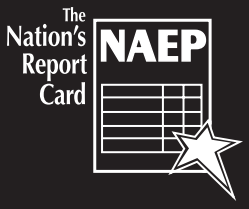


U.S. Department of Education
Institute of Education Sciences
NCES 2005–484

NAEP 1999 Long-Term Trend Technical Analysis Report

Three Decades of Student Performance

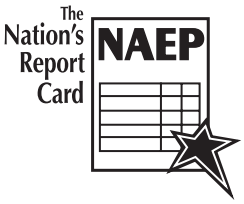
What is The Nation's Report Card?

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



U.S. Department of Education
Institute of Education Sciences
NCES 2005–484

NAEP 1999 Long-Term Trend Technical Analysis Report

Three Decades of Student Performance

April 2005

Nancy L. Allen
Catherine A. McClellan
Joan J. Stoeckel

In collaboration with:
Steven P. Isham
Bruce A. Kaplan
Venus Leung
Jo-Lin Liang
Norma A. Norris
Ingeborg U. Novatkoski
Spencer S. Swinton
Yuxin Tang
Lois H. Worthington
Educational Testing Service

Nancy W. Caldwell
Jean A. Fowler
Andrea R. Piesse
Keith F. Rust
Mark M. Waksberg
Leslie S. Wallace
Westat

Connie R. Smith
NCS Pearson

Arnold A. Goldstein
Project Officer
National Center for Education Statistics

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Statistics
Grover J. Whitehurst
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
U.S. Department of Education
1990 K Street NW
Washington, DC 20006–5651

April 2005

The NCES World Wide Web Home Page is: http://nces.ed.gov
The NCES World Wide Web electronic catalog is: http://nces.ed.gov/pubsearch

SUGGESTED CITATION
Allen, N.L., McClellan, C.A., and Stoeckel, J.J. (2005). *NAEP 1999 Long-Term Trend Technical Analysis Report: Three Decades of Student Performance* (NCES 2005–484). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

For ordering information for this report, write:
U.S. Department of Education
ED Pubs
P.O. Box 1398
Jessup, MD 20794–1398
or call toll-free 1-877-4ED-PUBS, or order online at http://www.edpubs.org

Content contact:
Arnold A. Goldstein
202–502–7344

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

THE NAEP 1999 LONG-TERM TREND TECHNICAL ANALYSIS REPORT

♦ TABLE OF CONTENTS ♦

Appendix D Westat Report: NAEP 1999 Long-Term Trend Data Collection, Sampling and Weighting Report—Continued

THIS PAGE INTENTIONALLY LEFT BLANK.

THE NAEP 1999 LONG-TERM TREND TECHNICAL ANALYSIS REPORT

♦ LIST OF TABLES AND FIGURES ♦

PART ONE INTRODUCTION TO THE NAEP 1999 LONG-TERM TREND ASSESSMENT: DESIGN AND IMPLEMENTATION

PART THREE DATA ANALYSIS FOR THE NAEP 1999 LONG-TERM TREND READING ASSESSMENT

APPENDIX A STATISTICAL SUMMARY OF THE 1999 NAEP SAMPLES—CONTINUED

APPENDIX D WESTAT REPORT: NAEP 1999 LONG-TERM TREND DATA COLLECTION, SAMPLING AND WEIGHTING REPORT–CONTINUED

ACKNOWLEDGMENTS

The design, development, administration, analysis, and reporting of the 1999 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), Educational Testing Service (ETS), Westat, and National Computer Systems (NCS Pearson). This report documents the technical analysis procedures for the 1999 NAEP long-term trend assessment, 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 1999 NAEP long-term trend assessment was funded through NCES, Institute of Education Sciences, in 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: James Carlson, Chris Chapman, Arnold Goldstein, Brent Mast, David Grissmer, Andrew Kolstad, Drew Malizio, Marilyn Seastrom, and Leslie Scott.

ETS management has encouraged high quality work on all NAEP activities. Thanks go to several members of ETS management: President of ETS, Kurt Landgraf; Paul Ramsey, formerly Vice President for the School and College Services Division; Drew Gitomer, formerly Senior Vice President for Research and Development; John Barone, Senior Research Director, Center for Data Analysis Research; and John Mazzeo, Senior Research Director, Center for Large Scale Assessment Research.

The NAEP program development and reporting areas within ETS's Government Research and Assessment Division have been very supportive of NAEP's technical work. Special thanks go to the following staff members in the NAEP program area who provided direct leadership for the NAEP project: Steve Lazer, Executive Director for NAEP; John Mazzeo, formerly Center Director, Large-Scale Assessment; Jay Campbell, Director of NAEP Reporting; and Jeff Haberstroh, Director of NAEP Test Development. Significant contributions to the project were also received from Loretta Casalaina**,** NAEP Publications Manager**.**

The design and data analysis of the 1999 national long-term trend assessment was primarily the responsibility of the NAEP Research and Development staff at ETS with significant contributions from NAEP management, Westat, and NCS staffs. In addition to managing day-to-day data analytic operations, NAEP Large Scale Assessment Research staff members have made many innovative statistical and psychometric contributions. The activities necessary to report results for the assessment were directed by Nancy Allen, John Donoghue, Catherine McClellan, Frank Jenkins, Jo-lin Liang, and Spencer Swinton. Jiahe Qian had responsibility for the 1999 long-term trend assessment of writing for which special analyses were completed, but as mandated by the NAGB, results were not reported. Catherine McClellan (formerly Hombo) not only contributed to the success of this document, but was also a co-author for the *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance* (Campbell, Hombo, and Mazzeo [2000]), the report that contains the results of the analyses described in this document.

The Center for Data Analysis Research at ETS, under the leadership of John Barone, was responsible for developing the operating systems and carrying out the data analyses. David Freund coordinated the analyses presented in this report with assistance from Steve Isham, Bruce Kaplan, Venus Leung, Norma Norris, Ingeborg Novatkoski, Tatyana Petrovicheva, Yuxin Tang, and Lois Worthington. 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. Many other members of this center made important contributions of their time and talent to NAEP data analyses and analysis software and data products, including Jim Ferris, Laura Jerry, Debbie Kline, Gerry Kokolis, Edward Kulick, Phillip Leung, Youn-Hee Lim, Mei-jang Lin, Duanli Yan, and Fred Yan.

The staff at Westat contributed their talents and efforts in all areas of the sample design and data collection. These activities were directed by Nancy Caldwell, Keith Rust, Debra Vivari, and Dianne Walsh. Renee Slobasky was the corporate officer for the project. Particular thanks are due to Yuki Carnes, Rob Dymowski, Jean Fowler, Brice Hart, Sharon Hirabayashi, Prakash Padmanabhan, and Mark Waksberg.

Critical to the program was the contribution of NCS, responsible for the printing, distribution, scoring, and processing activities. The leadership roles of Brad Thayer, Patrick Bourgeacq, Charles Brungardt, Matilde Kennel, Linda Reynolds, and Connie Smith are especially acknowledged.

Special recognition and appreciation go to Joan Stoeckel, editor of this report. She has been responsible for organizing, scheduling, editing, motivating, and ensuring the cohesiveness and correctness of the final report. Jinny Lieberman and Sharon Stewart are acknowledged for their editorial and administrative assistance during the preparation of this report.

There are numerous subject-area, technical advisory, policy-related, and state assessment groups that steer all aspects of the NAEP project. Their work has benefited the project enormously. Most importantly, NAEP is grateful to the students and school staff whose participation made the assessment possible.

Introduction

This report provides an update to the technical analysis procedures documenting the 1996 National Assessment of Educational Progress (NAEP) as presented in *The NAEP 1996 Technical Report* (Allen, Carlson, and Zelenak, 1999). It describes how the 1999 long-term trend data were incorporated into the trend analyses. Since no national main or state assessments were administered in 1999, this report does not contain the comprehensive details related to the general design and analysis issues that arise in NAEP assessments and that are included in the 1996 report.

Parts one and two provide an overview of the NAEP 1999 long-term trend assessment design and analysis, and parts three, four, and five include subject-area specific information. The appendices A, B, and C include statistical sample summaries, IRT parameters, and conditioning variables. Appendix D includes Westat's *NAEP 1999 Long-Term Trend Data Collection, Sampling and Weighting Report* (Caldwell, Fowler, Waksberg, and Wallace, 2002). Appendix E includes sections of the National Computer Systems' report on processing and professional scoring, *NAEP Report of Processing and Professional Scoring Activities: Long-Term Trend 1998-99 Mathematics/Science and Reading/Writing* (National Computer Systems, 2000).

THIS PAGE INTENTIONALLY LEFT BLANK.

Part One

Overview of the NAEP 1999 Long-Term Trend Assessment: Design and Implementation

Nancy L. Allen and Joan J. Stoeckel
Educational Testing Service

1.1 Overview of the NAEP 1999 Long-Term Trend Assessment

As the nation's only long-term assessment of students' educational progress, the National Assessment of Educational Progress (NAEP) is the resource for understanding what students know and can do. Since 1969, NAEP has conducted ongoing nationwide assessments of student achievement in various subject areas including reading, writing, mathematics, science, U.S. history, and world geography. Based on assessment and background questionnaire results, NAEP reports student achievement and relates student achievement to instructional, institutional, and demographic variables.

NAEP has two major goals. First, NAEP must measure student progress over time. Second, NAEP must measure student achievement using assessment instruments that reflect current curriculum content. In order to achieve both goals, the NAEP project encompasses two separate assessment programs. The NAEP long-term trend assessments in reading, writing, mathematics, and science are intended to measure student progress over time; consequently, the long-term trend assessments use assessment instruments and procedures that are as similar as possible across assessment years. The NAEP long-term trend assessments make use of questions (items) from previous assessments beginning in 1969 for science, 1971 for reading, 1973 in mathematics, and 1984 in writing. The long-term trend assessments are different from more recently developed assessments in the same subject areas, referred to as NAEP's ***main*** assessments. The ***main*** assessments reflect changes in educational priorities and advances in assessment methodology. The curriculum frameworks for the ***main*** assessments are developed and updated by the National Assessment Governing Board (NAGB).

The long-term trend assessments, as they were administered in 1999, were developed in the 1980's using items that were first administered during the period from 1969 through the early 1980's. In 1984, Educational Testing Service (ETS) began analyzing the data from the NAEP assessments using item response theory (IRT) and multiple imputations (see section 2.4). At this time, the assessment booklets were fixed as the permanent instruments for the long-term trend assessments so that trends in student achievement could be measured without bias due to different assessment items or different arrangements of assessment items within the booklets. Identical assessment booklets were presented to students six times in science and mathematics (1986, 1990, 1992, 1994, 1996, and 1999), and seven times in reading and writing (1984, 1988, 1990, 1992, 1994, 1996, and 1999). The data from these stable long-term trend booklets were linked (using IRT) with the data from previous NAEP assessments through the items that were common to the earlier assessments. The earliest assessments of mathematics and science had too few items in common with the current long-term trend booklets to link through IRT. Instead, they were connected to the current long-term trend scales using the methodologies described in sections 4.6 and 5.6 respectively.

