice items and 166 unique constructed-response items. Some of these items were used at more than one grade level. As a result, the sum of the items that appear at each grade level is greater than the total number of unique items.

Figure 2-3
Total Number of Items for the 1996 Mathematics Main Assessment



[1] These item counts include 15 multiple-choice questions that were presented at all three grades.
[2] These item counts include 7 short constructed-response questions that were presented at all three grades.

In the development process, every effort was made to meet the content and process targets specified in the assessment framework. Tables 2-3, 2-4, and 2-5 show the approximate percentage of aggregate assessment items devoted to each content strand at each grade level. Percentages are based on the classifications agreed upon by NAEP's 1996 Mathematics Instrument Development Committee, and confirmed by independent reviewers.

Table 2-3

Distribution of Assessment Items
for the Mathematics Assessment, Grade 4

	1990		1992		1996	
	Number of items	**Percentage Of items**	**Number of items**	**Percentage of items**	**Number of items**	**Percentage of items**
Content Strand						
Number Sense, Properties, and Operations	52	48%	63	40%	59	41%
Measurement	21	19%	31	20%	25	17%
Geometry and Spatial Sense	14	13%	27	17%	25	17%
Data Analysis, Statistics, and Probability	8	7%	20	12%	17	12%
Algebra and Functions	14	13%	17	11%	18	13%
Total	109	100%	158	100%	144	100%
Mathematical Ability						
Conceptual Understanding	42	39%	64	40%	61	42%
Procedural Knowledge	31	28%	31	20%	32	22%
Problem Solving	36	33%	63	40%	51	35%
Total	109	100%	158	100%	144	99%

Table 2-4

Distribution of Assessment Items
for the Mathematics Assessment, Grade 8

	1990		1992		1996	
	Number of items	**Percentage Of items**	**Number of items**	**Percentage of items**	**Number of items**	**Percentage of items**
Content Strand						
Number Sense, Properties, and Operations	45	33%	58	32%	48	29%
Measurement	21	15%	32	17%	27	16%
Geometry and Spatial Sense	26	19%	36	20%	31	19%
Data Analysis, Statistics, and Probability	19	14%	28	15%	25	15%
Algebra and Functions	26	19%	29	16%	34	21%
Total	137	100%	183	100%	165	100%
Mathematical Ability						
Conceptual Understanding	59	43%	67	37%	57	35%
Procedural Knowledge	41	30%	45	24%	46	28%
Problem Solving	37	27%	71	39%	62	38%
Total	137	100%	183	100%	165	101%

Table 2-5
Distribution of Assessment Items
for the Mathematics Assessment, Grade 12

	1990		1992		1996	
	Number of items	**Percentage Of items**	**Number of items**	**Percentage of items**	**Number of items**	**Percentage of items**
Content Strand						
Number Sense, Properties, and Operations	37	26%	43	24%	41	25%
Measurement	23	16%	29	16%	23	14%
Geometry and Spatial Sense	24	17%	31	18%	27	16%
Data Analysis, Statistics, and Probability	22	15%	29	16%	34	20%
Algebra and Functions	38	26%	47	26%	42	25%
Total	144	100%	179	100%	167	101%
Mathematical Ability						
Conceptual Understanding	53	37%	70	39%	66	40%
Procedural Knowledge	48	33%	52	29%	51	31%
Problem Solving	43	30%	57	32%	50	30%
Total	144	100%	179	100%	167	101%

Before proceeding, it is worth saying a few words about the other assessment components. The estimation blocks were given to nationally representative samples of students at each grade. Estimation was not given at the state level. Students in the estimation sample took one block from the main BIB. In addition, they took two estimation blocks. Both estimation blocks were administered by a paced-tape method.[7] The first estimation block was all multiple-choice, and was carried forward intact from 1992. It contained 20 items at grade 4, and 22 items at grades 8 and 12. The second estimation block was newly developed for 1996. It contained 13 items at grade 4 (10 multiple choice and 3 constructed-response), 15 at grade 8 (9 multiple-choice), and 16 at grade 12 (all multiple-choice). Trend was measured on this assessment component.

Theme blocks were also given to nationally-representative samples at all grades. Students in the theme sample first completed one block from the main assessment BIB. Then they were administered one 30-minute theme block. These blocks called for in-depth examination of mathematical problems, and for the use of a variety of mathematics skills. At grade 4, the two theme blocks included a total of one multiple-choice question and 13 constructed-response items. At grade 8, the total was 8 multiple-choice items and 13 constructed-response items. At grade 12, the total was 8 multiple-choice items and 10 constructed-response exercises.

Finally, the advanced special studies were administered to samples of students meeting certain course-taking criteria at grades 8 and 12 (algebra or higher at grade 8, and calculus, precalculus, or algebra three at grade 12). Students took a linking block composed of exercises from the main assessment and then two 20-minute blocks (at grade 8) or two 30-minute blocks (grade 12) of advanced exercises. These blocks contained 9 multiple-choice and 13 constructed-response items at grade 8, and 7 multiple-choice and 15 constructed-response items at grade 12.

[7] The blocks were paced so students had to estimate rather than calculate the answer.

2.4.8 Questionnaires

As part of the national assessment (as well as the State Assessment), a series of questionnaires was administered to students, teachers, and school administrators. Similar to the development of the cognitive items, the development of the policy issues and questionnaire items was a consensual process that involved staff work, field testing, and review by external advisory groups. A Background Questionnaire Panel drafted a set of policy issues and made recommendations regarding the design of the items. They were particularly interested in capitalizing on the unique properties of NAEP and not duplicating other surveys (e.g., the National Survey of Public and Private School Teachers and Administrators, the School and Staffing Survey, and the National Educational Longitudinal Study).

The Panel recommended a focused study that addressed the relationship between student achievement and instructional practices. The policy issues, items, and field test results were reviewed by the group of external consultants who identified specific items to be included in the final questionnaires. In addition, the Mathematics Instrument Development Panel and state representatives were consulted on the appropriateness of issues addressed in the questionnaires as they relate to mathematics instruction and achievement. The items underwent internal ETS review procedures to ensure fairness and quality and were then assembled into questionnaires.

⇒ Student Questionnaires

In addition to the cognitive items, the 1996 assessment included three student questionnaires; two sets of general and mathematics background items designed to gather contextual information about students, their instructional experiences in mathematics, and their attitudes toward mathematics, and one set of background items, given to students at the end of each booklet to determine their motivation in completing the assessment and their familiarity with assessment tasks. In order to ensure that all fourth-grade students understood the items and had every opportunity to respond to them, the three questionnaires were read aloud by administrators as fourth-grade students read along and responded in their booklets. Background questionnaires were not read aloud to eighth- and twelfth-grade students.

The **Student Demographics (common background) Questionnaire** included items about race/ethnicity, language spoken in the home, mother's and father's level of education, reading materials in the home, homework, attendance, which parents live at home, and which parents work. This questionnaire was the first section in every booklet. In many cases the items used were continued from prior assessments, so as to document changes in contextual factors that occur over time.

The second section of background items was the **Mathematics Background Questionnaire**. Categories of information represented in this section include:

> *Time Spent Studying Mathematics*: Students were asked to describe both the amount of instruction they received in mathematics and the time spent on mathematics homework.
>
> *Instructional Practices*: Students were asked to report their instructional experiences related to mathematics in the classroom, including group work, special projects, and writing in response to mathematics. In addition, they were asked about the instructional practices of their mathematics teachers and the extent to which the students themselves discussed what they did in class and demonstrated use of skills and strategies.

Attitudes Towards Mathematics: Students were asked a series of questions about their attitudes and perceptions about mathematics.

The **Student Motivation Questionnaire** asked students to describe how hard they tried on the NAEP mathematics assessment, how difficult they found the assessment, how many items they thought they got right, how important it was for them to do well, and how familiar they were with the assessment format.

⇒ Teacher, School, and SD/LEP Student Questionnaires

To supplement the information on instruction reported by students, the mathematics teachers of the students participating in the mathematics assessment were asked to complete a questionnaire about their instructional practices, teaching backgrounds, and characteristics. The teacher questionnaire contained two parts. The first part pertained to the teachers' background and general training. The second part pertained to specific training in teaching mathematics and the procedures the teacher uses for *each class* containing an assessed student, as well as collecting information on teachers' awareness and knowledge of the NCTM *Standards*.

The **Teacher Questionnaire, Part I: Background and General Training** included questions pertaining to gender, race/ethnicity, years of teaching experience, certification, degrees, major and minor fields of study, course work in education, course work in specific subject areas, amount of in-service training, extent of control over instructional issues, and availability of resources for their classroom.

The **Teacher Questionnaire, Part II: Training in Mathematics and Classroom Instructional Information** included questions on the teacher's exposure to various issues related to mathematics and teaching mathematics through pre- and in-service training, ability level of students in the class, whether students were assigned to the class by ability level, time on task, homework assignments, frequency of instructional activities used in class, methods of assessing student progress in mathematics, instructional emphasis given to the mathematics abilities covered in the assessment, and use of particular resources.

A **School Characteristics and Policies Questionnaire** was given to the principal or other administrator of each school that participated in the assessment. This information provided an even broader picture of the instructional context for students' mathematics achievement. This questionnaire included questions about background and characteristics of school principals, length of school day and year, school enrollment, absenteeism, dropout rates, size and composition of teaching staff, policies about grouping students, curriculum, testing practices and uses, special priorities and school-wide programs, availability of resources, special services, community services, policies for parental involvement, and school-wide problems.

The **SD/LEP Student Questionnaire** was completed by the teachers of students who were selected to participate in the assessment sample who were also identified as students with a disability (SD) or categorized as being of limited English proficiency (LEP). Some of these students were determined by the school to be ineligible to be assessed. In order to be excluded from the assessment, a student must have been identified as SD and must not have been mainstreamed at least 50 percent of the time, or was categorized as LEP. In addition, the school staff would have needed to determine that it was inappropriate to include the student in the assessment. This questionnaire asked about the nature of the student's disability or about the student's native language, and the special programs in which the student participated.

2.5 DEVELOPING THE SCIENCE ASSESSMENT

2.5.1 Overview

The science framework for the 1996 National Assessment of Educational Progress (NAEP) was produced under the auspices of the National Assessment Governing Board (NAGB). The consensus process was managed by the Council of Chief State School Officers (CCSSO) who worked with the National Center for Improving Science Education and the American Institutes for Research. Items were developed that were aligned with the specifications described in the framework and were extensively reviewed by specialists in science, measurement, and bias/sensitivity, as well as by government officials and state representatives.

The following sections include a detailed description of the development of the framework, objectives, and items for the 1996 NAEP science assessment. Section 2.5.8 describes the student background questionnaires and the science teacher questionnaire. Additional information on the structure and content of assessment booklets can be found in Chapter 4. Various committees worked on the development of the framework, objectives, and items for the mathematics assessment. The list of committee members and consultants who participated in the 1996 development process is provided in Appendix A.

2.5.2 Development of the Assessment Framework

NAGB is responsible for setting policy for NAEP; this policymaking role includes the development of assessment frameworks and test specifications. Appointed by the Secretary of Education from lists of nominees proposed by the Board itself in various statutory categories, the 26-member board is composed of state, local, and federal officials, as well as educators and members of the public.

The science framework for the NAEP 1996 assessment was developed over a 10-month period between October 1990 and August 1991. The following sections discuss how the specifications and items for science assessment were developed. The assessment instrument, the student assessment booklets, and the student, teacher, school, and SD/LEP questionnaires are also described.

A consensus process run by CCSSO was used to produce the science framework. This process involved two committees: a Planning Committee that conducted much of the actual framework development and a steering committee that provided policy and general guidance for the project. As general guidelines for the Planning Committee, the Steering Committee that recommended that the framework and ensuing science assessment have the following five characteristics:

- The framework should reflect the best thinking about the knowledge, skills, and competencies needed for a high degree of scientific understanding among all students in the United States. Accordingly it should encompass knowledge and use of organized factual information, relationships among concepts, major ideas unifying the sciences, and thinking and laboratory skills. In addition, the framework should be based on current understandings from research of teaching, learning, and students' performance in science.

- The framework and the assessment should address the nature and practices of knowing in science, as different from other ways of knowing; reflect the quantitative aspects of science as well as the concepts of life, earth, and physical sciences; deal

with issues raised by the role of science and technology in society; include practical problem solving in science; take into account the developmental levels of students; and ensure that students with diverse backgrounds are assessed in ways that provide them with equal and fair opportunities to reflect their knowledge and performance.

- Assessment formats should be used that are consistent with the objectives being assessed. A variety of strategies for assessing student performance are advocated, including performance tasks that allow students to manipulate physical objects and draw scientific understandings from the materials before them; constructed-response items that provide insights into students' levels of understanding and ability to communicate in the sciences, as well as their ability to generate, rather than simply recognize information related to scientific concepts and their interconnections; and multiple-choice items that probe students' conceptual understanding and ability to connect ideas in a scientifically sound way.

- The assessment should contain a broad enough range of items at different levels of proficiency for identifying three achievement levels for each grade.

- Information on students' demographic and other background characteristics should be collected. Additional information should be collected from students, teachers and administrators about instructional programs and delivery systems, so that their relationships with student achievement can be ascertained and used to inform program and policy decisions.

A Planning Committee was established to identify goals and objectives and to produce the framework. This Planning Committee met monthly from November 1990 through April 1991 and was joined in the first meeting and final meeting by the Steering Committee, which reviewed and reacted to all framework drafts. During this development process, input and reactions were continually sought from a wide range of committee members both within the field of science and external to it. A list of committee members who participated in the developmental process is provided in Appendix A.

2.5.3 Framework for the 1996 Assessment

The framework for the 1996 science assessment is represented as a matrix with two dimensions represented by three fields of science (*earth science*, *physical science*, and *life science*) and three elements of knowing and doing science (*conceptual understanding*, *scientific investigation*, and *practical reasoning*). In addition, there are two overarching domains that describe science and nature of science and themes. Figures 2-4 to 2-6, respectively, describe the three fields of science, the elements of knowing and doing science, and the overarching domains.

Figure 2-4
Descriptions of the Three Fields of Science

Earth Science

The earth science component assessed centers on objects and events that are relatively accessible or visible. The concepts and topics covered are solid earth (lithosphere), water (hydrosphere), air (atmosphere), and the earth in space. The solid earth consists of composition; forces that alter its surface; the formation, characteristics, and uses of rocks; the changes and uses of soil; natural resources used by humankind; and natural forces within the earth. Concepts and topics related to water consist of the water cycle; the nature of oceans and their effects on water and climate; and the location of water, its distribution, characteristics, and effect of and influence on human activity. The air is broken down into composition and structure of the atmosphere (including energy transfer); the nature of weather; common weather hazards; and air quality and climate. The earth in space consists of setting of the earth in the solar system; the setting and evolution of the solar system in the universe; tools and technology that are used to gather information about space; apparent daily motions of the sun, the moon, the planets and the stars; rotation of the earth about its axis, the earth's revolution around the sun; and tilt of the earth's axis that produces seasonal variations in the climate.

Physical Science

The physical science component relates to basic knowledge and understanding concerning the structure of the universe as well as the physical principles that operate within it. The major sub-topics probed are matter and its transformations, energy and its transformations, and the motion of things. Matter and its transformations are described by diversity of materials (classification and types and the particulate nature of matter); temperature and states of matter; properties and uses of material (modifying properties, synthesis of materials with new properties); and resource management. Energy and its transformations involve different forms of energy; energy transformations in living systems, natural physical systems, and artificial systems constructed by humans; and energy sources and use, including distribution, energy conversion, and energy costs and depletion. Motion is broken down into an understanding of frames of reference; forces and changes in position and motion; action and reaction; vibrations and waves as motion; general wave behavior; electromagnetic radiation; and the interactions of electromagnetic radiation with matter.

Life Science

The fundamental goal of life science is to attempt to understand and explain the nature and function of living things. The major concepts assessed in life science are change and evolution, cells and their functions, organisms, and ecology. Change and evolution includes diversity of life on earth; genetic variation within a species; theories of adaptation and natural selection; and changes in diversity over time. Cells and their functions consists of information transfer; energy transfer for the construction of proteins; and communication among cells. Organism are described by reproduction, growth and development; life cycles; and functions and interactions of systems within organisms. The topic of ecology centers on the interdependence of life—populations, communities, and ecosystems.

Figure 2-5
Descriptions of Knowing and Doing Science

Conceptual Understanding

Conceptual understanding includes the body of scientific knowledge that students draw upon when conducting a scientific investigation or engaging in practical reasoning. Essential scientific concepts involve a variety of information including facts and events the student learns from science instruction and experiences with the natural environment and scientific concepts, principles, laws, and theories that scientists use to explain and predict observations of the natural world.

Scientific Investigation

Scientific investigation probes students' abilities to use the tools of science, including both cognitive and laboratory tools. Students should be able to acquire new information, plan appropriate investigations, use a variety of scientific tools, and communicate the results of their investigations.

Practical Reasoning

Practical reasoning probes students' ability to use and apply science understanding in new, real-word applications.

Figure 2-6
Descriptions of Overarching Domains

The Nature of Science

The nature of science incorporates the historical development of science and technology, the habits of mind that characterize these fields, and methods of inquiry and problem-solving. It also encompasses the nature of technology that includes issues of design, application of science to real-world problems, and trade-offs or compromises that need to be made.

(continued)

Figure 2-6 (continued)
Descriptions of Overarching Domains

Themes

Themes are the "big ideas" of science that transcend the various scientific disciplines and enable students to consider problems with global implications. The NAEP science assessment focuses on three themes: systems, models, and patterns of change.

- Systems are complete, predictable cycles, structures or processes occurring in natural phenomena. Students should understand that a system is an artificial construction created to represent, or explain a natural occurrence. Students should be able to identify and define the system boundaries, identify the components and their interrelationships and note the inputs and outputs to the system.

- Models of objects and events in nature are ways to understand complex or abstract phenomena. As such they have limits and involve simplifying assumptions but also possess generalizability and often predictive power. Students need to be able to distinguish the idealized model from the phenomenon itself and to understand the limitations and simplified assumptions that underlie scientific models.

- Patterns of change involve students' recognition of patterns of similarity and differences, and recognition of how these patterns change over time. In addition, students should have a store of common types of patterns and transfer their understanding of a familiar pattern of change to a new and unfamiliar one.

Table 2-6 summarizes the distribution of assessment time across the three fields of science—*earth*, *physical*, and *life*. These fields provide the basis for the content area scales.

Table 2-6
Percentage Distribution of Items by Grade and Field of Science as Specified in the NAEP Science Framework

Field of Science	Grade 4	Grade 8	Grade 12
Earth Science	33%	30%	33%
Physical Science	33%	30%	33%
Life Science	33%	40%	33%

Table 2-7 shows the distribution of assessment time by *knowing and doing science*.

Table 2-7
Percentage Distribution of Items by Grade and Knowing and Doing Science as Specified in the NAEP Science Framework

Knowing and Doing Science Elements	Grade 4	Grade 8	Grade 12
Conceptual Understanding	45%	45%	45%
Scientific Investigation	45%	30%	30%
Practical Reasoning	10%	25%	25%

A number of items that assess each of the fields of science and each of the ways of *knowing and doing science* also probe *nature of science* and *themes* (*systems, models,* and *patterns of change*). Table 2-8 shows the recommended and actual percentages of assessment time for these two overarching domains.

Table 2-8
Percentage Distribution of Items Devoted to Nature of Science and Themes as Specified in the NAEP Science Framework

Overarching Domains	Grade 4	Grade 8	Grade 12
Nature of Science	≥15%	≥15%	≥15%
Themes	33%	50%	50%

In addition to calling for coverage of the content and cognitive domains described above, the framework instructed that all students participating in NAEP take part in a scientific investigation or hands-on experiment. In addition, it indicated that at least 30 percent of students should complete portions of the assessment involving in-depth examination of certain themes or topics in science.

2.5.4 Developing the Cognitive Items

The 1996 assessment was designed to meet the content specifications elaborated in the framework. Because of the broad content and skills definitions included in the framework, and the need to assess hand-on and theme-based science skills, the exercise development effort was extensive. At each grade, enough blocks were field tested to support 13 operational blocks; each block was 20 minutes at grade 4 and 30 minutes at the older two levels.[8] In addition, other portions of the 1993 field test were used to develop exercises for a special study of students at grade twelve who had received advanced science training. Given the fact that the science assessment was delayed from 1994 until 1996, a supplemental field test was held in 1995 (exercises were developed for this field test in 1994). This round of exercise development was used to build extra exercises for the general science and advanced components of the assessment.

The development of cognitive items involved careful field testing, both locally and nationally, of grade-appropriate questions and tasks for the assessment. Items were selected from a pool of questions

[8] Most of the development and field testing were conducted during 1992 and 1993; supplemental development and field testing of general science blocks was conducted during 1994 and 1995.

that were written by teachers from across the country as well as by science assessment specialists on staff at ETS. The framework stated that the assessment should include some performance-based questions and tasks that require students to reason and solve problems representing real-life applications of science. Final selections of items used in the 1996 assessment were approved by the Science Instrument Development Committee.

The assessment included constructed-response (short and extended) and multiple-choice items. The decision to use a specific item type was based on a consideration of the most appropriate format for assessing the particular objective. Both types of constructed-response items were designed to provide an in-depth view of students' ability to communicate their understanding of important concepts in science. Short constructed-response items (scored with either a 2- or 3-level scoring rubric) were used when students needed to respond briefly in order to demonstrate full comprehension. Extended constructed-response items (scored with a 4- or 5-level scoring rubric) were used when the task required more thoughtful consideration of the problem and engagement in more complex reasoning processes. Some items also required diagrams, graphs, or calculations. It was expected that students could adequately answer the short constructed-response items in about two to three minutes and the extended constructed-response items in about five minutes. In addition, blocks of items were developed that required the manipulation of equipment (hands-on tasks) and others were developed that assessed each of the three themes: *systems*, *models*, and *patterns of change*. In the case of some of the hands-on blocks, compound items were created in which student responses to a variety of items were scored as a single item. Multiple-choice items were used when a straightforward, single correct answer was required.

A carefully developed and proven series of steps was used to create the assessment items. These steps are described earlier in the chapter under Section 2.3.

2.5.5 Development of the Operational Forms

Most of the items for the 1996 science assessment were field tested in February and March 1993; however, since the assessment was delayed from 1994 to 1996 an opportunity was afforded for further items to be field tested in February and March 1995. Each of these field tests involved students in many states and were intended to try out the cognitive items and hands-on-tasks and to give jurisdictions and contractors practice and experience with the proposed materials and tasks. Approximately 500 responses were obtained for each item in each field test.

The field test data were collected, scored, and analyzed in preparation for meetings with the Science Instrument Development Committee. The objectives that guided the review of these items were:

- to determine which items were most suitable for assessing understanding in science in accordance with the framework;
- to select items that displayed appropriate statistical attributes;
- to determine the need for revisions of items that lacked clarity, or had ineffective item formats;
- to determine appropriate timing for assessment items.

Committee members, ETS assessment staff, and NAEP staff reviewed the materials. Item analyses (which provided the percentage of correct responses, the biserial correlations for multiple-

choice and items with a two-level scoring guides, and percentages of responses in each category or at each level of the scoring guide and the polyserial correlations for other constructed-response items) were used as a guide in identifying and flagging for further review those test items that were not measuring the intended objective well.

Once the committees had selected the items, they were rechecked for content, measurement, and sensitivity concerns. In addition, a meeting of representatives from state education agencies was convened to review the items chosen for the components of the operational assessment that were to be administered at the state level (that is, the grade 8 assessment). The federal clearance package containing 13 blocks of cognitive items was submitted to NCES in August 1993. A further clearance package containing two blocks of items was submitted to NCES in 1995. Throughout the clearance process, revisions were made in accordance with changes required by the government. Upon approval, the 15 blocks (assembled into booklets) and questionnaires were ready for printing in preparation for the assessment.

2.5.6 Distribution of Assessment Items

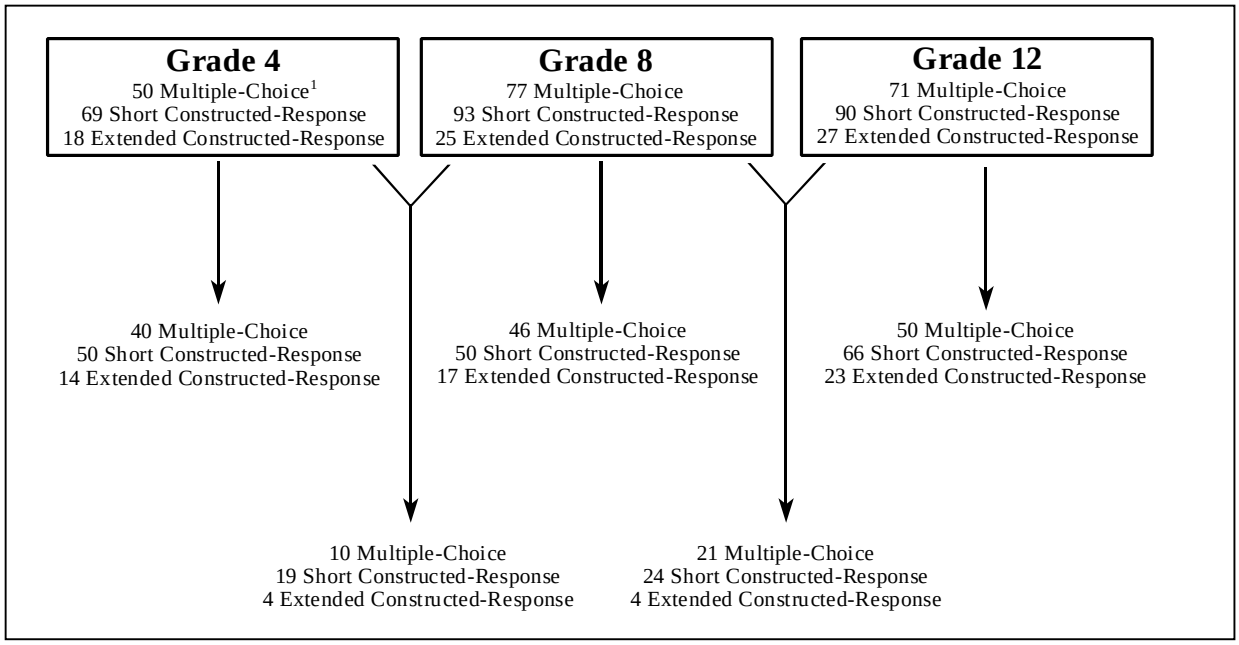
The science assessment developed for use in 1996 was organized according to a series of blocks, each containing a set of items. Some of the blocks were unique to a particular grade level. Other blocks were designed to be given to students in two grades (either 4 and 8 or 8 and 12).

At each grade, the main component of the assessment used for the creation of reporting scales included fifteen different blocks of multiple-choice and constructed-response items (some with both regular and extended constructed-response items). These blocks each were 20-minutes long at grade four and thirty minutes at the other levels. Three 30-minute special study blocks targeted students in advanced mathematics courses. Of the 15 blocks at each grade that were used for the main assessment:

- four blocks at each grade level required students to complete a hands-on science task
- three blocks at each grade required students to work in-depth in a single thematic or topical area
- eight blocks at each grade covered general knowledge and concepts in the fields of science

Figure 2-7 lists the total number of items at each grade level in the 1996 assessment. Of the total of 438 items, there were 167 unique multiple-choice items and 271 unique constructed-response items that made up the 1996 science assessment. Some of these items were used at more than one grade level. As a result, the sum of the items that appear at each grade level is greater than the total number of unique items.

Figure 2-7
Total Number of Items for the 1996 Science Main Assessment



[1] The percentage of time for multiple-choice items is low because the hands-on tasks and theme blocks contain very few multiple-choice items and take up 47% of the assessment time at each grade level.

Table 2-9 summarizes the distribution of assessment time across the three fields of science. Since these fields make up the core of the *Science Framework*, care has been taken to ensure the greatest possible congruence between the proportions used in the assessment and those indicated in the assessment specifications.

Table 2-9
Distribution of Assessment Time by Field of Science

	Earth Science	Physical Science	Life Science
Grade 4	33%	34%	33%
Grade 8	30%	30%	40%
Grade 12	33%	33%	34%

Table 2-10 summarizes the assessment in terms of percentage of time devoted to different cognitive domains. The classification of items into these domains was conducted by both ETS staff and members of the instrument development committee. Every effort was made to meet the specified targets.

Table 2-10
Distribution of Assessment Time Across Cognitive Domains (Knowing and Doing)

	Conceptual Understanding	Scientific Investigation	Practical Reasoning
Grade 4	45%	38%	17%
Grade 8	45%	29%	26%
Grade 12	44%	28%	28%

Exercises assessing each of the fields of science and each of the cognitive domains also probe three themes, (models, systems, and patterns of change), and the students' knowledge of the nature of science. Table 2-11 shows the distribution of assessment time devoted to themes, and Table 2-12 shows the distribution of assessment time devoted to the nature of science.

Table 2-11
Percentage of Assessment Time Devoted to Themes

	Actual	Recommended
Grade 4	53%[1]	30%
Grade 8	49%	50%
Grade 12	55%	50%

[1]Several of the hands-on tasks were classified as themes.

Table 2-12
Percentage of Assessment Time Devoted to Nature of Science

	Actual	Recommended
Grade 4	19%	≥15%
Grade 8	21%	≥15%
Grade 12	31%	≥15%

In addition to the main components of the assessment, an advanced study was conducted at grade 12. Students who were in their fourth year of high-school science were sampled. Each participant took a linking block made up of 17 exercises from the main assessment, and three special advanced blocks: one in biology, one in chemistry, and one in physics. Each of the advanced blocks contained seven multiple-choice and nine constructed-response items.

2.5.7 Questionnaires

As part of the national assessment (as well as the State Assessment), a series of questionnaires was administered to students, teachers, and school administrators. Similar to the development of the cognitive items, the development of the policy issues and questionnaire items was a consensual process that involved staff work, field testing, and review by external advisory groups. A Background Questionnaire Committee drafted a set of policy issues and made recommendations regarding the design of the items. They were particularly interested in capitalizing on the unique properties of NAEP and not duplicating other surveys (e.g., the National Survey of Public and Private School Teachers and Administrators, the School and Staffing Study, and the National Educational Longitudinal Study). The policy issues, items, and field test results were reviewed by the group of external consultants who identified specific items to be included in the final questionnaires. In addition, the Science Instrument Development Committee and state representatives were consulted on the appropriateness of issues addressed in the questionnaires as they relate to science instruction and performance. The items underwent internal ETS review procedures to ensure fairness and quality and were then assembled into questionnaires. The questionnaires were then submitted to the Office of Management and Budget (OMB) for approval.

⇒ **Student Questionnaires**

In addition to three blocks of cognitive items, each booklet in the assessment included three student questionnaires. Two of these were sets of general and science background questionnaires designed to gather contextual information about students, their instructional experiences in science, and their attitudes toward science. The third questionnaire was given to students at the end of each booklet to determine students' motivation in completing the assessment and their familiarity with assessment tasks.

The **Student Demographics (common background) Questionnaire** included questions about race/ethnicity, mother's and father's level of education, types of reading materials in the home, and school attendance.

The **Science Background Questionnaire** included questions that addressed the following.

Attitudes Towards Sciences: Students were asked a series of questions about their attitudes and perceptions about science.

Time Spent Studying Science: Students were asked to describe both the amount of instruction they received in science and the time spent on science homework.

Instructional Practices: Students were asked to report their instructional experiences related to science in the classroom, including group work, special projects, and writing in response to science. In addition, they were asked about the instructional practices of their science teachers.

The **Student Motivation Questionnaire** asked students how many questions they thought they got right on the NAEP science assessment, how difficult they found it, how hard they tried, how important it was for them to do well, and how often they wrote long answers on tests or assignments for science.

⇒ Teacher, School, and SD/LEP Student Questionnaires

To supplement the information on instruction reported by students, the science teachers of the students participating in the assessment were asked to complete a questionnaire that addressed teachers' background and general training as well as their science preparation and information concerning science instruction.

The **Teacher Questionnaire, Part I: Background and General Training** included questions about gender, race/ethnicity, years of teaching experience, certification, degrees, major and minor fields of study, course work in education, course work in specific subject areas, amount of in-service training, professional development activities, and availability of resources for their classroom.

The **Teacher Questionnaire, Part II: Science Preparation and Science Instructional Information** included questions on the number and types of science courses taken over the past two years, membership in science organizations, frequency of instructional activities such as asking students to prepare a written science report or an oral science report, emphasis on objectives such as developing science problem-solving skills, methods used to assess student progress in science, and ability level of students in class.

A **School Characteristics and Policies Questionnaire** was given to the principal of each school that participated in the assessment program. This questionnaire asked about background and characteristics of school principals, length of school day and year, school enrollment, absenteeism, drop-out rates, size and composition of teaching staff, policies about grouping students, curriculum, testing practices and uses, special priorities and school-wide programs, availability of resources, special services, community services, policies for parental involvement, and school-wide problems.

The **SD/LEP Student Questionnaire** was completed by the teachers of those students who were selected to participate in the assessment sample and were identified as students with a disability (SD) or were categorized as being of limited English proficiency (LEP). Some of these students were determined by the school to be ineligible to be assessed. In order to be excluded from the assessment, a student must have been identified as SD and must not have been mainstreamed at least 50 percent of the time, or was categorized as LEP. In addition, the school staff would have needed to determine that it was inappropriate to include these students in the assessment. This questionnaire asked about the nature of the student's disability or about the students' native language, and the special programs in which the student participated.

Chapter 3

SAMPLE DESIGN[1]

Leslie Wallace and Keith F. Rust
Westat, Inc.

3.1 INTRODUCTION

The samples for the 1996 NAEP assessment were selected using a complex multistage sample design involving the sampling of students from selected schools within 94 selected geographic areas, called primary sampling units (PSUs), across the United States.

The long-term trend sample design had four steps in the selection process and the main sample design had five steps in the selection process:

1. selection of geographic PSUs (counties or groups of counties),
2. selection of schools within PSUs,
3. assignment of sample type to schools (main samples only),
4. assignment of session types to schools, and
5. selection of students for session types within schools.

The samples were drawn for the three different age classes,[2] and for each age class the samples were of two distinct types. The first type consisted of the cross-sectional or "main" samples, while the second type consisted of the long-term trend samples. The populations surveyed with each of these sample types are defined in Table 1-1 in Chapter 1. Separate samples of schools were required for the long-term trend samples and main samples, because of various differences in the calendar period for test administration, the format of the administration, the fact that the trend samples include age-based samples, whereas main samples do not and, in the case of age class 17, the grade definition of the population of interest. (See the description of Table 1-1 in Chapter 1.)

In addition to representing the respective populations as a whole, for the main samples there was oversampling of nonpublic schools, and of public schools with moderate or high enrollment of Black or Hispanic students (see Section 3.3). This oversampling was undertaken to increase the sample sizes of nonpublic-school students and minority students, so as to increase the reliability of estimates for these groups of students. These oversampling rates have been used in the past several rounds of NAEP. The oversampling rates were based on experience, after attempting to report results for these groups in assessments where no oversampling was used.

[1] Ralph DiGaetano, Keith F. Rust, and Leslie Wallace were responsible for the design and implementation of the sampling process for the 1996 NAEP assessments.

[2] The term "age class" is used in this report when it is appropriate to discuss one of the three student cohorts in a general way (not necessarily in reference to a specific sample). For the 1996 assessment, age class 9 refers to age 9 or grade 4 long-term trend, or grade 4 main sample students; age class 13 refers to age 13 or grade 8 long-term trend, or grade 8 main sample students; and age class 17 refers to age 17 or grade 11 long-term trend, or grade 12 main sample students.

The overall assessment period fell into three time periods—fall, winter, and spring. Not all assessment components were conducted in each time period. Table 3-1 shows the relationship between the various sample components and the assessment periods. The sizes of the PSU and school samples and the procedures for their selection were determined by the assessment period, as well as by the population to be surveyed and the method of administration in each case.

Table 3-1
Assessment Type by Age Class and Assessment Period

Age Class/Assessment	Fall	Winter	Spring
9			
Main	—	1/3/96 - 3/29/96	—
Long-Term Trend	—	1/3/96 - 3/8/96	—
13			
Main	—	1/3/96 - 3/29/96	—
Long-Term Trend	10/9/95 - 12/22/95	—	—
17			
Main	—	1/3/96 - 3/29/96	—
Long-Term Trend	—	—	3/11/96 - 5/10/96

Special trend samples were required because:

- The long-term trend samples had different school and student eligibility requirements than the main samples. Both grade- and age-eligible students were targeted in long-term trend, and only grade-eligible students were targeted in the main samples. This meant that schools with any of several grades were eligible for long-term trend (grades 2 to 6 for age class 9, grades 6 to 9 for age class 13, and grades 9 to 12 for age class 17), while only schools with grades 4, 8, or 12 were eligible for the main samples.

- The conditions for administration of the assessment varied considerably between the main sample and long-term trend sample sessions.

- The need in the long-term trend samples for four distinct session types for age class 9 and 13 and three for age class 17, together with the need for up to six distinct session types for the main samples, made it not feasible to conduct both main sample sessions and long-term trend sessions in a given school. For long-term trend, the session types were spiral booklets 51-56 and tape booklets 91-93 for ages 9 and 13, and spiral booklets 51-56 and tape booklets 84-85 for age 17. For the main samples, the session types were mathematics, science, mathematics estimation, and mathematics theme at all grades (4, 8, and 12), advanced mathematics at grades 8 and 12, and advanced science at grade 12.

- For age classes 13 and 17, the main sample administrations were conducted in the winter; while the long-term trend sample administrations were conducted in the fall

and spring respectively. The fall and spring administration periods match administration periods used in NAEP as far back as 1969-71.

This chapter gives details of the sample selection procedure, and information on the results of the sampling process. Further details are given in the report *The 1996 NAEP Sampling and Weighting Report* (Wallace & Rust, 1999).

3.2 PRIMARY SAMPLING UNITS

In the first stage of sampling, the United States (the 50 states and the District of Columbia) was divided into geographic primary sampling units (PSUs). The PSUs are those that were used beginning in 1994 and incorporate 1990 U.S. Census information. With a few exceptions, each PSU met a minimum size requirement (a 1990 U.S. Census population of at least 60,000 in the Northeast and Southeast and 45,000 in the Central and West regions) and comprised either a consolidated metropolitan statistical area (CMSA), a metropolitan statistical area (MSA), a single county, or (more likely in the case of nonMSA PSUs) a group of contiguous counties. In the case of New England MSAs, which are not formed from whole counties, the corresponding New England County Metropolitan Areas (NECMAs), which are defined in terms of whole counties, were designated as PSUs. The PSUs were designed to serve as the PSUs for NAEP samples from 1994 until 2002. Thus 1990 total population was used as a size measure, rather than 1990 school age population, as this was considered likely to correlate more highly with school age population over this period. Each PSU was contained entirely within one of the four NAEP regions defined in Table 3-2. These NAEP regions were used to stratify the PSUs, ensuring that each region was adequately represented in the various assessment samples.