Despite the use of the same long-term trend booklets for almost a decade, there are differences in the conditions of the long-term trend assessments that could threaten the validity of comparisons made over time. For instance, federal legislation regarding the identification and testing of students with disabilities (SD) and students with limited English proficiency (LEP) has changed over the last decade. Although the criteria used to exclude students from NAEP long-term trend assessments has stayed the same (see section 1.5), the proportions of students who were actually excluded may have changed over time. For this reason, student exclusion rates are reported in table 1–8 so that the reader can evaluate the impact on the reported long-term trend results.

Although every effort has been made to provide information about any factors that could bias the long-term trend results, several possible sources of bias are not described in this document. The administration of the long-term trend assessments took place during comparable time windows each assessment year, and efforts are made to balance the timing of assessment sessions within the testing windows. However, no special examination of variations in test administration timing within the testing windows was undertaken. There are also specific aspects of the scaling of the assessments across the years that are not documented in this report. Most often, items in the assessments were treated in the same way each time they were scaled, but some items were treated differently in the analysis of data from different assessment years An evaluation of the treatment of items from previous assessments could be made by comparing the items that were deleted from the scales and the items that were not treated as trend items across the years, as reported in previous technical reports (Beaton, 1987; Beaton, 1988; Johnson and Zwick, 1990; Johnson and Allen, 1992; Johnson and Carlson, 1994; Allen, Kline and Zelenak, 1996; Allen, Carlson, and Zelenak, 1999).

1.2 The NAEP 1999 Long-Term Trend Assessment Design

In 1999, NAEP conducted national long-term trend assessments in reading, writing, mathematics, and science at three age groups: 9, 13, and 17. Although long-term trend writing assessments have also been administered since 1984, the results from these assessments are undergoing evaluation. Therefore, the **analysis of the long-term trend writing assessment data is not described in this document.**

The assessments were funded by the U.S. Department of Education and conducted by ETS for the National Center for Education Statistics (NCES). ETS was responsible for overall management of the program, development of the overall design, development of the items and questionnaires, data analysis, and reporting. Westat was responsible for all aspects of sampling and field operations. National Computer Systems (NCS) carried out the printing, distribution, and receipt of materials, as well as the scanning of assessment data, and professional scoring of constructed responses.

Results from the NAEP 1999 long-term trend assessments can be found in the report, *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance* (Campbell, Hombo, and Mazzeo, 2000). Many of the NAEP reports are available on the Internet at http://nces.ed.gov/nationsreportcard. For information about ordering printed copies of these reports,

go to the U.S. Department of Education Web Page at http://www.ed.gov/about/ordering.jsp, 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

1.2.1 The 1999 NAEP Student Samples

Only NAEP long-term trend assessments were administered in 1999; no main or state assessments were administered. The student samples for the 1999 long-term trend assessment are summarized in table 1–1. Each row of the table 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.

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 printed booklets, thus pacing the students through exercises at a fixed rate. The same method is currently in practice for mathematics and science; however, the reading assessments were administered in print form only from 1988 to 1999. (See sections 1.2.3 and 1.2.4.)

4. The *cohort assessed* denotes the age/grade or age of the population being sampled. For the reading and writing assessments, the age/grade classification is defined as students either in grade 4 or age 9, grade 8 or age 13, and grade 11 or age 17. The mathematics and science assessments use the age only classification—age 9, age 13, or age 17. (See sections 1.2.3 and 1.2.4.)

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; therefore, those assessment seasons were used for the 1999 long-term trend assessment.

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.[1]

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.

[1]See *Expanding the New Design: The NAEP 1985-86 Technical Report,* (pp. 6-7), (Beaton, 1988).

8. The *number assessed* 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.

Table 1–1. NAEP long-term trend student samples: 1999

Sample	Book ID	Mode	Cohort assessed	Time of testing	Age definition	Modal grade	Number assessed
Total							32,782
9 [RW–LTTrend]	51–56	Print	Age 9/Grade 4	1/3/99 – 3/8/99 (Winter)	CY	4	5,793
13 [RW–LTTrend]	51–56	Print	Age 13/Grade 8	10/9/98 – 12/22/98 (Fall)	CY	8	5,933
17 [RW–LTTrend]	51–56	Print	Age 17/Grade 11	3/11/99 – 5/10/99 (Spring)	Not CY	11	5,288
9 [MS–LTTrend]	91–93	Tape	Age 9	1/3/99 – 3/8/99 (Winter)	CY	4	6,032
13 [MS–LTTrend]	91–93	Tape	Age 13	10/9/98 – 12/22/98 (Fall)	CY	8	5,941
17 [MS–LTTrend][1]	84–85	Tape	Age 17	3/11/99 – 5/10/99 (Spring)	Not CY	11	3,795

[1]The number assessed for the 17[MS–LTTrend] sample is less than that for the other samples because only two booklets, rather than three, were presented to students in this sample. At age 17, booklets 84 and 95 contained 3 blocks of mathematics and/or science items, while at the other ages each booklet contained one mathematics and one science block.

LEGEND

MS	Mathematics and science
RW	Reading and writing
LTTrend	Long-term trend assessment booklets are identical to the 1986 (mathematics/science) or 1984 (reading/writing) long-term trend assessments
Tape	Audiotape administration
Print	Print administration
CY	Calendar year: birthdates in 1989 and 1985 for ages 9 and 13, respectively
Not CY	Age 17 only: birthdates between October 1, 1981, and September 30, 1982

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Each sample was defined in the same way as equivalent samples in several previous assessments and generally used the same assessment technology. 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 1999 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, 1994, and 1996 (many of the items were also used in pre–1984 assessments). As in 1984, 1988, 1990, 1992, 1994, and 1996 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 nine previous reading assessments (1971, 1975, 1980, 1984, 1988, 1990, 1992, 1994 and 1996). Information about how the estimates of achievement from these samples were linked to one another is provided in sections 1.7 and 3.7.

[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, 1994, and 1996 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 nine previous science assessments (1970, 1973, 1977, 1982, 1986, 1990, 1992, 1994, and 1996); the estimates of mathematics achievement link to eight previous assessments (1973, 1978, 1982, 1986, 1990, 1992, 1994, and 1996). Information about how the estimates of achievement from these samples were linked to one another is provided in sections 1.7, 4.5, and 5.5.

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 1999 subject areas of reading, mathematics, and science, several other subject areas have been assessed over the years—civics, social studies, U.S. history, citizenship, geography, literature, music, career development, art, and computer competence. Many subject areas are reassessed periodically to measure trends over time.

THIS PAGE INTENTIONALLY LEFT BLANK.

Table 1–2. NAEP subject areas, grades, and ages assessed: 1969–1999

Assessment year	Subject area(s)	Grades/ages assessed										
		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 dvlpt.			X			X			X	X	X
	Writing			X			X			X	X	
1974–75	Reading			X			X			X	X	
	Art			X			X			X	X	
1975–76	Citizenship/social studies			X			X			X	X	
	Mathematics[2]						X			X	X	
1976–77	Science			X			X			X		
	Basic life skills[2]									X		
	Health[2]										X	
	Energy[2]										X	
	Reading[2]										X	
	Science[2]										X	
1977–78	Mathematics			X			X			X		
	Consumer skills[2]									X		
1978–79	Art			X			X			X		
	Music			X			X			X		
	Writing			X			X			X		
1979–80	Reading			X			X			X	X	
	Literature			X			X			X	X	
1983–84	Reading		X	X		X	X			X		
	Writing		X	X		X	X			X		
1985	Adult literacy[2]											X

See notes at the end of table →

Table 1–2. NAEP subject areas, grades, and ages assessed: 1969–1999—Continued

Assessment year	Subject area(s)	Grades/ages assessed: Grade 3	Grade 4	Age 9	Grade 7	Grade 8	Age 13	Grade 11	Grade 12	Age 17	Age 17OS[1]	Adult
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		
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		
	Geography[2]								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		
1990	Mathematics (long-term trend)			X			X	X		X		
	Science (long-term trend)			X			X	X		X		
	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			X		
	Science (long-term trend)			X		X	X			X		
	Trial state mathematics					X						

See notes at the end of table →

Table 1–2. NAEP subject areas, grades, and ages assessed: 1969–1999—Continued

Assessment year	Subject area(s)	Grades/ages assessed										
		Grade 3	Grade 4	Age 9	Grade 7	Grade 8	Age 13	Grade 11	Grade 12	Age 17	Age 17OS[1]	Adult
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									
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[3]					X						
1997	Music					X						
	Theatre					X						
	Visual arts					X						
1998	Reading		X			X			X			
	Writing		X			X			X			
	Civics		X			X			X			
	State reading		X			X						
	State writing					X						

See notes at the end of table →

Table 1–2. NAEP subject areas, grades, and ages assessed: 1969–1999—Continued

Assessment year	Subject area(s)	Grades/ages assessed										
		Grade 3	Grade 4	Age 9	Grade 7	Grade 8	Age 13	Grade 11	Grade 12	Age 17	Age 17OS[1]	Adult
1999	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		

[1]Age 17 students who had dropped out of school or had graduated prior to assessment.

[2] Small, special–interest assessments conducted on limited samples at specific grades or ages

[3]Department of Defense Education Activity (DoDEA) schools were assessed at both grades 4 and 8. All other states and jurisdictions in the 1996 state science assessment were assessed at grade 8 only.

NOTE: Somewhat different age definitions were used in the 1984, 1986, and 1988 assessments. In the 1984 assessments, 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 national 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 et al., 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 defined 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 math assessments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

1.2.3 The Design of the 1999 Reading Long-Term Trend Assessment

Because students' ages vary within each grade level, the overall sample from which the reading results are derived contains students in grade 4 or at age 9, in grade 8 or at age 13, and in grade 11 or at age 17. For example, age 9 students may not all be in grade 4, but may be in grade 3 or grade 5. The NAEP assessments in reading and writing are administered to the same sample of students, but the results for the two subject areas are based on different subsamples of these students. For historical reasons, the writing assessment results are based on a subsample of students in grades 4, 8, and 11, and the reading assessment results are based on a subsample of students of ages 9, 13 and 17.