Table 3-2
Geographic Regions Used for Stratification

Northeast	Southeast	Central	West
Connecticut	Alabama	Illinois	Alaska
Delaware	Arkansas	Indiana	Arizona
District of Columbia	Florida	Iowa	California
Maine	Georgia	Kansas	Colorado
Maryland	Kentucky	Michigan	Hawaii
Massachusetts	Louisiana	Minnesota	Idaho
New Hampshire	Mississippi	Missouri	Montana
New Jersey	North Carolina	Nebraska	Nevada
New York	South Carolina	North Dakota	New Mexico
Pennsylvania	Tennessee	Ohio	Oklahoma
Rhode Island	Virginia[1]	South Dakota	Oregon
Vermont	West Virginia	Wisconsin	Texas
Virginia[1]			Utah
			Washington
			Wyoming

[1] That part of Virginia that is part of the Washington, DC-MD-VA metropolitan statistical area at the time of the 1990 Census, is included in the Northeast region; the remainder of the state is included in the Southeast region.

In a few cases an MSA crossed region boundaries. Such MSAs were split into two or more PSUs as necessary (e.g., the Cincinnati OH-KY-IN MSA was split into the Cincinnati OH-IN PSU in the Central region and the Cincinnati KY PSU in the Southeast). Ninety-four PSUs were selected for the main samples and 52 PSUs were selected for the long-term trend samples, as described below.

For the main samples, the 22 largest PSUs were included with certainty. The inclusion of these PSUs in the sample with certainty provided an approximately optimum, cost-efficient sample of schools and students when samples were drawn within them at the required national sampling rate. The 22 largest PSUs by region were:

22 Largest PSUs by Region

Northeast:
- Baltimore, MD MSA
- Boston-Lawrence-Salem-Lowell-Brockton, MA NECMA
- New York-Northern New Jersey-Long Island, NY-NJ CMSA (excluding that part in CT)
- Philadelphia-Wilmington-Trenton, PA-NJ-DE-MD CMSA
- Pittsburgh-Beaver Valley, PA CMSA
- Washington, DC-MD-VA MSA

Southeast:
- Atlanta, GA MSA
- Miami-Fort Lauderdale, FL CMSA
- Tampa-St. Petersburg-Clearwater, FL MSA

Central:
- Chicago-Gary-Lake County, IL-IN-WI CMSA
- Cleveland-Akron, OH CMSA
- Detroit-Ann Arbor, MI CMSA
- Minneapolis-St. Paul, MN-WI MSA
- St. Louis, MO-IL MSA

West:
- Dallas-Fort Worth, TX CMSA
- Denver-Boulder, CO CMSA
- Houston-Galveston-Brazoria, TX CMSA
- Los Angeles-Anaheim-Riverside, CA CMSA
- Phoenix, AZ MSA
- San Diego, CA MSA
- San Francisco-Oakland-San Jose, CA CMSA
- Seattle-Tacoma, WA CMSA

The remaining smaller PSUs were not guaranteed to be selected for the sample. These were grouped into a number of noncertainty strata (so called because the PSUs in these strata were not included in the sample with certainty), and one PSU was selected from each stratum. The PSUs were classified into four regions, each containing about one-fourth of the U.S. population. These regions were defined primarily by state (Table 3-2). In each region, noncertainty PSUs were classified as MSA (metropolitan) or nonMSA (nonmetropolitan) according to 1990 definitions. The resulting major strata are shown in Table 3-3.

Table 3-3
The Sampling Major Strata and the Number of Noncertainty Strata in Each

Region	Number of Strata for MSA PSUs	Number of Strata for NonMSA PSUs	Total Strata
Northeast	6	4	10
Southeast	12	12	24
Central	8	12	20
West	10	8	18
Total	36	36	72

Within each major stratum, further stratification was achieved by ordering the noncertainty PSUs according to several additional socioeconomic characteristics, yielding 72 strata. The number of such strata formed within each major stratum is shown in Table 3-3. The strata were defined so that the aggregate of the measures of size of the PSUs in a stratum was approximately equal for each stratum. The size measure used was the population from the 1990 Census. The characteristics used to define strata were the percent minority population, the percentage change in total population since 1980, the per capita income, the percent of persons age 25 or over with college degrees, the percent of persons age 25 or over who completed high school, and the civilian unemployment rate. Up to four of these characteristics were used in any one major stratum. For each major stratum, the characteristics used were chosen by modeling PSU-level mean reading proficiency scores for 1988, 1990, and 1992. The characteristics chosen were the best predictors of PSU-level mean reading proficiency scores in these models. One PSU was selected with probability proportional to size from each of the 72 noncertainty strata. That is, within each stratum, a PSU's probability of being selected was proportional to its population. The PSUs were selected with probability proportional to size (PPS) with the twin aims of obtaining approximately self-weighting samples of students, and having approximately equal workloads in each PSU.

Samples of 94 PSUs each were drawn for the 1994, 1996, 1998, 2000, and 2002 main samples simultaneously. They were drawn to minimize overlap of the PSUs from one assessment to the next, except that certainty PSUs were retained in each assessment year, and some of the larger noncertainty PSUs are in the sample for more than one of these assessment years. Each main sample of 94 PSUs was drawn from a population of about 1,000 PSUs. Primarily because of the use of MSAs as PSUs, PSUs varied considerably as to their probability of selection, since they varied greatly in size. In 1996, the 36 selected noncertainty MSA PSUs had probabilities of selection ranging from 0.04 to 0.58, while the 36 selected nonMSA PSUs had probabilities ranging from 0.03 to 0.08. Parts of 43 states were included in the main sample PSUs.

For the long-term trend samples, 52 PSUs were selected. The long-term trend samples were much smaller than the main samples and used separate field staff. Fewer PSUs were used for the long-term trend samples to avoid having the sample spread too thinly across PSUs. The long-term trend PSUs were drawn for 1994, 1996, 1998, 2000, and 2002 to minimize overlap of the trend PSUs from one assessment to the next, and to minimize overlap between trend and main samples within the same assessment.

The 10 largest main sample certainty PSUs were also included with certainty in the long-term trend samples. Six additional PSUs were selected systematically and with probability proportional to the 1990 population from the 12 remaining main sample certainties. Finally, 36 PSUs were selected from the 72 noncertainty strata so that the overall procedure was equivalent to systematic sampling with probabilities proportional to the 1990 population. The 72 noncertainty strata from the main sample design were paired, and one PSU per pair was selected for the trend samples. Note that the noncertainty long-term trend PSUs are not a subsample of the noncertainty main sample PSUs, in order to minimize the burden on a given school district in any one year.

3.3 SELECTION OF SCHOOLS

In the second stage of sampling, the public schools (including Bureau of Indian Affairs (BIA) schools and Department of Defense Education Activity (DoDEA) schools) and nonpublic schools (including Catholic schools) were listed according to the grade ranges associated with the three age classes. The lists of schools were obtained from two sources. Regular public, BIA, and DoDEA schools were obtained from the 1994 list of schools maintained by Quality Education Data, Inc. (QED). Regular public schools are schools with students who are classified as being in a specific grade (as opposed to

schools having only "ungraded" classrooms). This includes statewide magnet schools and charter schools. Catholic and other nonpublic schools were obtained from both QED and the Private School Survey (PSS) developed for the National Center for Education Statistics' 1993-1994 School and Staffing Survey. The majority of the PSS list comes from complete enumeration of schools, but a small portion of the PSS list was restricted to a sample of counties selected for the survey. Certain PSS counties, generally large in population, were also included, independently by chance, in the NAEP sample PSUs. The schools from such counties were added to the NAEP sampling frame after steps were taken to eliminate duplicates with the QED list of nonpublic schools. In previous years, nonpublic schools were also obtained from telephone directories. This process was not repeated in 1996 because the use of the PSS files supplanted the need for this supplement.

Table 3-4 shows the numbers of schools included in the various sampling frame components. The population of eligible schools for each age class was restricted to the selected PSUs. Main sample schools were selected from the 94 main sample PSUs and long-term trend schools were selected from the 52 long-term trend PSUs. Note that there are relatively large numbers of nonpublic schools that are listed in the QED or PSS only. The discrepancy between the schools contained in the PSS dataset versus those in the QED dataset is primarily due to two factors: (1) the relative outdatedness of the two school lists, and (2) PSS's inclusion of a special area supplement designed to find schools not normally available on lists.

Table 3-4
Grade Definition of School Eligibility for Sampling Frame Inclusion and Frame Sizes, Main and Long-Term Trend Samples[1]

Sample[2]	Sampling Frame Included Schools With Any of Grades	Public[3]	Nonpublic from QED Only[4]	Nonpublic from PSS Only[4]	Nonpublic from QED and PSS[4]
Main Samples					
Grade 4	4	18,046	1,308	1,716	7,300
Grade 8	8	6,093	1,029	1,322	6,409
Grade 12	12	4,357	578	869	2,474
Long-Term Trend Samples					
Age Class 9	2 - 5	15,873	1,147	1,600	6,019
Age Class 13	6 - 9	13,667	1,089	1,474	6,592
Age Class 17	9 - 12	3,495	491	829	2,106

[1] The numbers in this table reflect the full samples, including all sample types (see Section 3.4).
[2] "Age Class" is a term that refers to either an age or a grade definition of the samples. For the 1996 main assessments, unlike for previous main assessments, only grade samples were drawn. Long-term trend age class definitions vary by subject area. They are explained more fully in Chapters 14 through 17.
[3] Public, BIA, and DoDEA schools
[4] Catholic and other nonpublic schools

Any school having one or more of the eligible grades, and located within an appropriate PSU, was included in the sampling frame of schools (the list of schools from which the samples of schools were drawn) for a given sample. For each age class in the long-term trend samples, only a fraction of one percent of age-eligible students was enrolled in ineligible schools. An independent sample of schools was selected for each of the age classes, separately for main and long-term trend. Thus, some schools were selected for assessment of two age classes, and a few were selected for all three. For all three age classes, a sample of schools was first drawn for the long-term trend assessments. The schools selected for

long-term trend at a particular age class were excluded from the sampling frame when the samples of schools for the corresponding grade were drawn for the main assessments. In addition, the schools selected for the 1996 NAEP State Assessment program at grade 8 were excluded from the sampling frame for the main samples at that grade. In regard to both of these situations, adjustments were made to the sampling weights to reflect the appropriate probabilities of selection to yield unbiased estimates for both long-term trend and main samples.

For each NAEP sample, schools were selected (without replacement) across all PSUs with probabilities proportional to assigned measures of size. In those certainty PSUs included in both main and trend samples, the probability of selection for long-term trend for any school in a given age class was capped at 0.5, to ensure that adequate schools remained to be selected for the main sample. For long-term trend samples, the measure of size used for each school was the estimated number of age eligible students in the school, since for each age class the large majority of students selected were assigned to sessions for which only students of the appropriate age were eligible. In most schools having the modal grade, some additional students were selected who were in the modal grade but not age-eligible, so that the maximum sample size of students within a school was about 80 grade- and age-eligible students. Equal measures of size were assigned to schools containing estimates of age-eligible students ranging from 20 to 60 for each age class. Schools with more than 60 age-eligible students were selected with probabilities proportional to the measure of size. Schools with fewer than 20 estimated age-eligible students were assigned somewhat lower measures of size, and thus lower probabilities of selection, since assessment in these schools involved substantially higher per-student administrative costs.

For the main samples, equal measures of size were assigned to schools containing estimates of grade-eligible students ranging from 20 to 120 (for grade 4), 20 to 150 (for grade 8), or 20 to 180 (for grade 12). Schools larger than the indicated maximum size were selected with probabilities proportional to the measure of size. This procedure was used so as to obtain approximately self-weighting samples of students (i.e., students selected with approximately equal overall probabilities) at each grade. Three variations to the overall goal of self-weighting samples were implemented. Schools with fewer than 20 estimated grade-eligible students were assigned somewhat lower measures of size, and thus lower probabilities of selection. This was designed to increase cost efficiency. Each public school designated as high-minority (with over 10 percent Black and/or Hispanic enrollment for grades 4 and 8, or over 15 percent Black and/or Hispanic enrollment for grade 12) was given double the probability of selection of a public school, not designated high-minority, of similar size in the same PSU. Such high-minority schools were oversampled in order to enlarge the sample of Black and Hispanic students, thereby enhancing the reliability of estimates for these groups. For a given overall size of sample, this procedure reduces somewhat the reliability of estimates for all students as a whole and for those not Black or Hispanic. Each nonpublic school was given triple the probability of selection of a public school not designated high-minority of similar size in the same PSU. These greater probabilities of selection were used to ensure adequate samples of nonpublic-school students in order to allow the derivation of reliable estimates for such students. No subgroups (high minority schools or nonpublic schools) were oversampled in the long-term trend samples.

The total number of schools selected for each age class in both the long-term trend and main samples was such that the predesignated student sample sizes would be achieved by selecting all eligible students in a selected school, up to the maximum sizes indicated above. The target sample size also allowed for losses due to nonparticipation of selected schools and students and the exclusion of students from the assessment. This design, with the important exceptions described above, had the goal of yielding a sample of students in a given age class or grade with approximately uniform probabilities of selection. The efforts to oversample nonpublic-school students and minority students in the main samples and the practical constraints on the sample size within each school resulted in some substantial violations

of this general goal. The distributions of selection probabilities of the selected students, as reflected in their sampling weights, are mentioned in Chapter 10.

The QED files do not contain schools that opened between 1994 and the time of the assessments. Therefore, special procedures were implemented to be sure that the NAEP assessment represented students in new public schools. Small school districts, which generally contained only one eligible school for a given age class, were handled differently from large school districts, which generally contained more than one eligible school for a given age class. In small school districts, the schools selected for a given age class were thought to contain all students in the district that were eligible for the assessment. Districts containing such schools in the school sample were asked if other schools with the appropriate grades for the assessment existed, and if so, they were automatically included in the assessment. For large school districts, a district-level sampling frame was constructed from the schools on the QED file that were eligible for one of the national assessments. Then districts were sampled systematically with probabilities proportional to a measure of size. In most cases, the measure of size was total district enrollment, but a minimum measure of size was used in districts below a certain cutoff. Each sampled district was asked to update lists of eligible schools according to information on the QED files. Sampling frames of eligible new schools for these large districts were then constructed separately for both main and long-term trend samples at each age class, and separate samples of new schools were selected systematically with probability proportional to eligible enrollment using the same sampling rates as for the original sample. Seven new schools were added to the main samples: two at grade 4, three at grade 8, and two at grade 12. Four new schools were added to the long-term trend samples: one at age class 9, three at age class 13, and none at age class 17. Although new school sampling procedures were applied at age class 17, no new schools were selected since schools with the necessary characteristics were not available. All new schools added to the sample were obtained from large districts.

In a few PSUs where school refusals were relatively heavy for a particular sample, substitute school selections were made, replacing the refusals (to the extent feasible) with schools from within the same PSU and similar in size, affiliation (public, Catholic, or other nonpublic), grade span, and minority composition. The goal of this procedure was to maintain the student sample sizes needed, while keeping variance and nonresponse bias at acceptable levels. For the main samples, 31 substitute schools were selected using this procedure (10 at grades 4 and 8, and 11 at grade 12), and 28 substitute schools were selected for the long-term trend samples (15 at age class 9, 7 at age class 13, and 6 at age class 17). Tables 3-5 and 3-6 show the number of in-scope schools selected, cooperating, and substituted, in the main and long-term trend samples, respectively. The participation rates given are based on the original sample of schools (excluding substitutes). School participation rates for grade 8 and nonpublic schools in the main samples appear lower compared to those achieved in 1994, while the rates for public schools appear higher. The other response rates are comparable for the two years. Note that since the response rates quoted do not include the substitute selections, the potential for nonresponse bias is likely to be a little less than these rates would indicate. This is because the substitute selections were chosen based on their similarity to the initially refusing selections.

For the main samples, the schools that were participating with no eligible students left for testing had all of their eligible students tested for State NAEP, so that no students were left for the main samples. These schools were accounted for by treating them as nonrespondents for weighting. For long-term trend at age class 13, the considerable numbers of schools selected with no eligible students enrolled resulted primarily from the fact that some schools with grades 6, 7, or 9, but no grade 8, were sampled. Such schools had a reasonable chance of containing some age 13 students. Often they did have a number of eligible students, but sometimes they had none. Because of the grade structure of schools, this occurred most often for age class 13.

Table 3-5

School Sample Sizes, Refusals, and Substitutes for the Main Samples[1]

Status	Grade 4	Grade 8	Grade 12	Total	Public[2]	Nonpublic[3]
Selected, in scope	723	761	779	2,263	1,392	871
Refusals	99	127	160	386	212	174
Participation rate of originally selected schools	86%	83%	79%	83%	85%	80%
1994 participation rate	86%	86%	79%	83%	82%	85%
Participating, no eligible students enrolled	0	0	1	1	0	1
Participating, no eligible students left for testing[4]	20	42	27	89	30	59
Substitutes participating	1	1	2	4	4	0
Final assessed sample	605	593	593	1,791	1,154	637

[1] The numbers in this table reflect the full samples, consisting of schools in the S1, S2, and S3 samples. These sample types are described in Section 3.4.
[2] Public, BIA, and DoDEA schools
[3] Catholic and other nonpublic schools
[4] No students were left for testing because they had been tested in the State Assessment

Table 3-6

School Sample Sizes, Refusals, and Substitutes for the Long-Term Trend Samples[1]

Status	Age Class 9	Age Class 13	Age Class 17	Total
Selected, in scope	291	316	237	844
Refusals	43	51	44	138
Participation rate of originally selected schools	85%	84%	81%	84%
1994 participation rate	87%	82%	81%	83%
Participating, no eligible students enrolled	8	27	2	37
Substitutes participating	8	4	0	12
Final assessed sample	248	242	191	681

[1] The numbers in this table reflect the full samples, consisting of schools in the S1, S2, and S3 samples. These sample types are described in Section 3.4.

3.4 ASSIGNMENT OF SAMPLE TYPE TO SCHOOLS

In order to determine the effect of using different criteria for including students with disabilities and limited English proficient students in the assessment, three different sample types were assigned to the schools selected for the main assessment. In sample type 1 (S1) schools, the inclusion criteria for the main samples were identical to those used in 1990 and 1992 and were intended to be somewhat more rigorously defined than those used in the long-term trend samples. In sample type 2 (S2) schools, new 1996 inclusion criteria were used. In sample type 3 (S3) schools, the new 1996 inclusion criteria were used and accommodations were offered to students with disabilities (SD) and students of limited English proficiency (LEP). For more details of the inclusion criteria and their implementation, and the accommodations offered students, see Chapter 5. The information in this chapter and in Chapter 5 applies to all three sample types or subsamples.

Sample type was assigned to schools separately for each grade. For schools that were not also selected for the State Assessment program, sample type was assigned as follows. At grade 4, 20 percent of the schools were assigned sample type 1, 45 percent were assigned sample type 2, and 35 percent were assigned sample type 3. At grade 8, one-sixth of the schools were assigned sample type 1, and five-twelfths each were assigned sample types 2 and 3. At grade 12, two-thirteenths of the schools were assigned sample type 1, six-thirteenths were assigned sample type 2, and five-thirteenths were assigned sample type 3. Sample type was assigned so that a variety of schools with respect to region, school type, urbanization, and size were in each sample type at each grade.

For schools selected for both the main samples and State Assessment program, sample type was initially assigned as described above, and then reassigned for the main samples as follows. Schools retained their initial sample type assignment for the State Assessment. For the national assessment, schools were ultimately assigned the same sample type as for the State Assessment, with one exception. Schools that were initially assigned to sample type 3 for the national assessment and sample type 2 for the State Assessment, retained these different respective sample types for each assessment. For all other schools, the sample type for the main samples was switched to match the state sample type. The effect of this procedure was to assign sample type 1 to somewhat more schools and sample types 2 and 3 to somewhat fewer schools at grades 4 and 8 than initially assigned.

3.5 ASSIGNMENT OF SESSIONS TO SCHOOLS

Sessions were assigned to the selected schools found to be in-scope at the time of session assignment in the following manner. First, the number of sessions per school was established (three sessions per school were specified for the long-term trend samples and for sample type 1 for the main samples. Five sessions per school were specified for sample types 2 and 3 for the main samples at grades 4 and 8, and six sessions per school were specified for sample types 2 and 3 for the main samples at grade 12). This was the maximum number of sessions that could be administered without creating unduly small session sizes with few eligible students. Thus, in most long-term trend schools, for example, three sessions were conducted. However, schools with fewer than 20 eligible students were asked to conduct only a single session.

Session types associated with each sample are listed in Table 3-7. In the main samples, four to six different session types were conducted at each grade (mathematics, science, mathematics estimation, and mathematics theme at all grades; advanced mathematics at grades 8 and 12; and advanced science at grade 12). All of the session types were not offered for all samples. For long-term trend, four session types were conducted at age classes 9 and 13 (spiral plus three tape sessions), and three session types were conducted at age class 17 (spiral plus two tape sessions). Schools could be assigned multiple sessions of the same type (for example, two spiral and one tape sessions in long-term trend, or three mathematics and three science sessions in the main samples).

Sessions were assigned to schools with three aims in mind. The first was to distribute students to the different session types across the whole sample for each age class so that the target numbers of assessed students would be achieved (in each sample type separately in the main assessments). The second was to maximize the number of different session types that were administered within a given selected school, without violating the minimum session sizes discussed above. The third was to give each student an equal chance of being selected for a given session type regardless of the number of sessions conducted in the school.

3.6 SAMPLING STUDENTS

To facilitate the sampling of students, a consolidated list was prepared for each school of all grade-eligible and age-eligible students (for long-term trend) or all grade-eligible students (for the main assessments) for the age class for which the school was selected. A systematic selection of eligible students was made from this list (unless all students were to be assessed) to provide the target sample size. For schools assigned more than a single session type (the vast majority), students were assigned by Westat district supervisors to one of the various session types using specified procedures.

For each age class, separately for the long-term trend and main samples, maxima were established as to the number of students who would be selected for a given school. In those schools that, according to information on the sampling frame, had fewer eligible students than the established maxima, each eligible student enrolled at the school was selected in the sample for one of the sessions assigned to the school. In other schools, a sample of students was drawn, and then students were assigned to sessions as appropriate. For the main samples, the maximum sample sizes were established by sample type in terms of the number of grade-eligible students: 72 at grades 4 and 8, and 90 at grade 12 for sample type 1; 120 at grade 4, 160 at grade 8, and 180 at grade 12 for sample types 2 and 3. For the long-term trend samples, the maximum at each age class was 60 age-eligible students (about 80 grade- plus age-eligible students in most schools). Note that the number of students actually selected for assessment in a long-term trend sample school generally fell somewhat below 80, because students who were selected for one of the long-term trend tape-administered sessions and were in the modal grade but not age-eligible were subsequently dropped from the sample. Similarly, in the main assessments, at grades 8 and 12 in sample types 2 and 3, students selected for the advanced mathematics and advanced science assessments, who were subsequently found to be ineligible on the basis of their courses taken, were dropped from the samples. This reduced the sample size somewhat in these schools.

The sample of students to be selected in each school was derived in the following manner, both for main and for long-term trend samples. On the basis of data obtained from the school characteristics and policies questionnaire (or the sample frame when the questionnaire data were not obtained in time) an estimate of the number of eligible students was established for each school. For the main samples, the estimated number of grade-eligible students was used; for the long-term trend samples, the number of age-eligible students was used. A Session Assignment Form was generated for each school, showing the line numbers (described below) of the students to be selected, indicating the type of session to be taken by each such student. These line numbers were generated using a sampling interval designed to give the appropriate sample size for each school. Thus, the overall sampling interval was 1.0 for schools in which all eligible students were to be assessed. The appropriate sampling interval was specified for schools with larger numbers of eligible students, such as to give the appropriate maximum sample size (described above for each age class) in the case that the school had an enrollment of eligible students exactly equal to that predicted.

If the Westat supervisor found that, when applied to the numbered list of eligible students assembled in the field for each school, the line numbers generated gave rise to a sample in excess of 120 percent of the appropriate maximum sample size limit specified above, he or she called Westat's central office. By use of a personal computer, new line numbers based on the actual number of eligible students were generated and relayed to the supervisor. A similar revision to the line numbers was made in the case of a school with a sampling interval in excess of 1.0, and eligible enrollment less than 80 percent of that initially estimated. In this latter case, the sample size was increased to the appropriate level. This procedure gave a suitable compromise between control over the sampling rate within each school and operational autonomy and flexibility for Westat field supervisors. Note that in all cases, sampling

intervals were generated in Westat's central office, and stored for use in sample weighting. Supervisors were not required to derive or record within-school sampling rates.

Table 3-7 shows the number of students per school who were assessed for each assessment. Note that, for the various print samples, the number of students assessed per item per school is quite low, even though typically dozens of students were assessed in total in a particular school. Thus, the extent of clustering of the sample is in general quite modest, because most sampled schools conducted a few different assessments with a moderate number of students in each, and more importantly because the use of BIB-spiraling in the print-administered sessions greatly alleviated the effects of clustering the samples of students within schools, for item-level data.

Table 3-7
Number of Students Per School for Each Session Type[1]

Sample	Session Type	Number of Assessed Students	Number of Schools	Mean Number of Students Per Assessment Per School	Mean Number of Students Per Item Per School
Age Class 9	Print Booklets 51-56	5,019	215	23.3	3.9 - 7.8[2]
Long-Term	Tape Booklet 91	1,852	127	14.6	14.6
Trend	Tape Booklet 92	1,721	116	14.8	14.8
	Tape Booklet 93	1,840	125	14.7	14.7
Grade 4	Print Mathematics	10,830	445	24.3	5.6
Main	Print Science	11,578	421	27.5	4.5 - 7.4[2]
	Tape Mathematics Estimation	2,115	120	17.6	17.6
	Print Mathematics Theme	4,004	230	17.4	8.7
Age Class 13	Print Booklets 51-56	5,493	221	24.9	4.1 - 8.3[2]
Long-Term	Tape Booklet 91	1,928	128	15.1	15.1
Trend	Tape Booklet 92	1,866	125	14.9	14.9
	Tape Booklet 93	1,864	124	15.0	15.0
Grade 8	Print Mathematics	11,521	411	28.0	6.5
Main	Print Science	11,971	346	34.6	5.6 - 9.4[2]
	Tape Mathematics Estimation	2,244	104	21.6	21.6
	Print Mathematics Theme	4,227	175	24.2	12.1
	Print Advanced Mathematics	2,365	253	9.3	9.3
Age Class 17	Print Booklets 51-56	4,669	186	25.1	4.2 - 8.4[2]
Long-Term	Tape Booklet 84	1,848	133	13.9	13.9
Trend	Tape Booklet 85	1,691	122	13.9	13.9
Grade 12	Print Mathematics	10,660	430	24.8	5.7
Main	Print Science	11,481	401	28.6	4.6 - 7.7[2]
	Tape Mathematics Estimation	1,889	96	19.7	19.7
	Print Mathematics Theme	3,860	196	19.7	9.8
	Print Advanced Mathematics	2,965	207	14.3	14.3
	Print Advanced Science	2,431	222	11.0	11.0

[1] The numbers in this table reflect the full samples, consisting of schools in the S1, S2, and S3 samples. These sample types are described in Section 3.4.
[2] This number varied because some item blocks appeared more than others in the set of booklets used for this sample.

3.7 OVERSAMPLING OF SD/LEP STUDENTS FOR MAIN SAMPLE MATHEMATICS AND SCIENCE ASSESSMENTS

As noted earlier, in the main assessments for mathematics and science, the procedures for assessing SD and LEP students varied by sample type. The inclusion procedure used in sample type 1 differed somewhat from that used in sample types 2 and 3. SD/LEP students in sample type 3 were offered accommodations not available to other students or to SD/LEP students in other sample types.

As a measure to ensure an adequate sample size of SD/LEP students from both sample types 2 and 3, oversampling procedures were applied for SD/LEP students at all three grades, in sample types 2 and 3 for mathematics, and in sample type 3 for science. In this way, comparisons of the effect of offering accommodations to students would have enhanced power to detect effects.

The procedure for carrying out the oversampling was somewhat different for grade 4 than for grades 8 and 12. This was because of the presence of the advanced mathematics and advanced science samples at grades 8 and 12 only, which offered an opportunity to oversample in a way not possible at grade 4. The general intent of the oversampling was, within each school assigned sessions of regular mathematics in sample types 2 and 3, and regular science in sample type 3, to select SD/LEP students for these assessments at twice the rate at which non SD/LEP students were sampled (or to include all SD/LEP students if there were not sufficient numbers to permit sampling at twice the rate). There was no oversampling of schools as part of the procedure.

At grade 4, the procedure was as follows. In each school where oversampling was to occur, the initial desired sample of students was drawn for each session assigned, from the full list of eligible students. Among those students not selected for any sessions in this way, the SD/LEP students were identified. A sample from among these was drawn, using a sampling rate that would achieve the double sampling rate required overall. In most cases, this involved selecting all such SD/LEP students in the school. If the school was a sample type 3 school assigned to assess both mathematics and science, the extra SD/LEP students so selected were split among mathematics and science in the same proportion as the initial student sample for the school. Thus, if the school was assigned two sessions of science and one of mathematics, two-thirds of these extra SD/LEP students were assigned to science, and one-third to mathematics.

The sampling of additional SD/LEP students was carried out using designated line numbers, indicated on the session assignment form used to generate the samples of students in each school. In this way, the necessary information as to the selection probability of each student was retained for use in weighting. No reliance was placed on information generated in the field. Field supervisors had only to follow the prespecified sampling instructions.

At grades 8 and 12, a different approach was taken. As a result of the pattern of assigning sessions to schools, it was the case that in every school in which there were students remaining who were not selected for assessment in the initial sampling phase (so that there were in fact SD/LEP students available for oversampling), a session of either advanced mathematics, or advanced science (at grade 12), was assigned. This was the result of the scheme for assigning sessions to schools efficiently; it was not a condition imposed in order to facilitate oversampling. The SD/LEP students assigned as an oversample for the regular mathematics and (in the case of sample type 3) science assessments were those SD/LEP students who were initially selected for the advanced mathematics and science samples, but who were not eligible for those assessments because they had not taken the appropriate set of courses. Thus, for grades 8 and 12, the oversampling of SD/LEP students took place only among that subpopulation that was not eligible for the advanced assessments.

It was assumed that there would be relatively few SD/LEP students who would qualify for the advanced sessions, since nationally about 20 percent of all students were so eligible. To the extent to which there were SD/LEP students who qualified for the advanced mathematics and science assessments, however, the oversampling procedure for regular mathematics and science was not biased, because this was taken into account in the weighting of the regular mathematics and science assessments. This was possible because, for the students in the regular mathematics and science samples, those who were SD/LEP were identified, and those who qualified for the advanced assessments were also identified.

All such additional SD/LEP students were included unless this would have led to the sampling rate of SD/LEP students within the school being more than twice the rate of other students. In such cases, a random subsample of the extra SD/LEP students was selected. As for grade 4, all information needed in the field to carry out the oversampling was contained on the preprogrammed Session Assignment Form, so that the complex weighting process could be carried without the possibility of error being introduced in sampling information obtained from the field. Also, as in grade 4, in sample type 3 schools that were assigned both regular mathematics and regular science sessions, the extra SD/LEP students sampled were assigned in the appropriate proportions.

Since the aim was to oversample by a factor of two where possible, the overall rate of oversampling of SD/LEP students was instead less than two. That is because in smaller schools there were no students remaining who had not already been assigned to a session. Again, the weighting procedures ensured that the results were not biased as a result of the relative under representation of SD/LEP students from smaller schools.

Table 3-8 shows the results of the oversampling efforts for each grade and sample type for mathematics and science. The weighted results show the proportion of the sample that would have been SD/LEP students had no oversampling been attempted. The focus should be on sample types 2 and 3 for mathematics and sample type 3 for science, since this is where the oversampling of SD/LEP students occurred. The extent to which the unweighted percentage of SD/LEP students exceeds the weighted percentage is a measure of the effectiveness of the oversampling. As can be seen, the procedure was effective in increasing the sample of SD/LEP students considerably at grades 8 and 12 for both subjects, but was not very effective at grade 4 for either subject. To have increased the sample of SD/LEP students further at grade 4 would have required the assessment of additional schools.

Table 3-8
Percentage of Sampled Students Who Were Specified as SD/LEP in the 1996 Main Samples - Mathematics and Science

Subject/ Sample Type[1]	Grade 4		Grade 8		Grade 12	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Mathematics/S1	12.4%	13.2%	9.5%	10.2%	6.4%	6.8%
Mathematics/S2	17.9%	15.7%	19.9%	11.5%	15.3%	8.1%
Mathematics/S3	17.0%	15.4%	18.9%	12.1%	15.7%	8.5%
Total	15.8%	14.7%	16.1%	11.3%	12.7%	7.8%
Science/S2	16.5%	15.8%	13.1%	11.9%	9.5%	9.2%
Science/S3	17.7%	16.7%	18.9%	10.9%	13.6%	7.4%
Total	16.9%	16.2%	15.2%	11.4%	10.9%	8.3%

[1] The numbers in this table reflect the full samples, consisting of schools in the S1, S2, and S3 samples. These sample types are described in Section 3.4.

3.8 EXCLUDED STUDENTS

School staff completed an SD/LEP questionnaire for each sampled student identified as IEP (with an individualized education program) or LEP. Some of these students were deemed unassessable by school authorities and were excluded from the assessment. For the long-term trend samples, a distinct sample of excluded students was identified at each age class. For the main samples, a distinct sample of excluded students was identified for each subject and sample type combination.

The inclusion criteria for the main samples differed somewhat from those used for the long-term trend samples. In addition, the inclusion criteria for the main samples differed by sample type. In sample type 1 schools, the inclusion criteria for the main samples were identical to those used in 1990 and 1992, and were intended to be somewhat more rigorously defined than those used in the long-term trend samples. In sample type 2 schools, new 1996 inclusion criteria were used. In sample type 3 schools, the new 1996 inclusion criteria were used and accommodations were offered to SD/LEP students.

For the long-term trend samples, the inclusion criteria were the same as in past long-term trend assessments, dating back to the early 1980s.

For all samples, students were selected for specific sessions, and the school was then asked to identify those to be excluded. Thus, only age-eligible students were considered for inclusion in the long-term trend tape-administered sessions, whereas both age- and grade-eligible students were considered for inclusion in the print-administered long-term trend samples. The samples of excluded students for the long-term trend samples were weighted in such a way as to account for this procedure appropriately (see Chapter 10).

Table 3-9

Student Exclusion Rates by Age Class and School Type and Sample Type, Weighted

Subject/ Sample Type[1]	Age Class 9			Age Class 13			Age Class 17		
	Public	Non-Public	Total	Public	Non-Public	Total	Public	Non-Public	Total
Main Samples									
Mathematics/S1	5.6%	1.0%	5.2%	4.5%	0.2%	4.1%	3.4%	0.1%	3.0%
Mathematics/S2	9.1%	1.3%	8.1%	4.7%	0.1%	4.3%	3.5%	0.1%	3.2%
Mathematics/S3	4.4%	0.0%	3.9%	3.4%	0.0%	3.1%	3.1%	0.2%	2.8%
Science/S2	9.2%	0.3%	8.2%	4.7%	0.2%	4.3%	4.3%	0.4%	3.9%
Science/S3	6.5%	0.1%	5.9%	3.7%	0.3%	3.4%	2.8%	0.2%	2.6%
Estimation/All	6.7%	0.0%	5.8%	5.1%	0.0%	4.7%	4.0%	0.0%	3.5%
Theme/All	7.6%	0.6%	6.8%	4.4%	0.0%	4.0%	3.6%	1.4%	3.4%
Advanced Mathematics/All	—	—	—	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Advanced Science/All	—	—	—	—	—	—	0.0%	0.0%	0.0%
Long-Term Trend Samples									
Reading/Writing Print	6.6%	0.3%	5.9%	5.9%	0.4%	5.4%	5.3%	0.2%	5.0%
Mathematics/Science Tape	6.3%	0.6%	5.6%	4.8%	0.2%	4.3%	3.5%	0.4%	3.3%

[1] The numbers in this table reflect the full samples, consisting of schools in the S1, S2, and S3 samples. These sample types are described in Section 3.4.

Table 3-9 shows the rates of exclusion for each age class, subject, sample type (for math and science), and school type for the long-term trend and main samples. The most marked effects are the much higher rates of exclusion in public schools than in nonpublic, and the higher rates of exclusion at lower grades. The former phenomenon is no doubt a function of the greater prevalence of special education and language minority programs in public schools. The higher exclusion rates at lower ages, which occurred also in other years, result from the greater proportion of students at these grades who are excluded for reasons of limited English proficiency. In certain areas of the United States, fourth-grade public-school students whose native language is Spanish are taught predominantly in Spanish, and in these schools a very high proportion of sampled students are excluded. Factors that may limit the comparability of these rates to those in previous years are the different inclusion criteria, oversampling of SD/LEP in some subjects, the different subjects assessed, and the inclusion of only grade eligible students in the main samples in 1996.

3.9 STUDENT PARTICIPATION RATES

Table 3-10 summarizes the rates of participation of invited students. The set of invited students consists of the selected students, after removing the excluded students and, in the case of long-term trend samples, removing those students selected for tape-administered sessions who were not age-eligible. For a given session, a makeup session was called for when, for various reasons, more than a predetermined tolerable number of invited students failed to attend the originally scheduled session to which they were invited. The participation rates given in the table express the number finally assessed as a percentage of those initially invited in the participating schools. Participation rates are shown for the main and long-term trend samples and for public and nonpublic schools separately in the case of the main samples. Overall participation rates are also shown for comparable samples from the 1994 NAEP assessment. The table shows that student participation rates in 1996 are similar to those experienced in 1994. The rates increased slightly at age class 9 for both samples, and remained fairly steady for the other samples. At all age classes, the participation rate of nonpublic-school students exceeds that of public-school students, with the difference, both relative and absolute, increasing with age class.

Table 3-10

Student Participation Rates by Age Class and School Type, Unweighted[1]

	1996 Public		1996 Nonpublic		1996 Combined		1994
Samples	Number Invited	Participation Rate	Number Invited	Participation Rate	Number Invited	Participation Rate[2]	Participation Rate[2]
Age Class 9							
Long-Term Trend	9,715	95.4%	1,204	96.8%	10,919	95.5%	94.2%
Main	24,082	95.1%	5,834	96.6%	29,916	95.4%	93.2%
Age Class 13							
Long-Term Trend	10,980	91.6%	1,152	95.0%	12,132	91.9%	92.2%
Main	28,351	92.3%	6,368	96.7%	34,719	93.1%	91.0%
Age Class 17							
Long-Term Trend	9,051	83.4%	717	91.9%	9,768	84.0%	84.1%
Main	34,199	77.3%	7,473	91.5%	41,672	79.9%	81.1%

[1] The numbers in this table reflect the full samples, including all sample types (see Section 3.4).

[2] Somewhat different inclusion criteria were used for the main samples than for the long-term trend samples in each year, and for the main samples in 1994 versus 1996. The total rates for the main samples are based on a relatively greater contribution from nonpublic-school students. Nonpublic-school students constitute about 18% of the 1996 main samples, 16% of the 1994 main samples, and 11% of the 1994 and 1996 long-term trend samples.

3.10 OVERALL STUDENT PARTICIPATION RATES

The combined impact of school nonparticipation and student absenteeism from sessions within participating schools is summarized in Table 3-11. The table shows the percentages of students assessed, from among those who would have been assessed if all initially selected schools had participated, and if all invited students had attended either an initial or make-up session. The results show that, consistent with earlier rounds of NAEP, the overall level of participation decreases substantially with the increase in age and grade of the students.

So far in this chapter, only unweighted participation rates by age class and school type have been presented. However, analysis is typically performed separately by age class and session type, and NCES standards regarding acceptable potentials for bias are expressed in terms of weighted participation rates. Therefore, Tables 3-12 and 3-13 show weighted participation rates by age class and session type for the main and long-term trend samples, respectively. The main sample rates are for students in the reporting populations. Note that for the main samples, the student participation rates are similar for different session types at grades 4 and 8, but the student participation rates at grade 12 and the school participation rates at all grades vary by session type. The differential school participation rates reflect the fact that, more so than in previous years, different session types include different schools. This is due to the assignment of schools to sample type, the fact that all session types were not assessed in all sample types, and the specific sample types included in the reporting populations for each session type (see Chapter 10). For long-term trend, the participation rates are similar for different session types in the same age class. They are also similar, in general, to the unweighted rates.

The procedures for substituting for nonparticipating schools or imputing for them through weighting and the procedures for imputing for absent students through weighting were designed (so far as feasible) to reduce the biases resulting from school and student nonparticipation. These procedures are discussed in Chapter 10.

Table 3-11

Overall Participation Rates (School and Student Combined) by Age Class, Unweighted[1]

1996 Samples	Age Class 9	Age Class 13	Age Class 17	Overall
Main Samples				
School participation	86.3%	83.3%	79.5%	82.9%
Student participation	95.4%	91.5%	79.9%	88.6%
Overall student participation	82.3%	76.2%	63.5%	73.4%
Number of participating students	28,527	32,328	33,286	94,141
Long-Term Trend Samples				
School participation	85.2%	83.9%	81.4%	83.6%
Student participation	95.5%	91.9%	84.0%	90.8%
Overall student participation	81.4%	77.1%	68.4%	75.9%
Number of participating students	10,432	11,151	8,208	29,791
Overall				
School participation	86.0%	83.5%	79.9%	83.1%
Student participation	95.4%	92.8%	80.7%	89.1%
Overall student participation	82.0%	77.5%	64.5%	74.0%
Number of participating students	38,959	43,479	41,494	123,932

[1] The numbers in this table reflect the full samples, including all sample types (see Section 3.4).

Table 3-12

Weighted Participation Rates by Age Class and Session Type, 1996 Main NAEP Reporting Samples

Participation (Sample Type)	Mathematics Print	Science Print	Mathematics Estimation Print	Mathematics Theme Print	Advanced Mathematics Print	Advanced Science Print
Grade 4						
School participation	82.3%	77.8%	93.5%	77.9%	—	—
Student participation	95.3%	94.9%	96.7%	95.4%	—	—
Overall participation	78.4%	73.8%	90.4%	74.4%	—	—
Grade 8						
School participation	81.5%	79.7%	85.3%	86.8%	77.0%	—
Student participation	92.9%	93.1%	93.8%	92.7%	95.6%	—
Overall participation	75.7%	74.3%	80.0%	80.4%	73.6%	—
Grade 12						
School participation	76.2%	77.4%	63.9%	78.4%	77.6%	77.7%
Student participation	82.3%	77.5%	81.0%	78.2%	85.8%	86.5%
Overall participation	62.7%	60.0%	51.7%	61.3%	66.6%	67.2%

Table 3-13

Weighted Participation Rates by Age Class and Session Type
1996 Long-Term Trend Samples

Participation	Reading/Writing Print	Mathematics/Science Tape
Age Class 9		
School participation	83.5%	82.6%
Student participation	95.6%	95.4%
Overall participation	79.9%	78.8%
Age Class 13		
School participation	82.0%	80.8%
Student participation	92.2%	92.6%
Overall participation	75.6%	74.8%
Age Class 17		
School participation	81.7%	75.6%
Student participation	83.8%	84.1%
Overall participation	68.5%	63.6%

3.11 SAMPLING TEACHERS

The teacher questionnaire was administered to teachers of fourth- and eighth-grade students assessed in mathematics and science and twelfth-grade students assessed in advanced mathematics. Teachers were given the questionnaire if they taught the student the subject in which the student was assessed. The purpose of drawing these samples was not to estimate the attributes of the teacher

population, but to estimate the number (proportion) of students whose teachers had various attributes and to correlate student characteristics and performance with the characteristics of their teachers.

The selected teachers were asked to complete a questionnaire concerning themselves and their teaching practices, with specific references to each individual class period containing a student included in the main assessment.

Chapter 4

ASSESSMENT INSTRUMENTS[1]

Stephen Lazer
Educational Testing Service

4.1 INTRODUCTION

In the 1996 assessment, four types of instruments were used to collect data about students, teachers, and schools. Each assessed student received an **assessment booklet** containing both cognitive and background questions. An **SD/LEP Student Questionnaire** was completed by school officials for each sampled student identified as having a disability (SD) or classified as Limited English Proficient (LEP), whether or not the students were able to participate in the assessment. The teachers of fourth-, eighth-, and twelfth-grade students participating in the assessment were asked to complete a **Teacher Questionnaire.** A **School Characteristics and Policies Questionnaire** was distributed to each participating school.

This chapter begins with a discussion of the characteristics of the student booklets used for the 1996 main and long-term trend assessments and how the booklets were assembled. The contents of each booklet and item block is presented in detail in a set of tables. Section 4.4 describes the student, teacher, SD/LEP, and school questionnaires that were part of the 1996 assessment.

4.2 STUDENT BOOKLETS—MAIN ASSESSMENTS

4.2.1 Mathematics

Each student assessed in mathematics received a booklet containing a set of general background questions, content questions, subject-specific background questions, and questions about his or her motivation and familiarity with the assessment materials. The content questions were assembled into sections or blocks. Students in the main assessment were given three 15-minute blocks. Those sampled for the theme assessment completed one 15-minute block and one 30-minute block. Those sampled for the advanced study at grade 8 completed three 20-minute blocks; at grade 12 advanced sample students completed three 30-minute blocks. Students in the estimation sample completed one 15-minute block from the main assessment and two paced-tape sections. The overall assessment time for each student was approximately 63 minutes.

The assembly of blocks into booklets for the main assessment and their subsequent assignment to sampled students was determined by a *balanced incomplete block* (BIB) design with *spiraled* administration. The student booklets contained two five-minute background sections, a one-minute background section, and three 15-minute blocks of items according to a BIB design.

The BIB design for the 1996 national mathematics assessment was *focused* by subject area, so that students received booklets containing only blocks of mathematics questions (not science). The BIB design also balances the order of presentation of the 15-minute blocks of items—every 15-minute block

[1] Stephen Lazer manages assessment development activities for the NAEP program at ETS.

appears as the first cognitive block in two booklets, as the second cognitive block in two other booklets, and as the third cognitive block in another two booklets.

The design used in 1996 required that 13 blocks of mathematics items at each grade be assembled into 26 booklets. Theme blocks were placed in two other booklets, and estimation blocks in one other booklet. At grades 8 and 12, the advanced study was placed in one additional booklet.[2] Once assembled, the main assessment booklets were then *spiraled* and bundled. Spiraling involves interweaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in each position in a bundle.

The final step in the BIB-spiraling procedure was the assigning of the booklets to the assessed students. The students within an assessment session were assigned booklets in the order in which the booklets were bundled. Thus, most students in an assessment session received different booklets. In the assessment design, representative and randomly equivalent samples of students responded to each item at a given grade level.

Tables 4-1, 4-2, and 4-3 provide the composition and number of booklets administered in the 1996 mathematics assessment. Table 4-4 gives details of the item blocks used in the main assessment, including the number of cognitive items in each block and the booklets in which each block appeared; Table 4-5 gives the same information for blocks in the special components of the assessment. Table 4-6 gives pertinent information about the background sections.

[2] See Chapter 2 for descriptions of these types of assessment blocks.

Table 4-1
Main Sample Booklet Configuration
Grade 4, Mathematics

Booklet Number	Background		Cognitive Blocks			Motivation Background Block
	Common	Mathematics				
101	C13	M2	M3	M4	M7	MX
102	C13	M2	M4	M5	M8	MX
103	C13	M2	M5	M6	M9	MX
104	C13	M2	M6	M7	M10	MX
105	C13	M2	M7	M8	M11	MX
106	C13	M2	M8	M9	M12	MX
107	C13	M2	M9	M10	M13	MX
108	C13	M2	M10	M11	M14	MX
109	C13	M2	M11	M12	M15	MX
110	C13	M2	M12	M13	M3	MX
111	C13	M2	M13	M14	M4	MX
112	C13	M2	M14	M15	M5	MX
113	C13	M2	M15	M3	M6	MX
114	C13	M2	M3	M5	M10	MX
115[1]	C13	M2	M4	M6	M11	MX
116	C13	M2	M5	M7	M12	MX
117	C13	M2	M6	M8	M13	MX
118	C13	M2	M7	M9	M14	MX
119	C13	M2	M8	M10	M15	MX
120	C13	M2	M9	M11	M3	MX
121[2]	C13	M2	M10	M12	M4	MX
122	C13	M2	M11	M13	M5	MX
123	C13	M2	M12	M14	M6	MX
124	C13	M2	M13	M15	M7	MX
125	C13	M2	M14	M3	M8	MX
126	C13	M2	M15	M4	M9	MX
127[3]	C13	M2	M4	M16	M17	MX
128[4]	C13	M2	M4	M21	—	MX
129[4]	C13	M2	M4	M22	—	MX
921[5]	C13	M2	M10	M12	M4	MX

[1] This booklet was a large print version.
[2] This booklet was also used for SD/LEP students who took a regular-print version.
[3] This was an estimation booklet (involved paced audiotapes).
[4] This was a theme booklet.
[5] This was a bilingual booklet presented to some SD/LEP students. It contained the same blocks as Booklet Number 121.

Table 4-2
Main Sample Booklet Configuration
Grade 8, Mathematics

Booklet Number	Background		Cognitive Blocks			Motivation Background Block
	Common	Mathematics				
101	C13	M2	M3	M4	M7	MX
102	C13	M2	M4	M5	M8	MX
103	C13	M2	M5	M6	M9	MX
104	C13	M2	M6	M7	M10	MX
105	C13	M2	M7	M8	M11	MX
106	C13	M2	M8	M9	M12	MX
107	C13	M2	M9	M10	M13	MX
108	C13	M2	M10	M11	M14	MX
109	C13	M2	M11	M12	M15	MX
110	C13	M2	M12	M13	M3	MX
111	C13	M2	M13	M14	M4	MX
112	C13	M2	M14	M15	M5	MX
113	C13	M2	M15	M3	M6	MX
114	C13	M2	M3	M5	M10	MX
115[1]	C13	M2	M4	M6	M11	MX
116	C13	M2	M5	M7	M12	MX
117	C13	M2	M6	M8	M13	MX
118	C13	M2	M7	M9	M14	MX
119	C13	M2	M8	M10	M15	MX
120	C13	M2	M9	M11	M3	MX
121[2]	C13	M2	M10	M12	M4	MX
122	C13	M2	M11	M13	M5	MX
123	C13	M2	M12	M14	M6	MX
124	C13	M2	M13	M15	M7	MX
125	C13	M2	M14	M3	M8	MX
126	C13	M2	M15	M4	M9	MX
127[3]	C13	M2	M4	M16	M17	MX
128[4]	C13	M2	M4	M21	—	MX
129[4]	C13	M2	M4	M22	—	MX
130[5]	C13	M2	M20	M18	M19	MX
921[5/6]	C13	M2	M10	M12	M4	MX

[1] This booklet was a large print version.
[2] This booklet was also used for SD/LEP students who took a regular-print version.
[3] This was an estimation booklet (involved paced audiotapes).
[4] This was a theme booklet.
[5] This was an advanced booklet.
[6] This was a bilingual booklet presented to some SD/LEP students. It contained the same blocks as Booklet Number 121.

Table 4-3

Main Sample Booklet Configuration

Grade 12, Mathematics

Booklet Number	Background		Cognitive Blocks			Motivation Background Block
	Common	Mathematics				
101	C13	M2	M3	M4	M7	MX
102	C13	M2	M4	M5	M8	MX
103	C13	M2	M5	M6	M9	MX
104	C13	M2	M6	M7	M10	MX
105	C13	M2	M7	M8	M11	MX
106	C13	M2	M8	M9	M12	MX
107	C13	M2	M9	M10	M13	MX
108	C13	M2	M10	M11	M14	MX
109	C13	M2	M11	M12	M15	MX
110	C13	M2	M12	M13	M3	MX
111	C13	M2	M13	M14	M4	MX
112	C13	M2	M14	M15	M5	MX
113	C13	M2	M15	M3	M6	MX
114	C13	M2	M3	M5	M10	MX
115[1]	C13	M2	M4	M6	M11	MX
116	C13	M2	M5	M7	M12	MX
117	C13	M2	M6	M8	M13	MX
118	C13	M2	M7	M9	M14	MX
119	C13	M2	M8	M10	M15	MX
120	C13	M2	M9	M11	M3	MX
121[2]	C13	M2	M10	M12	M4	MX
122	C13	M2	M11	M13	M5	MX
123	C13	M2	M12	M14	M6	MX
124	C13	M2	M13	M15	M7	MX
125	C13	M2	M14	M3	M8	MX
126	C13	M2	M15	M4	M9	MX
127[3]	C13	M2	M4	M16	M17	MX
128[4]	C13	M2	M4	M21	—	MX
129[4]	C13	M2	M4	M22	—	MX
130[5]	C13	M2	M20	M18	M19	MX

[1] This booklet was a large print version.
[2] This booklet was used for SD/LEP students who took a regular-print version.
[3] This was an estimation booklet (involved paced audiotapes).
[4] This was a theme booklet.
[5] This was an advanced booklet.

Table 4-4

1996 Mathematics Assessment, Main BIB

Block	Designation	Grade	Multiple-Choice	Short Constructed-Response	Extended Constructed-Response	Total	Comments
3	S1M3	4	9	4	0	13	Trend (92)
	S2M3	8	9	3	1	13	Trend (92)
	S3M3	12	10	4	0	14	Trend (92)
4	S123M4a	4	14	0	0	14	Trend (90)
	S123M4b	8	21	0	0	21	Trend (90)
	S123M4c	12	22	0	0	22	Trend (90)
5	S12M5a	4	4	5	1	10	New
	S12M5b	8	6	4	1	11	New
	S3M5	12	4	5	1	10	New -- Calc
6	S123M6a	4	0	11	0	11	Trend (90)
	S123M6b	8	0	16	0	16	Trend (90)
	S123M6c	12	0	17	0	17	Trend (90)
7	S12M7a	4	3	4	1	8	New -- Manip
	S12M7b	8	5	4	1	10	New -- Manip
	S3M7	12	4	5	1	10	New -- Manip
8	S123M8a	4	14	1	0	15	Trend Calc (90)
	S123M8b	8	16	2	0	18	Trend Calc (90)
	S123M8c	12	17	4	0	21	Trend Calc (90)
9	S1M9	4	9	2	1	12	Trend (92)
	S23M9b	8	5	3	1	9	Trend (92)
	S23M9c	12	6	2	1	9	Trend (92)
10	S123M10a	4	0	6	0	6	Manipulatives (92)
	S123M10b	8	0	7	0	7	Manipulatives (92)
	S123M10c	12	3	6	1	10	Manipulatives (92)
11	S12M11a	4	11	5	0	16	Trend (92)
	S12M11b	8	13	6	0	19	Trend (92)
	S3M11	12	11	3	0	14	Trend (92)
12	S1M12	4	5	3	1	9	New Calculator
	S23M12b	8	4	4	1	9	New Calculator
	S23M12c	12	4	5	1	10	New Calculator
13	S1M13	4	6	5	1	12	Trend (92)
	S2M13	8	6	4	1	11	Trend (92)
	S3M13	12	3	5	1	9	Trend Prot (92)
14	S1M14	4	4	5	1	10	New Calculator
	S23M14b	8	5	3	1	9	New Calculator
	S23M14c	12	5	3	1	9	New Calculator
15	S1M15	4	3	6	1	10	New -- Ruler
	S2M15	8	4	4	1	9	New -- Calculator/Protractor
	S3M15	12	4	5	1	10	New -- Calculator

Table 4-5

1996 Mathematics Assessment, Estimation and Targeted Assessment

Block	Designation	Grade	Multiple-Choice	Short Constructed-Response	Extended Constructed-Response	Total	Comments
16	S123M16a	4	20	0	0	20	Trend Estimation (90)
	S123M16b	8	22	0	0	22	Trend Estimation (90)
	S123M16c	12	22	0	0	22	Trend Estimation (90)
17	S1M17	4	10	3	0	13	New Estimation
	S2M17	8	9	6	0	15	New Estimation
	S3M17	12	16	0	0	16	New Estimation
18	S2M18	8	3	6	1	10	New Algebra
	S3M18	12	3	6	2	11	New Advanced Mathematics
19	S2M19	8	6	5	1	12	New Algebra
	S3M19	12	4	5	2	11	New Advanced Mathematics
21[1]	S1M21	4	1	4	3	8	New Theme
	S23M21b	8	4	5	2	11	New Theme
	S23M21c	12	4	5	2	11	New Theme
22	S1M22	4	0	3	3	6	New Theme
	S2M22	8	4	4	2	10	New Theme
	S2M23	12	4	2	1	7	New Theme

[1] Block 20 was a block composed of exercises from the main assessment used for linking.

Table 4-6

Background Sections of Student Mathematics Booklets

	Number of Questions	Placement in Student Booklet
Grade 4		
General Background	24	Section 1
Mathematics Background	25	Section 2
Motivation	5	Section 6[1]
Grade 8		
General Background	26	Section 1
Mathematics Background	31	Section 2
Motivation	5	Section 6[1]
Grade 12		
General Background	35	Section 1
Mathematics Background	44	Section 2
Motivation	5	Section 6[1]

[1] Or Section 5 in theme booklets.

4.2.2 Science

Each student assessed in science received a booklet containing general background questions, content questions, subject-specific background questions, and questions about his or her motivation and familiarity with the assessment materials. The content questions were assembled into sections or blocks. Students in the main assessment were given three 20-minute blocks at grade 4, and three 30-minute blocks at grades 8 and 12. The last block in every book was a hands-on block. Those sampled for the advanced study at grade 12 completed four 30-minute blocks. The overall assessment time for each student was, on average, 120 minutes.

The assembly of blocks into booklets for the main assessment and their subsequent assignment to sampled students was determined by a complex design with *spiraled* administration. The student booklets contained two five-minute background sections, a one-minute background section, and three blocks of items.

The design for the 1996 national assessment was *focused* by subject area, so that students received booklets containing only blocks of science questions (not mathematics). The design also balances the order of presentation of the blocks of items, except for the hands-on blocks, which always appear in position three of a booklet. All other blocks appear an equal number of times in position one and position two. Further, the design was set up to ensure that no student answered more than one theme-based block (though some students did not receive any). This design allows for some balancing of the impact of context and fatigue effects to be measured and reported, but makes allowance for the difficulties and disruption of administering hands-on blocks. It also takes into account the limited breadth of content coverage included in the theme blocks.

The design used in 1996 required that 15 blocks of science items at each grade be assembled into 37 booklets. At grade 12, the advanced study was composed of three additional booklets. Once assembled, the main assessment booklets were then *spiraled* and bundled. Spiraling involves interweaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in each position in a bundle.

The final step in the spiraling procedure was the assigning of the booklets to the assessed students. The students within an assessment session were assigned booklets in the order in which the booklets were bundled. Thus, most students in an assessment session received different booklets. In the assessment design, representative and randomly equivalent samples of about 1,200 students responded to each item at a given grade level.

Tables 4-7, 4-8, and 4-9 provide the composition and number of booklets administered in the main 1996 science assessment. Table 4-10 provides the composition of booklets in the advanced science study. Table 4-11 gives details of the item blocks used in the main assessment, including the number of cognitive items in each block and the booklets in which each block appeared; Table 4-12 gives the same information for blocks in the special components of the assessment. Table 4-13 gives pertinent information about the background sections.

Table 4-7
Main Sample Booklet Configuration
Grade 4, Science

Booklet Number	Cognitive Blocks		Hands-On Task[1]	Background		Motivation Background Block
				Common	Science	
201	S7	S10	S3	C19	S2	SX
202	S7	S11	S4	C19	S2	SX
203	S7	S12	S5	C19	S2	SX
204	S7	S13	S6	C19	S2	SX
205	S10	S11	S3	C19	S2	SX
206	S12	S8	S4	C19	S2	SX
207	S10	S13	S5	C19	S2	SX
208	S10	S8	S6	C19	S2	SX
209	S11	S12	S3	C19	S2	SX
210	S13	S14	S4	C19	S2	SX
211	S11	S8	S5	C19	S2	SX
212	S11	S14	S6	C19	S2	SX
213	S13	S8	S3	C19	S2	SX
214	S8	S15	S4	C19	S2	SX
215	S12	S14	S5	C19	S2	SX
216	S12	S15	S6	C19	S2	SX
217	S8	S14	S3	C19	S2	SX
218	S14	S20	S4	C19	S2	SX
219	S8	S20	S5	C19	S2	SX
220[2]	S13	S20	S6	C19	S2	SX
221	S14	S15	S3	C19	S2	SX
222	S15	S21	S4	C19	S2	SX
223	S15	S9	S5	C19	S2	SX
224	S8	S21	S6	C19	S2	SX
225	S20	S21	S3	C19	S2	SX
226	S20	S9	S4	C19	S2	SX
227	S20	S7	S5	C19	S2	SX
228	S14	S9	S6	C19	S2	SX
229	S21	S9	S3	C19	S2	SX
230	S21	S7	S4	C19	S2	SX
231	S21	S10	S5	C19	S2	SX
232	S15	S7	S6	C19	S2	SX
233	S9	S13	S3	C19	S2	SX
234	S9	S10	S4	C19	S2	SX
235	S9	S11	S5	C19	S2	SX
236	S9	S12	S6	C19	S2	SX
237	S14	S7	S3	C19	S2	SX

[1] Hands-on task blocks: Block S3 uses “A” kit - seeds; Block S4 uses “B” kit - unknown powders; Block S5 uses “C” kit - floating pencil; and Block S6 uses “D” kit - markers.
[2] This booklet was also used for SD/LEP students who took a regular-print version.

Table 4-8
Main Sample Booklet Configuration
Grade 8, Science

Booklet Number	Cognitive Blocks		Hands-On Task[1]	Background		Motivation Background Block
				Common	Science	
201	S7	S10	S3	C19	S2	SX
202	S7	S11	S4	C19	S2	SX
203	S7	S12	S5	C19	S2	SX
204	S7	S13	S6	C19	S2	SX
205	S10	S11	S3	C19	S2	SX
206	S12	S8	S4	C19	S2	SX
207	S10	S13	S5	C19	S2	SX
208	S10	S8	S6	C19	S2	SX
209	S11	S12	S3	C19	S2	SX
210	S13	S14	S4	C19	S2	SX
211	S11	S8	S5	C19	S2	SX
212	S11	S14	S6	C19	S2	SX
213	S13	S8	S3	C19	S2	SX
214	S8	S15	S4	C19	S2	SX
215	S12	S14	S5	C19	S2	SX
216	S12	S15	S6	C19	S2	SX
217	S8	S14	S3	C19	S2	SX
218	S14	S20	S4	C19	S2	SX
219	S8	S20	S5	C19	S2	SX
220[2]	S13	S20	S6	C19	S2	SX
221	S14	S15	S3	C19	S2	SX
222	S15	S21	S4	C19	S2	SX
223	S15	S9	S5	C19	S2	SX
224	S8	S21	S6	C19	S2	SX
225	S20	S21	S3	C19	S2	SX
226	S20	S9	S4	C19	S2	SX
227	S20	S7	S5	C19	S2	SX
228	S14	S9	S6	C19	S2	SX
229	S21	S9	S3	C19	S2	SX
230	S21	S7	S4	C19	S2	SX
231	S21	S10	S5	C19	S2	SX
232	S15	S7	S6	C19	S2	SX
233	S9	S13	S3	C19	S2	SX
234	S9	S10	S4	C19	S2	SX
235	S9	S11	S5	C19	S2	SX
236	S9	S12	S6	C19	S2	SX
237	S14	S7	S3	C19	S2	SX

[1] Hands-On task blocks: Block S6 uses "D" kit - markers; Block S3 uses "E" kit - powders; Block S4 uses "F" kit - salt solutions; and Block S5 uses "G" kit - soil tests.
[2] This booklet was also used for SD/LEP students who took a regular-print version.

Table 4-9
Main Sample Booklet Configuration
Grade 12, Science

Booklet Number	Cognitive Blocks		Hands-On Task[1]	Background		Motivation Background Block
				Common	Science	
201	S7	S10	S3	C19	S2	SX
202	S7	S11	S4	C19	S2	SX
203	S7	S12	S5	C19	S2	SX
204	S7	S13	S6	C19	S2	SX
205	S10	S11	S3	C19	S2	SX
206	S12	S8	S4	C19	S2	SX
207	S10	S13	S5	C19	S2	SX
208	S10	S8	S6	C19	S2	SX
209	S11	S12	S3	C19	S2	SX
210[2]	S13	S14	S4	C19	S2	SX
211	S11	S8	S5	C19	S2	SX
212	S11	S14	S6	C19	S2	SX
213	S13	S8	S3	C19	S2	SX
214	S8	S15	S4	C19	S2	SX
215	S12	S14	S5	C19	S2	SX
216	S12	S15	S6	C19	S2	SX
217	S8	S14	S3	C19	S2	SX
218	S14	S20	S4	C19	S2	SX
219	S8	S20	S5	C19	S2	SX
220	S13	S20	S6	C19	S2	SX
221	S14	S15	S3	C19	S2	SX
222	S15	S21	S4	C19	S2	SX
223	S15	S9	S5	C19	S2	SX
224	S8	S21	S6	C19	S2	SX
225	S20	S21	S3	C19	S2	SX
226	S20	S9	S4	C19	S2	SX
227	S20	S7	S5	C19	S2	SX
228	S14	S9	S6	C19	S2	SX
229	S21	S9	S3	C19	S2	SX
230	S21	S7	S4	C19	S2	SX
231	S21	S10	S5	C19	S2	SX
232	S15	S7	S6	C19	S2	SX
233	S9	S13	S3	C19	S2	SX
234	S9	S10	S4	C19	S2	SX
235	S9	S11	S5	C19	S2	SX
236	S9	S12	S6	C19	S2	SX
237	S14	S7	S3	C19	S2	SX

[1] Hands-On task blocks: Block S5 uses “G” kit - soil tests; Block S3 uses “H” kit - antacid; Block S4 uses “I” kit - separation; and Block S6 uses “J” kit - pendulum.
[2] This booklet was also used for SD/LEP students who took a regular-print version.

Table 4-10
Main Sample Booklet Configuration
Grade 12, Advanced Science

Booklet Number	Cognitive Blocks			Hands-On Task	Background		Motivation Background Block
					Common	Science	
238	S19	S18	S17	S16	C19	S2	SX
239	S19	S17	S16	S18	C19	S2	SX
240	S19	S16	S18	S17	C19	S2	SX

Table 4-11
1996 Science Assessment, Main BIB

Block	Designation	Grade	Multiple-Choice	Short Constructed-Response	Extended Constructed-Response	Total	Comments
3	S1S3	4	0	7	0	7	Hands-on
	S2S3	8	0	4	2	6	Hands-on
	S3S3	12	0	4	2	6	Hands-on
4	S1S4	4	1	0	7	7	Hands-on
	S2S4	8	3	4	3	10	Hands-on
	S3S4	12	0	1	3	4	Hands-on
5	S1S5	4	0	5	6	11	Hands-on
	S23S5	8/12	0	6	0	6	Hands-on
6	S12S6A	4	0	0	4	4	Hands-on
	S12S6B	8	0	5	2	7	Hands-on
	S3S6	12	0	6	2	8	Hands-on
7	S1S7	4	0	10	0	10	Theme-based
	S2S7	8	4	10	0	14	Theme-based
	S3S7	12	5	7	3	15	Theme-based
8	S1S8	4	1	6	1	8	Theme-based
	S23S8A	8	5	5	0	10	Theme-based
	S23S8B	12	6	7	1	14	Theme-based
9	S12S9A	4	2	6	1	9	Theme-based
	S12S9B	8	3	9	1	13	Theme-based
	S3S9	12	4	8	2	14	Theme-based
10	S1S10	4	6	4	1	11	Concept/Problem-Solving
	S2S10	8	8	7	1	16	Concept/Problem-Solving
	S3S10	12	7	7	1	15	Concept/Problem-Solving
11	S1S11	4	6	5	0	11	Concept/Problem-Solving
	S2S11	8	8	7	1	16	Concept/Problem-Solving
	S3S11	12	7	5	3	15	Concept/Problem-Solving
12	S1S12	4	6	4	1	10	Concept/Problem-Solving
	S23S12	8/12	8	6	2	16	Concept/Problem-Solving
13	S1S13	4	6	4	1	11	Concept/Problem-Solving
	S23S13	8/12	8	7	1	16	Concept/Problem-Solving
14	S12S14A	4	5	5	0	10	Concept/Problem-Solving
	S12S14B	8	7	11	0	18	Concept/Problem-Solving
	S3S14	12	8	6	2	16	Concept/Problem-Solving
15	S12S15A	4	3	5	1	9	Concept/Problem-Solving
	S12S15B	8	7	7	2	16	Concept/Problem-Solving
	S3S15	12	0	5	2	7	In-depth
20	S1S20	4	6	2	3	11	Concept/Problem-Solving
	S2S20	8	8	6	2	16	Concept/Problem-Solving
	S3S20	12	7	6	3	16	Concept/Problem-Solving
21	S1S21	4	5	3	2	10	Concept/Problem-Solving
	S2S21	8	7	7	2	16	Concept/Problem-Solving
	S3S21	12	8	4	4	16	Concept/Problem-Solving

Table 4-12
1996 Science Assessment, Advanced Assessment

Block	Designation	Grade	Multiple-Choice	Short Constructed-Response	Extended Constructed-Response	Total	Comments
16	S3S16	12	7	8	1	16	Advanced Block
17	S3S17	12	7	5	4	16	Advanced Block
18	S3S18	12	7	4	5	16	Advanced Block
19	S3S19	12	9	7	2	18	Advanced Block

Table 4-13
Background Sections of 1996 Student Science Booklets

	Number of Questions	Placement in Student Booklet
Grade 4		
General Background	24	Section 4
Science Background	39	Section 5
Motivation	5	Section 6
Grade 8		
General Background	26	Section 4
Science Background	42	Section 5
Motivation	5	Section 6
Grade 12		
General Background	36	Section 4
Science Background	53	Section 5
Motivation	5	Section 6

4.3 STUDENT BOOKLETS—LONG-TERM TREND ASSESSMENTS

There were several long-term trend samples in the 1994 assessment (see Chapter 1), each of which required the use of special booklets. Tables 4-14, 4-15, and 4-16 summarize the contents of each trend assessment booklet and show how many of each booklet were administered. Tables 4-20, 4-21, and 4-22 give details of the item blocks used in the long-term trend assessments, including the number of cognitive and constructed-response items in each block and the booklets in which each block appeared.

Reading and Writing Long-Term Trend. Six booklets (numbered 51 to 56) containing reading and writing items were administered to each age class. These booklets were identical to booklets used in previous assessments of reading and writing and were spiraled for administration. Each booklet consisted of a common background block (BZ) and three cognitive blocks (at least one reading block and at least one writing block). In addition to cognitive items, the cognitive blocks also contained subject-related background questions.

Mathematics and Science Long-Term Trend. Three booklets (91, 92, and 93) at ages 9 and 13 and two booklets (84 and 85) at age 17, containing mathematics and science items, were identical to those used in previous assessments to measure trends. Each booklet contained a common background block (C1 or BZ) and three cognitive blocks. At ages 9 and 13, these booklets contained one reading block (R1,

R2, or R3), one mathematics block (M1, M2, or M3) and one science block (S1, S2, or S3). At age 17, each booklet contained at least one mathematics block (M1 to M3) and at least one science block (S1 - S3). Mathematics block M3 contained items that required the use of a calculator. All cognitive blocks also contained subject-related background questions.

Table 4-14
Long-Term Trend Sample Booklet Contents and Number of Booklets Administered
Age Class 9

Subject Area	Booklet Number	Common Background Block	Subject Area Background Block[1]	Cognitive Blocks			Number of Booklets Administered
Reading and Writing	51	BZ	—	BC	BL	BQ	1,186
	52	BZ	—	BH	BE	BR	1,165
	53	BZ	—	BC	BK	BJ	1,178
	54	BZ	—	BG	BO	BE	1,180
	55	BZ	—	BM	BG	BN	1,169
	56	BZ	—	BV	BR		1,184
Mathematics and Science	91	C1	—	R1	M1	S1	2,388
	92	C1	—	S2	R2	M3[2]	2,512
	93	C1	—	M2	S3	R3	2,435

[1] Subject area background questions are included in cognitive blocks for this booklet.
[2] Calculator needed for this block.

Table 4-15

Long-Term Trend Sample Booklet Contents and Number of Booklets Administered
Age Class 13

Subject Area	Booklet Number	Common Background Block	Subject Area Background Block[1]	Cognitive Blocks			Number of Booklets Administered
Reading and Writing	51	BZ	—	BM	BK	BD	919
	52	BZ	—	BC	BL	BQ	906
	53	BZ	—	BH	BE	BR	923
	54	BZ	—	BN	BC	BD	905
	55	BZ	—	BG	BO	BE	928
	56	BZ	—	BG	BJ	BP	933
Mathematics and Science	91	C1	—	R1	M1	S1	1,928
	92	C1	—	S2	R2	M3[2]	1,976
	93	C1	—	M2	S3	R3	2,005

[1] Subject area background questions are included in cognitive blocks for this booklet.
[2] Calculator needed for this block.

Table 4-16

Long-Term Trend Sample Booklet Contents and Number of Booklets Administered
Age Class 17

Subject Area	Booklet Number	Common Background Block	Subject Area Background Block[1]	Cognitive Blocks			Number of Booklets Administered
Reading and Writing	51	BZ	—	BM	BK	BD	927
	52	BZ	—	BC	BL	BQ	924
	53	BZ	—	BH	BE	BR	917
	54	BZ	—	BN	BC	BD	951
	55	BZ	—	BG	BO	BE	939
	56	BZ	—	BG	BJ	BP	911
Mathematics and Science	84	C1	—	M1	M2	S3	2,207
	85	C1	—	S1	S2	M3[2]	2,152

[1] Subject area background questions are included in cognitive blocks for this booklet.
[2] Calculator needed for this block.

Table 4-17

Long-Term Trend Sample Block Information, Age Class 9

Block	Type	Total Number of Items	Number of Cognitive Items	Number of Open-Ended Items Cognitive	Number of Open-Ended Items Noncognitive	Booklets Containing Block
BZ	Common Background	37	0	0	1	51 - 56
C1	Common Background	28	0	0	0	91 - 93
BC	Writing Background/Cognitive	23	1	1	0	51, 53
BE	Writing Background/Cognitive	11	2	2	0	52, 54
BG	Writing Background/Cognitive	8	2	2	0	54, 55
BH	Reading Background/Cognitive	15	11	1	0	52
BJ	Reading Background/Cognitive	24	13	1	0	53
BK	Reading Background/Cognitive	19	11	0	0	53
BL	Reading Background/Cognitive	26	7	1	1	51
BM	Reading Background/Cognitive	16	12	1	0	55
BN	Reading Background/Cognitive	25	14	1	0	55
BO	Reading Background/Cognitive	22	11	0	0	54
BQ	Reading Background/Cognitive	21	12	0	0	51
BR	Reading Background/Cognitive	16	12	0	0	52, 56
BV	Reading and Writing Background/Cognitive	36	7 Rd. 1 Wr.	1 Rd. 1 Wr.	0	56
R1	Reading Background/Cognitive	20	9	0	0	91
R2	Reading Background/Cognitive	20	11	0	0	92
R3	Reading Background/Cognitive	17	10	1	0	93
M1	Mathematics	26	26	9	0	91
M2	Background/Cognitive	26	26	9	0	93
M3	Mathematics Background/Cognitive Mathematics Background/Cognitive (Calc.)	19	16	10	0	92
S1	Science Background/Cognitive	23	18	0	0	91
S2	Science Background/Cognitive	25	25	0	0	92
S3	Science Background/Cognitive	31	20	0	0	93

Table 4-18

Long-Term Trend Sample Block Information, Age Class 13

Block	Type	Total Number of Items	Number of Cognitive Items	Number of Open-Ended Items Cognitive	Number of Open-Ended Items Noncognitive	Booklets Containing Block
BZ	Common Background	37	0	0	1	51 - 56
C1	Common Background	30	0	0	0	91 - 93
BC	Writing Background/Cognitive	23	1	1	0	52, 54
BD	Writing Background/Cognitive	25	1	1	0	51, 54
BE	Writing Background/Cognitive	11	2	2	0	53, 55
BG	Writing Background/Cognitive	8	2	2	0	55, 56
BH	Reading Background/Cognitive	18	13	1	1	53
BJ	Reading Background/Cognitive	24	14	2	0	56
BK	Reading Background/Cognitive	17	9	1	0	51
BL	Reading Background/Cognitive	27	6	1	1	52
BM	Reading Background/Cognitive	16	12	1	0	51
BN	Reading Background/Cognitive	23	12	1	0	54
BO	Reading Background/Cognitive	21	10	2	0	55
BP	Reading Background/Cognitive	15	9	1	0	55
BQ	Reading Background/Cognitive	23	17	0	0	52
BR	Reading Background/Cognitive	19	15	0	0	53
R1	Reading Background/Cognitive	31	12	1	0	91
R2	Reading Background/Cognitive	19	10	0	0	92
R3	Reading Background/Cognitive	28	13	0	0	93
M1	Mathematics	51	37	9	0	91
M2	Background/Cognitive	44	37	8	0	93
M3	Mathematics Background/Cognitive Mathematics Background/Cognitive (Calc.)	32	24	10	0	92
S1	Science Background/Cognitive	36	25	0	0	91
S2	Science Background/Cognitive	40	27	0	0	92
S3	Science Background/Cognitive	36	27	0	0	93

Table 4-19
Long-Term Trend Sample Block Information, Age Class 17

Block	Type	Total Number of Items	Number of Cognitive Items	Number of Open-Ended Items Cognitive	Number of Open-Ended Items Noncognitive	Booklets Containing Block
BZ	Common Background	48	0	0	1	51 - 56
C1	Common Background	48	0	0	0	84, 85
BC	Writing Background/Cognitive	23	1	1	0	52, 54
BD	Writing Background/Cognitive	25	1	1	0	51, 54
BE	Writing Background/Cognitive	11	2	2	0	53, 55
BG	Writing Background/Cognitive	8	2	2	0	55, 56
BH	Reading Background/Cognitive	19	13	1	2	53
BJ	Reading Background/Cognitive	17	6	2	1	56
BK	Reading Background/Cognitive	17	9	1	0	51
BL	Reading Background/Cognitive	32	6	1	2	52
BM	Reading Background/Cognitive	16	12	1	0	51
BN	Reading Background/Cognitive	32	12	1	1	54
BO	Reading Background/Cognitive	24	13	1	0	55
BP	Reading Background/Cognitive	25	11	1	0	56
BQ	Reading Background/Cognitive	17	11	1	0	52
BR	Reading Background/Cognitive	20	9	0	0	53
M1	Mathematics Background/Cognitive	49	35	10	0	84
M2	Mathematics Background/Cognitive	49	35	5	0	84
M3	Mathematics Background/Cognitive (Calculator)	35	24	14	0	85
S1	Science Background/Cognitive	38	27	0	0	85
S2	Science Background/Cognitive	41	32	0	0	85
S3	Science Background/Cognitive	32	23	0	0	84

4.4 STUDENT, TEACHER, AND SCHOOL QUESTIONNAIRES

4.4.1 Student Questionnaires

Each booklet in the main assessment included three student background questionnaires. The first, consisting of general background questions, included questions about race/ethnicity, mother's and father's level of education, reading materials in the home, homework, attendance, academic expectations, and which parents lived at home. The second, consisting of subject-area background questions, included questions about instructional activities, courses taken, use of specialized resources such as calculators in mathematics class, and views on the utility and value of the subject matter. Students were given five minutes to complete each of these questionnaires, with the exception of the fourth graders, who were given more time because the items in the general questionnaire were read aloud for them. The third

questionnaire followed the three cognitive blocks and contained five questions about students' motivation to do well on the assessment, their perceptions concerning the difficulty of the assessment, and their familiarity with types of questions included.