The reading long-term trend scale was established in 1984 using data from that year and from earlier assessments. Although reading long-term trend results are only reported for age samples, both age and grade samples are used in scaling. NAEP reports student reading performance at age 9, at age 13, and at age 17 in 10 reading assessments conducted during the school years ending in 1971, 1975, 1980, 1984, 1988, 1990, 1992, 1994, 1996, and 1999. For each assessment, 13–year–olds and eighth graders were assessed in the fall, 9–year–olds and fourth graders were assessed in the winter, and 17–year–olds and eleventh graders were assessed in the spring of the assessment school year. The same assessment booklets, containing blocks of reading, writing, and background questions, were used in 1984, 1988, 1990, 1992, 1994, 1996, and 1999. The reading assessments were administered in printed form only from 1988 to 1999. Previous to 1984, audiotapes were used in conjunction with the printed booklets directing students taking the assessment to adhere to a fixed time period. In 1984, both methods of administration were used to provide a link between the two administration methods.[2]

The reading tasks required students to read and answer questions based on a variety of materials, including informational passages, literary text, and documents. Although some tasks required students to provide written responses, most questions were multiple–choice questions. The assessment was designed to evaluate students' ability to locate specific information, make inferences based on information in two or more parts of a passage, or identify the main idea in a passage. For the most part, these questions measured students' ability to read either for specific information or for general understanding. Although the reading assessments conducted through the 1970s underwent some changes from test administration to test administration, the set of reading passages and questions included in the long-term trend assessments has been kept essentially the same since 1984, and most closely reflects the objectives developed for that assessment and identified in *NAEP Reading Objectives: 1983–84 Assessment* (NAEP, 1984).

At each of the three cohorts assessed, the reading and writing long-term trend assessment booklets consisted of three different segments or "blocks" of content questions. The blocks were assembled three to a booklet, together with a general background questionnaire that was common to all booklets. This section included questions about demographic information and home environment, and a set of questions pertaining to students' experiences and instruction related to reading and writing.

The reading long-term trend assessment administered at age 9/grade 4 included 45 passages and 105 questions, including 8 that required students to construct written responses. At age 13/grade 8, the assessment included 43 passages and 107 questions, 7 of them requiring constructed responses. At age 17/grade 11, the assessment contained 36 passages and 95 questions, 8 of them requiring constructed responses.

1.2.4 The Design of the 1999 Science and Mathematics Long-Term Trend Assessment

At each of the three ages assessed (9, 13, and 17), both the science and mathematics long-term trend assessment booklets consisted of three different 15–minute segments or "blocks" of content

[2]See *Marginal Estimation Procedures* (Mislevy and Sheehan, 1987).

questions. The blocks were assembled three to a booklet, together with a general background questionnaire that was common to all booklets. This section included questions about demographic information and home environment, and a set of questions pertaining to students' experiences and instruction related to the particular subject area being assessed. (i.e., either science or mathematics).

At ages 9 and 13, the blocks were placed in three booklets, each containing one block of mathematics questions, one block of science questions, and one block of reading questions. The reading block in these booklets is not used in the reading long-term trend assessment, but is included in order to preserve the context of the science and mathematics questions and replicates booklets from the original 1986 design. At age 17, two booklets were administered—one contained two mathematics blocks and one science block, while the other contained two science blocks and one mathematics block and replicates the 1986 design.

At all three ages, the science and mathematics questions were administered using a paced audiotape. The tape recording that accompanied the booklets standardized timing, and was intended to help students with any difficulty they might have in reading the questions. Thus, in an administration session, all students were being paced through the same booklet.

1.3 Instrument Design

1.3.1 Student Assessment Booklets

Students received different blocks of exercises in their booklets according to a procedure called "partially balanced incomplete block (PBIB) spiraling." The term PBIB spiral refers to the method used to assemble NAEP assessment exercises into booklets for administration. Spriraling refers to the method by which test booklets are assigned to students; it ensures that any group of students will be assessed using approximately equal numbers of the different booklets. This method was developed to allow for the study of the interrelationships among exercises within a subject area. As a result of this design, all exercises are given to approximately the same number of students, but no student responds to all exercises. The exercise blocks, along with sections of background questions, were assembled into booklets according to the design shown in tables 1–3, 1–4, and 1–5, respectively, for ages 9, 13, and 17.

Student Questionnaires

Two sets of multiple–choice background questions were included in separate sections of each student booklet:

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

Subject–area Background: Students were asked to report their instructional experiences related to the relevant subject area (e.g., science, mathematics, reading or writing) in the classroom, including questions about instructional activities, and their views on the utility and value of the subject matter.

Tables 1–3, 1–4, and 1–5 show the configuration of booklets for each age/grade. Each booklet contains a section of background questions, followed by the cognitive blocks.

Table 1–3. NAEP long-term trend, age 9/grade 4 booklet configuration: 1999

Subject area	Booklet number	Section 1 Common background questions	Section 2[1] Cognitive block 1	Section 3[1] Cognitive block 2	Section 4[1] Cognitive block 3
Reading and writing	51W	CC	C[2]	L	Q
	52W	CC	H	E[2]	R
	53W	CC	C[2]	K	J
	54W	CC	G[2]	O	E[2]
	55W	CC	M	G[2]	N
	56W	CC	—V[2,3]—		R
Mathematics and science	91T	B1	R1	M1	S1
	92TC	B1	S2	R2	M3[4]
	93T	B1	M2	S3	R3

[1] Subject area background questions are included in cognitive blocks.
[2] Writing blocks
[3] Block V contained one writing task, in addition to reading questions.
[4] Calculator needed for this block.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 1–4. NAEP long-term trend, age 13/grade 8 booklet configuration: 1999

Subject area	Booklet number	Section 1 Common background questions	Section 2[1] Cognitive block 1	Section 3[1] Cognitive block 2	Section 4[1] Cognitive block 3
Reading and writing	51W	CC	M	K	D[2]
	52W	CC	C[2]	L	Q
	53W	CC	H	E[2]	R
	54W	CC	N	C[2]	D[2]
	55W	CC	G[2]	O	E[2]
	56W	CC	G[2]	J	P
Mathematics and science	91T	B1	R1	M1	S1
	92TC	B1	S2	R2	M3[3]
	93T	B1	M2	S3	R3

[1] Subject area background questions are included in cognitive blocks.
[2] Writing blocks
[3] Calculator needed for this block.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 1–5. NAEP long-term trend, age 17/grade 11 booklet configuration: 1999

Subject area	Booklet number	Section 1 Common background questions	Section 2[1] Cognitive block 1	Section 3[1] Cognitive block 2	Section 4[1] Cognitive block 3
Reading and writing	51W	CC	M	K	D[2]
	52W	CC	C[2]	L	Q
	53W	CC	H	E[2]	R
	54W	CC	N	C[2]	D[2]
	55W	CC	G[2]	O	E[2]
	56W	CC	G[2]	J	P
Mathematics and science	84T	B1	M1	M2	S3
	85TC	B1	S1	S2	M3[3]

[1] Subject area background questions are included in cognitive blocks.
[2] Writing blocks
[3] Calculator needed for this block.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

1.3.2 Other Questionnaires

In addition to the student assessment booklets two other instruments provided data relating to the assessment: 1) a school questionnaire, and 2) the Students with Disabilities/Limited English Proficiency (SD/LEP) questionnaire. A school questionnaire was completed by school principals or their representatives, and provided information about school administration, staffing patterns, special programs, subject requirements, and school resources. Specific guidelines for exclusion were provided for all samples in the 1999 assessment; these guidelines were the same as those used in previous long-term trend assessments. For each student who was excluded, school staff that had knowledge of the student's capabilities completed a (SD/LEP) questionnaire, listing the reason for exclusion and providing some background information.

1.4 Sampling and Data Collection

This section summarizes the sampling and data collection activities conducted by Westat for the 1999 long-term trend assessments. A detailed report describing the sampling, data collection, and weights is available in appendix D.

Based on procedures used since the inception of NAEP, the data collection schedule was: 13–year–olds/eighth graders in the fall (October to December, 1998), 9–year–olds/fourth graders in the winter (January to mid–March, 1999), and 17 year–olds/eleventh graders in the spring (mid–March to May, 1999). Although only 9, 13, and 17–year–olds were assessed in science and mathematics, both age– and grade–eligible students were assessed in reading and writing. Age eligibility was defined by calendar year for 9– and 13–year olds, while by birth date range for 17–year olds (from October 1, 1981 through September 30, 1982). In conjunction with the development of the national main assessments, changes in sampling, analysis, and reporting by age, grade or age/grade samples were made sample–by–sample and subject–by–subject with the purpose of reporting more detailed information about a specific subject area curriculum during each assessment year.

As with all NAEP long-term trend national assessments, students attending both public and nonpublic schools were selected for participation using a stratified, three–stage, random sampling procedure. The first stage of sampling involved defining geographic primary sampling units (PSUs), which are typically groups of contiguous counties, but sometimes a single county; classifying the PSUs into strata defined by region and community type; then selecting PSUs with probability proportional to size. In the second stage, within each PSU that was selected at the first stage, both public and nonpublic schools were selected from a list of public and nonpublic schools with probability proportional to the number of age–eligible students in the school. Each school selected was assigned at least one substitute school with similar characteristics that could be included in the sample if the school administration chose not to allow the original school to participate in the assessment. The third stage involved systematically selecting students from a list of students within each school, using a random starting point.

The student sample sizes for the long-term trend assessments, as well as the school and student participation rates, are presented in the following tables. The numbers in the tables are based on the full age/grade samples of students, at the time the samples were collected. Students within schools were randomly assigned to either mathematics/science or reading assessment sessions subsequent to their selection for participation in the 1999 assessments. The student sample sizes for the 1999 long-term trend assessments are presented in table 1–6, and the school and student participation rates are shown in table 1–7. In order to meet reporting requirements of 62 students per reporting group and scaling requirements of 2,000 students per item, the target sample sizes of 11,200 in age classes 9 and 13, and 9,200 in age class 17 were selected (see section D.3.1.2).

Table 1–6. NAEP long-term trend assessments, student sample sizes: 1999

Age	Mathematics/Science[1]	Reading	Total
Total	15,768	17,014	32,782
Age 9	6,032	5,793	11,825
Age 13	5,941	5,933	11,874
Age 17	3,795	5,288	9,083

[1]These totals reflect the same sample of students for mathematics and science.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 1–7. NAEP long-term trend assessments, school and student participation rates: 1999

Subject	Age	Weighted percentage of schools participating[1]	Weighted percentage of students participating	Overall participation
Mathematics/Science[2]				
	9	83.5	93.7	78.3
	13	79.3	92.5	73.4
	17[3]	72.1	81.3	58.6
Reading				
	9	84.9	94.4	80.2
	13	80.8	92.1	74.4
	17[3]	74.0	80.2	59.4

[1]Participation rates in this column were calculated prior to the substitution of replacement schools.
[2]These totals reflect the same sample of students for mathematics and science.
[3]Since the overall participation rate at age 17 for both reading and mathematics/science was below 70 percent, a nonresponse bias study was conducted; the results are reported in appendix D, section D.4.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

1.5 Student Exclusion Rates

Some students selected for participation in the NAEP assessments are identified as special needs students. The term "special needs students" is generally used to describe both students with limited English proficiency (LEP) and students with disabilities (SD). If, in accordance with guidelines provided by NAEP, it is decided that a special needs student cannot meaningfully participate in the NAEP assessment for which he or she was selected, then that student is excluded from the assessment.

The criteria for excluding students for the long-term trend assessments differ from those for the main assessments. In order to maintain the common testing conditions of the long-term trend assessments, the guidelines and criteria that were established previously are followed. Three types of students could be excluded under these guidelines: 1) all non–English speaking students, 2) students who are educable but who were judged incapable of meaningfully responding to exercises appropriate to their age level, and 3) students so functionally disabled that they could not perform in the NAEP assessment situation.

In recent years, a number of policy, legislative, and civil rights issues have caused the NAEP program to look more closely at its administration and assessment procedures regarding increasing participation among SD and LEP students. Thus, in 1996 the inclusion criteria for the **main** assessments were revised with the intention of making them clearer, more inclusive, and more likely to be applied consistently. However, the long-term trend assessments retain the same criteria as stipulated above. In addition in 1996, for the first time in NAEP, a variety of assessment accommodations were offered to: 1) students with disabilities whose Individualized Education Plan (IEP) specified such accommodations for testing; and 2) LEP students, who in the opinion of their instructors, required an accommodation in order to take the English assessment. **Accommodations are not provided for the long-term trend assessments,** and criteria from previous long-term trend assessments were used to identify students to be excluded from these assessments. In light of current trends in the identification of students with disabilities and LEP students, exclusion rates should be evaluated with caution.

The exclusion rates for the 1990s are presented in table 1–8. In reading, mathematics, and science the exclusion rates appear to be slightly higher in 1999 than in 1990 for all age groups. However, only at ages 9 and 17 are the rates significantly higher in 1999 than in 1990.

Table 1–8. Student exclusion percentage rates by subject and age for the NAEP long-term trend assessments: 1990–1999

Subject and Age	1990	1992	1994	1996	1999
Reading					
Age 9	5.54(0.45)*	6.56(0.37)	7.38(0.56)	8.12(0.88)	7.94(0.73)
Age 13	5.27(0.47)	5.73(0.40)	6.45(0.53)	6.88(0.53)	6.45(0.64)
Age 17	4.49(0.28)*	5.33(0.33)	5.19(0.45)	7.30(0.53)	6.02(0.58)
Mathematics/Science[1]					
Age 9	5.30(0.44)*	6.71(0.38)	7.76(0.57)	7.78(0.88)	7.35(0.66)
Age 13	5.28(0.47)	6.04(0.40)	6.19(0.54)	6.52(0.52)	6.09(0.64)
Age 17	4.47(0.27)*	5.44(0.34)	5.27(0.45)	7.38(0.53)	6.12(0.59)

*Significantly different from 1999.

[1]These totals reflect the same sample of students for mathematics and science.

NOTE: Accommodations were not provided as part of the long-term trend assessments. Standard errors of the exclusion rates appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

1.6 Scoring

Materials from the 1999 long-term trend assessment were shipped to National Computer Systems (NCS) in Iowa City, Iowa for processing and scoring; these activities were reported in NCS (2000). (See appendix E for detailed information from this report pertaining to the long-term trend assessment.)

Receipt and quality control were managed through a sophisticated bar coding and tracking system. After all appropriate materials were received from a school, they were forwarded to the professional scoring area, where trained staff using guidelines prepared by NAEP evaluated the responses to constructed–response (e.g., written response) questions. Each constructed–response question had a unique scoring rubric that defined the criteria used to evaluate students' responses. Subsequent to the professional scoring, the booklets were scanned, and all information was transcribed to the NAEP database at Educational Testing Service (ETS). Detailed information describing the steps involved in the creation of the database, quality control of data entry, and creation of the database products can be found in chapter 8 of *The NAEP 1996 Technical Report* (Ferris, Pashley, Freund, and Rogers, 1999).

An overview of the professional scoring for mathematics and reading follows. No constructed–response questions were scored for science. Most of the constructed–response mathematics long-term trend questions were scored on a correct/incorrect basis. Those that had several categories of responses were later dichotomized into correct or incorrect categories. The scoring guides identified the correct or acceptable answers for each question in each block. The scores for these questions included a 0 for no response, a 1 for a correct answer or a 2 for an incorrect or "I don't know" response. Because of the straightforward nature of the scoring, lengthy training was not required. In an orientation period, the readers were trained to follow the procedures for scoring the mathematics questions and given an opportunity to become familiar with the scoring guides, which listed the correct answer for the questions in each of the blocks. During the scoring, every tenth booklet in a session was scored by a second reader to provide a quality check.

The 1999 reading long-term trend assessment included eight constructed–response items at age 9, (three of these were scored dichotomously), seven constructed–response items at age 13, and eight such items at age 17. Some of the items were administered to more than one age group.

The scoring guides for the constructed–response reading questions focused on students' ability to perform various reading tasks—for example, identifying the author's message or mood and substantiating their interpretations, making predictions based on given details, supporting an interpretation, and comparing and contrasting information. Scoring guides for the reading questions varied somewhat, but typically included a distribution of five rating categories. Some of the scoring guides included secondary scores, which typically involved categorizing the kind of evidence or details the student used as support for an interpretation.

The training program for the reading long-term trend assessment scoring was carried out on all assessment questions one at a time for each age group and covered the range of student responses. Because the purpose of the scoring was to measure trends from the 1984 assessment, preparation for training included rereading hundreds of 1984 responses and compiling training sets. In order to ensure continuity with the past scoring of the trend questions, at least half of the sample papers in the training sets were taken from the 1984 training sets, and previously scored 1984 booklets were masked to ensure that scoring for training and the subsequent trend reliability scoring would be done without knowledge of the previous scores given.

The actual training was conducted by ETS staff assisted by NCS's scoring director and team leaders. Training began with each reader receiving a photocopied packet of materials consisting of a scoring guide, a set of 15 to 20 scored samples and an additional 20 to 40 response samples to be scored. The trainers reviewed the scoring guide, explained all the applicable score points, and elaborated on the rationale used to arrive at a particular score. The readers then reviewed the 15 to 20 scored samples, as the trainers clarified and elaborated on the scoring guide. After this explanation, the additional samples were scored and discussed until the readers were in agreement. If necessary, additional packets of 1984 responses were used for practice scoring. As a further step to achieve reliability with 1984, a 25 percent sample of the 1984 responses was scored on separate scoring sheets following the formal training session. These sheets were key entered, and a computerized report was generated comparing the new scores with those assigned in 1984. After some further discussion, scoring of the 1999 responses began.

Three reliability studies were conducted as part of this scoring. For the 1999 material, 25 percent of the constructed responses were scored by a second reader to produce interreader reliability statistics. In addition, a trend reliability study was conducted by rereading 20 percent of the 1984 responses. Finally, another trend reliability study was conducted by rereading 20 percent of the 1996 responses. The reliability information from these studies is shown in table 1–9.

Table 1–9. NAEP reading long-term trend assessment scoring, percent exact agreement between readers: 1999

	1984 Responses rescored in 1999		1996 Responses rescored in 1999		1999 Responses scored twice	
Age	Mean percent agreement	Range of agreement	Mean percent agreement	Range of agreement	Mean percent agreement	Range of agreement
9	89.4	86.7–91.7	86.1	78.9–91.9	91.7	88.1–95.7
13	85.9	83.7–88.8	86.8	66.7–95.7[1]	88.6	84.1–92.7
17	92.6	87.0–96.5	92.4	89.4–96.4	91.9	85.2–96.9

[1]Only one of the items had a percent agreement lower than 81.7% and that item was deleted from the age 13 long-term trend reading scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

1.7 Data Analysis and Item Response Theory (IRT) Scaling

After the assessment information had been compiled in the NAEP database, the data were weighted according to the sample design and the population structure. The weighting for the samples reflected the probability of selection for each student as a result of the sampling design, adjusted for nonresponse (further information is detailed in appendix D, section D.4). Through poststratification, the weighting assured that the representation of certain subpopulations corresponded to figures from the U.S. Census and the Current Population Survey.

Analyses were then conducted to determine the percentage of students who gave various responses to each cognitive and background question. Item response theory (IRT)[3] was used to estimate average proficiency for the nation and various subgroups of interest within the nation. IRT scaling was performed separately within each age/grade level for each of the three long-term trend assessments (science, mathematics, and reading). Each of the three assessments employs slightly different steps in data analysis and IRT scaling.

IRT models the probability of answering a question correctly as a mathematical function of proficiency or skill. The main purpose of IRT analysis is to provide a common scale on which performance can be compared across groups, such as those defined by age, assessment year, or subpopulations (e.g., race/ethnicity or gender).

Students do not receive enough questions about a specific topic to permit reliable estimates of individual performance. Traditional test scores for individual students, even those based on IRT, would

[3]See *Applications of Item Response Theory to Practical Testing Problems* (Lord, 1980).

contribute to misleading estimates of population characteristics, such as subgroup averages and percentages of students at or above a certain proficiency level. Instead, NAEP constructs sets of plausible values designed to represent the distribution of proficiency in the population.[4] A plausible value for an individual is not a scale score for that individual, but may be regarded as a representative value from the distribution of potential scale scores for all students in the population with similar characteristics and identical patterns of item response. Statistics describing performance on the NAEP scales are based on these plausible values. These statistics estimate values that would have been obtained had individual proficiencies been observed—that is, had each student responded to a sufficient number of cognitive questions so that his or her proficiency could be precisely estimated.

For the 1999 mathematics, reading, and science long-term trend assessments, separate IRT scales were constructed within each grade. These scales were linked to the previously established scales within each subject area via a common population linking procedure using data from the 1996 and 1999 assessments. The reading long-term trend scale was first constructed after the 1984 assessments and links all previous reading assessments to the same scale. The science and mathematics assessments long-term trend scales were first developed after the 1986 science and mathematics assessments, respectively, and links all previous assessments in each subject area to the long-term trend scales. The initial long-term trend scaling, however, did not include the 1969–70 or 1973 science assessments or the 1973 mathematics assessment because these assessments had too few questions in common with subsequent assessments. To provide a link to the early assessment results for the nation and for subgroups defined by race/ethnicity and gender at each of three age levels, estimates of average scale scores were extrapolated from previous analyses.