The student questionnaires are described in detail in Chapter 2.

4.4.2 Teacher Questionnaires

To supplement the information on instruction reported by students, the mathematics teachers of the students participating in the mathematics assessment were asked to complete a questionnaire about their instructional practices, teaching backgrounds, and characteristics. The teacher questionnaire contained two parts. The first part pertained to the teachers' background and general training. The second part pertained to specific training in teaching mathematics and the procedures the teacher uses for *each class* containing an assessed student, as well as collecting information on teachers' awareness and knowledge of the NCTM *Standards*.

The **Teacher Questionnaire, Part I: Background and General Training** included questions pertaining to gender, race/ethnicity, years of teaching experience, certification, degrees, major and minor fields of study, course work in education, course work in specific subject areas, amount of in-service training, extent of control over instructional issues, and availability of resources for their classroom.

The **Teacher Questionnaire, Part II: Training in Mathematics and Classroom Instructional Information** included questions on the teacher's exposure to various issues related to mathematics and teaching mathematics through pre- and in-service training, ability level of students in the class, whether students were assigned to the class by ability level, time on task, homework assignments, frequency of instructional activities used in class, methods of assessing student progress in mathematics, instructional emphasis given to the mathematics abilities covered in the assessment, and use of particular resources.

Because the sampling for the teacher questionnaires was based on participating students, the responses to a particular teacher questionnaire do not necessarily represent all teachers of that subject area at that grade level in the nation. Rather, they are teachers of the representative sample of students assessed. It is important to note that in all NAEP reports, the student is always the unit of analysis, even when information from the teacher or school questionnaire is being reported. Using the student as the unit of analysis makes it possible to describe the instruction received by representative samples of students. Although this approach may provide a different perspective from other studies simply reporting information about teachers or schools, it is consistent with NAEP's goal of providing information about the educational context and performance of students.

The teacher questionnaires are described in detail in Chapter 2.

4.4.3 School Questionnaires

A **School Characteristics and Policies Questionnaire** was given to the principal or other administrator of each school that participated in the assessment. This information provided an even broader picture of the instructional context for students' mathematics achievement. This questionnaire included questions about background and characteristics of school principals, length of school day and year, school enrollment, absenteeism, dropout rates, size and composition of teaching staff, policies about grouping students, curriculum, testing practices and uses, special priorities and school-wide

programs, availability of resources, special services, community services, policies for parental involvement, and school-wide problems.

School Characteristics and Policies questionnaires are described in detail in Chapter 2.

4.4.4 SD/LEP Student Questionnaires

The SD/LEP Student Questionnaire was completed by the teachers of those students who were selected to participate in the assessment sample who were classified as Students with Disabilities (SD) or were classified as Limited English Proficient (LEP). The questionnaire was completed for all SD or LEP students, whether or not they actually participated in the assessment. This questionnaire asked about the nature of the student's disability and the special programs in which the student participated.

Schools were permitted to exclude certain students from the assessment. The same exclusion criteria and rules used in the national assessment were also applied to the Trial State Assessment. Although the intent was to assess all sampled students, students who were identified by school staff as not capable of participating meaningfully were excluded. The NAEP guidelines for exclusion were intended to assure uniformity of exclusion criteria from school to school as well as from state to state.

More information about the SD/LEP questionnaire and exclusion criteria are provided in Chapters 2 and 5.

Chapter 5

FIELD OPERATIONS AND DATA COLLECTION[1]

Lucy M. Gray, Mark M. Waksberg, and Nancy W. Caldwell
Westat, Inc.

5.1 INTRODUCTION

This chapter describes the field operations and data collection activities for the national assessment component of the 1996 National Assessment of Educational Progress (NAEP). The national assessment is comprised of main samples and long-term trend samples. Main NAEP samples typically involve new assessment items, and may include new subject areas and innovative features; in long-term trend, the procedures and items from previous years are carried forward so that trends in student achievement can be measured over time. Both the main and long-term trend assessments are based on probability samples of schools and students that allow for regional and national reporting only. The State Assessment, the second major component of NAEP, comprises the state program that uses main assessment materials and involves much larger sample sizes per state (or jurisdiction), so that results can be reported for each participating state (for further technical information on the State Assessment, see the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics*, Allen, Jenkins, Kulick, and Zelenak, 1997, and the *Technical Report of the NAEP 1996 State Assessment Program in Science*, Allen, Swinton, Isham, and Zelenak, 1997).

The design of the national assessment component of NAEP is described in the remaining sections of this chapter. For all components, NAEP guarantees the anonymity of participants, and student or teacher names are never recorded on assessment booklets nor removed from the schools. NAEP results are reported on the national level and by region of the country, not by school district, school, or individual student. Only group statistics are reported, broken down by gender, race/ethnicity, and a host of variables that illuminate teachers' instructional practices.

5.1.1 Field Organization

The 1996 main assessment involved some new items and components including many innovative features. For example, the science assessment differed from previous NAEP science assessments in that every student performed an experiment. The mathematics assessment involved the use of mathematical tools and a larger number of constructed-response items than in the past. Much of the mathematics assessment has been used since 1990, thus providing "short-term trend" data. All students in a particular assessment session received a booklet in the same subject (i.e., mathematics or science). Even though many different booklets were used in a particular session, they were all for the same subject.

In most schools sampled for the main assessment, more than one session type was conducted. In about one-third of the schools at each grade, only two session types, mathematics and mathematics estimation, were possible. In the remaining two-thirds of schools, up to a maximum of four session types

[1] Lucy M. Gray and Mark M. Waksberg assist in survey operations and field activities for the NAEP assessments, under the direction of Nancy W. Caldwell.

for grade 4 or five session types at grades 8 or 12 could have been conducted. At grade 12, only one advanced session—either mathematics or science—was conducted.

Historically, a small proportion (less than 10%) of the sampled students have been "excluded" from NAEP assessment sessions because, according to school records, they are students with either disabilities or limited English language proficiency who have been determined to be incapable of participating meaningfully in the assessment. More recently, especially with the passage of the Individuals with Disabilities Education Act, increased attention has been given to these students and to including as many of them as possible in NAEP sessions (and in other testing situations as well). NAEP has addressed these concerns, first in the 1995 field test and continuing with the 1996 operational assessment, through a Special Study that uses both old and new "inclusion" criteria and (in some schools) offers accommodations for testing students with disabilities and/or limited English proficiency (SD/LEP). For the 1996 main assessment, a split-sample design was used, placing the sampled schools into three subsamples, so that the impact of both the new SD/LEP criteria and the provision of accommodations could be evaluated, while also collecting data with the old criteria to maintain comparability with previous NAEP data bases. This Special Study was incorporated in the main assessment but was not a part of the 1996 long-term trend assessment. The information in this chapter and in Chapter 3 applies to all three sample types or subsamples.

For administrative purposes, the main and long-term trend assessments were conducted in different schools. Responsibility for the assessments in long-term trend schools was given to one group of assessment supervisors, while responsibility for the main assessment was assigned to another group of supervisors. Since these supervisors worked in some of the same areas and sometimes in the same school districts, careful coordination was required.

In order to reduce the burden on the participating schools, national assessment field staff performed most of the work associated with the assessments. Introductory contacts and meetings were held in the fall (1995) to enlist cooperation and explain the assessment procedures to district and school representatives and to set a mutually agreed-upon assessment date for each school. The assessment supervisor visited the school to select the sample of students a week or two before the assessment. The assessment sessions were conducted by national assessment field staff, called exercise administrators, under the direction of the assessment supervisor. At the conclusion of the assessment in a school, field staff coded demographic information on the booklet covers and shipped the completed materials to National Computer Systems (NCS), the processing subcontractor for NAEP (see Chapter 6 for more detailed information on processing assessment materials).

5.2 PREPARING FOR THE ASSESSMENTS

5.2.1 Gaining the Cooperation of Sampled Schools

The process of gaining cooperation of the schools selected for the national assessment began in late August 1995 with a series of letters and contacts with state and district-level officials. The National Center for Education Statistics (NCES) first sent each jurisdiction a letter announcing NAEP plans for 1996. Westat then contacted the State Test Directors or NAEP State Coordinators in each sampled state to notify them of the districts and schools selected in their states. In the 40 jurisdictions participating in the State Assessment that also had schools sampled for the national assessment, the state received the list of districts and schools sampled for both the national and state assessments.

From September through early December 1995, Westat sent lists of schools sampled for the national assessment component and other NAEP materials to district superintendents, diocesan superintendents of Catholic schools, and principals or heads of schools in other nonpublic schools, inviting their participation. These initial mailings paved the way for telephone contacts by NAEP field supervisors who were assigned the task of gaining cooperation and scheduling assessment dates.

The schedule for project activities for the 1996 main and long-term trend assessments was as follows:

Date	Activity
Mid-August 1995	Department of Education sent first letter to Chief State School Officers about the 1996 assessment.
August 27 - 31, 1995	Training sessions were held for long-term trend assessment supervisors.
September 7 - 9, 1995	Training sessions were held for main assessment schedulers.
Mid-September 1995	Westat sent state coordinators a list of their schools initially selected for either or both major components.
Mid-to-Late September 1995	Westat sent samples and informational materials to districts if not already sent by state coordinators.
Mid-September - December 1, 1995	Supervisors contacted districts and schools to secure cooperation and to schedule assessments.
	Supervisors conducted introductory meetings for the national assessment, by telephone (or in person if requested by districts/schools). Westat selected substitutes for refusals.
	Supervisors recruited, hired, and trained exercise administrators.
October 9 - December 22, 1995	Fall long-term trend assessments were administered.
Early December 1995	Supervisors sent informational materials to principals and school coordinators. Letter confirming assessment schedule sent to each school from Westat.
December 9 - 13, 1995	Main assessment supervisor training session was held.
January 3 - March 29, 1996	Main assessments were administered.[2]
January 3 - March 8, 1996	Winter long-term trend assessments were administered.
March 11 - May 10, 1996	Spring long-term trend assessments were administered.

5.2.2 Supervisor Training

Training for assessment supervisors was multi-phased and involved separate sessions conducted in August, September, and December 1995. All training was conducted by the Westat project director,

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

field director, and home office staff. Also in attendance were representatives from Educational Testing Service (ETS), NCS, and NCES.

The first of these training sessions was held August 27-31, 1995 in Baltimore, Maryland for field staff assigned to the long-term trend program for 1996. Attending the session were the long-term trend field manager, the 11 field supervisors responsible for conducting the long-term trend assessments, and 4 troubleshooters.

After an introduction to the study, which included the background and history of NAEP, an overview of the long-term trend assessments, and the 1995-1996 assessment schedule, the training continued with a thorough (2 half-day sessions) presentation of NAEP contact/gaining cooperation activities. This is a lengthy process of contacting states, districts, and schools regarding their participation in and scheduling for NAEP; several demonstration phone calls, role plays, and exercises were used to provide some practical experience during this part of the training. The long-term trend staff was also trained on setting assessment schedules, recruiting/hiring/training exercise administrators, and sample selection and preparation of Administration Schedules and other assessment materials. Several practice exercises were used to demonstrate these topics. The training concluded with: discussions of conducting the session and using the session script; preparing school worksheets and holding makeup sessions; post-assessment activities; and Westat administrative procedures.

After an overview of NAEP and introductory remarks on the study schedule, the main assessment group of about 25 supervisors received extensive training (similar to the August training for long-term trend) in contacting the schools, gaining cooperation and scheduling the assessments; numerous demonstrations, role plays and exercises were used. Other training topics included: supervisory responsibilities; setting the assessment schedule; recruiting and training exercise administrators; and administrative forms and procedures. The scheduling supervisors also received a full day of training on using the reporting system installed on the laptop computers assigned to each of them for the gaining cooperation/scheduling phase.

The 75 NAEP supervisors who were responsible for main assessment activities were trained during a third session, held December 9-13, 1995. Training focused on a review of the preliminary activities during the fall including results of initial contacts with districts and schools, scheduling of assessments, the status of exercise administrators' recruitment, and a thorough discussion of assessment activities: sampling procedures; inclusion of SD/LEP students; teacher surveys; providing testing accommodations; conducting science sessions; and administrative forms and procedures. Westat's classroom management videotape was also shown at this training session.

The main and state assessment field managers were present at the December session to support training activities and answer questions concerning districts and schools that fell into the samples for more than one component of the assessment. Each supervisor also met with the person who completed the scheduling in their area, as a first step in preparing for the new supervisors' contacts with each school (and district, if needed).

5.2.3 Contacting Districts and Nonpublic Schools

Once the supervisors were trained in August and September, they began working on obtaining cooperation. In states participating in the State Assessment, the assessment supervisor first spoke with the State field manager to determine what contacts, if any, had already been made with districts about the national assessment. The approach the supervisors took when calling superintendents depended on whether the district had been notified about NAEP by the State Coordinator and whether the district also

had schools selected for the State Assessment. For districts that had been contacted by the State Coordinator, the supervisor began by referring to that contact.

In previous NAEP assessments, the supervisors offered and usually held "introductory meetings" with representatives from the superintendents' offices and the selected schools, typically the superintendent and the principals. These served as both an introduction to NAEP and a presentation on what would be asked of the school. The meetings were also used to establish a schedule for the sampling visits and the assessments in the schools.

However, over the years, these meetings have become somewhat redundant since many districts have fallen into the NAEP sample more than one time. It has also become more and more difficult to schedule these meetings, as district and school officials find it harder to allot time away from their offices. Thus, beginning with the fall 1995 preparation for the 1996 study, the material was almost always presented to the superintendents and principals during telephone calls rather than in formal meetings. Generally, only if an in-person meeting was specifically requested by the district or school officials, or if the supervisor felt that there was a better chance of convincing a district to participate in person, was such a meeting held.

As the supervisors contacted superintendents, principals, and nonpublic school officials to introduce NAEP and determine the schools' cooperation status, they completed two forms and entered the school status in the receipt control system installed on their laptop computers. The Results of Contact Form was completed to document the discussion the supervisor had with each administrator concerning the district's willingness to participate and any special circumstances regarding the schools' cooperation or assessments.

The supervisor also completed portions of a School Control Form. This form was preprinted with the number and types of assessment sessions assigned to the school, so that this information could then be shared with the district/school official. Information gathered during the phone call, including the name of the person designated to be the school coordinator, the number of students in the designated grade, tentative dates for the sampling visit and assessment, and other information that could have some bearing on the assessment, was recorded on the form. This information was used to update records in the home office. In December, the forms were provided to the supervisors who would be conducting the assessments.

A small number of in-person introductory meetings were held. The New York City and Los Angeles City school districts have previously used these meetings to present information about the NAEP assessments to the officials of all the selected schools and to encourage their participation, and wished to continue that practice for the current assessment. A small number of other school districts also requested such a meeting, involving representatives from their selected schools so that they would have a full understanding of what the assessments entailed.

During the telephone presentation or the introductory meeting, the supervisor discussed arrangements for the assessments with representatives from each school. Within the weeks scheduled for the PSU, the supervisor had the flexibility to set each school's assessment date in coordination with school staff. The staff sometimes expressed preferences for a particular day or dates or had particular times when the assessment could not be scheduled. Their preferences or restrictions depended on the events that had already been scheduled on their school calendar. Using this information from the schools, the supervisors set up the assessment schedule for each PSU.

The supervisor usually learned during the introductory contact whether a school required some form of parental notification or permission. Three versions of standard NAEP letters were offered for the

school's use, and each letter could be produced for selected students only or for all eligible students. The first version informs parents about the assessment. The second assumes parental consent unless parents send the form back stating that they do not want their child to participate in the assessment. The third version requires that parents sign and return the form before students can be assessed. All versions of the letter were available to the schools, although when the issue of parental permission came up in discussion, supervisors offered the least restrictive version that met the requirements of the school or district. In addition, Spanish language versions of the parent information letter were made available to the schools. Schools could also send out their own letters and notices if they preferred not to use those offered through NAEP. Information on whether the school required parent letters and the type of letter used was recorded on the School Control Form.

5.2.4 Recruiting, Hiring, and Training Exercise Administrators

During the fall, while the supervisors were contacting their schools and scheduling assessments, their other major responsibility was to recruit and hire exercise administrators, who would administer the assessment sessions. Exercise administrators were recruited from many sources. Each supervisor was given a PSU-by-PSU computerized list of exercise administrators and other field staff who had worked previously on education studies for Westat. People who had served as exercise administrators before, with good evaluations from their previous supervisors, were usually the first considered for hiring. Subsequently, during contacts with the schools, the supervisors asked the school principals and other staff to recommend potential exercise administrators. These referrals were frequently retired teachers or substitutes. Finally, where necessary, ads were placed in local newspapers and the employment service was notified.

Supervisors were told that, in general, four to five exercise administrators should be hired for each PSU, although a variety of factors might influence the actual number. The number of schools in a PSU, the size of the student sample in each school, distances to be traveled, the geography of the area, and weather conditions during the assessment period were all factors taken into consideration by supervisors in developing their plan for hiring exercise administrators.

A few supervisors, whose NAEP assignments contained contiguous PSUs, hired the same exercise administrators to work in all their PSUs. Other supervisors, whose assignments comprised PSUs that were not geographically connected, tended to hire teams of exercise administrators for each PSU. Supervisors were encouraged to hire locally and to hire individuals with teaching experience and the ability to handle classroom situations.

The scheduling supervisors, all of whom were experienced NAEP supervisors, had complete responsibility for recruiting, hiring, and training all of the exercise administrators, including ones who would report to different assessment supervisors. The training was standardized so that all supervisors used a prepared script and exercises to train the exercise administrators.

Each exercise administrator received an exercise administrator manual, which covered the full range of their job responsibilities. After studying the manual, they attended a half-day training session. During the training, the supervisor reviewed all aspects of the exercise administrators' job including preparing materials, booklets, and Administration Schedules for assessments; the actual conduct of the session; post-assessment collection of materials; coding booklet covers; recordkeeping; and administrative matters. In January 1996, each exercise administrator attended a shorter, refresher training session, conducted by the assessment supervisor, to gain further experience with the auxiliary materials, such as mathematics manipulatives and science kits, to be used in specific assessment sessions.

5.3 SELECTING THE STUDENT SAMPLE FOR MAIN NAEP

5.3.1 Grade-Eligible Sample

After securing cooperation from the school, the first scheduled visit to the school was made to select the sample of students to take part in the main assessments, and to conclude the arrangements for the actual testing. This visit was made in January by the supervisor responsible for the assessments in the school. Upon arriving at the school (rarely, sampling was done at the district office instead of in the school), the supervisor first reviewed the list of grade-eligible students and confirmed with the school coordinator that all eligible students were listed. If any eligible students were omitted, sampling could not proceed until the list was completed.

Using the computer-generated Session Assignment Form (SAF) for the main assessment, which was specific to the school, the supervisor selected the sample of students to be assessed. The SAF documented the types of sessions to be administered, the anticipated number of students to be assessed, the expected number of students eligible for the assessment, and a series of line numbers designating the students to be sampled. Those eligible students on the school's master list whose line numbers were shown on the SAF were selected for the assessment. After making sure that all eligible students had been listed, the supervisor numbered the students on the master list. If the total number of eligible students was within the minimum and maximum limits indicated on the SAF, the supervisor could proceed to select the sample. If the number was outside the limits, the supervisor called Westat for additional sampling instructions. With either the original instructions or revised line numbers, the supervisor proceeded to select the sample of students. The SAFs provided step-by-step instructions for sampling, indicating not just the line number of each student to be selected, but the type of assessment session for which each student was selected.

Once students were assigned to sessions, the supervisor and exercise administrators filled out an Administration Schedule for each session. The Administration Schedule is the primary control document for the assessment. It is used to list each sampled student and is the only link between booklets and students. The sample was designed so that about 30 students were assigned to each session. The supervisor discussed the final schedule of the sessions with the school coordinator and the date, time, and location of each session were filled in on the Administration Schedules. Because student names were recorded on the Administration Schedules, those forms remained in the schools after the sample was drawn.

The supervisor then asked the school coordinator to identify any students in the sample with an Individualized Education Program (IEP) (for reasons other than being gifted and talented) and/or who were designated as LEP. Any student with either (or both) of these designations was to be indicated on the Administration Schedules. The school was asked to complete an SD/LEP Student Questionnaire for each student with this designation. This was to be completed by a teacher, counselor or other school official who knew the designated student well.

The school coordinator was also asked to determine whether any of these students should be excluded from NAEP based on the criteria for assessing SD/LEP students (discussed in detail in Section 5.3.2).[3] Preliminary results indicate that less than half of the students with SD and/or LEP designations were excluded from the assessment. If the school coordinator could not identify the excluded students while the supervisor was at the school, the instructions were left with the Coordinator along with blank copies of the SD/LEP Student Questionnaire. In those cases, the Coordinator consulted with other school

[3] The criteria differs for the main and long-term trend assessments.

officials and informed the supervisor as to who was to be excluded when he/she returned for the assessment.

At the end of the sampling visit, if requested by the school, the supervisor and/or exercise administrators made lists of the sampled students for the teachers and/or completed appointment cards notifying students about their assessment schedule. Teacher notification letters were also prepared in some schools, which explained the assessment and listed the students who had been selected.

5.3.2 Sampling for Special Studies

Two special studies, requiring added steps in the sampling process, were included in the main assessment for 1996. (The special studies were not a part of the long-term trend component.) One of these special studies involved students eligible for advanced mathematics or science sessions. The other involved applying two versions of the SD/LEP "inclusion" criteria for NAEP assessments and, in some schools, offering accommodations for testing students designated as SD/LEP.

➪ Advanced Sessions

Samples of advanced mathematics or advanced science students were designated by separate series of line numbers on the SAF as was done for the other session types scheduled for a school. However, before these students could be listed on the Administration Schedule for the advanced session, it was necessary to check each selected student's eligibility using lists of students in advanced courses prepared by the school. The definitions of advanced students were as follows:

- Grade 8 Mathematics: students enrolled in Algebra 1 or beyond at anytime during the 1995-96 school year;

- Grade 12 Mathematics: students enrolled in Algebra 3, Pre-Calculus, Calculus and Analytic Geometry, Calculus or AP Calculus; and

- Grade 12 Science: students enrolled in AP Biology, Chemistry 2 (AP), Physics 1, Physics 2 without Calculus, and Physics 2 (AP).

The advanced sessions were available only at grade 8 in mathematics and grade 12 in mathematics and science. Of the students designated for advanced sessions as per the line numbers on the SAF, only those who were also eligible (according to the definitions above) were actually listed on the Administration Schedules for the advanced sessions.

One further sampling step was applied in grade 8 and grade 12 schools regardless of whether the school was actually scheduled for an advanced session. After the samples for all sessions were selected, the supervisors compared the school's lists of advanced students against the students selected for each session to determine those students who were eligible for an advanced session. This did not mean that all eligible students would take the advanced session (only those selected and eligible for an advanced session were actually assessed at the advanced level), but the eligibility status was then recorded on the Administration Schedules for all sessions to provide a source of information on the extent to which the school offered advanced mathematics and science courses.

➪ SD/LEP Sampling and Inclusion Criteria

Because of increased interest throughout the education community in assessing as many students as possible, NAEP has begun to evaluate the effects of using revised criteria for inclusion of SD/LEP students and providing testing accommodations that are usually offered to these students by their schools. For the 1996 study, the school sample was divided into three subsamples by the statisticians at Westat. The purpose of the subsamples was to: collect data under the same conditions as previous studies in order to maintain trend in mathematics within NAEP; evaluate the impact of a revised, more specific set of inclusion criteria; and evaluate the combined effect of the new criteria and the use of accommodations for testing students. The three subsamples of schools were defined as follows:

S1 (Sample 1)	These schools used the criteria from 1990 and 1992, and accommodations were not provided. Since the issue of "trend" applies to mathematics and not to science, only mathematics was assessed in these schools.
S2 (Sample 2)	These schools used the new 1996 criteria, but accommodations were not offered. This was designed to evaluate only the impact of changing the criteria.
S3 (Sample 3)	These schools applied the new 1996 criteria and the accommodations most commonly used for achievement testing were offered. For LEP students, the adaptations included a bilingual (English-Spanish) version of the mathematics assessment and a Spanish-language glossary for the science assessment. For IEP students, the accommodations included: small group or one-on-one assessments, untimed assessments, and reading aloud. Large-print and Braille booklets were also offered in some mathematics sessions.

During the sampling visit, after the samples of students were selected for each session, the schools were asked to identify any sampled students whom the school considered to be SD/LEP. The school was then asked to complete an SD/LEP Questionnaire regarding each of these students. These basic steps were consistent with previous studies and did not vary among the three types of schools.

Further, the schools were asked to indicate which of the SD/LEP students should be included in the assessment and which should not. Again, this step is consistent with previous studies, but in 1996, the specific criteria and availability of accommodations varied among the three types of schools. Each school based its decisions about assessing SD/LEP students on the specific criteria provided to that school.

To produce as large a sample as possible of students from which to evaluate the new criteria and procedures, SD/LEP students were oversampled in certain session types in some schools. This was the case for mathematics sessions at grades 4, 8, and 12 in S2 schools and in both mathematics and science sessions at all three grades in S3 schools. The oversampled students were added to the appropriate Administration Schedules (according to the instructions on the SAF) as part of the sampling process.

The SAF contained specific instructions on oversampling for the NAEP supervisor if it was required in a particular school. Oversampling of SD/LEP students was performed only in S2 and S3 schools, and the "pool" from which the oversample was selected varied according to the grade to be assessed and the types of sessions scheduled.

The pool was defined as:

- Grade 4 - any IEP and/or LEP students from the grade-eligible list who were not selected for any session type; or

- Grade 8 or 12 with an advanced session - any IEP and/or LEP students preselected for an advanced session but not eligible for that session; or

- Grade 8 or 12 with no advanced session - any IEP and/or LEP students from the grade-eligible list who were not selected for any session type.

For the supervisors to select the oversample correctly, it was necessary to first complete the sampling for every session in the school and to be sure that the school had provided IEP and/or LEP status for every student on the grade-eligible list (i.e., not just for the selected students).

Once the oversample pool was established for a particular school, the supervisor numbered (consecutively) the students in the pool. (This was essentially a renumbering and was done separately from the original numbering of all students on the list). The oversample was then selected using the oversampling line numbers specified on the SAF. Students were added to either regular mathematics or regular science sessions according to the specifications on the SAF. The names of the oversampled students were inserted at the end of the appropriate Administration Schedule. The school was asked to complete an SD/LEP Questionnaire on each such student and to determine whether the student should be included in the assessment and, for S3 schools, what specific testing accommodation(s), if any, are called for in the students IEP and/or are normally provided for each student by the school.

The unweighted results of the 1996 assessments show that the sampling process generated, in total, 15,871 students to be assessed in S1 schools, 48,769 to be assessed to S2 schools, and 41,513 in S3 schools. These counts include the SD/LEP students that the schools determined should participate in the assessments. Accommodations were used in just over 200 S3 schools (about 30%) for approximately 1,050 students. The most frequently provided accommodations were small group, extended time (untimed testing), and bilingual assessment booklets. These results are very preliminary, however, because they are unweighted and cannot be used to compare results for S1, S2, and S3 schools (without applying the weighting process). Detailed information and results of the SD/LEP special study will be provided in a separate report.

5.4 CONDUCTING THE ASSESSMENT SESSIONS

The primary responsibility for conducting assessment sessions was given to the exercise administrators. Supervisors were required to observe the first session each exercise administrator conducted to ensure that they followed the procedures properly. Supervisors were also required to be present in all schools with more than one small session to be conducted. The supervisor plays an important role as the liaison between the national assessment and school staff ensuring that the assessments go smoothly.

To ensure that sessions were administered in a uniform way, the exercise administrator was provided with scripts for each session type. The scripts were to be read verbatim. The scripts began with a brief introduction to the study. The exercise administrator was then directed to distribute the booklets, being careful to match the student with the preassigned booklet.

After the booklets were distributed, some additional, scripted directions were read. Students were asked to write in the NAEP school ID (except in grade 4, where NAEP staff entered the ID) and their home ZIP code on the cover of the booklet, and given some general directions in completing the assessment. For fourth grade students, all of the background questions were then read aloud by the exercise administrator; at the upper grades, the first question, which asks the students' race/ethnicity, was read by the exercise administrator, and the students read the rest to themselves. After the background questions were completed, the students were told that any further questions they might have could not be answered by the exercise administrator, and that they were to begin the first cognitive section of the assessment. This process (along with the script) was modified somewhat for science where the background questions were at the end of the assessment booklet, and none of the items was read aloud at grades 8 or 12.

During the sessions, the exercise administrators walked around the room monitoring the students to make sure they were working in the correct section of their booklet and to discourage them from looking at a neighbor's booklet.

At the end of each assessment session, booklets were collected and students dismissed according to the school's policy. The exercise administrator was then responsible for completing the information at the top of the Administration Schedule, totaling the number of participating students, and coding the covers of all booklets, including those booklets assigned to absent students.

5.5 RESULTS OF THE MAIN NAEP ASSESSMENT

5.5.1 School and Student Participation

The unweighted school response rate for the main assessments in 1996 was 82 percent. The final sample of cooperating schools included 604 schools at grade 4, 592 schools at grade 8, and 591 schools at grade 12. Table 5-1 shows comparative response rates for the last four assessment periods.

Unlike the student response rates, there has been a small but steady decline in the main assessment school response rates over the last several assessment periods. This has occurred despite persistent efforts to convert schools and districts that indicate that they are not interested in participating in the assessments. Both Westat field managers and ETS staff have been employed in these conversion efforts.

Table 5-1
Comparison of Student and School Response Rates for Main NAEP, 1990-1996

	1990	1992	1994	1996
Student response				
Grade 4	92.9	93.4	93.2	95.5
Grade 8	89.0	88.8	91.0	93.2
Grade 12	80.8	80.8	81.1	80.1
School response				
Grade 4	88.3	86.4	86.0	85.8
Grade 8	86.7	85.3	85.5	81.9
Grade 12	81.3	81.5	78.6	78.7

The most frequently stated reason for school and district refusals has been the increase in testing throughout the jurisdictions and the resulting difficulty in finding time in the school schedule to conduct the NAEP assessments. With so many states now mandating their own testing, school schedules are becoming tighter, and administrators are finding it increasingly difficult to accommodate outside testing. Despite the increased visibility and publicity surrounding NAEP, schools are reluctantly finding it necessary to decline participation as a result of the increasing demands on their students' time.

Of the 113,846 students sampled for the 1996 assessment, roughly five percent overall were excluded by schools. Altogether, 94,157 students were assessed across all three grades: 28,528 students were assessed at fourth grade, 32,339 were assessed at eighth grade, and 33,290 students were assessed at twelfth grade. The overall student participation rate was 88.7 percent (after eliminating any withdrawn and excluded students).

The response rate at which supervisors were required to conduct a makeup session was raised from the standard that had been used in previous main assessments. The previous rates of 75 percent and 85 percent (in 1994 only) were changed to 90 percent for 1996, that is, any session (or group of sessions within the same subject area) at which fewer than 90 percent of the eligible students were assessed would require a makeup session (assuming that the school was willing to schedule one). This change resulted in 129 schools conducting makeup sessions that would not have been required to do so under the 85 percent rule. In these schools, an additional 595 students were assessed. These 595 students served to increase the overall response by less than one percentage point (0.6%). The greatest increase was at the grade 12 level, where 413 students were assessed in these additional sessions, which raised the response rate for this group of students by about one percentage point.

5.5.2 Assessment Questionnaires

Westat provided each school with a School Characteristics and Policies Questionnaire a few weeks before the assessment was scheduled to be conducted (i.e., at the time of sampling). At the same time, supervisors prepared an SD/LEP Student Questionnaire for each sampled student with either an IEP and/or LEP designation, with the request that it be completed by someone at the school knowledgeable about that student.

Selected teachers of fourth- and eighth-grade mathematics and science were asked to fill out Teacher Questionnaires. The teachers asked to participate were the mathematics or science teachers of those students selected for the assessment so that the teacher data could be linked to student performance data. The Teacher Questionnaire for grade 4 was combined into one form since it is recognized that at this grade level, the same teacher would probably teach all of the subjects. For grade 8, there were two distinct questionnaires, one for mathematics teachers and the other for science teachers. At grade 12, a teacher questionnaire was used only for advanced mathematics sessions.

The supervisor requested that the Teacher Questionnaires be distributed as quickly as possible after the sampling so that they could be returned by the day of the assessment. Additional introductory materials were included with the Teacher Questionnaires in response to questions that teachers have had in the past about the importance of completing the questionnaires and about NAEP in general. Teachers received a letter explaining the purpose of the Questionnaire, along with background materials about NAEP.

If the Teacher Questionnaires were not complete at the time of the assessment, the supervisor left a postage-paid envelope to NCS to be used to return the questionnaires. Table 5-2 shows the number of questionnaires distributed and the number completed.

Table 5-2
Background Questionnaires Received for Schools, Teachers, and SD/LEP Students in the 1996 Main Assessment[1]

Grade Assessed	School Characteristics and Policy Questionnaire	Teacher Questionnaires			SD/LEP Student Questionnaire
		Mathematics/Science (grade 4 only)	Mathematics	Science	
Grade 4					
Number expected	605	1,601	NA	NA	5,116
Number received	577	1,601	NA	NA	4,885
Percent received	95%	100%	NA	NA	95%
Grade 8					
Number expected	592	NA	1,400	844	5,048
Number received	554	NA	1,365	802	4,770
Percent received	94%	NA	98%	95%	94%
Grade 12[2]					
Number expected	593	NA	475	NA	4,147
Number received	546	NA	475	NA	3,806
Percent received	92%	NA	100%	NA	92%

[1] The numbers in this table reflect the full samples, including S1, S2, and S3.
[2] At grade 12, teacher questionnaires were used only for teachers of advanced mathematics. Thus, no data were collected from science teachers, and the data shown here represent teachers of advanced mathematics only.

5.6 LONG-TERM TREND ASSESSMENTS

5.6.1 Overview

To provide continuity and comparability with past NAEP studies, the long-term trend component (formerly referred to as the "bridge" assessments) replicates procedures and materials that have been used since the inception of NAEP. Student eligibility in long-term trend is always based on criteria used in years prior to 1988 (when the modal grade for students aged 17 changed from the grade 11 to grade 12). The 1996 schedule for long-term trend assessments was as follows: the fall assessment of age 13/grade 8 students was held in the 11-week period from October 9 through December 22, 1995; the winter assessment of age 9/grade 4 students was held during the 10-week period from January 3 through March 8, 1996; and the spring trend assessment of students who were age 17/grade 11 was conducted in the 9-week period from March 11 through May 10, 1996. Students were assessed in reading, writing, mathematics, and science.

Paced tape sessions were conducted with samples of age-eligible students only, as was done in all previous years. Additional samples of age- and grade-eligible students were assessed with spiral (print-administered) booklets, following procedures initiated in 1984. Six different types of sessions were conducted: one print-administered and five separate tape-administered sessions. Depending on the size of the school, up to four different session types, involving a total of about 80 students, might have been conducted in a participating school.

5.6.2 Selecting the Student Sample

Procedures for sampling in long-term trend schools were very similar to those employed in the schools selected for the main assessment. One to two weeks before the assessment, the supervisor visited the school to select the sample. Lists of students were reviewed to ensure that all age- and grade-eligible students were listed. The SAF for long-term trend schools specified a range for the expected number of eligible students. If the total number of students was within this allowable range, the sampling could proceed. Otherwise, the supervisor called Westat for additional sampling instructions. The SAF directed the supervisor to assign students to long-term trend session types based on their line numbers from the student list that the school had prepared. (The SAFs for the long-term trend sample were like those used in long-term trend sampling for previous years, and were distinct from the SAFs for the main assessment that were described earlier in this report.)

The only major variation within the sampling for long-term trend assessments was that, for the tape sessions, only age-eligible students were selected. For these sessions, the supervisor selected from the entire list of students (age- and grade-eligible), but then deleted those who were only grade-eligible before recording the names of the students to be assessed on the Administration Schedules.

The criteria for excluding students were also different for the long-term trend schools (compared to the main assessment), and again followed the criteria that were established previously for long-term trend. For those students who were excluded, the school was asked to complete an Excluded Student Questionnaire. If the school coordinator could not identify the excluded students while the supervisor was at the school, a set of instructions for excluding students was left with the coordinator along with the estimated number of questionnaires that would be needed.

5.6.3 Conduct of the Assessment

The conduct of the assessments in schools selected for the long-term trend program is essentially the same as in the schools selected for the main assessment. Scripts are provided for the supervisors and exercise administrators to use in administering the sessions. The major difference compared to main assessment is that most of the sessions are tape administered. In these sessions, after the distribution of the test booklets, the administrator is instructed to turn on the tape recorder. The remainder of the instructions are contained on the tape, and the timing is determined by the length of time that the tape runs.

5.6.4 Results of Long-Term Trend Assessments

The unweighted school response rate for the 1996 long-term trend assessments was 83 percent. The final sample of cooperating schools included 240 schools at age 9/grade 4, and 238 schools at age 13/grade 8, and 191 schools at age 17/grade 11. Nearly 30,000 students were assessed in long-term trend, or 91 percent of those eligible to be assessed.

Of the 36,371 students sampled for long-term trend assessments, eight percent were excluded by the schools. Overall, 29,791 students were assessed across all three age/grade groups: 10,432 students were assessed at age 9/grade 4, 11,151 students were assessed at age 13/grade 8, and 8,208 were assessed at age 17/grade 11. Table 5-3 shows comparative response rates for the last four long-term trend assessments.

Table 5-3
Comparison of Student and School Response Rates for Long-Term Trend NAEP, 1990-1996

	1990	1992	1994	1996[1]
Student response				
Grade 4	92.4	94.0	94.2	95.5
Grade 8	90.4	90.8	92.2	91.9
Grade 12	81.2	82.8	84.1	84.0
School response				
Grade 4	88.1	87.4	86.7	84.8
Grade 8	90.5	84.7	81.7	82.4
Grade 12	80.7	81.3	81.1	81.3

[1] The numbers in this table reflect the full samples, including S1, S2, and S3.

5.6.5 Assessment Questionnaires

The School Characteristics and Policies Questionnaire and the Excluded Student Questionnaire are forms that were distributed in the schools to be completed by school personnel. The School Characteristics and Policies Questionnaire was provided to the school by the assessment supervisor at the time of the sampling visit. This form was to be filled out by the principal or other staff member knowledgeable about the school's administrative policies and staff characteristics. The supervisors collected the completed questionnaire when they returned to the school for the assessment.

An Excluded Student Questionnaire was to be filled out for every student who was sampled for the assessment but excluded by the school. Following exclusion criteria used in previous long-term trend assessments, schools could exclude students with limited English-speaking ability, those who were educable mentally retarded, or functionally disabled students, if in the judgment of school staff or as indicated in school records, they were unable to "participate meaningfully" in the assessment. After the sample of students was drawn and Administration Schedules prepared, the supervisor requested that the school coordinator identify any students who should be excluded. The supervisor then gave an Excluded Student Questionnaire to the coordinator for every excluded student, with the request that it be completed by someone in the school knowledgeable about the student. (Note that this varies somewhat from the main assessment where questionnaires are assigned for all sampled students with an IEP and/or LEP, not just for those who are excluded from the assessment.)

The supervisor attempted to collect all completed questionnaires (School Characteristics and Policies Questionnaire and Excluded Student) on the assessment day. If the questionnaires were not ready, and it was convenient for the supervisor or an exercise administrator to return to the school later to pick them up, they would do so. Otherwise, the supervisor gave the coordinator a postage-paid envelope to use to mail the forms to NCS. All (100%) of the School Characteristics and Policies Questionnaires were completed and returned, and 95.9 percent of the Excluded Student Questionnaires were returned.

Once the assessments were completed in a school, the supervisor and exercise administrators completed the coding of the front covers of the assessment booklets, filled out the necessary forms, and shipped the booklets and forms to NCS. A copy of all forms was sent to Westat so that progress in the field could be closely monitored.

The School Worksheet was used by the supervisor to summarize the results of the assessment sessions in each school. The number of students to be assessed, the number actually assessed, and the

number absent were entered so that the supervisor could calculate whether a makeup session was required. Attendance of less than 90 percent required a makeup. If a makeup was required for one or more session types, the supervisor discussed the scheduling of the makeup with the school coordinator.

In long-term trend assessments prior to 1994, the percentage of students attending that would necessitate a makeup session was 75 percent or below. For 1994, this rate was increased to 85 percent, and it was raised again to 90 percent for 1996. By raising the rate to 90 percent, an additional 144 students were assessed in 73 schools, compared to the 85-percent-rule, increasing the overall student response rate for 1996 by less than one-half of one percentage point (0.44%).

The top (original) copy of the School Worksheet, any Excluded Student Questionnaires completed by the school, and the Administration Schedules (with the students' names removed and left at the school) were included with the booklets in the shipment to NCS. In addition, the supervisor included a packing list with the materials, which inventoried the assessment materials assigned to and returned from the school.

5.7 FIELD MANAGEMENT

Two field managers monitored the work of 25 scheduling supervisors who worked during fall 1995 to gain cooperation of districts and schools for the main assessment. During the assessment period, these staff were expanded to about 75 supervisors and 6 field managers (4 of whom were located in Westat's home office). An additional field manager was assigned exclusively to long-term trend NAEP throughout the gaining cooperation and assessment periods. All supervisors reported directly to their field managers who, in turn, reported to Westat's field director. All contacts were made at least weekly.

An automated management system was developed and maintained in Westat's home office. The scheduling supervisors working to contact schools during the fall used this system on their portable computers. The system contained a record for each sampled school. A disposition code structure was developed to indicate the status of each school's participation (e.g., school cooperating, decision pending, school refusal, district refusal, school closed, etc.). As a school's status was determined, the scheduling supervisors entered the status of the school onto their computers, and this information was downloaded onto the home office system on a weekly basis. Disposition reports were then generated from the receipt system once a week so that home office staff could review the progress of securing cooperation from the sampled schools.

These reports were an invaluable tool for the sampling statisticians as well as for the field director and field management staff. They provided the statisticians with the information needed to determine whether the sample of schools was adequate to produce representative results. Based on the information contained in these reports, the sampling statisticians selected substitute schools to replace some of the non-cooperating schools.

After assessments were completed, the system was used to enter data from the School Worksheets (for both Main NAEP and long-term trend) on the number of students to be assessed, the number assessed, and the number absent for each school. Data on completed questionnaires received was provided by NCS. The system was also used to alter school assessment dates, particularly when bad weather required a change in schedule, and to monitor plans for and progress in conducting makeup sessions. Reports were generated weekly during the assessment period that allowed the project staff to monitor the progress of the assessments both in terms of checking that the schools were assessed on schedule as well as assuring that a high response rate was achieved. The sampling statisticians used these reports to monitor the sample yield by school, PSU, and age/grade level.

Progress of the assessments was constantly monitored through telephone reports held between NAEP supervisors, field managers, and home office staff. During these phone conversations, the supervisors' schedules were reviewed and updated, and any problems that the supervisors were experiencing were discussed. Much of the attention this year was focused on maintaining the schedule in light of the many postponements due to severe winter weather.

The supervisors who traveled filled out a Work Schedule for a one- to two-week period, showing their whereabouts, so that they could be contacted if necessary. It also allowed field managers and project staff to review the supervisors' schedules and the distribution of work.

Progress of the field work was also monitored during quality control visits made to the field by Westat and ETS office staff.

Chapter 6

PROCESSING ASSESSMENT MATERIALS[1]

Patrick B. Bourgeacq, Bradley Thayer, and Timothy Robinson
National Computer Systems

6.1 INTRODUCTION

This portion of the report reviews the activities conducted by National Computer Systems (NCS) for the NAEP 1996 main assessments in mathematics and science, and long-term trend assessments in reading, mathematics, science, and writing. As a subcontractor to Educational Testing Service (ETS), NCS was responsible for printing all of the NAEP student booklets and control documents; distributing the assessment materials to the field; receiving, tracking, processing and editing the assessment materials as they returned from the field; scoring all of the constructed-response items (in conjunction with ETS); and delivering the assessment data files to ETS for analysis and reporting.

For this assessment, NCS was charged with processing and scoring the largest assessment in the history of NAEP in the shortest amount of time. Further, image scanning processes, eliminating almost all paper handling during scoring and improving monitoring and reliability scoring, increased to nearly twice that of the 1994 assessment. Materials management and distribution of over one quarter of a million science kits, receipt control processing for all receipts within 48 hours, image scanning throughput increased nearly twice that of 1995, and professional scoring of over nine million constructed responses highlight the challenges met by NCS for the 1996 NAEP national and state assessments.

NCS processed more than 134,000 booklets for the NAEP 1996 national assessment, as shown in Tables 6-1 and 6-2. NCS also received and processed a total of over 2,300 school characteristics and policies questionnaires, over 4,700 teacher questionnaires, and over 16,274 SD/LEP questionnaires for the three grades, as shown in Table 6-3. Table 6-4 lists key events and dates in the NAEP schedule.

This chapter of the report reviews the activities conducted by NCS for the NAEP 1996 main and long-term trend assessments.

[1] Bradley Thayer was the NCS project manager for 1996 NAEP, Patrick Bourgeacq was the NCS project director for 1996 NAEP scoring, and Timothy Robinson was the NCS senior processing coordinator for 1996 NAEP.

Table 6-1

Processing and Scoring Totals for the 1996 NAEP Assessment

	Booklets Processed	Number of Constructed-Response Items[1]	Number of Responses Scored[2]	Number of Scorers and Team Leaders[3]	Length of Training and Scoring
Fall Trend Reading/Writing	5,287	22	42,741	7 / 1	11/29/95 - 1/5/96
Fall Trend Mathematics	5,727	28	88,782	11 / 0	12/18/95 - 1/5/96
Winter Trend Reading/Writing	4,988	20	34,341	6 / 1	1/30/96 - 4/3/96
Winter Trend Mathematics	5,442	29	86,322	9 / 1	3/25/96 - 3/29/96
Spring Trend Reading/Writing	4,669	25	43,094	6 / 1	4/10/96 - 5/24/96
Spring Trend Mathematics	3,570	29	86,478	5 / 1	5/13/96 - 5/22/96
Long-Term Trend Writing Holistic	N/A	6	63,793	56 / 7	6/10/96 - 6/16/96
Long-Term Trend Writing Mechanics	N/A	3	2,329	33 / 6	7/15/96 - 7/31/96
1990 Rescore 4th Grade Mathematics	749	12	8,988	198 / 17	3/13/96 - 5/6/96
1990 Rescore 8th Grade Mathematics	730	18	13,374	198 / 17	3/13/96 - 5/6/96
1990 Rescore 12th Grade Mathematics	725	21	15,225	198 / 17	3/13/96 - 5/6/96
1992 Rescore 4th Grade Mathematics	2,498	36	27,031	198 / 17	3/13/96 - 5/6/96
1992 Rescore 8th Grade Mathematics	2,498	44	33,077	198 / 17	3/13/96 - 5/6/96
1992 Rescore 12th Grade Mathematics	2,498	44	32,387	198 / 17	3/13/96 - 5/6/96
National 4th Grade Mathematics Spiral	10,816	64	170,219	198 / 17	3/13/96 - 5/6/96
National 4th Grade Mathematics Estimation	2,128	3	7,756	198 / 17	3/13/96 - 5/6/96
National 4th Grade Mathematics Theme	4,038	13	29,750	198 / 17	3/13/96 - 5/6/96
National 8th Grade Mathematics Spiral	11,554	69	195,764	198 / 17	3/13/96 - 5/6/96

[1] This is the number of discrete constructed-response items in assessment booklets.

[2] This is the number of student responses to the constructed-response items. These scored responses include those that were rescored for reliability estimation.

[3] Because readers scored items from all grades and all types of booklets, it is not possible to break the numbers down by how many scored each classification of items.

(continued)

Table 6-1 (continued)
Processing and Scoring Totals for the 1996 NAEP Assessment

	Booklets Processed	Number of Constructed-Response Items[1]	Number of Responses Scored[2]	Number of Scorers and Team Leaders[3]	Length of Training and Scoring
National 8th Grade Mathematics Estimation	2,267	6	17,027	198 / 17	3/13/96 - 5/6/96
National 8th Grade Mathematics Theme	4,259	13	34,613	198 / 17	3/13/96 - 5/6/96
National 8th Grade Mathematics Advanced	2,382	14	41,693	198 / 17	3/13/96 - 5/6/96
National 12th Grade Mathematics Spiral	10,740	73	225,540	198 / 17	3/13/96 - 5/6/96
National 12th Grade Mathematics Estimation	1,883	0	0	198 / 17	3/13/96 - 5/6/96
National 12th Grade Mathematics Theme	3,892	12	29,210	198 / 17	3/13/96 - 5/6/96
National 12th Grade Mathematics Advanced	2,987	15	56,101	198 / 17	3/13/96 - 5/6/96
Bilingual 4th Grade Mathematics National	91	10	2,280	0 /4	5/2/96 - 5/2/96
Bilingual 8th Grade Mathematics National	36	12	1,080	0 /4	5/3/96 - 5/3/96
National 4th Grade Science Spiral	11,677	94	275,339	306 / 24	3/18/96 - 5/28/96
National 8th Grade Science Spiral	12,079	125	322,261	306 / 24	3/18/96 - 6/7/96
National 12th Grade Science Spiral	11,579	120	342,104	306 / 24	3/18/96 - 6/2/96
National 12th Grade Science Advanced	2,449	36	110,207	306 / 24	3/18/96 - 6/2/96

[1] This is the number of discrete constructed-response items in assessment booklets.
[2] This is the number of student responses to the constructed-response items. These scored responses include those that were rescored for reliability estimation.
[3] Because readers scored items from all grades and all types of booklets, it is not possible to break the numbers down by how many scored each classification of items.

Table 6-2

Student Participation and Session Information for the 1996 NAEP Assessment

	Number of Sessions	Number of Booklets for Assessed Students	Number of Booklets for Absent Students	Number of Booklets for Excluded Student	Number of Scanned Sheets
Long-Term Trend					
Fall	639	11,150	981	974	288,624
Winter	623	10,406	486	1,172	215,679
Spring	539	8,209	1,560	758	243,452
Main					
Grade 4 total	1,693	28,531	1,354	2,293	711,291
Mathematics	982	16,953	796	1,198	418,509
Science	711	11,578	558	1,095	292,782
Grade 8 total	1,698	32,339	2,359	1,823	922,892
Mathematics	819	17,992	1,338	1,062	509,064
Advanced Mathematics	345	2,375	94	4	73,268
Science	534	11,972	927	757	340,560
Grade 12 total	2,196	33,306	8,266	1,704	1,143,332
Mathematics	848	16,424	4,103	969	536,256
Advanced Mathematics	380	2,965	456	11	99,112
Science	613	11,486	3,332	715	387,504
Advanced Science	355	2,431	375	9	120,460
Other					
Rosters					23,535
Administration Schedules					24,575
Rescore Mathematics 1992 Grade 4					45,286
Rescore Mathematics 1992 Grade 8					61,311
Rescore Mathematics 1990					46,507

Table 6-3

Questionnaire Totals for the 1996 NAEP Assessment

	Expected	Received	Percent
Main Assessment			
Grade 4			
SD/LEP Questionnaire	5,116	4,885	95%
School Characteristics Questionnaire	597	577	97%
Grade 8			
SD/LEP Questionnaire	5,048	4,770	94%
School Characteristics Questionnaire	580	554	96%
Grade 12			
SD/LEP Questionnaire	4,147	3,806	92%
School Characteristics Questionnaire	582	546	94%
Long-Term Trend			
Fall			
Excluded Students Questionnaire	979	947	97%
School Characteristics Questionnaire	239	224	94%
Winter			
Excluded Students Questionnaire	1,187	1,145	96%
School Characteristics Questionnaire	247	235	95%
Spring			
Excluded Students Questionnaire	768	721	94%
School Characteristics Questionnaire	186	173	93%
Main Teacher Questionnaires			
Grade 4			
Mathematics/Science	1,599	1,609	101%
Grade 8			
Mathematics	1,401	1,359	97%
Science	1,310	1,270	97%
Grade 12			
Advanced Mathematics	476	487	102%

Table 6-4

NCS Schedule for the 1996 NAEP Assessment

Task	Planned Start	Planned Finish	Actual Start	Actual Finish
FALL LONG-TERM TREND				
Printing	6/15/95	9/19/95	6/15/95	9/19/95
Pre-packaging (barcoding, spiraling, and quality control)	8/14/95	9/7/95	8/14/95	9/7/95
Session file to NCS from Westat	8/25/95	8/25/95	8/25/95	8/25/95
Westat supervisor training	8/27/95	8/30/95	8/27/95	8/30/95
Print Administration Schedule	9/1/95	9/5/95	9/1/95	9/1/95
Supervisor add file from Westat	9/5/95	9/5/95	9/5/95	9/5/95
Final packing specifications to packaging	9/5/95	9/5/95	9/5/95	9/5/95
Packing list and labels to packaging	9/5/95	9/6/95	9/6/95	9/6/95
Grade 8 school characteristics and policies questionnaires arrive	9/7/95	9/7/95	9/7/95	9/19/95
Ship Administration Schedules	9/7/95	9/7/95	9/7/95	9/7/95
Information to Key Entry for screen set up	9/11/95	9/11/95	9/11/95	9/11/95
Bulk/session packaging	9/11/95	9/18/95	9/11/95	9/15/95
Packaging visit by ETS	9/13/95	9/14/95	9/14/95	9/14/95
Ship materials to supervisors	9/18/95	9/18/95	9/14/95	9/14/95
Materials due to supervisors	9/22/95	9/22/95	9/22/95	9/22/95
Processing specifications to Operations Department	10/2/95	10/2/95	10/2/95	10/2/95
Processing kick-off meeting	10/3/95	10/3/95	10/3/95	10/3/95
Requisitions for table leaders to HR	10/9/95	10/9/95	10/9/95	10/9/95
Requisitions for scorers to HR	10/9/95	10/9/95	10/9/95	10/9/95
Photocopy training materials	10/9/95	10/13/95	10/9/95	10/13/95
Test administration	10/9/95	12/19/95	10/10/95	12/19/95
Blue dot	10/10/95	10/20/95	10/20/95	10/20/95
Receiving	10/10/95	12/23/95	10/11/95	1/15/96
HR extends offer to table leaders	10/13/95	10/13/95	10/13/95	10/13/95
HR extends offers to scorers	10/16/95	10/16/95	10/16/95	10/16/95
General - Network Meeting	10/20/95	10/20/95	10/20/95	10/20/95
Processing	10/20/95	12/28/95	10/16/95	1/9/96
Scoring training preparation	10/9/95	11/17/95	10/9/95	10/17/95
Scoring training - writing	11/29/95	11/30/95	11/29/95	11/30/95
Scoring training - reading	12/4/95	12/5/95	12/4/95	12/5/95
Scoring reading/writing	11/29/95	1/5/96	11/29/95	1/5/96
Weights file shipped	1/12/96	1/15/96	1/15/96	1/15/96
Tape delivered	1/12/96	1/15/96	1/25/96	1/25/96
WINTER LONG-TERM TREND				
Printing	6/15/95	9/19/95	6/15/95	9/19/95
Pre-packaging (barcoding, spiraling)	9/29/95	10/10/95	9/6/95	9/14/95
Bundle sheets delivered to packaging	10/25/95	10/25/95	9/1/95	9/1/95
All bundles through clean quality control	11/3/95	11/3/95	9/1/95	9/1/95

(continued)

Table 6-4 (continued)
NCS Schedule for the 1996 NAEP Assessment

Task	Planned Start	Planned Finish	Actual Start	Actual Finish
WINTER LONG-TERM TREND (continued)				
Final packing specifications to packaging	11/17/95	11/17/95	11/16/95	11/16/95
HR extends offers to scorers	11/20/95	11/22/95	11/20/95	11/22/95
Session data file to NCS from Westat	11/22/95	11/22/95	11/22/95	11/22/95
Administration Schedule address file from Westat	11/28/95	11/28/95	11/22/95	11/22/95
Print Administration Schedules	12/1/95	12/1/95	11/30/95	11/30/95
Bulk/session address file from Westat	12/1/95	12/1/95	11/29/95	11/29/95
Ship Administration Schedules	12/4/95	12/4/95	12/1/95	12/1/95
Packing list and labels to packaging	12/4/95	12/5/95	12/1/95	12/1/95
Final packaging	12/6/95	12/8/95	12/6/95	12/7/95
Ship session materials	12/6/95	12/8/95	12/6/95	12/7/95
Processing specifications to Operations	12/18/95	12/18/95	12/18/95	12/18/95
Requisitions for scorers to HR	12/20/95	12/20/95	12/20/95	12/20/95
Materials due to supervisors	12/22/95	12/22/95	12/15/95	12/18/95
Test administration	1/3/96	3/15/96	1/2/96	3/15/96
Blue dot (s)	1/4/96	1/8/96	1/15/96	1/22/96
Receiving	1/4/96	3/20/96	1/9/96	3/20/96
Image definition ready	1/8/96	1/8/96	1/5/96	1/5/96
Processing	1/8/96	3/22/96	1/8/96	3/22/96
Photocopy training materials	1/9/96	1/12/96	1/9/95	10/13/95
Image test data ready	1/9/96	1/12/96	1/9/96	1/12/96
Image application ready	1/11/96	1/11/96	1/11/96	1/11/96
Scoring training preparation	1/29/96	1/31/96	1/29/96	1/29/96
Scoring training - writing	1/31/96	2/2/96	1/30/96	1/31/96
Scoring training - reading	1/31/96	2/2/96	2/6/96	2/6/96
Scoring reading/writing	2/5/96	4/5/96	2/7/96	4/2/96
Requisitions for mathematics scorers to HR	3/1/96	3/1/96	3/11/96	3/11/96
Scoring training - mathematics	3/25/96	3/25/96	3/25/96	3/25/96
Scoring - mathematics	3/25/96	4/5/96	3/25/96	3/29/96
Tape delivered	4/15/96	4/15/96	4/12/96	4/12/96
Weights file shipped	4/15/96	4/15/96	4/12/96	4/12/96
SPRING LONG-TERM TREND				
Printing	6/15/95	9/19/95	6/15/95	9/19/95
Pre-packaging (barcoding, spiraling)	7/21/95	9/25/95	7/21/95	9/25/95
HR extends offers to table leaders	10/13/95	10/13/95	10/13/95	10/13/95
HR extends offers to scorers	11/20/95	11/22/95	11/20/95	11/22/95
Session file to NCS from Westat	2/5/96	2/5/96	2/5/96	2/5/96
Bulk/session address file from Westat	2/8/96	2/8/96	2/5/96	2/5/96
Administration Schedule address file from Westat	2/8/96	2/8/96	2/7/96	2/7/96
Packing list and labels to packaging	2/8/96	2/8/96	2/9/96	2/9/96
Final packing specifications to packaging	2/8/96	2/8/96	2/12/96	2/12/96

(continued)

Table 6-4 (continued)
NCS Schedule for the 1996 NAEP Assessment

Task	Planned Start	Planned Finish	Actual Start	Actual Finish
SPRING LONG-TERM TREND (continued)				
Print Administration Schedules	2/12/96	2/12/96	2/9/96	2/9/96
Ship Administration Schedules	2/16/96	2/16/96	2/13/96	2/13/96
Package/ship session materials	2/23/96	2/23/96	2/23/96	2/23/96
Material due to supervisors	2/28/96	2/28/96	3/1/96	3/1/96
Processing specifications to Operations	3/8/96	3/8/96	3/8/96	3/8/96
Rescore booklets delivered to PSC	3/8/96	3/8/96	3/8/96	3/8/96
Photocopy training materials	3/9/96	3/12/96	10/9/95	10/13/95
Test administration	3/11/96	5/10/96	3/11/96	5/17/96
PSC approval of scoring sheets	3/15/96	3/15/96	3/13/96	3/13/96
Blue dot (s)	3/18/96	3/20/96	4/2/96	4/19/96
Processing	3/21/96	5/17/96	4/19/96	5/17/96
Requisitions for mathematics scorers to HR	4/1/96	4/1/96	4/1/96	4/1/96
Scoring training preparation	4/8/96	4/9/96	4/3/96	4/3/96
Scoring training - writing	4/10/96	4/11/96	4/10/96	4/11/96
Scoring reading/writing	4/10/96	5/24/96	4/11/96	4/11/96
Scoring training - reading	4/11/96	4/11/96	4/11/96	4/11/96
Scoring training - mathematics	5/13/96	5/13/96	5/13/96	5/13/96
Scoring mathematics	5/13/96	5/24/96	5/13/96	5/24/96
Project through clean post	5/17/96	5/17/96	5/17/96	5/17/96
Tape delivered	5/31/96	5/31/96	6/6/96	6/6/96
Weights file shipped	5/31/96	5/31/96	6/6/96	6/6/96
WRITING LONG-TERM TREND - HOLISTIC SCORING				
Requisition for scorers to HR	5/15/96	5/15/96	5/15/96	5/15/96
Requisition for table leaders to HR	5/15/96	5/15/96	5/15/96	5/15/96
PSC approves score sheet	5/15/96	5/15/96	5/15/96	5/15/96
HR makes offers to table leaders	5/20/96	5/31/96	5/28/96	6/4/96
HR makes offers to scorers	5/20/96	5/31/96	5/28/96	6/7/96
Samples drawn	5/24/96	5/24/96	5/20/96	5/24/96
Scoring preparation with J. Kennedy	6/4/96	6/7/96	6/5/96	6/7/96
Training and Scoring	6/10/96	6/14/96	6/10/96	6/16/96
Data tape delivered	7/1/96	7/1/96	6/27/96	6/27/96
WRITING LONG-TERM TREND - MECHANICS SCORING				
Requisition for scorers to HR	6/14/96	6/14/96	7/9/96	7/9/96
Requisition for table leaders to HR	6/14/96	6/14/96	7/9/96	7/9/96
HR makes offers to scorers	6/24/96	6/28/96	7/9/96	7/12/96
HR makes offers to table leaders	6/24/96	6/28/96	7/9/96	7/10/96
Samples drawn	6/24/96	6/24/96	6/24/96	6/24/96

(continued)

Table 6-4 (continued)
NCS Schedule for the 1996 NAEP Assessment

Task	Planned Start	Planned Finish	Actual Start	Actual Finish
WRITING LONG-TERM TREND - MECHANICS SCORING (continued)				
Responses copied for scoring	6/25/96	7/12/96	6/25/96	7/12/96
Training and scoring	7/15/96	7/26/96	7/15/96	7/31/96
Transcribe and proofread essays and scores	7/18/96	7/31/96	7/18/96	7/31/96
Data tape delivered	8/9/96	8/9/96	8/7/96	8/7/96
MAIN ASSESSMENT				
Printing	9/2/95	12/11/95	9/2/95	12/11/95
Administration Schedule approved	9/15/95	9/15/95	9/21/95	9/21/95
Pre-packing specifications to packaging	10/2/95	10/2/95	10/2/95	10/2/95
Grade 8 teacher questionnaire roster delivered to NCS	10/12/95	10/12/95	10/16/95	10/16/95
Grade 4 mathematics/science teacher questionnaire roster to NCS	10/12/95	10/12/95	10/17/95	10/17/95
PSC obtains copies of final blocks	10/16/95	10/16/95	10/16/95	10/16/95
Administration Schedule delivered to NCS	10/18/95	10/18/95	10/23/95	10/23/95
Grade 12 teacher questionnaire roster delivered to NCS	10/20/95	10/20/95	10/16/95	10/16/95
Grade 8 school characteristics and policies questionnaires at NCS	10/20/95	10/20/95	10/23/95	10/23/95
SD/LEP roster delivered to NCS	10/20/95	10/20/95	10/24/95	10/24/95
Grade 4 school characteristics and policies questionnaires at NCS	10/20/95	10/20/95	10/25/95	10/25/95
Grade 12 school characteristics and policies questionnaires at NCS	10/20/95	10/23/95	10/23/95	10/23/95
Grade 8 mathematics spiral material at NCS	10/23/95	11/2/95	10/18/95	11/3/95
Pre-packaging begins	10/23/95	12/20/95	10/16/95	12/1/95
Grade 4 mathematics spiral material at NCS	10/26/95	11/1/95	11/1/95	11/1/95
Grade 8 mathematics teacher questionnaire at NCS	10/30/95	10/30/95	10/25/95	10/25/95
Grade 8 science teacher questionnaire at NCS	10/30/95	10/30/95	11/1/95	11/1/95
Final valid score range for each item	11/1/95	11/1/95	10/24/95	11/17/95
PSC obtains rubrics from ETS	11/1/95	11/1/95	10/24/95	11/17/95
ETS/PSC define non-scorable codes	11/1/95	11/1/95	11/1/95	11/1/95
NCS/ETS meet to review items and schedule	11/2/95	11/3/95	11/2/95	11/3/95
Grade 4 mathematics/science teacher questionnaire delivered to NCS	11/3/95	11/3/95	11/2/95	11/2/95
Grade 8 science spiral material at NCS	11/6/95	11/13/95	11/13/95	11/21/95
Grade 12 mathematics spiral material at NCS	11/14/95	11/21/95	11/21/95	12/1/95
General - sub-contractor's meeting	11/16/95	11/17/95	11/16/95	11/17/95
SD/LEP questionnaire delivered to NCS	11/22/95	11/22/95	12/5/95	12/5/95
Grade 4 science spiral material at NCS	11/22/95	11/30/95	11/21/95	12/1/95

(continued)

Table 6-4 (continued)
NCS Schedule for the 1996 NAEP Assessment

Task	Planned Start	Planned Finish	Actual Start	Actual Finish
MAIN ASSESSMENT (continued)				
All materials at NCS for packaging	11/29/95	12/1/95	12/1/95	12/15/95
Grade 12 science spiral material at NCS	12/1/95	12/11/95	12/1/95	12/4/95
NCS receive 95% session data from Westat	12/4/95	12/4/95	12/6/95	12/6/95
Westat training for main supervisors	12/9/95	12/13/95	12/9/95	12/13/95
Westat send Administration Schedule home address file	12/11/95	12/11/95	12/11/95	12/11/95
Westat send NCS Wave 1 address file	12/11/95	12/11/95	12/11/95	12/11/95
Print Administration Schedule	12/11/95	12/13/95	12/13/95	12/13/95
WAVE 1 packing list and labels to packaging	12/13/95	12/13/95	12/14/95	12/14/95
Format/content Interrater Agreement Report	12/15/95	12/15/95	11/1/95	12/15/95
PSC obtains sample booklets	12/15/95	12/15/95	11/2/95	11/13/95
Purpose/use of T-Test and bridge reliability	12/15/95	12/15/95	11/2/95	11/3/95
PSC submit requisition for mathematics and science scorers	12/15/95	12/15/95	1/6/96	1/16/96
PSC submit requisition for mathematics and science team leaders	12/15/95	12/15/95	1/16/96	1/16/96
PSC submit requisition for mathematics and science table leaders	12/15/95	12/15/95	1/30/96	1/30/96
Ship Administration Schedule, teacher questionnaire, SD/LEP questionnaires, rosters	12/15/95	12/18/95	12/14/95	12/14/95
Ship bulk/Wave 1 material	12/20/95	12/20/95	12/15/95	12/18/95
Bulk/Wave 1 materials due to supervisors	12/26/95	12/26/95	12/26/95	12/29/95
Receiving	1/5/96	4/8/96	1/5/96	4/8/96
Test administration	1/3/96	4/5/96	1/3/96	4/5/96
Blue dot Grades 4, 8 and 12 mathematics	1/4/96	1/7/96	1/15/96	1/25/96
Blue dot Grades 4, 8, and 12 science	1/4/96	1/7/96	1/15/96	1/25/96
Blue dot advanced science	1/4/96	1/7/96	1/15/96	1/26/96
Blue dot Grade 12 estimation, theme, and advanced mathematics	1/4/96	1/7/96	1/15/96	1/31/96
Processing	1/8/96	4/14/96	1/25/96	4/29/96
Blue dot SD/LEP questionnaires	1/9/96	1/11/96	2/20/96	2/27/96
Assignments of mathematics items to teams	1/15/96	1/15/96	1/2/96	2/23/96
Scoring calendar item-by-item for mathematics	1/15/96	1/15/96	1/2/96	2/15/96
Wave 2 addresses from Westat	1/17/96	1/17/96	1/17/96	1/17/96
Packaging Wave 2 materials (2 shifts)	1/18/96	1/26/96	1/19/96	1/23/96
Wave 2 packing and mailing labels to packaging	1/17/96	1/19/96	1/19/96	1/19/96
Blue dot school characteristics and policies questionnaires and teacher questionnaires	1/22/96	1/25/96	2/9/96	2/26/96
Blue dot short-term rescore	1/25/96	1/25/96	3/5/96	3/11/96
Assignment of science items to teams	2/1/96	2/1/96	1/15/96	5/31/96

(continued)

Table 6-4 (continued)
NCS Schedule for the 1996 NAEP Assessment

Task	Planned Start	Planned Finish	Actual Start	Actual Finish
MAIN ASSESSMENT (continued)				
Mathematics and science table leaders hired	2/1/96	2/1/96	1/15/96	2/28/96
Plan for staff range finding	2/1/96	2/1/96	1/15/96	2/1/96
PSC selects science team leaders	2/1/96	2/1/96	1/15/96	1/16/96
PSC selects mathematics team leaders	2/1/96	2/1/96	1/16/96	1/16/96
Blue dot rosters (all types)	2/1/96	2/1/96	1/26/96	2/12/96
Wave 2 materials due to supervisors	2/2/96	2/2/96	2/1/96	2/2/96
Scoring calendar item by item (science)	2/3/96	2/6/96	1/15/96	5/31/96
Wave 3 addresses from Westat	2/7/96	2/7/96	2/7/96	2/7/96
Segment 1 day 25% scorers hired	2/12/96	2/12/96	2/7/96	2/7/96
Wave 3 packing list/mailing labels to packaging	2/12/96	2/12/96	2/9/96	2/13/96
Packaging/ship Wave 3 materials	2/12/96	2/19/96	2/19/96	2/20/96
Pre-range finding paper selection – mathematics	2/12/96	3/8/96	2/5/96	3/8/96
Pre-range finding paper selection – science	2/12/96	3/15/96	2/5/96	3/8/96
Segment 1 day 50% scorers hired	2/19/96	2/19/96	2/9/96	2/9/96
Segment 1 day 75% scorers hired	2/26/96	2/26/96	2/15/96	2/15/96
Wave 3 materials due in supervisors	2/26/96	2/26/96	2/26/96	2/27/96
PSC selects mathematics table leaders	3/1/96	3/1/96	2/1/96	2/28/96
PSC selects science table leaders	3/1/96	3/1/96	2/1/96	2/28/96
Segment 2 day 25% scorers hired	3/1/96	3/1/96	2/8/96	2/8/96
Segment 3 day 25% scorers hired	3/1/96	3/1/96	2/8/96	2/8/96
Segment 3 evening 25% scorers hired	3/1/96	3/1/96	2/20/96	2/20/96
Segment 2 evening 25% scorers hired	3/1/96	3/1/96	2/22/96	2/22/96
Segment 1 day 100% scorers hired	3/4/96	3/4/96	3/6/96	3/6/96
Table leaders for mathematics hired	3/8/96	3/8/96	2/1/96	3/22/96
Segment 3 day 50% scorers hired	3/8/96	3/8/96	2/14/96	2/14/96
Segment 2 day 50% scorers hired	3/8/96	3/8/96	2/15/96	2/15/96
Segment 2 evening 50% scorers hired	3/8/96	3/8/96	3/6/96	3/6/96
Segment 3 evening 50% scorers hired	3/8/96	3/8/96	3/6/96	3/6/96
Mathematics and science scorers assigned to teams	3/11/96	3/11/96	3/11/96	3/11/96
Train/score mathematics - Segment 1(day shift)	3/13/96	4/3/96	3/13/96	4/5/96
Segment 2 day 75% scorers hired	3/15/96	3/15/96	3/5/96	3/14/96
Segment 3 day 75% scorers hired	3/15/96	3/15/96	3/6/96	3/6/96
Segment 2 evening 75% scorers hired	3/15/96	3/15/96	3/14/96	3/14/96
Segment 3 evening 75% scorers hired	3/15/96	3/15/96	3/14/96	3/14/96
Train/score science - Segment 1 (day shift)	3/18/96	4/5/96	3/18/96	4/5/96
Segment 2 evening 100% scorers hired	3/22/96	3/22/96	3/25/96	3/25/96
Segment 3 evening 100% scorers hired	3/22/96	3/22/96	3/25/96	3/25/96
Segment 2 day 100% scorers hired	3/22/96	3/22/96	4/1/96	4/1/96
Segment 3 day 100% scorers hired	3/22/96	3/22/96	4/1/96	4/1/96
Table leaders for science hired	4/5/96	4/5/96	2/12/96	2/28/96

(continued)

Table 6-4 (continued)
NCS Schedule for the 1996 NAEP Assessment

Task	Planned Start	Planned Finish	Actual Start	Actual Finish
MAIN ASSESSMENT (continued)				
Train/score mathematics - Segment 2 (evening shift)	4/8/96	5/2/96	4/8/96	5/2/96
Train/score science - Segment 2 (evening shift)	4/8/96	5/2/96	4/8/96	5/2/96
Train/score mathematics - Segment 2 (day shift)	4/8/96	5/3/96	4/8/96	5/6/96
Train/score science - Segment 2 (day shift)	4/8/96	5/3/96	4/8/96	5/3/96
Mathematics through clean post	4/14/96	4/14/96	4/14/96	4/29/96
Science through clean post edit	4/28/96	4/28/96	4/26/96	4/26/96
Bilingual mathematics scoring	5/2/96	5/3/96	5/2/96	5/3/96
Grade 8 mathematics weights	5/4/96	5/6/96	5/3/96	5/3/96
Grade 4 mathematics weights	5/4/96	5/6/96	5/9/96	5/9/96
Grade 4 science data tape sent to ETS	5/31/96	5/31/96	5/30/96	5/30/96
School characteristics and policies questionnaires data tape shipped to ETS	7/11/96	7/12/96	7/11/96	7/11/96
Teacher questionnaire data tape shipped to ETS	7/18/96	7/19/96	7/19/96	7/24/96
SD/LEP questionnaire data shipped to ETS	7/26/96	7/29/96	8/2/96	8/2/96
Grade 12 science data tape sent to ETS	6/7/96	6/7/96	6/4/96	6/4/96
Grade 12 advanced science data tape sent to ETS	6/7/96	6/7/96	6/10/96	6/10/96
Grade 8 science data tape sent to ETS	5/31/96	5/31/96	6/26/96	6/26/96

6.1.1 Innovations for 1996

Much of the information necessary for documentation of accurate sampling and for calculating sampling weights is collected on the Administration Schedules that, until 1993, were painstakingly filled out by hand by Westat administrative personnel. In 1994, for the first time, much of the work was computerized—booklets were preassigned and booklet ID numbers were preprinted on the Administration Schedule. When Westat personnel received the documents, they filled in only the "exception" information. This new method also permitted computerized updating of information when the Administration Schedules were received at NCS, eliminating the need to sort and track thousands of pieces of paper through the processing stream.

The introduction of image processing and image scoring further enhanced the work of NAEP. Image processing and scoring were successfully piloted in a side-by-side study conducted during the 1993 NAEP field test, and so became the primary processing and scoring methods for the 1994 and 1996 assessments. Image processing allowed the automatic collection of handwritten demographic data from the administrative schedules and the student test booklet covers through intelligent character recognition (ICR). This service was a benefit to the jurisdictions participating in NAEP because they were able to write rather than grid certain information—a reduction of burden on the schools. Image processing also made image scoring possible, eliminating much of the time spent moving paper as part of the scoring process. The images of student responses to be scored were transmitted electronically to the scoring center, located at a separate facility from where the materials were processed. This process enhanced the reliability and monitoring of scoring and allowed both NCS and ETS to focus attention on the intellectual process of scoring student responses.

6.2 PRINTING

For the 1996 assessment, 255 unique documents were designed. NCS printed more than 1,900,000 booklets and forms, totaling over 58 million pages. Printing preparations began with the design of the booklet covers in June 1995. This was a collaborative effort involving staff from ETS, Westat and NCS. Since the goal was to design one format for use with all of the booklets, necessary data elements to be collected for the different assessment types had to be agreed upon. After various iterations, the cover design was finalized.

In a similar collaboration with ETS and Westat, NCS prepared administration schedules and questionnaire rosters. The camera-ready copies for these documents were created and edited using NCS Design Expert™ software.

Printing of the NAEP documents began with the documents for the long-term trend assessments. These included the 26 long-term trend assessment booklets, the Administration Schedule, the excluded student questionnaire, three school characteristics and policies questionnaires, and the roster of questionnaires. All materials for the long-term trend assessments were printed by September 19, 1995. The printing of assessment booklets, questionnaires and tracking forms for the main and state assessments followed. Printing of these documents was complete by December 11, 1995.

Details of the printing procedures are given in the *Report of Processing and Professional Scoring Activities* (National Computer Systems, 1996).

6.3 PACKAGING AND SHIPPING

The distribution effort for the 1996 NAEP assessment involved packaging and mailing documents and associated forms and materials to the Westat supervisors for the main and long-term trend assessments. The NAEP Materials Distribution System (MDS), initially developed by NCS in 1990 to control shipments to the schools and supervisors, was utilized again in 1996. Files in the MDS system contained the names and addresses for shipment of materials, scheduled assessment dates, and a listing of all materials available for use by a participant in a particular subject area. Changes to any of this information were made directly in the MDS file either manually or via file updates provided by Westat. Details of the accountability system and on-line bundle assignment and distribution system utilized for NAEP are given in the *Report of Processing and Professional Scoring Activities*.