The extrapolated estimates were obtained by assuming that, within a given age level, the relationship between the logit transformation of a subgroup's average p–value (i.e., average proportion correct) for common questions and its respective scale score average was linear, and that the same line held for all assessment years and for all subgroups within the age level. Because of the necessity for the use of extrapolation of the average scale scores for these early assessments, caution should be used in interpreting the patterns of mathematics and science trends across those assessment years. The logit transformation is:

$$\textbf{logit}\ (p) = l\,\mathrm{n}\left[\frac{p}{1-p}\right].$$

As described earlier, the NAEP scales for all the subjects make it possible to examine relationships between students' performance and a variety of background factors measured by NAEP. The fact that a relationship exists between achievement and another variable, however, does not reveal the underlying cause of the relationship, which may be influenced by a number of other variables. Similarly, the assessments do not capture the influence of unmeasured variables. The results are most useful when they are considered in combination with other information about the student population and the educational system, such as trends in mathematics and science instruction, changes in the school–age population, and societal demands and expectations.

To facilitate interpretation of the NAEP results, the scales were divided into successive levels of performance and a "scale anchoring" process was used to define what it means to score in each of these levels. NAEP's scale anchoring follows an empirical procedure whereby the scaled assessment results are analyzed to delineate sets of questions that discriminate between adjacent performance levels on the scales. For the science, mathematics, and reading long-term trend scales, these levels are 150, 200, 250, 300, and 350. For these five levels, questions were identified that were highly likely to be answered

[4]For theoretical justification of the procedures employed, see *Randomization-Based Inferences About Latent Variables From Complex Samples* (Mislevy, 1991).

correctly by students performing at a particular level on the scale and much less likely to be answered correctly by students performing at the next lower level. The guidelines used to select such questions were as follows: students at a given level must have at least a specified probability of success with the questions (65 percent for math and science, 80 percent for reading), while students at the next lower level have a much lower probability of success (that is, the difference in probabilities between adjacent levels must exceed 30 percent). For each of the three curriculum areas, subject–matter specialists examined these empirically selected question sets and used their professional judgment to characterize each level. The reading scale anchoring was conducted on the basis of the 1984 assessment,[5] and the scale anchoring for mathematics and science long-term trend reporting was based on the 1986 assessment.[6]

1.8 Reporting Subgroups

Results for the 1999 long-term trend assessment were reported for student subgroups defined by gender, race/ethnicity, parents' level of education, and public/nonpublic school attendance. The following explains how each of these subgroups was derived.

Gender (DSEX)

The variable SEX is the gender of the student being assessed, as taken from school records. For a few students, data for this variable was missing and was imputed by ETS after the assessment. The resulting variable DSEX contains a value for every student and is used for gender comparisons among students.

Race/Ethnicity (DRACE)

The variable DRACE is an imputed definition of race/ethnicity, derived from up to three sources of information. This variable is used for race/ethnicity subgroup comparisons in the 1999 long-term trend assessments (reading, mathematics and science). Two items from the student demographics questionnaire were used in determining derived race/ethnicity:

[5]See *Implementing the New Design: The NAEP 1983-84 Technical Report* (Beaton, 1987).
[6]See *Expanding the New Design: The NAEP 1985-86 Technical Report* (Beaton, 1988).

Demographic Item Number 2:

2. If you are Hispanic, what is your Hispanic background?

- ⊂⊃ I am not Hispanic.
- ⊂⊃ Mexican, Mexican American, or Chicano
- ⊂⊃ Puerto Rican
- ⊂⊃ Cuban
- ⊂⊃ Other Spanish or Hispanic background

Students who responded to Item Number 2 by filling in the second, third, fourth, or fifth oval were considered Hispanic. For students who filled in the first oval, did not respond to the item, or provided information that was illegible or could not be classified, responses to item number 1 were examined in an effort to determine race/ethnicity. Item Number 1 read as follows:

Demographic Item Number 1:

1. Which best describes you?

- ⊂⊃ White (not Hispanic)
- ⊂⊃ Black (not Hispanic)
- ⊂⊃ Hispanic ("Hispanic" means someone who is Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or from some other Spanish or Hispanic background.)
- ⊂⊃ Asian or Pacific Islander ("Asian or Pacific Islander" means someone who is Chinese, Japanese, Korean, Filipino, Vietnamese, or from some other Asian or Pacific Island background.)
- ⊂⊃ American Indian or Alaskan Native ("American Indian or Alaskan Native" means someone who is from one of the American Indian tribes, or one of the original people of Alaska.)
- ⊂⊃ Other (What?) _____

Students' race/ethnicity was then assigned to correspond with their selection. For students who filled in the sixth oval (Other), provided illegible information or information that could not be classified, or did not respond at all, race/ethnicity as provided from school records was used. Derived race/ethnicity could not be determined for the few students who did not respond to background items 1 or 2 and for whom race/ethnicity was not provided by the school.

Parents' Education Level (PARED)

Parents' education was reported at five levels—did not finish high school, graduated high school, had some education after high school, graduated college, or "I don't know"—gathered from student responses to questions about the extent of schooling experienced by each of their parents. In the 1999 long-term trend assessments, this information was gathered from the

student background questionnaires. Students were asked to identify the highest level of education attained by their parents by choosing one of the following responses:

A. She/he did not finish high school.
B. She/he graduated from high school.
C. She/he went to another school after she graduated from high school.
D. She/he graduated from college.
E. I don't know.

The information was combined into one parental education reporting category (PARED) as follows: If a student indicated the extent of education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. For students who did not know the level of education for both parents or did not know the level of education for one parent and did not respond for the other, the parental education level was classified as unknown. If the student did not respond for both parents, the student was recorded as having provided no response.

Type of School (SCHTY98, SCHTYPE)

School type information was initially provided by Westat and was used to determine the type of school that a student attended. The values for the variable SCHTY98 were identified as:

1 Public
2 Other Religious
3 Other Nonpublic
4 Catholic
5 Bureau of Indian Affairs (BIA)
6 Department of Defense (DoDEA)
7 State Department of Education (Charter)

Students were defined as attending one of two types of schools: Public or nonpublic. Public schools are those schools funded by public money, received from the local school district, state and federal sources. Such schools must comply with all rules regulations, and laws from the local, state, and federal regulatory bodies. Nonpublic schools primarily derive their funding from private sources, such as tuition, private donations, and religious organizations. Such schools are subject to some regulation of the local, state, and federal level, but do not have to comply with all such rules. The SCHTY98 values were collapsed into a five–level variable called SCHTYPE:

1 Public (SCHTY98 categories 1 and 7)
2 Private (SCHTY98 categories 2 and 3)
3 Catholic
4 Bureau of Indian Affairs (BIA)
5 Department of Defense (DoDEA)

Part Two

Overview of the Analysis of 1999 NAEP Data

Nancy L. Allen
Educational Testing Service

2.1 Introduction

The purpose of part two is to summarize some information that is integral to the analysis of NAEP data and analysis steps used for all subjects. The overview of the analyses conducted on the 1999 NAEP data focuses on the common elements of the analyses used across the subject areas of the assessment.

Because the analysis methods are not identical across subject areas, separate detailed descriptions for each major assessment are included in subsequent parts of this document (part three—reading; part four—mathematics, and part five—science). The procedures used depended on whether assessment items were scored dichotomously (two possible responses, one correct and one incorrect) or polytomously (more than two possible ordered categories of response, e.g., items given full credit, partial credit, or no credit). Basic procedures common to most or all of the subject area analyses are summarized here. The order is essentially that in which the procedures were carried out.

The following sections summarize the steps in analysis common to all subject areas. Some of this information is described in more detail in other parts of this document. The rest is included only within this section. The topics covered are as follows:

- Section 2.2 briefly describes the preparation of the final sampling weights. Detailed information about the weighting procedures and sampling design is provided in appendix D: Westat's *NAEP 1999 Long-term Trend Data Collection, Sampling, and Weighting Report* (Caldwell et al., 2002).

- Section 2.3 provides a description of the item properties examined for background questions and for cognitive items. It includes a description of the classical item statistics examined for both dichotomously (right versus wrong) and polytomously (more than two response categories) scored items. It also includes a description of the item–level results available from summary data tables. *The NAEP 1999 Long-term Trend Summary Data Tables* can be found on the NAEP Web Site at http://www.nces.ed.gov/naep3/tables/Ltt1999/, and are available for each sample. Tables are presented in three different file formats: HTML for

viewing and printing through your web browser, CSV (comma separated values) for use in spreadsheets and data analysis applications, and PDF for viewing and printing using Adobe Acrobat Reader. Section 2.5 contains additional information about the conventions used in creating these summary tables.

- Section 2.4 summarizes the steps used to scale NAEP data. The steps include item response theory (IRT) scaling of the items, generating plausible values to account for measurement error, transforming the results to the final reporting scale, and providing tables of reported statistics. Details of the theory behind these steps are available in chapter 12 of *The NAEP 1998 Technical Report* (Allen, Carlson, Johnson, and Mislevy, 2001).
- Finally, section 2.5 gives specific information about the conventions used in hypothesis testing and reporting NAEP results.

2.2 Preparation of Final Sampling Weights

Because NAEP uses a complex sampling design (see chapters 3 and 4 of *The NAEP 1998 Technical Report* [Allen, Donoghue, and Schoeps, 2001]) in which students in certain subpopulations have different probabilities of inclusion in the sample, the data collected from each student must be assigned a weight to be used in analyses. The weights reflect each student's probability of inclusion in the sample based on the school the student attends and the absences of students from that school on the day of the assessment administration. The 1999 NAEP weights were provided by Westat, the NAEP contractor in charge of sampling. Detailed information about the weighting procedures is available in appendix D.

2.3 Analysis of Item Properties: Background and Cognitive Items

The first step in the analysis of the 1999 data was item–level analysis of all instruments. Item analyses were performed separately for each age/grade on each item in each subject area. Each block of items was analyzed separately by age/grade, with the total score on the block (including the analyzed item) used as the criterion score for statistics requiring such a score. In the cases where final weights were not available, preliminary weights were used in these preliminary analyses. The item analysis of cognitive items was repeated after scaling of the items was completed.

2.3.1 Background Items

Each NAEP background item was examined by the weighted and unweighted percent of students who gave each response, the percent of students who omitted the item, the percent who did not reach the item, and the number of respondents tabulated. These preliminary analyses were conducted within age/grade cohorts and within major reporting categories. If unexpected results were found, the item data and the coding of responses were rechecked against similar data from previous years, and corrected if possible.

2.3.2 Cognitive Items

All NAEP cognitive items were subjected to analyses of item properties. The results of these analyses were used to screen items for incorrect coding or for changes in student responses across years that might effect scaling. These analyses included conventional item analyses and incorporated examinee sampling weights. Item analysis was conducted at the block level so that the "number correct" scores for students responding to an item, selecting each option of an item, omitting an item, or not reaching an item, is the average number of correct responses for the block containing that item. Because of the inclusion of polytomously scored items in the cognitive instruments, it was necessary to use special procedures for these items. The resulting statistics are analogous to those for the dichotomously scored items, as listed below.