The bar code technology introduced by NCS in the 1990 assessments continued to be utilized in document control. To identify each document, NCS utilized a unique ten-digit numbering system that consisted of the three-digit booklet number or form type, a six-digit sequential number, and a check digit. Each form was assigned a range of ID numbers. Bar codes reflecting this ID number were applied to the front cover of each document by NCS bar code processes and high-speed ink jet printers.

Once all booklets from a subject area were bar coded, they were spiraled and bundled into groups of 11 documents. For main samples in mathematics and science (done concurrently with the State Assessment samples in mathematics and science), NCS spiraled the booklets according to the pattern dictated by ETS in the bundle maps. Booklets were spiraled in such a manner that each booklet appeared in the first position in a bundle approximately the same number of times and the booklets were evenly distributed across the bundles. This assured that sample sizes of individual booklet types would not be jeopardized if entire bundles were not used. Since the mathematics and science estimation and advanced

booklet bundles contained only one booklet type, these were bundled into groups of 11. The mathematics bilingual booklets were bundled in groups of three.

Initially 5,161 individual sessions were shipped for the 1996 NAEP assessments. Approximately 600 additional shipments of booklets and miscellaneous materials were sent. All outbound shipments were recorded in the NCS Outbound Mail Management system. This was accomplished by having a bar code containing the school number on each address label. This bar code was read into the system, which determined the routing of the shipment and the charges. Information was recorded in a file on the system which, at the end of each day, was transferred by a PC upload to the mainframe. A computer program could then access information to produce reports on all shipments sent, regardless of the carrier used. These reports helped NCS phone staff trace shipments for Westat supervisors and assessment administrators.

A toll-free telephone line was maintained for supervisors and school administrators to request additional materials for the National assessments. To process a shipment, NCS phone staff asked the caller for information such as primary sampling unit (PSU), school ID, assessment type, city, state, and ZIP code. This information was then entered into the on-line short shipment system and the school's mailing address would be displayed on the screen to verify with the caller. The system allowed NCS staff to change the shipping address for individual requests. The clerk proceeded to the next screen that displayed the materials to be selected. After the requested items, due date and method of shipment were entered, the system produced a packing list and mailing labels. Phone staff also took phone calls concerning initial shipment delivery dates, tracing a shipment, and questions concerning NAEP. Approximately 750 calls were received regarding the 1996 NAEP assessments.

Further information regarding packaging and shipping is provided in the *Report of Processing and Professional Scoring Activities* (National Computer Systems, 1996).

6.4 PROCESSING

6.4.1 Overview

The following describes the various stages of work involved in receiving and processing the documents used in the 1996 assessment. NCS staff created a set of predetermined rules and specifications for the processing departments within NCS to follow. Project staff performed a variety of procedures on materials received from the assessment administrators before releasing these materials into the NCS NAEP processing system. Control systems were used to monitor all NAEP materials returned from the field. The NAEP Process Control System (PCS) contained the status of sampled schools for all sessions and their scheduled assessment dates. As materials were returned, the PCS was updated to indicate receipt dates, to record counts of materials returned, and to document any problems discovered in the shipments. As documents were processed, the system was updated to reflect processed counts. NCS report programs were utilized to allow ETS, Westat, and NCS staff to monitor the progress in the receipt control operations.

An "alert" process was used to record, monitor, and categorize all discrepant or problematic situations. Throughout the processing cycle, alert situations were either flagged by computer programs or identified during clerical check-in procedures. Certain alerts, such as missing demographic information on the administration schedule, were resolved by opening staff retrieving the information from booklet covers. These alerts, known as "Information Alerts," were recorded directly into the PCS system by opening personnel, eliminating the need for paper documentation. Since these problem situations were

categorized and tallied as they were key-entered into the PCS system, project staff were able to provide timely reporting on clerical-type errors made during test administration. Alert situations that could not be resolved by opening personnel were described on alert forms that were forwarded to project personnel for resolution. Once resolved, the problems and resolutions were recorded online in the PCS system.

NCS's Work Flow Management System was used to track batches of student booklets through each processing step, allowing project staff to monitor the status of all work in progress. It was also used by NCS to analyze the current work load, by project, across all work stations. By routinely monitoring these data, NCS's management staff was able to assign priorities to various components of the work and to monitor all phases of the data receipt and processing.

6.4.2 Document Receipt

Shipments were to be returned to NCS packaged in their original boxes. As mentioned earlier, NCS packaging staff applied a bar code label to each box indicating the NAEP school ID number. When a shipment arrived at the NCS dock area, this bar code was scanned into a personal computer file, and the shipment was forwarded to the receiving area. The file was then transferred to the mainframe and the shipment receipt date was applied to the appropriate school within the PCS system, providing the status of receipts regardless of any processing delays. Each receipt was reflected on the PCS status report provided to the NCS receiving department and supplied to Westat via electronic file transfer and in hard-copy format. ETS also received a hard copy.

The PCS file could be manually updated to reflect changes. Receiving personnel also checked the shipment to verify that the contents of the box matched the school and session indicated on the label. Each shipment was checked for completeness and accuracy. Any shipment not received within two days of the scheduled assessment date was flagged in the PCS system and annotated on the PCS report. The administration status of these delayed shipments was checked and in some cases a trace was initiated on the shipment.

A new requirement for NCS was to open all shipments within forty-eight hours of their receipt and to key-enter preliminary processing information into the PCS system from the Administration Schedule. The preliminary information was written on the Administration Schedule by Westat assessment administrators and consisted of the following:

- School number
- Session number
- Original test date
- Total number assessed

This preliminary information, used to provide Westat with timely student response rates, was updated with actual data when materials passed through processing error free. A completeness flag was also applied to the PCS file by NCS opening staff if any part of the shipment was missing.

If multiple sessions were returned in one box, the contents of the package were separated by session. The shipment was checked to verify that all booklets preprinted or handwritten on the Administration Schedule were returned with the shipment and that all administration codes matched from booklet cover to the Administration Schedule. If discrepancies were discovered at any step in this process, the receiving staff issued an alert to facilitate tracking. If the administrator indicated that a make-up session was being held the documents were placed on holding carts until the make-up session

documents arrived. If no make-up session was indicated, Westat was contacted for the status of the missing materials. If the missing materials were to be returned, the documents already received were held until that time. If the materials were not being returned, processing continued and the appropriate administration code was applied to the Administration Schedule.

6.4.3 Batching and Scanning Documents

Once all booklets listed on the Administration Schedule for a session were verified as present, the entire session (both the Administration Schedule and booklets) was batched by grade level and session type. Each batch was assigned a unique batch number. This number, created on the Image Capture Environment system for all image-scannable documents and on the Work Flow Management system for all key-entry and OMR-scannable documents[2], facilitated the internal tracking of the batches and allowed departmental resource planning. All other scannable documents (School Characteristics and Policies Questionnaires, Teacher Questionnaires, Students with Disabilities/Limited English Proficiency [SD/LEP] Questionnaires, and rosters) were batched by document type in the same manner.

Because all assessment booklets were image-scannable, batch numbers for these documents were created on the Image Capture Environment system. Sessions were sorted by grade level and automatically uploaded to the Work Flow Management system after batch creation. The Administration Schedule for these document types was used as a session header within a batch.

When batching mathematics documents, NCS needed to allow for having both image-scannable and key-entry documents present in the same session, or having booklets listed on the Administration Schedule that would not be present in processing. This was due to the testing accommodations of large-print and Braille that were key-entry documents.

Large-print booklets had to be processed separately from the Administration Schedule and scannable booklets in their session. A key-entry session header was created for these booklets that contained the school ID number and session code from the Administration Schedule. Long-term trend reading/writing booklets were processed through key-entry with the same type of key-entry session header. The Administration Schedules from reading/writing sessions were processed in an Administration-Schedule-only batch through the image scanning system. After the session that a large-print booklet came with passed through image processing, session information was rejoined within the processing computer programs. The same computerized match occurred with Trend reading/writing materials once the Administration-Schedule-Only batch that contained a session's administration schedule passed through processing.

6.4.4 Questionnaires

The long-term trend assessments used one roster to account for all questionnaires. The roster of questionnaires recorded the distribution and return of the School Characteristics and Policies Questionnaires and the Excluded Student Questionnaires.

The main assessments utilized one roster to document and track the School Characteristics and Policies Questionnaires and the Students with Disabilities/Limited English Proficiency (SD/LEP) Questionnaire. In addition, the main and state assessments used the roster of Teacher Questionnaires to record the distribution and return of Teacher Questionnaires.

[2] OMR is the acronym for Optical Mark Reading.

Some questionnaires may not have been available for return with the shipment. These were returned to NCS at a later date in an envelope provided for that purpose. The questionnaires were submitted for scanning as sufficient quantities became available for batching.

Receipt of the questionnaires was entered into the system using the same process as was used for the Administration Schedules described in previous sections. The rosters were grouped with other rosters of the same type from other sessions, and a batch was created on the Image Capture Environment system. The batch was then forwarded to scanning where all information on the rosters was scanned into the system.

6.4.5 Booklet Accountability

NCS used a sophisticated booklet accountability system to track all distributed booklets. Prior to the distribution of NAEP materials, unique booklet numbers were read by bundle into a file. Specific bundles were then assigned to particular supervisors or schools. This assignment was recorded in the NAEP Materials Distribution System.

When shipments arrived at NCS from the field, all used booklets were submitted for processing and a "processed documents" file was maintained. Unused booklets were submitted for security scanning where booklet ID bar codes were read and recorded into a separate file. This file and the "processed documents" file were later compared to the original bundle security file for individual booklet matching. A list of unmatched booklet IDs was printed in a report used to confirm non-receipt of individual booklets. At the end of the assessment period, supervisors returned all unused materials. These booklet IDs were also read by the bar-code scanner and added to the bundle security file. All unused materials received were then inventoried and sent to the NCS warehouse for storage while awaiting authorization from ETS to salvage them.

6.4.6 Data Entry

The data entry process was the first point at which booklet-level data were directly available to the computer system. Depending on the NAEP document, one of three methods was used to transcribe NAEP data to a computerized form. The data on scannable documents were collected using NCS optical-scanning equipment that also captured images of the constructed-response items and ICR fields. Nonscannable materials were keyed through an interactive online system. In both of these cases, the data were edited and suspect cases were resolved before further processing.

All student booklets, questionnaires, and control documents were scannable. Throughout all phases of processing, the student booklets were batched by grade and session type. The scannable documents were then transported to a slitting area where the folded and stapled spine was removed from the document. This process utilized an "intelligent slitter" to prevent slitting the wrong side of the document. The documents were jogged by machine so that the registration edges of the NAEP documents were smoothly aligned, and the stacks were then returned to the cart to be scanned.

During the scanning process, each scannable NAEP document was uniquely identified using a print-after-scan number consisting of the scan batch number, the sequential number within the batch, and the bar code ID of the booklet. These numbers were printed on each sheet of each document as it exited the scanner. This permitted the data editors to quickly and accurately locate specific documents during

the editing phase. The print-after-scan number remained with the data record, providing a method for easy identification and quick retrieval of any document.

The data values were captured from the booklet covers and Administration Schedules and were coded as numeric data. Unmarked fields were coded as blanks and editing staff were alerted to missing or uncoded critical data. Fields that had multiple marks were coded as asterisks (*). The data values for the item responses and scores were returned as numeric codes. The multiple-choice single response format items were assigned codes depending on the position of the response alternative; that is, the first choice was assigned the code "1," the second "2," and so forth. The mark-all-that-apply items were given as many data fields as response alternatives; the marked choices were coded as "1" while the unmarked choices were recorded as blanks. The images of constructed-response items were saved as a digitized computer file. The area of the page that needed to be clipped was defined prior to scanning through the document definition process. The fields from unreadable pages were coded "X" as a flag for resolution staff to correct. In addition to capturing the student responses, the bar code identification numbers used to maintain process control were decoded and transcribed to the NAEP computerized data file.

As the scanning program completed scanning each stack, the stack was removed from the output hopper and placed in the same order they were scanned on the output cart. The next stack was removed from the input cart and placed into the input hopper, after which the scanning resumed. When the operator had completed processing the last stack of the batch, the program was terminated. This closed the dataset that automatically became available for the data validation (edit) process. The scanned documents were then forwarded to a holding area in case they needed to be retrieved for resolution of edit errors.

NCS again used the ICR engine to read various hand and machine printing on the front cover of the assessment and supervisor documents for the 1996 NAEP assessments. Some information from scannable student documents, such as the Administration Schedule, the roster of questionnaires, and some questions in the School Characteristics and Policies Questionnaires, were read by the ICR engine and verified by an online key-entry operator. In all, the ICR engine read approximately 15 million characters. The ICR engine saved NAEP field staff and school personnel a significant amount of time because they no longer had to enter these data by gridding rows and columns of data.

NCS also implemented new programs that allowed the scanners to read imprinted codes, known as 2-out-of-5 codes, that were printed via a Xerox 4280 printer on the Administration Schedule. These 2-out-of-5 codes were imprinted at the same time the booklet ID numbers were printed on the Administration Schedule and identified which booklet IDs were listed on that document. When the scanning programs were unable to translate the 2-out-of-5 codes (thereby identifying the booklet ID numbers on the document) image clips of the booklet ID numbers were displayed to online editing staff for verification. This eliminated a significant amount of online editing time needed to process the NAEP assessments.

To provide another quality check on the image scanning and scoring system, NCS staff stamped blank booklets with a rubber stamp and assigned these booklets mock scores from the valid range. Each unique item type scored via the image system had two quality control stamps per valid score. An example of the stamp used is given below.

IMAGE SCORING
QUALITY ASSURANCE
SAMPLE

The quality control booklets were batched and processed together with student documents of the same type. Because all of a specific item were batched together for transmission to the scoring facility, the quality control-stamped responses were integrated with the student responses and transmitted simultaneously to the scoring facility. During the scoring process, both student responses and the quality control items were randomly displayed so scores could be applied.

When a person who was scoring responses (reader) later saw the quality control sample on the monitor during scoring, he or she was to notify the team leader, who confirmed the score assigned by the reader was the score listed on the sample. The quality control booklets were included in the pool of all items to be drawn from for the 25 percent reliability rescore.

All image quality-assurance documents were created prior to the beginning of scoring and all pre-determined score points were used. Because during the process of scoring, valid score points can be changed or dropped completely, NCS provided ETS with documentation explaining what quality control documents were produced and which score points on these items were no longer valid. When an image quality control stamp was displayed to a reader that contained a score point that was no longer valid, the reader gave the response a score point of zero.

A process of key entry and verification was used to make corrections to the non-scannable long-term trend reading/writing documents and large print booklets. Teacher questionnaire and SD/LEP questionnaire information was also corrected using key-entry methods. NCS used the Falcon system to enter this data. The Falcon system is an on-line data-entry system designed to replace most methods of data input such as keypunch, key-to-disk, and many of the microcomputer data-entry systems. The terminal screens were designed to enhance operator speed and convenience. The fields to be entered were titled to reflect the actual source document. Therefore, all key-entry fields were specific to the NAEP student documents or questionnaire types being keyed.

6.4.7 Data Validation

Each dataset produced by the scanning system contained data for a particular batch. These data had to be validated (or edited) for type and range of response. The data-entry and resolution system used was able to simultaneously process a variety of materials from all age groups, subject areas, control documents, and questionnaires as the materials were submitted to the system from scannable and non-scannable media.

The data records in the scan file were organized in the same order in which the paper materials were processed by the scanner. A record for each batch header preceded all data records for that batch. The document code field on each record distinguished the header record from the data records.

When a batch-header record was read, a pre-edit data file and an edit log were generated. As the program processed each record within a batch from the scan file, it wrote the edited and reformatted data records to the pre-edit file and recorded all errors on the edit log. The data fields on an edit log record identified each data problem by the batch sequence number, booklet serial number, section or block code, field name or item number, and data value. After each batch had been processed, the program generated a listing or online edit file of the data problems and resolution guidelines. An edit log listing was printed at the termination of the program for all non-image documents. Image “clips” requiring editing were routed to online editing stations for those documents that were image scanned.

As the program processed each data record, it first read the booklet number and checked it against the session code for appropriate session type. Any mismatch was recorded on the error log and processing

continued. The booklet number was then compared against the first three digits of the student identification number. If they did not match, a message was written on the error log. The remaining booklet cover fields were read and validated for the correct range of values. The school codes had to be identical to those on the PCS record. All data values that were out of range were read “as is” but were flagged as suspect. All data fields that were read as asterisks (*) were recorded on the edit log or online edit file.

Document definition files described each document as a series of blocks that in turn were described as a series of items. The blocks in a document were transcribed in the order that they appeared in the document. Each block’s fields were validated during this process. If a document contained suspect fields, the cover information was recorded on the edit log along with a description of the suspect data. The edited booklet cover was transferred to an output buffer area within the program. As the program processed each block of data from the dataset record, it appended the edited data fields to the data already in this buffer.

The program then cycled through the data area corresponding to the item blocks. The task of translating, validating, and reporting errors for each data field in each block was performed by a routine that required only the block identification code and the string of input data. This routine had access to a block definition file that had, for each block, the number of fields to be processed, and, for each field, the field type (alphabetic or numeric), the field width in the data record, and the valid range of values. The routine then processed each field in sequence order, performing the necessary translation, validation, and reporting tasks.

The first of these tasks checked for the presence of blanks or asterisks (*) in a critical field. These were recorded on the edit log or online edit file and processing continued with the next field. No action was taken on blank fields for multiple-choice items because the asterisk code indicated a non-response. The field was validated for range of response, and any values outside of the specified range were recorded on the edit log or online edit file. The program used the item-type code to make a further distinction among constructed-response item scores and other numeric data fields.

Moving the translated and edited data field into the output buffer was the last task performed in this phase of processing. When the entire document was processed, the completed string of data was written to the data file. When the program encountered the end of a file, it closed the dataset and generated an edit listing for non-image and key-entered documents. Image-scanned items that required correction were displayed at an online editing terminal.

6.4.8 Editing for Non-Image and Key-Entered Documents

Throughout the system, quality procedures and software ensured that the NAEP data were correct. All student documents on the Administration Schedule were accounted for, as receipt control personnel checked that the materials were undamaged and assembled correctly. The machine edits performed during data capture verified that each sheet of each document was present and that each field had an appropriate value. All batches entered into the system, whether key-entered or machine-scanned, were edited for errors.

Data editing took place after these checks. This consisted of a computerized edit review of each respondent’s document and the clerical edits necessary to make corrections based upon the computer edit. This data-editing step was repeated until all data were correct.

The first phase of data editing was designed to validate the population and ensure that all documents were present. A computerized edit list, produced after NAEP documents were scanned or key entered, and all the supporting documentation sent from the field were used to perform the edit function. The hard-copy edit list contained all the vital statistics about the batch: number of students, school code, type of document, assessment code, suspect cases, and record serial numbers. Using these inputs, the data editor verified that the batch had been assembled correctly and that each school number was correct.

During data entry, counts of processed documents were generated by type. These counts were compared against the information captured from the Administration Schedules. The number of assessed and absent students processed had to match the numbers indicated on the PCS.

In the second phase of data editing, experienced editing staff used a predetermined set of specifications to review the field errors and record necessary corrections to the student data file. The same computerized edit list used in phase one was used to perform this function. The editing staff reviewed the computer-generated edit log and the area of the source document that was noted as being suspect or as containing possible errors. The composition of the field was shown in the edit box. The editing staff checked this piece of information against the NAEP source document. At that point, one of the following took place:

Correctable error. If the error was correctable by the editing staff as per the editing specifications, the correction was noted on the edit log for later correction via key-entry.

Alert. If an error was not correctable as per the specifications, an alert was issued to NAEP project staff for resolution. Once the correct information was obtained, the correction was noted on the edit log for key-entry correction.

Non-correctable error. If a suspected error was found to be correct as stated and no alteration was possible according to the source document and specifications, the programs were tailored to allow this information to be accepted into the data record. No corrective action was taken.

The corrected edit log was then forwarded to the key-entry staff for processing. When all corrections were entered and verified for a batch, an extract program pulled the corrected records into a mainframe dataset. At this point, the mainframe edit program was initiated. The edit criteria were again applied to all records. If there were further errors, a new edit listing was printed and the cycle was repeated.

When the edit process produced an error-free file, the booklet ID number was posted to the NAEP tracking file by age, assessment, and school. This permitted NCS staff to monitor the NAEP processing effort by accurately measuring the number of documents processed by form. The posting of booklet IDs also ensured that a booklet ID was not processed more than once.

6.4.9 Data Validation and Editing of Image-Processed Documents

The paper edit log for key-entered documents was replaced by online viewing of suspect data for all image-processed documents. For rapid resolution, the edit criteria for each item in question appeared on the screen along with the suspect item. Corrections were made immediately. The system employed an edit/verify system that ultimately meant that two different people viewed the same suspect data and operated on it separately. The “verifier” made sure the two responses (one from either the entry operator or the ICR engine) were the same before the system accepted that item as being correct. The verifiers

could either overrule or agree with the original correction made if the two did not match. If the editor could not determine the appropriate response, he or she escalated the suspect situation to a supervisor. For errors or suspect information that could not be resolved by supervisory staff, a product-line queue was created. This allowed supervisors to escalate edits to project staff for resolution. By having this product-line queue, project staff were able to quickly locate edit clips within the image system, speeding up the resolution process.

Once an entire batch was through the edit phase, it became eligible for the count-verification phase. The Administration Schedule data were examined systematically for booklet IDs that should have been processed (assessed administration codes). All documents under that Administration Schedule were then inspected to ensure that all of the booklets were included.

With the satisfactory conclusion of the count-verification phase, the edited batch file was uploaded to the mainframe, where it went through yet another edit process. A paper edit log was produced and, if errors remained, was forwarded to another editor. When this paper edit was satisfied, the PCS and Workflow Management system were updated. Because there was a possible time lag between a clean edit in the image system and a clean edit in the mainframe systems, the batch was not archived until 48 hours after the image edit phase was completed.

6.4.10 Data Transmission

Due to the rapid pace of scoring on an item-by-item basis, the NCS scoring specialists found it necessary to continually monitor the status of work available to the readers and plan the scoring schedule several weeks in advance. On Wednesday of each week, the NCS performance assessment specialist in charge of each subject area planned the next two weeks' schedule. That information was then provided to the person in charge of downloading data to the scoring center. By planning the scoring schedule two weeks in advance, the scoring specialists were able to ensure that readers would have sufficient work for at least one week, after which the next download would occur to supplement the volume of any unscored items and add an additional week's work to the pool of items to score. Additionally, by scheduling two weeks' data transmission, flexibility was added to the scoring schedule, making it possible to implement last-minute changes in the schedule once the items had been delivered to the scoring center. Depending on the number of items to be transmitted, the actual downloading was conducted on Friday or was divided into two smaller sessions for Thursday and Friday download. By the first week of May 1996, there was sufficient space on the scoring servers to load all remaining unscored items to the scoring center.

Delivery of data to the scoring center was accomplished via several T1 transmission lines linking the mainframe computers and the NAEP servers at the document-scanning site in the NCS main facility with the scoring servers dedicated to distributing work to the professional readers at the scoring center. The actual task of scheduling items for downloading was accomplished using a code written by the Image Software Development team. This code enabled the person scheduling the download to choose a team of readers and select the scheduled items from a list of all items that that team would be scoring throughout the scoring project. This process was repeated for all teams of readers until all anticipated work was scheduled. Once this task was completed, the scheduled job was tested to determine if there was sufficient free disk space on the servers at the scoring center. If for any reason sufficient disk space was not available, scheduled items could be deleted from the batch individually or as a group until the scheduled batch job could accommodate all items on the available disk space at the scoring center. Once it was determined that sufficient disk space was available, transmission of student responses commenced. Data transmission was typically accomplished during off-shift hours to minimize the impact on system-load capacity.

6.5 DATA DELIVERY

The 1996 NAEP data collection resulted in several classes of data files — student, school, teacher, weights, SD/LEP student, excluded student questionnaire data for long-term trends, student/teacher match, and student-response information. Student-response information included response data from all assessed students in 1996. Data resolution activities occurred prior to the submission of data files to ETS and Westat to resolve any irregularities that existed. This section details additional steps performed before creating the final data files to ensure capture of the most complete and accurate information.

An important quality-control component of the image-scoring system was the inclusion, for purposes of file identification, of an exact copy of the student edit record, including the student booklet ID number, with every image of a student's response to a constructed-response item. These edit files also remained in the main data files residing on the NCS mainframe computer. By doing this, exact matching of scores assigned to constructed-response items and all other data for each individual student was guaranteed, since the booklet ID for each image was part of every image file. This ensured scores were applied to the correct student's record on the mainframe.

When all of the responses for an individual item had been scored, the system automatically submitted all item scores assigned during scoring, along with their edit records, to a queue to be transmitted to the mainframe. Project staff then initiated a system job to transmit all scoring data to be matched with the original student records on the mainframe. A custom edit program matched the edit records of the scoring files to those of the original edit records on the mainframe. As matches were confirmed, the scores were applied to those individual files. After completion of this stage, all data collected for an individual student was located in one single and complete record/file identified by the edit record.

NCS processed the SD/LEP student questionnaires via OMR scanning. Edits performed on the questionnaires assured that responses to questions fell within the valid range for that question. SD/LEP questionnaires were then matched to a student record. SD/LEP Questionnaires that were not matched to a student document were cross-referenced with the corresponding Administration Schedule, roster of questionnaires, and student data files to correct, if necessary, the information needed to result in a match.

In 1996, NCS continued to use ICR technology to capture percentage figures written by school personnel directly in boxes on the school characteristics and policies questionnaires rather than requiring the school official to grid ovals in a matrix. The data were then verified by an edit operator.

To achieve the best possible student/teacher match rate, the same processes that were followed in previous cycles were used in 1996. The first step was to identify teacher questionnaires not returned to NCS for processing so as to exclude from the matching process the students of these teachers. Student identification numbers that were not matched to a teacher questionnaire were cross-referenced with the corresponding Administration Schedule and roster of teacher questionnaires to verify (and change, if necessary) the teacher number, teacher period, and questionnaire number recorded on these control documents. The NAEP school numbers listed on the roster of questionnaires and teacher questionnaire were verified and corrected, if necessary. Once these changes were made, any duplicate teacher numbers existing within a school were, if possible, cross-referenced for resolution with the roster of questionnaires. Since this information was located together on a single, central control document, the ability to match and resolve discrepant or missing fields was simplified.

After all data-processing activities were completed, data cartridges and/or diskettes were created and shipped via overnight delivery to ETS and/or Westat. A duplicate archive file is maintained at NCS for security and backup purposes.

6.6 MISCELLANEOUS

6.6.1 Storage of Documents

After batches of image-scanned documents had successfully passed the editing process, they were sent to the NCS warehouse for storage. The long-term trend reading/writing booklets were sent to the NCS scoring center to be scored on paper and to be used for holistic and mechanics scoring. After all scoring had been completed for the long-term trend booklets, they were also sent to the warehouse for storage. Due to the large number of rescore projects done with NAEP material, the documents were unspiraled and sequenced by grade and booklet type after all of the processing/scoring was completed. This will allow for efficient document retrieval to fill requests for specific booklet types for future scoring projects. Unspiraled and sequenced booklets were then assigned a new inventory number by grade and booklet type and sent back to the warehouse for storage. The storage locations of all documents were recorded on the inventory control system. Unused materials were sent to temporary storage to await completion of the entire assessment. Once the assessment was complete, NCS received authorization from ETS to salvage unused materials after determining that a sufficient quantity of each form type was retained in permanent archive.

6.6.2 Quality-Control Documents

ETS requires that a random sample of booklets and the corresponding scores/scoring sheets be pulled for an additional quality-control check. For image-scanned documents, a scoring sheet is not used, so ETS uses scores sent to them on a data tape to verify the accuracy of applied scores. During the scoring of mathematics, a selected number of image-processed booklets were paper scored. If any of the random sample of mathematics booklets used for paper scoring were selected as quality-control documents, the scoring sheet was also sent to ETS. For non-scannable trend reading/writing booklets and for the trend mathematics/science booklets that were scored via paper, both the booklet and its corresponding score sheet were sent. All of these documents were selected prior to sending the booklets to storage and were then sent to ETS to verify the accuracy and completeness of the data. A random sample of all the questionnaires used in the 1996 NAEP assessment was also sent to ETS.

6.6.3 Alert Analysis

Table 6-5 identifies the different types of alerts to problems that were encountered in the processing of NAEP data. For the 1996 main and long-term trend assessments, there was a total of 230 alerts; for the State Assessment, there was a total of 3,812 alerts.

Discrepancies were found in the receiving process that did not require an alert to be issued to Westat. They did require a great deal of effort by the opening staff to resolve in order to provide the most complete and accurate information. These are referred to as “info alerts.” These were categorized and codes were assigned to them. They are listed in the left-hand column of Table 6-5.

Even though receipt-control staff were well trained in the resolution of many situations, there were some problems that required resolution by NCS NAEP product line staff. These are referred to as "problem alerts." The various types of problem alerts were also categorized and coded. They are listed in the right-hand column of Table 6-5. For any unusual situations, Westat was contacted to help with the resolution of the alert.

Table 6-5

Alerts for the 1996 National and State Assessments

Information Alerts	Problem Alerts
Code 52 not written on Administration Schedules	Change of Administration Codes-A/S or Booklets
The yes/no box not gridded on Rosters	Incorrect Rosters/Questionnaires
Session Number not on Administration Schedules	Administration Notes/Writing on Covers
Administration Codes not on A/S; but on booklets	Duplicate Student / Booklet Number/ Administration Schedule
Administration Codes not on booklets; but on A/S	All material not returned
Items returned for Westat	Affected Testing - Problem
Writing on booklet covers	Transcribed page(s) for student booklet(s)
Other	Processed as is
	Involves Inclusion Check List
	Other

A/S = Administration Schedules

Chapter 7

PROFESSIONAL SCORING[1]

Patrick B. Bourgeacq, Bradley Thayer, and Timothy Robinson
National Computer Systems

7.1 INTRODUCTION

The 1996 national assessment required the scoring of constructed responses in mathematics and science at grades 4, 8, and 12. Long-term trend assessments for 9-, 13-, and 17-year-olds in reading, writing, mathematics and science continued at levels comparable to previous assessment years. In all, over nine million constructed responses for the 1996 national and state assessments were scored.

More than 300 professional readers split between a full-time day shift and part-time evening shift were hired for scoring. Veteran scorers were mixed with new hires screened for their ability to score constructed responses, providing the scoring center with excellent groups of qualified readers. Regular full-time staff, with the help of administrative assistants, bridged the shifts to ensure quality scoring between the two groups. For the first time in a National Assessment of Educational Progress (NAEP) assessment, National Computer Systems (NCS) provided a significant number of trainers who worked with Educational Testing Service (ETS) staff to train teams of scorers on many items. Also, NCS used lead scorers for the first time in a NAEP assessment to assist the table leaders with administrative duties and monitoring quality of scoring. The help of lead scorers made it possible to score greater volumes of responses with teams as large as 14 scorers without any apparent compromises in the quality of scoring.

As in previous years, the image system distributed responses among teams who scored them in an efficient manner to maximize consistency and reliability. The system also provided enhanced tools to display images, gather data, and monitor the quantity and quality of work.

The figures and tables on the following pages summarize the scope of professional scoring for the 1996 NAEP.

[1] Bradley Thayer was the NCS project manager for 1996 NAEP, Patrick Bourgeacq was the NCS project director for 1996 NAEP scoring, and Timothy Robinson was the NCS senior processing coordinator for 1996 NAEP.

Figure 7-1
Image Scoring Flow Chart

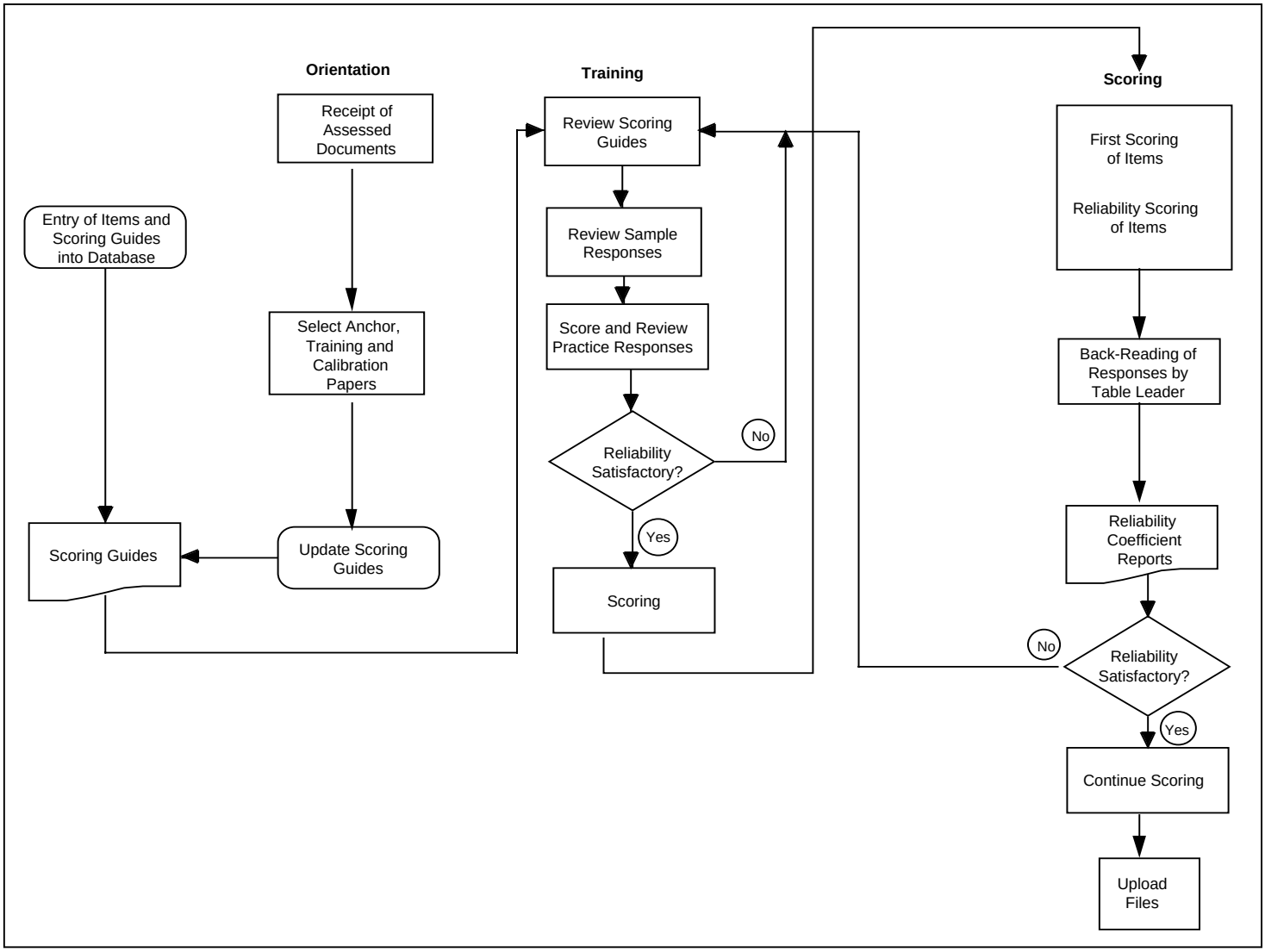
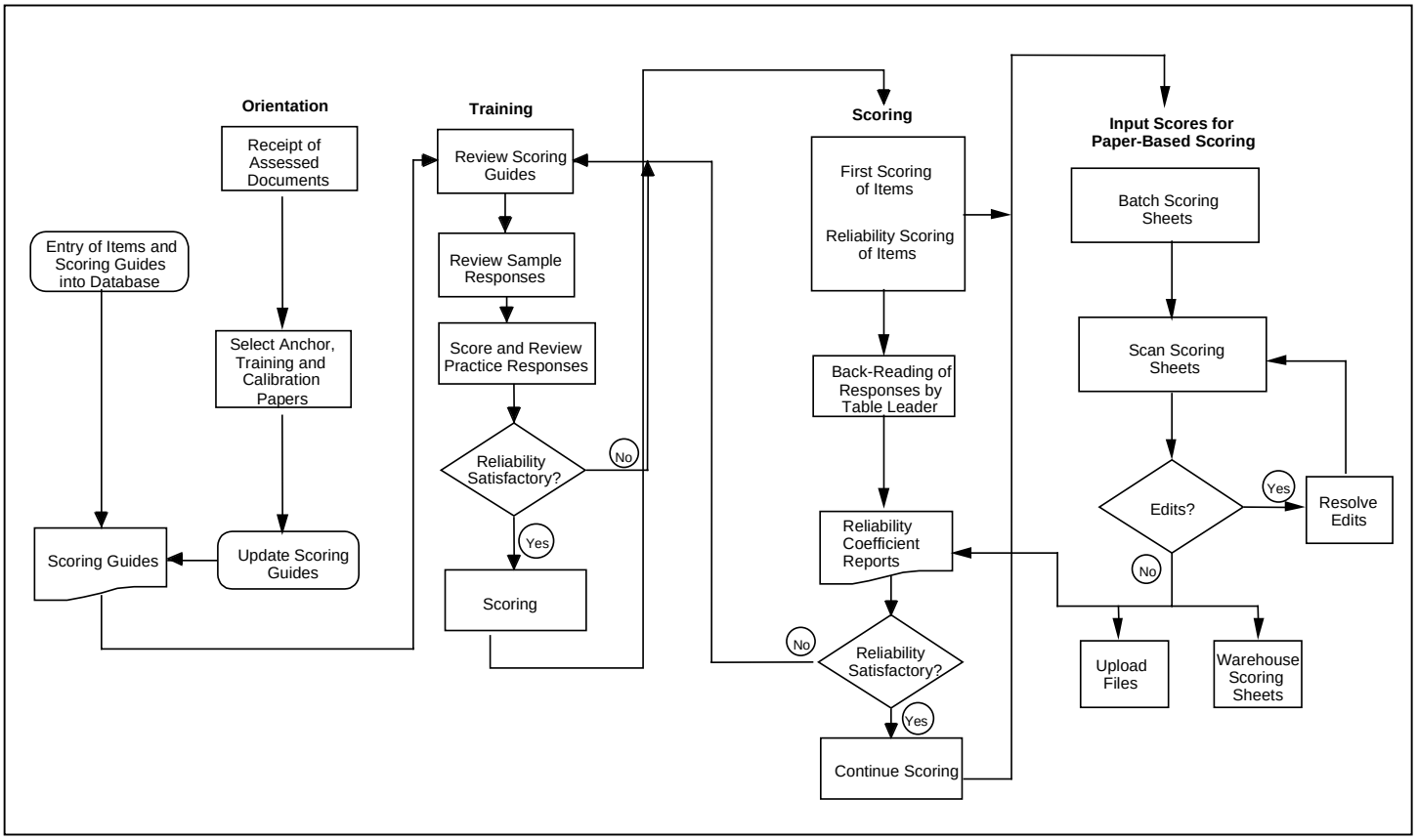


Figure 7-2
Paper Scoring Flow Chart



7.2 LONG-TERM TREND ASSESSMENTS

7.2.1 Mathematics

Items that contributed to long-term trend in mathematics items were scored as "right," "wrong," or "omitted." The scoring criteria identified the correct or acceptable answers for each item in each block. The scores for these items included a "0" for no response, a "1" for a correct answer, or a "2" for an incorrect or "I don't know" response. The reading items that appeared in the mathematics/science booklets were scored as "attempted," or "omitted." This scoring consisted of merely checking to see whether the student had responded in any way to that item, in which case the item was determined to have been reached or attempted. The scoring here was "0" for not attempting the item (blank) or "1" for any writing in the space provided. This includes one reading item in an age 9 math booklet and one reading item in an age 13 math booklet. The numbers of discrete constructed-response mathematics items can be found in Table 7-1.