Dichotomously Scored Item. Multiple–choice items and constructed–response items that were scored as correct or incorrect were analyzed using standard classical test theory procedures resulting in a report for each item that included:

- For each option of the item, for examinees omitting and not reaching the item, and for the total sample of examinees:
    - the number of examinees,
    - the percentage of examinees,
    - the mean of number–correct scores for the block in which the item appeared, and
    - the standard deviation of number–correct scores for the block in which the item appeared;
- The percentage of examinees providing a response that was "off–task,"[1] if the item was a constructed response item;
- p+, the proportion of examinees who received a correct score on the item (ratio of number correct to number correct plus wrong plus omitted);
- △, the inverse–normally transformed p+ scaled to mean 13 and standard deviation 4 (this transformation of the p+ is the standard practice followed at Educational Testing Service);
- The biserial correlation coefficient between the item and the number–correct score for the block in which the item appeared; and
- The point–biserial correlation coefficient (Pearson correlation coefficient) between the item and the number–correct score for the block in which the item appeared.

The number–correct block score for each examinee was calculated by adding a one for each dichotomously scored item answered correctly plus the credit assigned to the examinee's response category for each polytomously scored item.

Polytomously Scored Items. Enhanced procedures were employed for constructed–response items that were scored polytomously. Methods parallel to those used for dichotomously scored

[1] "Off–task" is a response that is unrelated to the question and considered inappropriate.

items resulted in values reported for each distinct response category for the item. Response categories for each item were defined in two ways—one based on the original codes for responses as specified in the scoring rubrics used by the scorers, and one used in defining the IRT model scales. The latter was based on a scoring guide developed by subject–area and measurement experts and it defines the treatment of each response category in scaling. The scoring guide could result in collapsing of some response categories and a new set of statistics corresponding to the new categories. The ordered categories would usually be mapped into a set of integers in the corresponding order. Using this procedure, for example, a constructed–response item that initially has seven categories (not reached, omitted, off–task, and the four valid response categories) can be mapped into four response categories, based on the final scoring guide developed by subject–area and measurement experts. The new response categories were used to calculate the polytomously scored item statistics. Each response category was assigned zero, partial or full credit.

The following statistics, analogous to those for dichotomously scored items, were computed:

- For each response category for the item, for examinees omitting and not reaching the item, and for the total sample of examinees:
    - the number of examinees,
    - the percentage of examinees,
    - the mean of number–correct scores for the block in which the item appeared, and
    - the standard deviation of number–correct scores for the block in which the item appeared.
- The percentage of examinees providing a response that was "off–task."
- In place of p+, the ratio of the mean item score to the maximum–possible item score was used.
- In place of Δ, the inverse–normally transformed ratio of the mean item score to the maximum–possible item score scaled to mean 13 and standard deviation 4 (this transformation of the p+ is the standard practice followed at Educational Testing Service).
- The polyserial correlation coefficient between the item and the number–correct score for the block in which the item appeared was used in place of the biserial.
- The Pearson correlation coefficient between the item and the number–correct score for the block in which the item appeared was used in place of the point–biserial.

The number–correct block score for each examinee was calculated by adding a one for each dichotomously scored item answered correctly plus the credit assigned to the examinee's response category for each polytomously scored item.

2.3.3 Tables of Item–Level Results

Tables were created of the percentages of students choosing each of the possible responses to each item within each of the samples administered in 1999. The results for each item were

cross–tabulated against the basic reporting variables such as region, gender, race/ethnicity, public/nonpublic school, and parental education. All percentages were computed using the sampling weights. These tables are referred to as the *NAEP 1999 Long-term Trend Summary Data Tables*[2] and are available for each sample. In the *summary data tables*, the sampling variability of all population estimates was obtained by the jackknife procedure[3] used in previous assessments.

2.3.4 Tables of Block–Level Results

Tables summarizing the item statistics for all of the items within each block are provided in parts three, four, and five. These tables contain statistics calculated using student weights to account for NAEP's complex sampling of students, as well as the unweighted sample size. Weighted summary statistics estimate the results for the whole population of students in the NAEP sampling frame.

- The **unweighted sample size** is the number of students in the reporting sample who receive each block in the assessment. It is the number of students contributing to the statistics presented in the tables.

- The **weighted average item score** for the block is the average, over items, of the mean item score for each of the items in the block. Missing responses to polytomous items before the last observed response in a block are also considered intentional omissions and scored so that the response is in the lowest category. Occasionally, extended constructed–response items are the last item in a block of items. Because considerably more effort is required of the student to answer these items, nonresponse to an extended constructed–response item at the end of a block is considered an intentional omission (and scored as the lowest category) unless the student also did not respond to the item immediately preceding that item. In that case, the extended constructed–response item is considered not reached and treated as if it had not been presented to the student.

- The ***weighted average polyserial correlation*** is the average, over items, of the item–level polyserial correlations (biserial correlations for dichotomous items) between the item and the number–correct block score. For each item–level polyserial, the block number–correct block score (including the item in question, and with students receiving zero points for all not–reached items) was used as the criterion variable for the correlation. The number–correct block score for each examinee is calculated by adding a one for each dichotomously scored item answered correctly plus the credit assigned to the examinee's

[2] *The NAEP 1999 Long-term Trend Summary Data Tables* can be found on the NAEP Web Site at http://www.nces.ed.gov/naep3/tables/Ltt1999/, and are available for each sample. Tables are presented in three different file formats: HTML for viewing and printing through your web browser, CSV (comma separated values) for use in spreadsheets and data analysis applications, and PDF for viewing and printing using Adobe Acrobat Reader.

[3] See *Introduction to Variance Estimation* (Wolter, 1985*),* and *Considerations and Techniques for the Analysis of NAEP Data* (Johnson, 1989).

response category for each polytomously scored item. Data from students classified as not reaching the item were omitted from the calculation of the statistic.[4]

- The ***weighted alpha reliability*** is Cronbach's coefficient alpha calculated using appropriate student weights for each block of items. Cronbach (1951) describes coefficient alpha when each student's responses are weighted equally in the calculation.

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

2.3.5 Differential Item Functioning Analysis of Cognitive Items

Differential item functioning (DIF) analysis refers to procedures that assess whether items are differentially difficult for different groups of examinees. DIF procedures typically control for overall between–group differences on a criterion, usually test scores. Between–group performance on each item is then compared within sets of examinees having the same total test scores.

DIF analyses were conducted for items in the long-term trend assessment in reading because of a change in the text for one block of items (see part three, section 3.2 for a description of this change and DIF results). Each set of analyses involved three reference group/focal group comparisons: male/female, White/Black, and White/Hispanic.

The Mantel–Haenszel Procedure. The DIF analyses of the dichotomous items were based on the Mantel–Haenszel chi–square procedure (Mantel and Haenszel, 1959), as adapted by Holland and Thayer (1988). The procedure tests the statistical hypothesis that the odds of correctly answering an item are the same for two groups of examinees that have been matched on some measure of proficiency (usually referred to as the matching criterion). The DIF analyses of the polytomous items were completed using the Mantel–Haenszel ordinal procedure which is based on the Mantel procedure (Mantel, 1963), (Mantel and Haenszel, 1959). These procedures compare proportions of matched examinees from each group in each polytomous item–response category.

For both types of analyses, the measure of proficiency used is typically the total item score on some collection of items. Since, by the nature of the BIB or PBIB design, booklets comprise different combinations of blocks, there is no single set of items common to all examinees. Therefore, for each student, the measure of proficiency used was the total item score on the entire booklet. These scores were then pooled across booklets for each analysis. This procedure is described by Allen and Donoghue (1994, 1996). In addition, because research results (Zwick and Grima, 1991) strongly suggest that sampling weights should be used in conducting DIF analyses, the weights were used.

[4] In almost all NAEP IRT analyses, missing responses at the end of each block of items a student was administered are considered "not reached," and are treated as if they had not been presented to the respondent. Missing responses to dichotomous items before the last observed response in a block are considered intentional omissions, and are treated as fractionally correct at the value of the reciprocal of the number of response alternatives, if the item was a multiple–choice item. With regard to the handling of not–reached items, Mislevy and Wu (1988) found that ignoring not–reached items introduces slight biases into item parameter estimation when not reached items are present and speed is correlated with ability. With regard to omissions, they found that the method described above provides consistent limited–information maximum likelihood estimates of item and ability parameters under the assumption that respondents omit only if they can do no better than responding randomly.

For each dichotomous item in the assessment, an estimate of the Mantel–Haenszel common odds ratio, α_{MH} , expressed on the ETS delta scale for item difficulty, was produced. The estimates indicate the difference between reference group and focal group item difficulties (measured in ETS delta scale units), and typically run between about +3 and –3. Positive values indicate items that are differentially easier for the focal group than the reference group after making an adjustment for the overall level of proficiency in the two groups. Similarly, negative values indicate items that are differentially harder for the focal group than the reference group. It is common practice at ETS to categorize each item into one of three categories (Petersen, 1988): "A" (items exhibiting no DIF), "B" (items exhibiting a weak indication of DIF), or "C" (items exhibiting a strong indication of DIF). Items in category "A" have Mantel–Haenszel common odds ratios on the delta scale that do not differ significantly from 0 at the alpha = .05 level or are less than 1.0 in absolute value. Category "C" items are those with Mantel–Haenszel values that are significantly greater than 1 and larger than 1.5 in absolute magnitude. Other items are categorized as "B" items. A plus sign (+) indicates that items are differentially easier for the focal group; a minus sign (–) indicates that items are differentially more difficult for the focal group.

The ETS/NAEP DIF procedure for polytomous items uses the Mantel–Haenszel ordinal procedure (Mantel and Haenszel, 1959). Polytomous items are identified as "AA," "BB," or "CC," generalizations of the dichotomous A, B, and C categories.

In order to assure that the Mantel–Haenszel significance tests were appropriate, all NAEP DIF analyses used sampling weights that were rescaled to reflect the size of the sample, rather than the size of the student population. A separate rescaled weight was defined for each comparison as

$$Rescaled\ Weight = Original\ Weight \bullet \frac{Total\ Sample\ Size}{Sum\ of\ the\ Weights}$$

where the total sample size is the total number of students for the two groups being analyzed (e.g., for the White/Hispanic comparison, the total number of White and Hispanic examinees in the sample at that grade), and the sum of the weights is the sum of the sampling weights of all the students in the sample for the two groups being analyzed. Three rescaled weights were computed for White examinees—one for the gender comparison and two for the race/ethnicity comparisons. Two rescaled overall weights were computed for the Black and Hispanic examinees—one for the gender comparison and another for the appropriate race/ethnicity comparison. The rescaled weights were used to ensure that the sum of the weights for each analysis equaled the number of students in that comparison, thus providing an accurate basis for significance testing. The use of weights rescaled in this way does not change the estimate of a percentage or scale score mean.

In the calculation of total item scores for the matching criterion, not–reached, off–task, and omitted items were considered to be wrong responses. Polytomous items were weighted according to the number of score categories. As a result, the polytomous items were weighted more heavily than dichotomous items in the formation of the matching criterion to reflect relative amounts of time spent on average for each type of item. For each item, calculation of the Mantel–Haenszel statistic did not include data from examinees who did not reach the item in question.

Each DIF analysis was a two–step process. In the initial phase, total item scores were formed and the calculation of DIF indices was completed. Before the second phase, the matching criterion was refined by removing all identified C or CC items, if any, from the total item score. The revised score was used in the final calculation of all DIF indices. Note that when analyzing an

item classified as C or CC in the initial phase, that item score is added back into the total score for the analysis of that item only. Adding the item score for the item of interest back into the total score makes the total score (the criterion) have a distribution that is most appropriate for the M–H statistical test (Holland and Thayer, 1988). See section 3.2 for further discusion of DIF analyses.

Following standard practice at ETS for DIF analyses conducted on final forms, all C or CC items were reviewed by a committee of trained test developers and subject–matter specialists. Such committees are charged with making judgments about whether or not the differential difficulty of an item is unfairly related to group membership. The committees assembled to review NAEP items include both ETS staff and outside members with expertise in the field. The committees carefully examine each identified item to determine if either the language or contents would tend to make the item more difficult for an identified group of examinees. As pointed out by Zieky (1993):

> It is important to realize that DIF is not a synonym for bias. The item response theory based methods, as well as the Mantel–Haenszel and standardization methods of DIF detection, will identify questions that are not measuring the same dimension(s) as the bulk of the items in the matching criterion . . . .Therefore, judgment is required to determine whether or not the difference in difficulty shown by a DIF index is unfairly related to group membership. The judgment of fairness is based on whether or not the difference in difficulty is believed to be related to the construct being measured . . . .The fairness of an item depends directly on the purpose for which a test is being used. For example, a science item that is differentially difficult for women may be judged to be fair in a test designed for certification of science teachers because the item measures a topic that every entry–level science teacher should know. However, that same item, with the same DIF value, may be judged to be unfair in a test of general knowledge designed for all entry–level teachers. (p. 340)

2.4 Scaling

Scales based on IRT were derived for each subject area. chapter 12 of *The NAEP 1998 Technical Report* (Allen, Carlson, et al., 2001) describes in detail the theoretical underpinnings of NAEP's scaling methods and the required estimation procedures. The basic analysis steps are outlined here.

1. Use the NAEP BILOG/PARSCALE computer program to estimate the parameters of the item response functions on an arbitrary provisional scale. This program uses an IRT model incorporating the two– and three–parameter logistic forms for dichotomously scored items and the generalized partial–credit form for polytomously scored items. In order to select starting values for the iterative parameter–estimation procedure for each dataset, the program is first run to convergence, imposing the condition of a fixed normal prior distribution of the scale score variable. Once these starting values are computed, the main estimation runs model examinee scale score ability as a multinomial distribution. That is, no prior assumption about the shape of the scale score distribution is made. In analyses involving more than one population, estimates of parameters are made with the overall mean and standard deviation of all subjects' proficiencies specified to be 0 and 1, respectively.

2. Use a version of the MGROUP program, which implements the method of Mislevy (Mislevy, 1991) to estimate predictive scale score distributions for each respondent on an arbitrary scale, based on the item parameter estimates and the responses to cognitive items and background questions.

3. Use random draws from these predictive scale score distributions (plausible values, in NAEP terminology) for computing the statistics of interest, such as mean proficiencies for demographic groups.

4. Determine the appropriate metric for reporting the results and transform the results as needed. This includes the linking of current scales to scales from the past or the selection of the mean and variance of new scales.

5. Use the jackknife procedure to estimate the standard errors of the mean proficiencies for the various demographic groups.

The plausible values obtained through the IRT approach are not optimal estimates of individual scale scores; instead, they serve as intermediate values to be used in estimating subpopulation characteristics. Under the assumptions of the scaling models, these subpopulation estimates are statistically consistent, which would not be true of subpopulation estimates obtained by aggregating optimal estimates of individual scale scores.

2.4.1 Scaling the Cognitive Items

The data from the long-term trend samples were scaled using IRT models. For dichotomously scored items two– and three–parameter logistic forms of the model were used (the two–parameter model was used for dichotomous constructed–response items; the three–parameter model was used for multiple–choice items, when guessing can be a factor), while for polytomously scored items the generalized partial–credit model form was used. These two types of items and models were combined in the NAEP scales. Item parameter estimates on a provisional scale were obtained using the NAEP BILOG/PARSCALE program. The fit of the IRT model to the observed data was examined within each scale by comparing the empirical item response functions with the theoretical curves, as described in chapter 12 of *The NAEP 1998 Technical Report* (Allen, Carlson, et al., 2001). Plots of the empirical item response functions and theoretical curves were compared across assessments for the long-term trend assessments. The DIF analyses previously described also provide information related to the model fit across subpopulations. The same long-term trend booklets have been used for almost a decade, and most often, items were treated exactly the same way in scaling as they were treated in previous assessment years (see previous NAEP technical reports: Beaton, 1987, 1988; Johnson and Allen, 1992; Johnson and Carlson, 1994; Allen, Kline and Zelenak, 1996; Allen, Carlson, and Zelenak, 1999).

Item parameters for reading, mathematics, and science trends were reestimated, separately for each age/grade group, using the data from the most recent previous assessment year (in this case 1996) as well as the 1999 assessment. The resulting scales, based on these reestimated item parameters, were then linked to the existing long-term trend scales.

2.4.2 Generation of Plausible Values for Each Scale

Plausible values were drawn from the predictive distribution of scale score values for each student (this process is called conditioning). For the long-term trend scales, the plausible values were computed separately for each age or age/grade group and year, and were based on the student's respones to the items going into the scale as well as on the values of a set of background variables that were important for the reporting of proficiency scores. All plausible values were later rescaled to the final scale metric using appropriate linear transformations.

The variables used to calculate plausible values for a given national assessment scale included a broad spectrum of background, attitude, and experiential variables and composites of such variables. All standard reporting variables were included. Trend scales used the same or similar sets of conditioning variables that were used when the scales were originally constructed. Details of the conditioning process and of the NAEP BGROUP and NAEP CGROUP (Thomas, 1994) computer programs that implement the process are presented in chapter 12 of *The NAEP 1998 Technical Report* (Allen, Carlson, et al., 2001). The variables used in conditioning along with their contrast codings are listed in appendix C.

2.4.3 Transformation to the Reporting Metric

Transformations were of the form

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

where

θ_{target} = scale level in terms of the system of units of the final scale used for reporting;

$\theta_{calibrated}$ = scale level in terms of the system of units of the provisional NAEP–BILOG/PARSCALE scale;

$A = SD_{target} / SD_{calibrated}$;

$B = M_{target} - A \bullet M_{calibrated}$;

SD_{target} = the estimated or selected standard deviation of the scale score distribution to be matched;

$SD_{calibrated}$ = the estimated standard deviation of the sample scale score distribution on the provisional NAEP–BILOG/PARSCALE scale;

M_{target} = the estimated or selected mean of the scale score distribution to be matched; and

$M_{calibrated}$ = the estimated mean of the sample scale score distribution on the provisional NAEP–BILOG/PARSCALE scale.

After the plausible values were linearly transformed to the new scale, any plausible value less than 0 was censored to 0 because they are so close to 0. Generally in NAEP, less than one percent of the plausible values is censored to zero. The final transformation coefficients for transforming each provisional scale to the final reporting scale are given in subsequent sections of this document.

2.4.4 Tables of Scale Score Means and Other Reported Statistics

Scale scores and trends in scale scores were reported by age/grade for a variety of reporting categories. Additionally, the percentages of the students within each of the reporting

groups who were at or above anchor levels were reported to provide information about the distribution of achievement within each subject area. All estimates based on scale score values have reported variances or standard errors based on scale score values, including the error component due to the latency of scale score values of individual students as well as the error component due to sampling variability. These tables are part of the electronically delivered *summary data tables*.

2.5 Conventions Used in Hypothesis Testing and Reporting NAEP Results

2.5.1 Minimum School and Student Sample Sizes for Reporting Subgroup Results

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

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

Here a decision was reached to report subgroup results only if the student sample size exceeded 61.[5] A design effect of two was assumed for this decision, implying a sample design–based variance twice that of simple random sampling. This assumption is consistent with previous NAEP experience (Johnson and Rust, 1992). In carrying out the statistical power calculations when comparing a subgroup to the total group, it was assumed that the total population sample size is large enough to contribute negligibly to standard errors. Furthermore, it was required that the students within a subgroup be adequately distributed across PSUs to allow for reasonably accurate estimation of standard errors. The degrees of freedom are determined by the number of PSUs. If the degrees of freedom are lower than five, too few PSUs contributed to the result (see discussion of PSUs in section 1.4). In consultation with Westat, a decision was reached to publish only those statistics that had standard error estimates based on five or more degrees of freedom. The same minimum student and PSU sample size restrictions were applied to proportions and to comparisons

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

of percentages or proportions as well as average scale scores and comparisons of average scale scores.

2.5.2 Identifying Estimates of Standard Errors with Large Mean Squared Errors

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

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

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

where $\hat{N}$ is a point estimate of N and $SE(\hat{N})$ is the jackknife standard error of $\hat{N}$ (described in chapter 10 of *The NAEP 1998 Technical Report* [Qian, Kaplan, Johnson, Krenzke, and Rust, 2001]).

Experience with previous NAEP assessments suggests that when this coefficient exceeds 0.2, the mean squared error of the estimated standard errors of means and proportions based on samples of this size may be quite large. In other words, when the coefficient of variation exceeds 0.2, the standard errors of means and proportions are not well estimated. (Further discussion of this issue can be found in Johnson and Rust, 1992.) Therefore, the standard errors of means and proportions for all subgroups for which the coefficient of variation of the population size exceeds 0.2 are flagged as described above. In the *summary data tables*, statistical tests involving one or more quantities that have standard errors, confidence intervals, or significance tests so marked should be interpreted with caution.