Since scoring of the long-term trend mathematics items was identical to previous years, no new training papers were needed. Preparation for scoring included copying the scoring guides from previous assessment years and drawing samples from previous years, retrieving the booklets listed in the samples, and printing and matching scoring sheets for those booklets.

Because the mathematics items were scored as "right," "wrong," or "omitted," lengthy training for scoring these items was unnecessary. For each component (fall, winter, and spring), a different team was trained to follow the procedures for scoring the mathematics items and became familiar with the scoring standards, which listed general guidelines and also the correct answer for the items in each of the blocks. Each season, the entire scoring was done in one or two weeks at the end of the administration period. The number of booklets processed and the number of constructed responses scored for each age level are reported in Table 6-1 in Chapter 6.

A different team scored each age level at the time of year the age level was assessed. The booklets arrived in sessions, so each reader scored all items in all mathematics booklet types during the course of the project. All scorers held the same qualifications as the readers for main and state assessments. The number of readers, table leaders, and dates of scoring are reported in Table 7-2.

To establish the consistency of scoring across years, the readers rescored a subset of the responses from previous assessments. Samples of 350 responses to each item from the 1990 assessment and 250 from the 1994 assessment were drawn. The Performance Assessment Scoring Center (PSC) score sheet scanning system gave real-time reports comparing the original scores to the scores assigned by this year's team. The team also second scored 33 percent of the current year sample to measure consistency of scoring. The table leaders monitored daily interreader agreement reports and t-tests to verify consistency of scores within year and across years. Summaries of the interreader agreement figures can be found in Table 7-3.

Table 7-1

Number of Constructed-Response Items by Score-Point Levels

Subject Age/Grade	2-Category	3-Category	4-Category	5-Category	6-Category	Total
Mathematics Long-Term Trend						
Age 9	29	0	0	0	0	29
Age 13	28	0	0	0	0	28
Age 17	29	0	0	0	0	29
Total	86	0	0	0	0	86
Reading Long-Term Trend						
Age Class 9	9	0	1	2	2	14
Age Class 13	8	0	2	4	1	15
Age Class 17	8	0	3	4	1	16
Total	25	0	6	10	4	45
Writing Long-Term Trend						
Age Class 9	0	2	4	0	0	6
Age Class 13	1	1	5	0	0	7
Age Class 17	3	0	6	0	0	9
Total	4	3	15	0	0	22
Writing Long-Term Trend (Holistic Scoring)						
Age Class 9	0	0	0	0	2	2
Age Class 13	0	0	0	0	2	2
Age Class 17	0	0	0	0	2	2
Total	0	0	0	0	6	6
Bilingual Mathematics						
Grade 4	1	7	1	1	0	10
Grade 8	1	9	2	0	0	12
Total	2	16	3	1	0	22
Mathematics New Items						
Grade 4	0	18	14	7	0	39
Grade 4/8	0	5	0	0	0	5
Grade 8	5	30	5	6	0	46
Grade 8/12	0	4	1	2	0	7
Grade 12	2	29	4	8	0	43
Grade 4/8/12	0	0	0	0	0	0
Total	7	86	24	23	0	140
Mathematics Short-Term Trend (Base Year 1990)						
Grade 4	0	18	14	7	0	39
Grade 4/8	0	5	0	0	0	5
Grade 8	5	30	5	6	0	46
Grade 8/12	0	4	1	2	0	7
Grade 12	2	29	4	8	0	43
Grade 4/8/12	0	0	0	0	0	0
Total	7	86	24	23	0	140

(continued)

Table 7-1 (continued)

Number of Constructed-Response Items by Score-Point Levels

Subject Age/Grade	2-Category	3-Category	4-Category	5-Category	6-Category	Total
Mathematics Short-Term Trend (Base Year 1992)						
Grade 4	5	4	2	2	0	13
Grade 4/8	1	1	2	5	0	9
Grade 8	5	4	1	3	0	13
Grade 8/12	1	1	0	0	0	2
Grade 12	2	3	4	4	0	13
Grade 4/8/12	0	3	1	0	0	4
Total	14	16	10	14	0	54
Science						
Grade 4	7	52	10	2	0	71
Grade 4/8	2	47	8	1	0	58
Grade 8	1	41	17	0	0	59
Grade 8/12	2	50	8	0	0	60
Grade 12	4	83	32	6	1	126
Total	16	273	75	9	1	374

Table 7-2

Professional Scoring - Readers and Dates

Assessment	Number of Table Leaders	Number of Scorers	Dates
Fall Long-Term Trend Reading/Writing	1	7	11/29/95-1/5/96
Fall Long-Term Trend Mathematics	0	11	12/18/95-1/5/96
Winter Long-Term Trend Reading/Writing	1	6	1/30/96-4/3/96
Winter Long-Term Trend Mathematics	1	9	3/25/96-3/29/96
Spring Long-Term Trend Reading/Writing	1	6	4/10/96-5/24/96
Spring Long-Term Trend Mathematics	1	5	5/13/96-5/24/96
Long-Term Trend Writing Holistic	7	56	6/10/96-61/6/96
Long-Term Trend Writing Mechanics	6	33	7/15/96-7/31/96
Mathematics Segment 1 Days	8	72	3/13/96-4/5/96
Mathematics Segment 2 Days	9	126	4/8/96-5/6/96
Mathematics Segment 2 Evenings	5	70	4/8/96-5/2/96
Bilingual Mathematics	0	4	5/2/96-5/3/96
Science Segment 1 Days	3	24	3/18/96-4/5/96
Science Segment 2 Days	3	27	4/8/96-5/3/96
Science Segment 2 Evenings	5	70	4/8/96-5/2/96
Science Segment 3 Days	12	156	5/6/96-6/7/96
Science Segment 3 Evenings	12	150	5/6/96-6/7/96

Table 7-3
Interreader Reliability Ranges

Assessment[1]	Number of Unique Items Total	Number of Items in Percentage Exact Agreement Range			
		60-69%	70-79%	80-89%	Above 90%
Fall Long-Term Trend Reading/Writing	13	0	0	3	10
Fall Long-Term Trend Mathematics	28	0	0	0	28
Winter Long-Term Trend Reading/Writing	11	0	0	0	11
Winter Long-Term Trend Mathematics	29	0	0	0	29
Spring Long-Term Trend Reading/Writing	14	0	0	6	8
Winter Long-Term Trend Mathematics	29	0	0	0	29
Long-Term Trend Writing Holistic[2]	6	0	0	1	5
Long-Term Trend Writing Mechanics	3	N/A	N/A	N/A	N/A
4th Grade Mathematics	80	0	0	2	78
8th Grade Mathematics	102	0	0	4	98
12th Grade Mathematics	100	0	1	9	90
4th Grade Science	94	0	0	13	81
8th Grade Science	125	0	0	20	105
12th Grade Science	156	0	0	26	130

[1] Not all long-term trend items received second scoring. Figures are included here only for those that were second scored.
[2] Figures for long-term trend writing holistic include adjacent scores.

7.2.2 Reading and Writing (Primary Trait)

All of the writing items for the three long-term trend assessments (fall, winter, and spring) were scored using the primary trait method. This method focused on the writer's effectiveness in accomplishing specific assigned tasks. The primary trait scoring criteria defined five levels of task accomplishment:

1. not rated,
2. unsatisfactory,
3. minimal,
4. adequate, and
5. elaborated.

The scoring standard for each item described these levels in detail. Some of these items were also scored for secondary traits, which involved indicating the presence or absence of elements that were of special significance to a particular item (e.g., whether notes were made before writing or whether critical information was filled out on a form).

The scoring guides for the constructed-response reading items focused on students' abilities to perform various reading tasks:

- identifying the author's message or mood and substantiating their interpretation,
- making predictions based on given details, and
- comparing and contrasting.

The guides for the reading items varied somewhat, but typically included a range of scores denoting inability to address the task, unsatisfactory responses, minimal ability in accomplishing the task, satisfactory ability in addressing the task, or elaborated responses addressing the task fully. Some of the reading items received scores for secondary traits based on what reactions or information the student gave (i.e., whether the response was mostly content based, form based, a subjective reaction, or some combination of the three).

The scoring guides for the constructed-response writing items focused on students' abilities to write in informative, persuasive, and narrative styles. The guides for the writing items were based on a range of scores denoting unsatisfactory writing to address the task, minimal writing to address the task, satisfactory writing to address the task, and elaborated writing to address the task.

The item known as "The Door" was scored for attemptedness only. The readers coded all blanks as "0" and any attempt to answer as a "1."

The numbers of discrete constructed-response reading and writing items can be found in Table 7-1.

As with mathematics, the scorers used the same training materials as in previous assessments for reading and writing. Thus, there was no need to select new training material from current year responses. Preparation for the three long-term trend scoring projects began with identifying samples from previous years as indicated below. Scores assigned in assessment booklets from 1984 (reading responses) and 1988 (writing responses) had been masked in previous years to ensure that scoring for training, and subsequent long-term trend reliability scoring, would be done without knowledge of the previous scores given. The 1994 booklets required no masking because scores had never been written directly in the booklets. Finally, clerical support staff members matched scoring sheets with the booklets selected for rescore after they had been pulled from the warehouse.

The formal training for the long-term trend assessments was divided into two parts to accommodate the reading and writing items. The reading/writing long-term trend scoring project started with one team of seven readers with one table leader for fall trend. Six of the same scorers and the same table leader continued for winter and spring trends. Dates for scoring the three seasons of trend booklets are given in Table 7-2. During training each reader received a photocopied packet of materials used in the 1984 scoring of the reading items and the 1988 scoring of the writing items.

Prior to scoring any 1996 reading and writing trend material, a training reliability report was generated using a 25 percent sampling of the 1984 assessment materials for reading and 1988 materials for writing. Following the formal training sessions, the readers scored this material on scannable scoring sheets produced for specific booklet types with the appropriate long-term trend items pre-printed on the scoring sheets. These sheets were then routed to scanning under a special job number to ensure that this material was labeled for training scoring only. The scoring coordinator was able to generate a computer report that listed the individual and group percent agreement by item. The system automatically compared the new score with the original score assigned in the 1984 or 1988 scoring and produced a report on the training reliability. T-tests were also generated for each item to verify comparability of scoring across years. The NCS scoring specialist then conferred with the appropriate ETS staff on this training reliability agreement report before proceeding with scoring.

All readers for this project were experienced scorers with a minimum of a bachelor's degree. One team member had scored the same long-term trend items for the 1994 assessment. The team read materials as they arrived at the scoring center, with occasional breaks in scoring when receipts were slow. The table leader monitored consistency within the current year as well as across years on a daily

basis as indicated below. The number of booklets processed and constructed responses scored for each age level can be found in Table 6-1 in Chapter 6.

Reliability studies were conducted for the scoring of the long-term trend reading and writing items. For the 1996 booklets, 33 percent of the constructed-response items were scored by a second reader to produce interreader reliability statistics. In addition, a long-term trend reliability study was conducted to ensure that the scoring procedures were consistent with those used in 1984, 1988, and 1994. For this study, 350 of the 1984 reading responses and 350 of the 1988 writing responses were sampled. Also, 250 of both reading and writing responses from 1994 were sampled for rescore. The scoring of these long-term trend samples was intermixed with the scoring of the 1996 reading and writing trend material. The readers selected a bundle of approximately five of each booklet type each day and gridded their scores on separate scannable scoring sheets for each item. These sheets were then scanned and cross-referenced with the original data tape to extract information for long-term trend reliability reporting. T-tests were generated daily to verify comparability of scoring across years. Note that only primary trait scores were compared in the across-year rescore. Secondary traits and items scored for attemptedness only were not second scored in the current year nor rescored in the long-term trend sample. Composite ranges of interreader reliability figures can be found in Table 7-3.

7.2.3 Holistic Writing Scoring

Certain writing items included in the long-term trend assessment were scored holistically. Holistic scoring is based on an overall judgment of writing fluency and structure. Because a six-point holistic scale was used, no scores fell exactly in the middle of the scale. This was done to force readers to categorize a student's response into either the top half or the bottom half of the scale. Within the upper and lower halves of the scale, the scores reflected the degree to which the student demonstrated fluency, structure, or both, in responding to the prompt.

In the 1996 holistic scoring session, two items were scored from each of the three grade levels. The items scored at grade 4 were

- *flashlight*, an imaginative task, and
- *spaceship*, a persuasive task.

At grades 8 and 11, the two items scored were

- *food on the frontier*, an informative task, and
- *recreation opportunities*, a persuasive task.

The responses scored were taken from six assessment years: 1984, 1988, 1990, 1992, 1994, and 1996. Fifty-six readers participated in the 1996 holistic scoring session. They were organized into seven teams of eight readers, with each team led by a table leader and assisted by one clerical aide.

The total number of responses scored, readers, table leaders, and dates are given in Table 6-1 in Chapter 6. Detailed lists of responses broken down by prompt, grade, and year are given in the *Report of Processing and Professional Scoring Activities* (NCS, 1996).

7.2.3.1 Materials Preparation

All grade-eligible student booklets with on-task scores were selected from each of the six assessment years and assembled into bundles of 25 booklets each. Bundle header sheets were then generated from the existing data file of student booklet identification numbers. The system assigned the first 25 grade-eligible booklet ID numbers for each specific booklet type to a bundle. The next 25 booklets for that booklet type were assigned the next consecutive bundle number, and so on until all booklets were assigned. The bundle header was placed on top of the bundle of booklets and these were then rubber-banded together. In addition to the bundle number, the bundle header showed the name of the item to be scored, the grade level, the assessment year, and the booklet ID numbers of the booklets in that bundle.

For each bundle, two sets of scoring sheets were generated—one for the first scoring and one for the second scoring. The scoring sheets indicated the bundle number, the name of the item to be scored, the grade level, the assessment year, and the booklet ID number. Second scoring sheets were generated for a random 25 percent of the booklets in each bundle. For both first and second scoring sheets, a separate sheet was generated for each booklet in the bundle.

Since some of the booklets for grades 8 and 11 contained two items to be scored during the holistic scoring, Recreation Opportunities and Food on the Frontier, two separate bundle headers and sets of scoring sheets were created for these booklets to accommodate the two scorings of the same booklets. The appropriate header and corresponding scoring sheets for both items were placed with the booklet bundle prior to scoring. Upon completion of the holistic scoring for the Food on the Frontier item, clerical staff removed the header and scoring sheets for this item and replaced them with the sheets for the Recreation Opportunities item; then the responses for the Recreation Opportunities item were scored.

All bundles were separated by item on carts. A sheet identifying the item and year was attached to each cart. Thus, all bundles could be readily identified and retrieved quickly and easily.

7.2.3.2 Training Preparation

Three days prior to the beginning of holistic scoring, the ETS writing coordinator met with the table leaders and NCS administrative assistant to prepare training materials. They first reviewed the training materials from the 1994 scoring session and augmented them with sample responses from the 1995-96 assessment. The training packets included a *familiarization and range finder* set each containing six examples, an *upper half and a lower half* set each containing three examples, and three *calibration* sets each containing four examples. The sets, together with their scoring guides and keys, can be found in the *Report of Processing and Professional Scoring Activities* (NCS, 1996).

Following standard ETS procedures, the training of the readers was conducted immediately prior to the beginning of scoring for each individual item.

7.2.3.3 Holistic Scoring Training

The training materials for the first item were distributed to each reader. Each reader read a copy of the prompt and the holistic scoring guide for that item. The scoring guide was explained by the writing long-term trend coordinator. A clerical aide was assigned to each of the seven teams to keep the materials

flowing in an organized and efficient manner. Training began on June 10, 1996. The general sequence of training for each of the six prompts was as follows:

Scoring Guide and Prompt - The scoring guide was introduced and briefly discussed, followed by a reading of the prompt and a general discussion of the scoring expectations.

Familiarization Set - Readers were asked to arrange the papers in the familiarization set into 'best to worst' order, assigning a score to each paper, with no score being used more than once. The writing long-term trend coordinator announced the scores given by the table leaders for each of the individual papers in the set. The table leaders discussed with their teams the rationale behind these scores, using the scoring guide.

Range Finders - The writing long-term trend coordinator read the six range finders (best to worst) and discussed the rationale behind the scores given. If there was too much discussion regarding 'why,' an alternate range finder for that score point was inserted later. It was at this point the writing long-term trend coordinator asked the readers to look only to their range finders and no longer to the scoring guide.

Upper Half - Readers first scored three papers representing the upper half. There was a sample of a four-, a five-, and a six-score point paper included. A tally was kept of the number of readers assigning each score point. After the tally, the writing long-term trend coordinator announced the scores assigned in the preparation process and compared them to the scores on the tally sheet.

Lower Half - The readers were then asked to score three lower-half papers. This sample included a one-, a two-, and a three-score point paper. Again, a tally of the scores assigned by the individual readers was kept. The table leaders discussed the rationale behind the reader-assigned scores including both the upper and lower half score point papers at this time and answered any team member questions.

Calibration Set 1 - Prior to the beginning of scoring, the writing long-term trend coordinator assigned Calibration Set 1 consisting of four papers. After scoring Calibration Set 1, a tally was kept and the writing long-term trend coordinator announced the scores. The table leader and team discussed the readers' scores and questions.

Calibration Sets 2 and 3 - A calibration set was used after a break of longer than 15 minutes (lunch, morning break, afternoon break, or overnight). The readers reviewed the range finders and the table leaders provided a one-sentence review of each of the score points. The readers scored the set of four papers and the scores given were tallied. The tallies were kept for each of these calibration sets and can be found with the training materials in the *Report of Processing and Professional Scoring Activities* (NCS, 1996).

7.2.3.4 Holistic Scoring

Holistic scoring occurred during the week of June 10, 1996. When scoring for an item began, each reader was given a bundle to score. The reader entered his or her reader ID number in the appropriate space on each scoring sheet and scored all the responses in the bundle. The readers were directed to check the booklet number listed on the scoring sheet against the booklet ID number printed on

the front of the assessment booklet. If any discrepancies were found, they were brought to the table leader for resolution.

When the first scoring of a bundle of booklets was completed, the reader handed the first scoring sheets to the clerical support staff who placed the bundle of booklets, with the bundle header and second scoring sheet still on top, back on the designated cart. NCS clerical staff distributed all bundles to the readers to ensure an even flow of material and a distribution of second scoring among teams.

Reliability scoring was handled in the same manner as first scoring. When the second scoring was completed, the second scoring sheets were placed on the table leader's desk. All bundles of booklets were placed on carts and removed from the scoring area. Clerical staff then sorted all scoring sheets and routed them to the scanning area to be entered onto the database.

The table leaders read through each reader's entire initial bundle of booklets and evaluated the scores assigned by the reader. This process is known as backreading. They also periodically backread the remaining bundles throughout the reading of each item. If discrepant scoring occurred, the table leader brought it to the attention of the reader. If the problem recurred, the reader was retrained by the table leader. If the problem continued, the writing long-term trend coordinator assisted with the re-training.

Twenty-five percent of the papers were read again by a reader from a different team. A preliminary reliability measure was taken by the administrative assistant approximately two to three hours after the beginning of scoring.

Scoring the booklets took between four and seven hours depending on the grade level and the type of writing. The readers were trained and scored the items in the following order:

Informative:	Food on the Frontier	Grade 11, followed by Grade 8
Persuasive:	Recreation Opportunities	Grade 11, followed by Grade 8
	Spaceship	Grade 4
Narrative:	Flashlight	Grade 4

The scoring was completed on Sunday, June 16, 1996. A more detailed log of daily activities can be found in the *Report of Processing and Professional Scoring Activities* (NCS, 1996).

Table 7-4
Holistic Writing Scoring Reliability Figures

Grade	Item	Exact Agreement	Adjacent Agreement
4	Flashlight	52.1%	89.6%
	Spaceship	58.3%	93.5%
8	Food on the Frontier	53.1%	94.5%
	Recreational Opportunities	58.2%	95.3%
11	Food on the Frontier	56.3%	94.5%
	Recreational Opportunities	57.1%	94.9%
	Total	**56.1%**	**94.0%**

After all the score sheets were scanned, the scanner operator produced a final report showing the n-counts scored and the interreader agreement rates. The figures are given at the item level in Table 7-4.

They are given in more detail, broken down by prompt, grade, and year, in the *Report of Processing and Professional Scoring Activities* (NCS, 1996).

7.2.4 Writing Mechanics

Mechanics scoring focused on the extent to which the writer can control the conventions of written English—grammar, spelling, capitalization and punctuation. In addition, the procedures include identifying sentence structures and word choice errors. A team of 33 readers and six table leaders scored selected essays for each age/grade group from the writing long-term trend assessments conducted in years 1988, 1990, 1992, 1994, and 1996. Spaceship, an imaginative task, was scored at age 9/grade 4 and recreation opportunities, a persuasive task, was scored at both age 13/grade 9 and age 17/grade 11.

7.2.4.1 Preparation

Essays to be scored from the 1996 assessment were selected according to ETS specifications as follows. For each booklet in which each item appeared:

1. select all grade-eligible booklets with primary trait scores greater than 0 and less than 7, in other words only on-task responses;
2. select every third booklet; and
3. select all Black students (based on the student's response in the background questions) not picked in Step 2.

As a result of this selection process, 1,593 essays were scored from the 1996 assessment. In addition, 10 percent of the essays previously scored for mechanics from the 1988, 1990, 1992, and 1994 assessments were rescored for reliability. This sample was selected by locating specific booklets from a list generated by ETS and resulted in a rescore of 736 essays for all four years.

In preparation for the scoring process, copies were made of each selected essay and its corresponding booklet cover. The booklet cover, containing assessment year, age/grade, primary sampling unit (PSU) and student ID information, was stapled to the essay. Papers were then grouped by assessment year and grade into packets of 20. Packets were numbered consecutively and were identified by headers. Three identical sets of packets (A, B, and C) were assembled since each essay had to be scored independently by two readers and discrepancies had to be resolved by a table leader. Packets A and B were used by the two readers and Packet C was used by the resolver. For prior year scoring, the same sample was used as in 1994. Therefore, copies were made from the 1994 master sets rather than returning to the original booklets. The master sets were labeled as D Packets and warehoused for potential future use.

7.2.4.2 Writing Mechanics Training

Training of the six table leaders and 33 readers was conducted by the writing long-term trend coordinator during the week of July 15, 1996. The training involved a detailed discussion of the scoring guide. The writing coordinator presented the main sections of the guide:

- type of sentence construction,
- faulty sentence construction,
- punctuation, and
- word level categorization.

Copies of pertinent resource information were distributed, briefly reviewed, and reference materials were identified. After discussion of each of the main sections of the guide, the group reviewed the scored papers from the training packet, paying special attention to scores reflecting the category under discussion.

To further train the readers, the trainer used a pool of responses from the1988 and 1990 assessments that has been scored for mechanics in 1990 but not used in the rescore sample in 1996. Copies were made of these training essays to be used for practice. Each reader then individually scored a selected group of essays. The scores were compared among the group, discussions were held when discrepancies occurred, and again references were made to resource materials or to the scored sample papers. When the group was comfortable with the decisions being made, the actual scoring began. Copies of these sets and the scoring guidelines can be found in the *Report of Processing and Professional Scoring Activities* (NCS, 1996).

7.2.4.3 Scoring

The actual scoring, resolution, data entry work and proofreading of writing mechanics began on July 18 and was completed on July 31, 1996. In selecting packets for scoring, readers alternated among the different grade levels and assessment years. The mechanics readers marked each paper with a series of symbols, addressing the elements of sentence type, sentence construction, word choice, spelling, punctuation, and capitalization. These symbols, written in red ink, designated each word or punctuation mark in error and indicated sentence type or faulty sentence construction. Each essay was scored independently by two different readers selecting either Packet A or B.

To track the movement of the packets, the NAEP internal tracking log was used. As readers and resolvers worked on particular packets, the appropriate columns were initialed and dated. This enabled NCS staff to see at a glance the status of each packet. The completed tracking logs have been warehoused with the training materials. A sample of a blank tracking log can be found in the *Report of Processing and Professional Scoring Activities* (NCS, 1996).

Resolution and quality control were conducted by table leaders who compared the scores marked on copies A and B of each unique packet and resolved any discrepancies. After determining the appropriate marks, the resolver used the unmarked copy C to record the final version. The copy with the resolved marks was sent to the word processing area in NCS's Creative Services department for transcription as described below. To avoid confusion, unused copies were discarded or returned to original readers with feedback information and with follow-up training, if deemed necessary.

To maintain a consistent scoring standard, the six table leaders, along with the NCS performance assessment specialist, met twice daily to resolve any questions that arose during the scoring. Resolution scoring allowed the table leaders to determine the accuracy of individual readers. If a reader was confused by a facet of the coding, the table leader would approach the reader individually. If the table leaders identified a trend in the coding, the issue would be broached at the re-calibration sessions. Twice a week, for approximately 30 minutes on Wednesday mornings and Friday afternoons, the six tables stopped scoring and re-calibrated; table leaders brought up pertinent scoring issues and readers asked for clarification on coding decisions.

Resolved packets were sent to the NCS Creative Services department where the text of the essays, along with the assigned marks and identification information, were entered into a data file. Essays were typed exactly as they were written. Each essay was typed on a separate page and double-spaced pages were printed for proofreading. The scoring team proofread the data entry work against the scored

papers. Student identification and PSU numbers were checked and discrepancies were resolved. Corrections were indicated in red on the typed copy. If corrections were needed, all pages belonging to a packet were returned to the Creative Services department for additional corrections. Complete and correct packets were uploaded to the NCS mainframe. The data were reformatted according to ETS specifications and a data file containing the scored information was sent to ETS on August 7, 1996.

7.3 MAIN NAEP ASSESSMENT

7.3.1 Selection of Training Papers

A pool of papers to be used for training for the NAEP main assessments was selected by NCS in February and March of 1996. Persons identified as potential mathematics and science table leaders were selected to copy student responses. Team leaders, with assistance of the potential table leaders, gave tentative scores to the responses and selected 50 student responses from each dichotomous item, 75 responses from each 3-point item, 100 responses from each 4-point item and 125 from each 5-point item. Because NCS staff screened the responses, the pool sent to the ETS test development specialists contained a full range of point values.

NCS staff numbered the papers sequentially and copied the sets. NCS retained and filed the originals and sent the copies to the appropriate subject area coordinators at ETS. ETS returned a list of the anchor and practice sets with scores to NCS staff, who used the master file copies to create training sets. NCS staff then masked the sequential reference numbers and wrote the actual scores on the anchor papers and a new sequential reference number for the training sets.

The NCS copy center and PSC clerical staff shared responsibility for making multiple copies of the sets for scorers. The master sets, team leader/trainer copy and table leader copies also had keys to the training sets. When copying was complete, the master copy was placed in the appropriate file.

7.3.2 Calibration Policies

During scoring, the teams used calibration sets to calibrate on a daily basis and to calibrate across longer periods of time. The table leader built pools of items for calibration, which were then distributed to the scorers in sets of five or ten, depending upon the complexity of the item, whenever a break of longer than 15 minutes occurred, such as after lunch or at the beginning of a new scoring day. All readers on the team scored the same calibration sets, and the system compared the scores of each reader to the scores assigned by the trainer and table leader. The table leader reviewed the interreader agreement report with the trainer and the ETS subject area coordinator, discussed any discrepancies that arose, and then proceeded with scoring.

Whenever a team returned to scoring an item after having worked on a different item in the meantime, the team scored a calibration set of 75 responses and analyzed the results before proceeding. This occurred in mathematics because many mathematics items were scored in two sessions since the first sweep through the items was done while booklets were still arriving from the field. If the item had fewer than 500 responses left out of a pool that contained both main and state samples, an extended calibration was waived. In science, the teams did not begin scoring most items until all responses to the item were available for scoring, so this type of calibration was not necessary.

Table leaders printed and archived hard copies of all calibration sets used for scoring. For more information on the functionality of the calibration tool, see Section 7.3.3.3.

7.3.3 Table Leader Utilities

Two of the significant advantages of the image-scoring system were the ease of regulating the flow of work to readers and the ease of monitoring scoring. The image system provided table leaders with tools to determine reader qualification, to backread scores, to determine reader calibration, to monitor interreader reliability, and to gauge the rate at which scoring was being completed. These various tools are described below.

7.3.3.1 Reader-Qualification Tool

One of the utilities at a table leader's disposal was a qualification algorithm used after training on extended constructed-response items. The table leader would give identical qualification packets to each reader. These packets contained 10 student responses to be independently scored by the readers. After the readers finished, the table leader would enter each reader's scores into the computer for tabulation. The computer would calculate each reader's percentage of exact, adjacent, and non-adjacent agreement with the master key. If a reader attained a percentage of exact agreement above a pre-determined threshold of 80%, the reader would be allowed to score. Readers not attaining the pre-determined threshold were handled on a case-by-case basis—typically receiving individual training by the trainer or the NCS table leader before being allowed to score. A table leader could cancel a reader's qualification to score an item if review of a reader's work indicated inaccurate scoring and that supplemental training was necessary. Note that reader qualification was required only on extended constructed-response items involving 4 or more score point levels.

7.3.3.2 Backreading Tool

After scoring began, NCS table leaders reviewed each reader's progress using a backreading utility that allowed the table leader to review papers scored by each reader on the team. Typically, a table leader reviewed responses scored by each reader in quantities similar to the amount second scored (i.e., 6% for items with both state and national samples and more for items with only a national sample). Table leaders noted the score the reader awarded each response as well as the score a second reader gave that same response. This was done as an interreader reliability check. Alternatively, a table leader could choose to review all student responses given a particular score to determine if the team as a whole was scoring consistently. Both of these review methods used the same display screen and showed the ID number of the reader and the scores awarded. If the table leader disagreed with the score given an item, he or she discussed it with the reader for possible correction. Replacement of scores by the table leader was done only with the knowledge and approval of the reader, thereby serving as a learning experience for the reader. Additionally, neither score was changed in the case where the response was second scored.

7.3.3.3 Calibration Tool

While backreading, a table leader could identify individual responses for inclusion in a pool of calibration papers. These papers could be selected because they exemplified criteria set down in the

scoring rubrics or because they were unusual and pointed out less obvious aspects of the scoring guidelines. After selecting a number of papers for inclusion in the calibration set, the table leader could decide to route any number of these calibration papers to the scorers. A typical number of papers routed to scorers during a mid-scoring calibration was 10, although the image system could accommodate as many or as few as the table leader and trainer determine necessary to check the accuracy of scoring. When all scorers had completed the calibration set, the table leader could then produce an interreader reliability report on the scoring of the calibration set.

7.3.3.4 Tool for Monitoring Interreader Reliability

During the scoring of an item or the scoring of a calibration set, the table leader could monitor progress using an interreader reliability tool. This display tool could be used in either of two modes:

1. to display information of all first readings versus all second readings, or
2. to display all readings of an individual that were also scored by another reader versus the scores assigned by the other readers.

The information was displayed as a matrix with scores awarded during first readings displayed in rows and scores awarded during second readings displayed in columns for Mode 1 and the individual's scores in rows and all other readers in columns for Mode 2. In this format, instances of exact agreement fell along the diagonal of the matrix. For completeness, data in each cell of the matrix contained the number and percentage of cases of agreement (or disagreement). The display also contained information on the total number of second readings and the overall percentage of exact agreement on the item. Since the interreader reliability reports were cumulative, a printed copy of the exact agreement of each item was made every day and compared to previously generated reports.

7.3.3.5 Tool for Monitoring the Rate of Scoring

The table leaders were able to monitor work flow using a status tool that displayed the number of items scored, the number of items first-scored that still needed to be second-scored, the number of items remaining to be second-scored, and the total number of items remaining to be scored. This allowed the team leaders and performance assessment specialists to accurately monitor the rate of scoring and to estimate the time needed for completion of the various phases of scoring.

7.3.3.6 Scoring Buttons

To assign a score, readers clicked the mouse over a button contained in the scoring window. Since buttons were included only for valid scores, there is no need to edit for out-of-range scores. Another recent development was the implementation of a tool that allowed the performance assessment specialist to label scoring buttons with key phrases, correct responses or certain incorrect responses that were to be tracked. This enhanced scoring as readers no longer had to mentally translate a student's response into a numerical value before choosing a scoring button on the image screen.

7.3.4 Main Mathematics Assessment

The mathematics portion of the 1996 main assessment included a total of 226 discrete constructed-response items. Table 7-1 shows the types and number of constructed responses for the mathematics assessment. A variety of constructed-response items were used to measure different elements of students' mathematical sophistication and understanding. These items were administered in scannable assessment booklets. The bilingual booklets were key-entry documents. The items scored included traditional computational items, short-answer constructed responses, extended constructed responses, diagrams, geometric figures, and graphs. Each constructed-response item had a unique scoring guide that identified the range of possible scores for the item and defined the criteria to be used in evaluating student responses. Long-term trend items that focused on the students' computational ability were typically scored on a right/wrong basis. The scoring guides for the more complex items were developed to be of diagnostic value, by including categories that reflected partial credit and/or different kinds of incorrect answers that indicated particular misunderstandings. New items developed for the 1993 and 1995 NAEP field tests were scored on a partial-credit scale (3-point or 4-point) or on an extended scale (5-point).

The operational assessment included, for the first time other than the field test, blocks that were based on a theme. There were two different theme blocks administered at each grade. General information on the number of constructed responses scored can be found in Table 6-1 in Chapter 6. Table 7-5 gives more detailed information by grade and booklet type (spiral, estimation, theme, and advanced).

7.3.4.1 Training

The training for each mathematics item was conducted by mathematics specialists from ETS and NCS just prior to the scoring of that item. Training and scoring began on Wednesday, March 13, and ended on Monday, May 6, 1996. The NCS mathematics performance assessment specialist and selected NCS team leaders conducted all the training of the short-term trend items scored in March. The ETS mathematics coordinator met with each trainer individually before the beginning of training to discuss any questions and/or discrepancies. He returned to NCS at the beginning of April to review individually the items assigned in April. The NCS training staff added another team leader to assist in training for the evening shift in April.

Training involved explaining the item and its scoring guide and discussing responses that represented the various score points in the guide. When this was completed, the readers scored and discussed from 5 to 35 selected "practice papers" for each item. Next, readers practiced scoring by gathering around a single image terminal and scoring several responses to the item. Once the trainer and the table leader determined the individuals on the team understood the scoring guide, the table leader qualified the scorers to enter the system using the reader-qualification tool, discussed in Section 7.3.3.1.

Table 7-5
Mathematics Constructed Responses Scored

Type	Data	1990	1992	1996	Grand Total	Assessment Proportions 1996 National	Assessment Proportions 1996 State
			Grade 4				
Regular	Unique items	12	36	64	112	7.9%	92.1%
	Responses first scored	0	0	2,037,866	2,037,866	160,584	1,877,282
	Responses second scored	8,988	27,031	122,272	158,291	9,635	112,637
	First and second scored	8,988	27,031	2,160,138	2,196,157	170,219	1,989,919
	Average % exact agreement	95.9	93.8	96.8	95.7		
Estimation	Unique items	0	0	3	3		
	Responses first scored	0	0	7,317	7,317		
	Responses second scored	0	0	439	439		
	First and second scored	0	0	7,756	7,756		
	Average % exact agreement	N/A	N/A	93.3	93.3		
Theme	Unique items	0	0	13	13		
	Responses first scored	0	0	23,800	23,800		
	Responses second scored	0	0	5,950	5,950		
	First and second scored	0	0	29,750	29,750		
	Average % exact agreement	N/A	N/A	95.8	95.8		
Total	Unique items	12	36	80	128		
	Responses first scored	0	0	2,068,983	2,068,983		
	Responses second scored	8,988	27,031	128,661	164,680		
	First and second scored	8,988	27,031	2,197,644	2,233,663		
	Average % exact agreement	95.9	93.8	96.5	95.7		
	Average % exact agreement	95.9	93.8	96.5	95.7		
			Grade 8				
Regular	Unique items	18	44	69	131	9.3%	90.7%
	Responses first scored	0	0	1,991,682	1,991,682	184,683	1,806,999
	Responses second scored	13,374	33,077	119,501	165,952	11,081	108.420
	First and second scored	13,374	33,077	2,111,183	2,157,634	195,764	1,915,419
	Average % exact agreement	95.3	94.4	96.6	95.7		
Estimation	Unique items	0	0	6	6		
	Responses first scored	0	0	13,622	13,622		
	Responses second scored	0	0	3,405	3,405		
	First and second scored	0	0	17,027	17,027		
	Average % exact agreement	N/A	N/A	97.3	97.3		
Theme	Unique items	0	0	13	13		
	Responses first scored	0	0	27,690	27,690		
	Responses second scored	0	0	6,923	6,923		
	First and second scored	0	0	34,613	34,613		
	Average % exact agreement	N/A	N/A	93.9	93.9		

(continued)

Table 7-5 (continued)
Mathematics Constructed Responses Scored

Type	Data	1990	1992	1996	Grand Total	Assessment Proportions 1996 National	Assessment Proportions 1996 State
			Grade 8				
Advanced	Unique items	0	0	14	14		
	Responses first scored	0	0	33,354	33,354		
	Responses second scored	0	0	8,339	8,339		
	First and second scored	0	0	41,693	41,693		
	Average % exact agreement	N/A	N/A	95.1	95.1		
Total	Unique items	18	44	102	164		
	Responses first scored	0	0	2,066,348	2,066,348		
	Responses second scored	13,374	33,077	138,168	184,619		
	First and second scored	13,374	33,077	2,204,516	2,250,967		
	Average % exact agreement	95.3	94.4	96.1	95.5		
			Grade 12[2]				
Regular	Unique items	21	44	73	138	N/A	N/A
	Responses first scored	0	0	212,774	212,774	N/A	N/A
	Responses second scored	15,225	32,387	12,766	60,378	N/A	N/A
	First and second scored	15,225	32,387	225,540	273,152	N/A	N/A
	Average % exact agreement	94.8	92.1	95.6	94.4		
Theme	Unique items	0	0	12	12		
	Responses first scored	0	0	23,368	23,368		
	Responses second scored	0	0	5,842	5,842		
	First and second scored	0	0	29,210	29,210		
	Average % exact agreement	N/A	N/A	94.5	94.5		
Advanced	Unique items	0	0	15	15		
	Responses first scored	0	0	44,881	44,881		
	Responses second scored	0	0	11,220	11,220		
	First and second scored	0	0	56,101	56,101		
	Average % exact agreement	N/A	N/A	95.1	95.1		
Total	Unique items	21	44	100	165		
	Responses first scored	0	0	281,022	281,022		
	Responses second scored	15,225	32,387	29,829	77,441		
	First and second scored	15,225	32,387	310,851	358,463		
	Average % exact agreement	94.8	92.1	95.4	94.5		
Grand Total - Unique items		51	124	282	457		

[1] 4th grade mathematics had no advanced booklets.
[2] 12th grade estimation block had no constructed-response items.

7.3.4.2 Scoring

Mathematics scoring took place in two segments over two shifts as outlined in Table 7-2. Most dichotomously scored short-term trend items were scored in Segment 1, with the remainder of the items scored in Segment 2.

During scoring, the team leaders continued to compile notes on scoring decisions for the readers' reference and guidance. Additionally, table leaders closely monitored interreader reliability using both team and individual statistics as a reference. Consistently throughout the scoring of each item, the table leaders also performed backreading duties in which they reviewed a sample of the responses scored by each reader on the team. Lead scorers selected for their experience and accuracy in scoring assisted the table leaders in backreading. The team leaders and performance assessment specialist continuously monitored the progress of each team and noted all scoring-related decisions to ensure that training and scoring progressed smoothly and in a timely manner.

The codes that were used for unscorable mathematics items were:

0 = Blank or random marks
8 = Completely crossed-out or erased
9 = "I don't know," refusal, off-task, illegible or language other than English

7.3.4.3 Reliability

A minimum of 25 percent of the mathematics responses for items involved only in the national sample and 6 percent of the responses for items involved in the state and national samples were scored by a second reader to obtain statistics on interreader reliability. Responses were automatically routed for second scoring so that the reader could not discern any difference between a response being presented for first or second scoring. The reliability figures were available to the table leader as soon as scoring began and could be viewed on demand. These figures included a frequency distribution of all second scores for the team and a frequency distribution for all second scores for individual members of the team who scored the item. Ranges for interreader reliability figures for mathematics are reported in Table 7-3. Average reliabilities, given in terms of percentage exact agreement, for each booklet type (spiral, estimation, theme, and advanced) are reported in Table 7-5. This reliability information was also used by the table leader in monitoring the capabilities of all readers and the uniformity of scoring across readers. When scoring was completed for an item, a hard copy of the report was printed for analysis by ETS project staff.

7.3.4.4 Short-Term Trend

The 1996 main assessment of mathematics included a number of items that had previously been used in the 1990 and 1992 assessments. A list of these items is included in the tables in Appendix J. For these items, the trainers used the same scoring guides and training sets as in 1992. Also, for those items that originated in the 1992 assessment, 750 responses from 1992 were scanned into the scoring system and rescored. For those items that originated in the 1990 assessment, the sample included 750 responses each from 1990 and 1992. Table leaders used the management tools to distribute the rescore responses out among the current-year material. Because the rescore responses could not be seeded into the current materials, the table leaders divided the rescore material into thirds, inserting about 25 responses per scorer at various intervals during scoring. Since the development group had loaded the original scores

from the previous years, the system was able to give real-time comparisons of scores. Table leaders, trainers, and the NCS and ETS subject area specialists monitored the interreader agreement rates and t-tests across years for all items used in the short-term trend.

7.3.4.5 Paper Scoring

Some mathematics items could not be scored on the image system because of printing problems or difficulties with overlay templates. When scoring these items, professional readers coded any response that could not be scored by viewing the image with a designated code such as “8” or “pull.” The development group created a list of all responses coded for pulling, and the project coordinator led the effort to pull these booklets from the warehouse. The booklets were then transported to the PSC, where professional readers scored them using the same scoring guides, but marking their scores on scannable scoring sheets. Clerical support staff then scanned the sheets and uploaded the file to the mainframe, where the development team merged the data with the image scoring data.

As soon as the last item on any score sheet was completed, the score sheets were collected and taken to a central clerical support area to be scanned on the NCS paper-based scoring system using OpScan 5 scanners. As each sheet was processed, the scanning system compared the incoming data with tables to ensure that all responses were scored with one and only one valid score, and that only raters who were qualified to score an item were allowed to score it. Any discrepancies (e.g., no score assigned, double gridding, out-of-range scores, or invalid scorer ID numbers) were flagged and resolved before the data from that sheet were accepted into the scoring system. Interreader agreement reports were generated twice a day.

All the scoring data were stored on personal computers at NCS after all the responses for a subject area had been scanned. Upon completion of scoring, the scanner operator ran a query which compared the sheets scanned with a table of records in the system to make sure that all score sheets were accounted for. Once all edits were corrected, the personal computer file was renamed and put into an export file, which automatically created the mainframe file. This file was then uploaded to the mainframe to be merged with the mainframe student files.

7.3.4.6 Large Print

To accommodate students with visual impairment, field administrators had Braille and large print versions of designated booklets available to sampled students who qualified for them. The scoring center received no Braille material for scoring. Two large print booklets were received back for scoring. One of the large print booklets was completely blank. The NCS performance assessment specialist in charge of mathematics scored all the items in the other booklet and gave the scores to the development team to enter into the database.

7.3.4.7 Bilingual Scoring

Some students who participated in the main fourth- and eighth-grade mathematics assessments received bilingual booklets. All students who used bilingual booklets received the equivalent of Booklet 121 for their grade. The same blocks with the same items appeared in the bilingual booklets as in the regular Booklet 121, with the exception that the bilingual booklets contained both Spanish and English versions of each item. The items appeared in Spanish on the left-hand pages and in English on the right-

hand pages. The instructions told students to answer according to their preference. Altogether, the PSC scored 91 bilingual booklets at grade 4 and 36 at grade 8.

Because of the small number of booklets involved, the PSC selected a team of four readers to score all of the booklets on paper. The team consisted of two males and two females. Two of the scorers were born in Chile, South America, and raised partly in Chile and partly in the United States. Both of them have traveled and recently lived extended periods of time in South America, while making the United States their permanent home. The third scorer comes from a Panamanian-American background. Born and raised in a bilingual family, he holds a bachelor's degree in Spanish, has lived extended periods of time in Panama and Mexico, and has experience teaching university-level Spanish as well as working several years in a business position that required daily telephone and correspondence contact with Spanish-speaking clientele, mainly from Puerto Rico. The fourth scorer, while not a native speaker of Spanish, holds a master's degree in Spanish, has traveled and studied extensively in Mexico, the Caribbean and South America, coordinated an adult education program for Spanish-speaking immigrants from Mexican and Central American backgrounds for three years, taught university-level Spanish for eight years, and worked in customer relations positions for four years dealing with Spanish-speaking clientele, mainly from Puerto Rico.

The NCS mathematics specialist trained all four readers on the fourth-grade items on Thursday, May 2, 1996, and the team scored all of the fourth-grade booklets the same day. On Friday, May 3, the team learned the eighth-grade items, many of which overlapped with the fourth-grade items, and scored all of the eighth-grade booklets on that day. Twenty-five percent of the booklets were scored by a second reader to measure interreader agreement.

The team applied the same scoring guides that were used for regular scoring. However, two scores were assigned for each item, one for the Spanish side and one for the English side. Since most students answered relatively consistently on either one side or the other, most booklets received on-task scores for one language and blanks for the other language. This procedure will allow analysts to separate the data of those students who answered in English from those who answered in Spanish. Several students wrote their answers in English on the side of the page where the item was written in Spanish. In these cases, the scorers coded the score as a Spanish answer since that is where the student read the item and wrote the answer. During the course of scoring, the team noted that over half of the fourth-grade booklets came from the same school, and not a single student from that school answered any questions in Section 3 of the booklet.

The same scanning procedures were performed as outlined in Section 7.3.4.5.

7.3.5 Main NAEP Science

The science portion of the 1996 NAEP included a total of 374 discrete constructed-response items (see Table 7-1). It was scored over three segments and two shifts (see Table 7-2). Many kinds of constructed-response items were utilized in the assessment to measure different elements of students' conceptual understanding of scientific material as well as their practical reasoning ability. The items scored included short-answer constructed responses and extended constructed responses. Each constructed-response item had a unique scoring guide that identified the range of possible scores for the item and defined the criteria to be used in evaluating student responses.

During the course of the project, each team scored short constructed-response items using a scale that allowed for partial credit as follows:

1 = incorrect response
2 = partial understanding
3 = correct response

The readers scored extended constructed-response items on a scale of "1" to "4" as follows:

1 = incorrect response
2 = minimal understanding
3 = satisfactory level of comprehension
4 = correct response

7.3.5.1 Training

The training on each item was conducted by science specialists from ETS and NCS. The first teams began training on March 18, 1996. Other teams were phased in throughout the project. Hands-on items were scored a block at a time with a unique scoring guide for each item because of the related nature of the items. The rest of the assessment was scored item-by-item so that each reader worked on only one set of rubrics at a time. After scoring all available responses, a team would then proceed with training and scoring the next item. Scoring was completed on June 7, 1996. Table 7-2 gives detailed information on the dates of scoring and the number of readers and table leaders.

Training involved explaining the item and its scoring guide to the team and discussing responses that represented the various score points in the guide. Typically, two or three anchor responses were chosen for each score point. During this stage, readers and the table leader kept notes of scoring decisions. The table leader was then responsible for compiling those notes and ensuring that all readers were in alignment. When review of the anchor packet was completed, the readers scored and discussed 10 to 20 pre-scored "practice papers" that represented the entire range of score points the item could receive. After the trainer and table leader determined that the team had reached consensus, the table leader then released work on the image-scoring system to the readers. The readers would initially take turns reading their first "live" responses to the team or work in pairs as a final check before beginning work individually. Once the practice session was completed, the formal scoring process began.

7.3.5.2 Scoring

All scoring for science was conducted via the image-based scoring system. During scoring, the team leaders continued to compile notes on scoring decisions for the readers' reference and guidance. Additionally, table leaders closely monitored interreader reliability using both team and individual statistics as a reference. Consistently throughout the scoring of each item, the table leaders also performed backreading duties in which they reviewed a sample of the responses scored by each reader on the team. Lead scorers selected for their experience and accuracy in scoring assisted the table leaders in backreading. The table leaders and performance assessment specialist continuously monitored the progress of each team and noted all scoring-related decisions to ensure that training and scoring progressed smoothly and in a timely manner.

Table 7-6

Science Constructed Responses Scored

Grade	Data	Type[1]				Assessment Proportions	
		Regular	Hands-On	Advanced	Grand Total	1996 National	1996 State
4	Unique items	70	24	0	94	74.8%	25.2%
	Responses first scored	200,319	94,004	0	294,323	220,271	74,052
	Responses second scored	50,080	23,501	0	73,581	55,068	18,513
	First and second scored	250,399	117,505	0	367,904	275,339	92,565
	Average % exact agreement	93.9	93.8	N/A	93.9		
8	Unique items	94	31	0	125	9.7%	90.3%
	Responses first scored	2,157,377	976,844	0	3,134,222	304,020	2,830,202
	Responses second scored	129,443	58,611	0	188,053	18,241	169,812
	First and second scored	2,286,820	1,035,455	0	3,322,275	322,261	3,000,014
	Average % exact agreement	93.4	95.0	N/A	93.8		
12	Unique items	94	26	36	156		
	Responses first scored	198,563	75,120	88,166	361,849		
	Responses second scored	49,641	18,780	22,041	90,462		
	First and second scored	248,204	93,900	110,207	452,311		
	Average % exact agreement	93.0	94.4	94.5	93.6		
Total	Unique items	258	81	36	375		
	Responses first scored	2,556,260	1,145,968	88,166	3,790,394		
	Responses second scored	229,163	100,892	22,041	352,096		
	First and second scored	2,785,423	1,246,860	110,207	4,142,490		
	Average % exact agreement	93.4	94.4	94.5	93.7		

[1] Regular and hands-on tasks include state and national constructed responses.

One advantage of utilizing an image-based scoring system is the ability to construct reader aids to simplify scoring, thus increasing reader reliability. Prior to the start of the project, the ETS subject area specialist and the NCS performance assessment specialist identified several items for the construction of overlays. Overlays serve as templates to define boundaries in which correct responses must be located or allow the placement of correct answers directly on the displayed image, and are displayed along with the student response. A schematic representation of each overlay was included with the scoring guide and sample papers for these items to familiarize readers with the use of the scoring aids during training.

General information on the number of constructed responses scored can be found in Table 6-1 in Chapter 6. Table 7-6 gives more detailed information by grade and booklet type (spiral and advanced). The codes that were used for unscorable science items were:

0 = Blank or random marks
8 = Completely cross-out or erased
9 = "I don't know," refusal, off-task, illegible or language other than English that could not be translated

7.3.5.3 Reliability

A minimum of 25 percent of the science responses for items involved only in the national sample and 6 percent of the responses for items involved in the state samples were scored by a second reader to obtain statistics on interreader reliability. Ranges for interreader reliability for science are reported in Table 7-3. Average reliabilities, given in terms of percentage exact agreement, for each booklet type (spiral and advanced) are reported in Table 7-6. This reliability information was also used by the team leaders to monitor the capabilities of all readers and maintain uniformity of scoring across readers. Reliability reports could be generated on demand by the table leader, scoring specialist, or performance assessment specialist when needed, and they were displayed at a computer workstation. In addition to the immediate feedback provided by the on-line reliability reports, each table leader could also review the actual responses scored by a reader by using the backreading tool. In this way, the table leader monitored each reader carefully and corrected difficulties in scoring almost immediately with a high degree of efficiency.

Chapter 8

CREATION OF THE DATABASE, QUALITY CONTROL OF DATA ENTRY, AND CREATION OF THE DATABASE PRODUCTS[1]

John J. Ferris, Katharine E. Pashley,
David S. Freund, and Alfred M. Rogers
Educational Testing Service

8.1 INTRODUCTION

The data processing, scoring and editing procedures described in Chapter 6 resulted in the generation of disk and tape files containing various data for students (assessed and excluded), teachers, schools, and SD/LEP (students with disabilities/students with limited English proficiency) information. The weighting procedures described in Chapter 10 resulted in the generation of data files that included the sampling weights required to make valid statistical inferences about the population from which the 1996 fourth-, eighth- and twelfth-grade NAEP samples were drawn. These files were merged into a comprehensive, integrated database. The creation of the database is described in Section 8.2.1.

Section 8.2.2 describes a central repository or master catalog of this information. The master catalog is accessible by all analysis and reporting programs and provides correct parameters for processing the data fields and consistent labeling for identifying the results of the analyses.

To evaluate the effectiveness of the quality control of the data entry process, the corresponding portion of the final integrated database was verified in detail against a sample of the original instruments received from the field. The results of this procedure are given in Section 8.3.

The integrated database was the source for the creation of the NAEP item information database and the NAEP secondary-use data files. These are described in Section 8.4.

8.2 CREATION OF THE DATABASE

8.2.1 Merging Files

The data processing conducted by National Computer Systems (NCS) resulted in the transmittal to ETS of four data files for each of fourth, eighth and twelfth grade: one for the student background and item response data and one file for each of the three questionnaires (Teacher Questionnaire, School Characteristics and Policies Questionnaire, and SD/LEP Questionnaire). The sampling weights, derived by Westat, Inc., comprised additional files for each grade—two sets for assessed students, two sets for excluded students and for schools four sets at grade 4, five at grade 8, and six at grade 12. (See Chapter 10 for a discussion of the sampling weights.) These files at each grade were the foundation for the

[1] John J. Ferris was responsible for the evaluation of the quality of the database and the data entry process; Katharine E. Pashley was responsible for database generation under the supervision of David S. Freund; Alfred M. Rogers created the secondary-use data files.

analysis of the 1996 NAEP data. Before data analyses could be performed, these data files had to be integrated into a coherent and comprehensive database.

The database ultimately comprised four files per cohort: three student files (mathematics, science, and long-term trend) and a single school file. The student files were separated by subject area to improve maintenance and efficiency of the databases and data analyses. Each record on the student file contained a student's responses to the particular assessment booklet the student was administered (in the case of excluded students, a booklet was assigned but the student response fields contain a special code indicating no response), and the information from the questionnaire that the student's teacher completed. Additionally, for a student (assessed or excluded) who was identified as a student with a disability (SD) or of limited English proficiency (LEP), the data from the SD/LEP Questionnaire are included. This questionnaire is filled out for all students identified as SD and/or LEP, both assessed and excluded. (See Chapter 2 for information regarding assessment instruments.) Also added to the student files were variables with school-level information supplied by Quality Education Department, Inc. (QED) including demographic information about schools such as race/ethnicity percentages. Since the teacher data are not from a representative sample of teachers and since the focus of NAEP is to report student level results, the teacher response data were added to the student records. The school data were on separate files that could be analyzed on their own and could also be linked to the student files through the unique school ID code.

The creation of the student data files for fourth, eighth, and twelfth grade began with the reorganization of the data files received from NCS. This involved two major tasks:

1. the files were restructured, eliminating unused (blank) areas to reduce the size of the files; and

2. in cases where students had chosen not to respond to an item, the missing responses were recoded as either "omit" or "not reached," as discussed in Part II of this report.

Next, the student response data were merged with the student weights files. The resulting file was then merged with the SD/LEP and teacher data. In all merging steps, the 10-digit booklet ID (the three-digit booklet number common to every booklet with the same block of items, a six-digit serial number unique to the booklet a student was given and a single check digit, distinguishing bilingual booklets) was used as the matching criterion.

The school file for each grade was created by merging the School Characteristics and Policies Questionnaire file with the file of school weights and school variables, supplied by Westat. The primary sampling unit (PSU) and school codes were used as the matching criteria. Since some schools did not return a questionnaire, some of the records in the school file contained only school-identifying information and sampling weight information.

When the student and school files for each grade had been created, the database was ready for analysis. In addition, whenever new data values, such as composite background variables or plausible values, were derived, they were added to the appropriate database files using the same matching procedures described above.

For archival purposes and to provide data for outside users, restricted-use data files and codebooks for each jurisdiction were generated from this database. The restricted-use data files contain all responses and response-related data from the assessment, including responses from the student booklets, Teacher Questionnaires, and School Characteristics and Policies Questionnaires, scale scores, sampling weights, and variables used to compute standard errors.

8.2.2 Creating the Master Catalog

A critical part of any database is its processing control and descriptive information. Having a central repository for this information, which may be accessed by all analysis and reporting programs, will provide correct parameters for processing the data fields and consistent labeling for identifying the results of the analyses. The NAEP master catalog file was designed and constructed to serve these purposes for the NAEP database.

Each record of the master catalog contains the processing, labeling, classification, and location information for each data field in the NAEP database. The control parameters are used by the access routines in the analysis programs to define the manner in which the data values are to be transformed and processed.

Each data field has a 50-character label in the master catalog describing the contents of the field and, where applicable, the source of the field. The data fields with discrete or categorical response values (e.g., multiple-choice items and professionally scored items, but not weight fields) have additional label fields in the catalog containing 8- and 20-character labels for those response values. These short labels can be used for reporting purposes as a concise description of the responses for the cognitive items.

The classification area of the master catalog record contains distinct fields corresponding to predefined classification categories (e.g., mathematics content and process areas) for the data fields. For a particular classification field, a nonblank value indicates the code of the subcategory within the classification category for the data field. This classification area permits the grouping of identically classified items or data fields by performing a selection process on one or more classification fields in the master catalog.

According to NAEP design, it is possible for item data fields to appear in more than one student sample and in more than one block within each sample. The location fields of the catalog record contain age, block and, where applicable, the sequence within the block for each appearance of the data field.

The master catalog file was constructed concurrently with the collection and transcription of the State Assessment data so that it would be ready for use by analysis programs when the database was created. As new data fields were derived and added to the database, their corresponding descriptive and control information were entered into the master catalog. The machine-readable catalog files are available as part of the secondary-use data files package for use in analyzing the data with programming languages such as SAS and SPSS-X (see the *NAEP 1996 Secondary-Use Data Files User Guide,* Rogers, Kline, & Schoeps, 1999).

8.3 QUALITY CONTROL OF NAEP DATA ENTRY FOR 1996

This section describes the evaluation of the data entry process for the 1996 national assessment. As in past years, the NAEP database was found to be more than accurate enough to support the analyses that were done. Overall, the observed error rates were comparable to those of past assessments, with the possible exception of the Teacher Questionnaire data (see discussion below); they ranged from three errors per 10,000 responses for the SD/LEP Student Questionnaire data to 33 errors per 10,000 responses for the Teacher Questionnaire data.

The purpose of the analysis reported in this sction is to assess the quality of the data resulting from the complete data entry system, beginning with the actual instruments collected in the field and ending with the final machine-readable database used in the analyses. The process involved the selection of instruments at random from among those returned from the field and the comparison of these instruments, character by character, with their representations in the final database. In this way, we were able to measure the error rates in the data as well as the success of the data entry system.

Of course the observed error rate cannot be taken at face value. For example, the sample of teacher questionnaires that happened to be selected for close inspection contained 22 errors out of a total of 6,741 characters. To conclude that the entire teacher questionnaire database has an error rate of 22/6741, or .0033, would be too optimistic; we may simply have been lucky (or unlucky) with this particular random sample. What is needed is an indication of how bad the true error rate might be, given what we observed. Such an indication is provided by confidence limits. Confidence limits indicate how likely it is that a value falls inside a specified range in a specified context or distribution. In our analysis, the specified range is an error rate between zero and some maximum value beyond which we are confident at a specified level (traditionally 99.8%) that the true error rate does not lie; the specified context or distribution turns out to be the cumulative binomial probability distribution. An example should demonstrate this technique:

> Let us say that 1,000 booklets were processed, each with 100 characters of data transcribed for a total of 100,000 characters. Let us say further that five of these characters were discovered to be in error in a random sample of 50 booklets that were completely checked; in other words, five errors were found in a sample of 5,000 characters. The following expression may be used to establish the probability that the true error rate is .0025 or less, rather than the single-value estimate of the observed rate of one in a thousand (.001):

$$\sum_{j=0}^{5} \binom{5000}{j} \times .0025^{j} \times (1-.0025)^{(5000-j)} = .0147$$

> This is the sum of the probability of finding five errors plus the probability of finding four errors plus. . . etc. . . plus the probability of finding zero errors in a sample of 5,000 with a true error rate of .0025; that is, the probability of finding five or fewer errors by chance when the true error rate is .0025. Notice that we did not use the size of the database in this expression. Actually, the assumption here is that our sample of 5,000 was drawn from a database that is infinite. The smaller the actual database is, the more confidence we can have in the observed error rate; for example, had there been only 5,000 in the total database, our sample would have included all the data, and the observed error rate would have been the true error rate. The result of the above computation allows us to say, conservatively, that .0025 is an upper limit on the true error rate with 98.53 percent (i.e., 1 - .0147) confidence; that is, we can be quite sure that our true error rate is no larger than .0025.

Virtually all of the data collected for this assessment were machine-scanned. The only exception was a set of six booklets used for the long-term trend reading and writing assessments; the format of these booklets was kept the same for comparability with earlier assessments, so these booklets had to be key-entered. As it happened, no errors at all were found in the sample of key-entered booklets selected for quality control.

In the 1994 and 1996 assessments, the selection of booklets for this comparison took place at the point of first entry into the recording process for data from the field. In earlier assessments, this selection took place only after data had reached the final database, in order to assure that only relevant booklets were involved in the quality control evaluation. The new selection process involves the risk that booklets will be selected that ultimately will not appear in the final database, however, as in 1994, sufficient numbers of booklets were in fact selected.

The individual instruments are briefly discussed in the following sections and a summary table (Table 8-1) gives the upper 99.8 percent confidence limit for the error rate for each of the instruments as well as the sampling information. The 99.8 percent confidence limit, and the selection rates noted, were chosen to make these results comparable to those of previous administrations when the same parameters were used.

Table 8-1
Summary of Quality Control Error Analysis for NAEP 1996 Data Entry

	Main Assessment Student Booklets	Long-Term Trend Assessment Student Booklets	SD/LEP Student Questionnaires	Teacher Questionnaires	School Characteristics and Policies Questionnaires
Selection Rate	1/392	1/372	1/112	1/120	1/57
Different Booklets	203	26	2	4	3
Number of Booklets Sampled	240	101	103	38	42
Number of Characters Sampled	30,134	12,082	13,098	6,741	6,633
Number of Errors	16	6	4	22	15
Observed Error Rate	.0005	.0005	.0003	.0033	.0023
Upper 99.8% Confidence Limit	.0011	.0015	.0011	.0058	.0045

8.3.1 Student Data

Data from about 94,000 students were processed across all samples in this assessment. Across all the student data, roughly one booklet in 392 was selected for close examination, which is comparable to the one in 400 target selection rate used in past assessments. The student data error rates were consistently low in all subject areas and across all three grades. The overall quality of the data was very high.

Data from some 30,000 additional students were also used in the study of long-term trends. These data showed the same consistently low error rates, as indicated in the accompanying table.

8.3.2 SD/LEP Student Questionnaire Data

In this assessment, 13,098 SD/LEP questionnaires were scanned. The quality control sampling rate was one in 112, a somewhat higher rate than that used in previous assessments. The data showed a somewhat lower error rate than in previous assessments—comparable to that for the student data. The few problems encountered involved the scanner's mistaking an erasure for a genuine response or failing to identify a multiple response as such.

8.3.3 Teacher Questionnaire Data

In this assessment, 4,585 teacher questionnaires were collected and scanned. About one percent of these questionnaires was sampled for the quality control procedure. The error rates for these questionnaires were higher than those of any other category of data, and also higher than those found for teachers in past assessments. There was some evidence that the questionnaire layout was confusing to at least some of the teachers in the part of the questionnaire that was intended to collect class period information. While the majority of teachers did not have difficulty with this, consideration is being given to a possible revision of the questionnaire, since a number of teachers also had similar problems in the NAEP State Assessment.

8.3.4 School Characteristics and Policies Questionnaire Data

In this assessment, 2,404 School Characteristics and Policies Questionnaires were collected. They were sampled at a rate of about 1 in 57. Fifteen scanning errors were found in these questionnaires, which included both regular and long-term trend schools. Most of these errors came from a single booklet that had been filled out in pen—something that frequently gives a scanner trouble.

8.4 NAEP DATABASE PRODUCTS

The NAEP database described to this point serves primarily to support analysis and reporting activities that are directly related to the NAEP contract. This database has a singular structure and access methodology that is integrated with the NAEP analysis and reporting programs. One of the directives of the NAEP contract is to provide secondary researchers with a nonproprietary version of the database that is portable to any computer system. In the event of transfer of NAEP to another client, the contract further requires ETS to provide a full copy of the internal database in a format that may be installed on a different computer system.

In fulfillment of these requirements, ETS provides two sets of database products: the item information database and the secondary-use data files. The contents, format and usage of these products are documented in the publications listed under the appropriate sections below.

8.4.1 The Item Information Database

The NAEP item information database contains all of the descriptive, processing, and usage information for every assessment item developed and used for NAEP since 1970. The primary unit of this database is the item. Each NAEP item is associated with different levels of information, including usage

across years and age cohorts, subject area classifications, response category descriptors, and locations of response data on secondary-use data files.

The item information database is used for a variety of essential NAEP tasks: providing statistical information to aid in test construction, determining the usage of items across assessment years and ages for trend and "main" analyses, labeling summary analyses and reports, and organizing items by subject area classifications for scaling analysis.

The creation, structure, and use of the NAEP item information database for all items used up to and including the 1996 assessment are fully documented in the NAEP publications *A Guide to the NAEP Item Information Database* (Rogers, Barone, & Kline, 1996) and *A Primer for the NAEP Item Information Database* (Rogers, Kline, Barone, Mychajlowycz, & Forer, 1989).

The procedures used to create the 1996 version of the item information database are the same as those documented in the guide. The version of the guide contains the subject area classification categories for the cognitive items.

8.4.2 The Secondary-Use Data Files

The secondary-use data files are designed to enable any researcher with an interest in the NAEP database to perform secondary analysis on the same data as those used at ETS. The data, documentation and supporting files are distributed on CD-ROM media. For each sample in the assessment, the following files are provided: the response data file; a printable codebook file; a file of control statements that will generate an SPSS system file; a file of control statements that will generate a SAS system file; and a machine-readable catalog file. Each codebook file is in portable document file (PDF) format, which may be browsed, excerpted and printed using the Adobe Acrobat Reader program on a variety of platforms. Each machine-readable catalog file contains sufficient control and descriptive information to permit the user who does not have either SAS or SPSS to set up and perform data analysis.

The remainder of this section summarizes the procedures used in generating the data files and related materials.

8.4.2.1 File Definition

The design of the 1996 assessment perpetuates two features of the 1990, 1992, and 1994 assessment design: the focused-BIB booklet design and the direct matching of teacher questionnaires to student assessment instruments. In addition, the sample of students who were excluded from the assessment is now incorporated into the appropriate assessed student subject area sample.

The focused-BIB design within the main assessment isolates the primary subject areas to separate groups of booklets. This permits the division of the main sample into subject-specific subsamples. The data files generated from these subsamples need only contain the data that are relevant to their corresponding subject areas and are therefore smaller and more manageable than their counterparts in previous assessments.

According to the design of the 1984, 1986, and 1988 assessments, only a sample of the teachers of the assessed students were asked to fill out the teacher questionnaires. The large size of the secondary-use main student files and the relatively low matching rate between students and teachers

made it impractical if not physically prohibitive to produce a complete file with student and teacher information. Both the 1984 and 1986 secondary-use data packages had separate teacher data files which could be linked to the student data files for analysis. The teacher file in the 1988 secondary-use data package contained not only the teacher response data, but also the data from the students who could be matched to teacher questionnaires. This type of file was more appropriate for the analysis of teacher data because it defined the student as the unit of observation.

The intent of the 1996 assessment design was to collect data from mathematics or science teachers of the main assessment students at specified grade levels who were administered mathematics or science booklets. A portion of the teacher questionnaire contained questions that were directly related to each matched student. This change in the design afforded a very high matching rate between student and teacher data. Therefore, for those subject areas in each grade cohort for which teacher data were collected, the teacher responses were appended to each student record in the secondary-use data files.

8.4.2.2 Definition of the Variables

The initial step in the variable definition process was the generation of a LABELS file of descriptors of the variables for each data file to be created. Each record in a LABELS file contains, for a single data field, the variable name, a short description of the variable, and processing control information to be used by later steps in the data generation process. This file could be edited for deletion of variables, modification of control parameters, or reordering of the variables within the file. The LABELS file is an intermediate file only; it is not included on the released data files.

The variables on all data files are grouped and arranged in the following order: identification information, weights, derived variables, proficiency scores (where applicable), and response data. On the student data files, these fields are followed by the teacher response data and the SD/LEP student questionnaire data, where applicable. The identification information is taken from the front covers of the instruments. The weight data include sample descriptors, selection probabilities, and replicate weights for the estimation of sampling error. The derived data include sample descriptions from other sources and variables that are derived from the response data for use in analysis or reporting.

For each subject area of the main assessment, the item response data within each block were left in their order of presentation. The blocks, however, were arranged according to the following scheme: common background, subject-related background, the cognitive blocks in ascending numerical order, and student motivation. The responses to cognitive blocks that were not present in a given booklet were left blank, signifying a condition of “missing by design.”

In order to process and analyze the spiral sample data effectively, the user must also be able to determine, from a given booklet record, which blocks of item response data were present and their relative order in the instrument. This problem was remedied by the creation of a set of control variables, one for each block, which indicated not only the presence or absence of the block but its order in the instrument. These control variables are included with the derived variables.

8.4.2.3 Data Definition

To enable the data files to be processed on any computer system using any procedural or programming language, it was desirable that the data be expressed in numeric format. This was possible, but not without the adoption of certain conventions for reexpressing the data values.

During creation of the NAEP database, the responses to all multiple-choice items were transcribed and stored in the database using the letter codes printed in the instruments. This scheme afforded the advantage of saving storage space for items with 10 or more response options, but at the expense of translating these codes into their numeric equivalents for analysis purposes. The response data fields for most of these items would require a simple alphabetic-to-numeric conversion. However, the data fields for items with 10 or more response choices would require "expansion" before the conversion, since the numeric value would require two column positions. One of the processing control parameters on the LABELS file indicates whether or not the data field is to be expanded before conversion and output.

The ETS database contained special codes to indicate certain response conditions: "I don't know" responses, multiple responses, omitted responses, not-reached responses, and unresolvable responses, which include out-of-range responses and responses that were missing due to errors in printing or processing. The scoring guides for the mathematics and science constructed-response items included additional special codes for ratings of "illegible," "I don't know," "off task," or non-rateable by the scorers. All of these codes had to be reexpressed in a consistent numeric format.

The following convention was adopted and used in the designation of these codes: The "illegible" response codes were converted to 5; the "off task" response codes were converted to 6; the "I don't know" and non-rateable response codes were converted to 7; the "omitted" response codes were converted to 8; the "not reached" response codes were converted to 9; and the multiple response codes were converted to 0; and. The out-of-range and missing responses were coded as blank fields, corresponding to the "missing by design" designation.

This coding scheme created conflicts for those multiple-choice items that had seven or more valid response options as well as the "I don't know" response and for those constructed-response items whose scoring guide had five or more categories. These data fields were also expanded to accommodate the valid response values and the special codes. In these cases, the special codes were "extended" to fill the output data field: the "I don't know" and non-rateable codes were extended from 7 to 77, omitted response codes from 8 to 88, etc.

Each numeric variable on the secondary-use files was classified as either continuous or discrete. The continuous variables include the weights, proficiency scores, identification codes, and item responses where counts or percentages were requested. The discrete variables include those items for which each numeric value corresponds to a response category. The designation of "discrete" also includes those derived variables to which numeric classification categories have been assigned. The constructed-response items were treated as a special subset of the discrete variables and were assigned to a separate category to facilitate their identification in the documentation.

8.4.2.4 Data File Catalogs

The CATALOG file is created by the GENCAT program from the LABELS file and the 1996 master catalog file. Each record on the LABELS file generates a CATALOG record by first retrieving the master catalog record corresponding to the field name. The master catalog record contains usage, classification, and response code information, along with positional information from the LABELS file: field sequence number, output column position, and field width. Like the LABELS file, the CATALOG file is an intermediate file and is not included on the released data files.

The information for the response codes, also referred to as “foils,” consists of the valid data values for the discrete numeric fields, and a 20-character description of each. The GENCAT program uses additional control information from the LABELS file to determine if extra foils should be generated and saved with each CATALOG record. The first flag controls generation of the “I don’t know” or non-rateable foil; the second flag regulates omitted or not-reached foil generation; and the third flag denotes the possibility of multiple responses for that field and sets up an appropriate foil. All of these control parameters, including the expansion flag, may be altered in the LABELS file by use of a text editor, in order to control the generation of data or descriptive information for any given field.

The LABELS file supplies control information for many of the subsequent secondary-use data processing steps. The CATALOG file provides detailed information for those and other steps.

8.4.2.5 Data File Layouts

The data file layouts were the first user product to be generated in the secondary-use data files process. The generation program, GENLYT, used a CATALOG file as input and produced a printable file. The LAYOUT file is little more than a formatted listing of the CATALOG file.

Each line of the LAYOUT file contains the following information for a single data field: sequence number, field name, output column position, field width, number of decimal places, data type, value range, key or correct response value, and a short description of the field. The sequence number of each field is implied from its order on the LABELS file. The field name is an 8-character label for the field that is to be used consistently by all secondary-use data files materials to refer to that field on that file. The output column position is the relative location of the beginning of that field on each record for that file, using bytes or characters as the unit of measure. The field width indicates the number of columns used in representing the data values for a field. If the field contains continuous numeric data, the value under the number of decimal places entry indicates how many places to shift the decimal point before processing data values.

The data type category uses five codes to designate the nature of the data in the field: Continuous numeric data are coded “C;” discrete numeric data are coded “D;” constructed-response item data are coded “OS;” if the item was dichotomized for scaling and “OE;” if it was scaled under a polytomous response model. Additionally, the discrete numeric fields that include “I don’t know” response codes are coded “DI.” If the field type is discrete numeric, the value range is listed as the minimum and maximum permitted values separated by a hyphen to indicate range. If the field is a response to a scorable item, the correct option value, or key, is printed; if the field is an assigned score that was scaled as a dichotomous item using cut point scoring, the range of correct scores is printed. Each variable is further identified by a 50-character descriptor.

8.4.2.6 Data Codebooks

The data codebook is a printed document containing complete descriptive information for each data field. Most of this information originates from the CATALOG file; the remaining data comes from the COUNTS file and the IRT parameters file.

Each data field receives at least one line of descriptive information in the codebook. If the data type is continuous numeric, no more information is given. If the variable is discrete numeric, the codebook lists the foil codes, foil labels, and frequencies of each value in the data file. Additionally, if

the field represents an item used in IRT scaling, the codebook lists the parameters used by the scaling program.

Certain blocks of cognitive items in the 1996 assessment that are to be used again in later assessments for trend comparisons have been designated as nonreleased. In order to maintain their confidentiality, generic labels have been substituted for the response category descriptions of these items in the data codebooks and the secondary-use files.

The frequency counts are not available on the catalog file, but must be generated from the data. The GENFREQ program creates the COUNTS file using the field name to locate the variable in the database, and the foil values to validate the range of data values for each field. This program also serves as a check on the completeness of the foils in the CATALOG file, as it flags any data values not represented by a foil value and label.

The IRT parameter file is linked to the CATALOG file through the field name. Printing of the IRT parameters is governed by a control flag in the classification section of the CATALOG record. If an item has been scaled for use in deriving the proficiency estimates, the IRT parameters are listed to the right of the foil values and labels, and the score value for each response code is printed to the immediate right of the corresponding frequency.

The LAYOUT and CODEBOOK files are written by their respective generation programs to print-image disk data files. Draft copies are printed and distributed for review before the production copy is generated. The production copy combines the LAYOUT and CODEBOOK files for each sample in a portable document file (PDF) format. This file may be browsed, excerpted and printed using the Adobe Acrobat Reader program on a variety of platforms and operating systems.

8.4.2.7 Control Statement Files for Statistical Packages

An additional requirement of the NAEP cooperative agreement is to provide, for each secondary-use data file, a file of control statements each for the SAS and SPSS statistical systems that will convert the raw data file into the system data file for that package. Two separate programs, GENSAS and GENSPX, generate these control files using the CATALOG file as input.

Each of the control files contains separate sections for variable definition, variable labeling, missing value declaration, value labeling, and creation of scored variables from the cognitive items. The variable definition section describes the locations of the fields, by name, in the file, and, if applicable, the number of decimal places or type of data. The variable label identifies each field with a 50-character description. The missing value section identifies values of those variables that are to be treated as missing and excluded from analyses. The value labels correspond to the foils in the CATALOG file. The code values and their descriptors are listed for each discrete numeric variable. The scoring section is provided to permit the user to generate item score variables in addition to the item response variables.

Each of the code generation programs combines three steps into one complex procedure. As each CATALOG file record is read, it is broken into several component records according to the information to be used in each of the resultant sections. These record fragments are tagged with the field sequence number and a section sequence code. They are then organized by section code and sequence number. Finally, the reorganized information is output in a structured format dictated by the syntax of the processing language.

The generation of the system files accomplishes the testing of these control statement files. The system files are saved for use in special analyses by NAEP staff. These control statement files are included on the distributed data files to permit users with access to SAS and/or SPSS to create their own system files.

8.4.2.8 Machine-Readable Catalog Files

For those NAEP data users who have neither SAS nor SPSS capabilities, yet require processing control information in a computer-readable format, the distribution files also contain machine-readable catalog files. Each machine-readable catalog record contains processing control information, IRT parameters, and foil codes and labels.

8.4.2.9 NAEP Data on Disk

The complete set of secondary-use data files described above are available on CD-ROM as part of the NAEP Data on Disk product suite. This medium can be ideal for researchers and policy makers operating in a personal computing environment.

The NAEP Data on Disk product suite includes two other components which facilitate the analysis of NAEP secondary-use data. The PC-based NAEP data extraction software, NAEPEX, enables users to create customized extracts of NAEP data and to generate SAS or SPSS control statements for preparing analyses or generating customized system files. The NAEP analysis modules, which currently run under SPSS® for Windows™, use output files from the extraction software to perform analyses that incorporate statistical procedures appropriate for the NAEP design.