---

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

2.5.3 Treatment of Missing Data From the Student and School Questionnaires

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

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

The estimates of proportions and proficiencies based on "missing completely at random" assumptions are subject to potential nonresponse bias if, as may be the case, the assumptions are not correct. The amount of missing data was small (usually less than 2%) for most of the variables obtained from the student and school questionnaires. For analyses based on these variables, reported results are subject to little, if any, nonresponse bias. However, for particular background items in these questionnaires, the level of nonresponse was somewhat higher, and so the potential for nonresponse bias is also somewhat greater. Results for background questions for which more than 10 percent of the responses were missing should be interpreted with caution. In the *NAEP 1999 Trends in Academic Progress* (Campbell, et al., 2000) there were no results reported with more than 10% missing responses defined in the subgroups of students. In the *NAEP 1999 Long-term Trend Summary Data Tables* (http://nces.ed.gov/nationsreportcard/tables/Ltt1999/), proportions and proficiencies data for background questions with more than 10% nonresponse were identified as, "****(****)" and footnoted as follows:

" ****(****) sample size is insufficient to permit a reliable estimate."

2.5.4 Hypothesis–Testing Conventions

2.5.4.1 Comparing Means and Proportions for Different Groups of Students

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

The set of comparisons is referred to as a "family," and the typical family involves all subgroups related by a certain background question. An example of a set of comparisons is the comparison of average science scale scores from 1999 and 1990 for male students and the comparisons of average science scale scores from 1999 and 1990 for female students. The text in the reports indicate that means or proportions from two groups were different only when the

[7] Information about survey participation rates (both school and student), as well as proportions of students excluded by each jurisdiction from the assessment, is given in tables 1–7 and 1–8, respectively. Sampling adjustments intended to account for school and student nonresponse are described in appendix D, section D.4 of this report; further details of methodology are given in chapters 10 and 11 of *The NAEP 1998 Technical Report* (Allen, et al, 2001).

[8] The term "missing completely at random" means that the mechanism generating the missing data is independent of the response to the particular background items and the scale score.

difference in the point estimates for the groups being compared was statistically significant at an approximate simultaneous α level of .05. A procedure was used for determining statistical significance NAEP staff judged to be statistically defensible, as well as being computationally tractable. Although all pairs of levels within a variable were tested and reported in the *summary data tables*, some text within the report was developed for only a subset of these comparisons, although the family size was maintained at that of the original tests. For example, text was included in the reports to compare the majority ethnic group and each minority group, but text for all possible comparisons of groups may not have been included. The procedure used to make statistical tests is described in the following paragraphs.

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

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

where T_α is the $(1 - \alpha)$ percentile of the *t* distribution with degrees of freedom, *df*, as estimated below, and *c* is the number of related comparisons being tested. See section 2.2.5.1 for a more specific description of multiple comparisons. In cases where group comparisons were treated as individual units, the value of *c* was taken as 1, and the test statistic was equivalent to a standard two–tailed *t*–test for independent samples. When *c* is greater than 1, this test is based on the Benjamini and Hochberg (1995) procedure of controlling the False Discovery Rate (FDR), described below.

The procedures in this section assume that the data being compared are from independent samples. Because of the sampling design in which PSUs, schools, and students within school are randomly sampled, the data from mutually exclusive sets of students may not be strictly independent. Therefore, the significance tests employed are, in many cases, only approximate. Another procedure, one that does not assume independence, could have been conducted. However, that procedure is computationally burdensome. A comparison of the standard errors using the independence assumption and the correlated group assumption was made using NAEP data. The estimated standard error of the difference based on independence assumptions was approximately 10 percent larger than the more complicated estimate based on correlated groups. In almost every case, the correlation of NAEP data across groups was positive. Because, in NAEP, significance tests based on assumptions of independent samples are only somewhat conservative, the approximate (assuming independence) procedure was used for most comparisons.

Because of clustering and differential weighting in the sample, the degrees of freedom are less than for a simple random sample of the same size. The degrees of freedom of this *t*–test is defined by a Satterthwaite (Johnson and Rust, 1992) approximation as follows:

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

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

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

Where m is the number of jackknife replicates (usually 62 in NAEP), t_{lk} is the l^{th} replicated estimate for the mean of a subgroup and t_k is the estimate of subgroup k mean using the overall weights and the first plausible value.

The number of degrees of freedom for the variance equals the number of independent pieces of information used to generate the variance. In the case of data from NAEP, the 62 pieces of information are the squared differences $(t_{lk}-t_k)^2$, each supplying at most one degree of freedom (regardless of how many individuals were sampled within PSUs). If some of the squared differences $(t_{lk}-t_k)^2$ are much larger than others, the variance estimate of m_k is predominantly estimating the sum of these larger components, which dominate the remaining terms. The effective degrees of freedom of S_{A_k} in this case will be nearer to the number of dominant terms. The estimate df_{A_k} reflects these relationships.

The two formulae above show us that when df_{A_k} is small, the degrees of freedom for the t–test, df, will also be small. This will tend to be the case when only a few PSU pairs have information about subgroup differences relevant to a t–test. It will also be the case when a few PSU pairs have subgroup differences much larger than other PSU pairs.

The procedures described above were used for testing differences of both means *and* nonextreme percentages. The approximation for the test for percentages works best when sample sizes are large, and the percentages being tested have magnitude relatively close to 50 percent. Hypotheses tests for "extreme" percentages cannot be accurately determined using the previously described procedures. Therefore, statements about group differences should be interpreted with caution if at least one of the groups being compared is small in size or if "extreme" percentages are being compared.

Differences in percentages were treated as involving "extreme" percentages if for either percentage, P:

$$P < P_{lim} = \frac{200}{N_{EFF} + 2},$$

where the effective sample size is

$$N_{EFF} = \frac{P(100-P)}{(SE_{JK})^2}, \text{ and}$$

SE_{JK} is the jackknife standard error of P. Similarly, at the other end of the 0 – 100 scale, a percentage is deemed extreme if $100 - P < P_{lim}$. In either extreme case, the normal approximation to the distribution is a poor approximation, and the value of P was reported, but no standard error was estimated and hence no significance tests were conducted.

2.5.4.2 Multiple Comparison Procedures

Frequently, groups (or families) of comparisons were made and were presented as a single set. The appropriate text, usually a set of sentences or a paragraph, was selected for inclusion in a report based on the results for the entire set of comparisons. For example, some reports contain a section that compared average scale scores for a predetermined group, generally the majority group (in the case of race/ethnicity, for example, White students) to those obtained by other minority groups. The entire set of tests was presented in the *summary data tables*. The procedures described above and the certainty ascribed to intervals (e.g., a 95 % confidence interval) are based on statistical theory that assumes that only one confidence interval or test of statistical significance is being performed. However, in some sections of a report, many different groups are compared (i.e., multiple sets of confidence intervals are being analyzed). In sets of confidence intervals, statistical theory indicates that certainty associated with the entire set of intervals is less than that attributable to each individual comparison from the set. To hold the significance level for the set of comparisons at a particular level (e.g., .05), adjustments—called "multiple comparison procedures"—must be made to the methods described in the previous section. One such procedure, the FDR procedure (Benjamini and Hochberg, 1995) was used to control the certainty level.

Unlike the other multiple comparison procedures, e.g., the Bonferroni procedure (Bickel and Doksum, 1977) that control the familywise error rate (i.e., the probability of making even one false rejection in the set of comparisons), the FDR procedure controls the expected proportion of falsely rejected hypotheses. Furthermore, familywise procedures are considered conservative for large families of comparisons (Williams, Jones, and Tukey, 1999). Therefore, the FDR procedure is more suitable for multiple comparisons in NAEP than other procedures.

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

2.5.4.3 Comparing Proportions Within a Group

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

graduating from college. For these types of analyses, NAEP staff determined that the dependencies in the data could not be ignored.

Unlike the case for analyses of the type described in section 2.5.4.1, the correlation between the proportion of students reporting a parent graduated from college and the proportion reporting that their parents did not finish high school is likely to be negative and large. For a particular sample of students, it is likely that the higher the proportion of students reporting "at least one parent graduated from college" is, the lower the proportion of students reporting "neither parent graduated from high school" will be. A negative dependence will result in underestimates of the standard error if the estimation is based on independence assumptions (as is the case for the procedures described in section 2.5.4.1). Such underestimation can result in an unacceptably large number of "nonsignificant" differences being identified as significant.

The procedures of section 2.5.4.1 were modified for analyses that involved comparisons of proportions within a group. The modification involved using a jackknife method for obtaining the standard error of the difference in dependent proportions. The standard error of the difference in proportions was obtained by first obtaining a separate estimate of the difference in question for each jackknife replicate (using the first plausible value only) then taking the standard deviation of the set of replicate estimates as the estimate. The procedures used for proportions within a group differed from the procedures of section 2.5.4.1 only with respect to estimating the standard error of the difference; all other aspects of the procedures were identical. In other words, let A_i and A_j be the statistics of interest for groups *i* and *j* and let S_{Ai-Aj} be the jackknife standard error of the difference. Then the text in reports identified the means or proportions for groups *i* and *j* as being different if:

$$\frac{\left|A_i - A_j\right|}{\sqrt{S^2_{A_i - A_j}}} \geq T_{\frac{.05}{2c}}.$$

THIS PAGE INTENTIONALLY LEFT BLANK.

Part Three

Data Analysis for the NAEP 1999 Long-Term Trend Reading Assessment[1]

Jo–Lin Liang, Lois H. Worthington, and Ingeborg U. Novatkoski
Educational Testing Service

3.1 Introduction

Part three describes the analyses performed on the responses to the cognitive and background items in the 1999 long-term trend reading assessment. The emphasis of part three is on the methods and results of procedures used to develop the IRT–based scale scores. However, some attention is given to the analysis of constructed–response items. The theoretical underpinnings of the IRT and plausible values methodology are given in part two.

The objectives of the reading long-term trend analysis were to prepare scale values and perform all analyses necessary to produce a long-term trend report in reading. The reading long-term trend results include the years 1971, 1975, 1980, 1984, 1988, 1990, 1992, 1994, 1996, and 1999. These analyses led to the results presented in the *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance* (Campbell et al., 2000).

The student samples that were administered reading items in the 1999 long-term trend reading assessment are shown in table 3–1. See part one, section 1.2.1 for descriptions of the target populations and the sample design used for the assessment.

The long-term trend reading results reported in Campbell et al. (2000) are based on print administrations and occur at all three age levels. The long-term trend booklets administered to the students in the long-term trend reading samples were of two types. One contained blocks of reading and writing[2] items in print form; the other contained blocks of reading items administered in print form or mathematics and science items administered by audiotape. All students received a block of common background questions, distinct for each age, and subject–area background questions that were presented in the cognitive blocks. The booklets are identical to those used for reading long-term trend assessments in 1984, 1988, 1990, 1992, 1994, and 1996. The booklets and the blocks within those booklets are listed in tables 1–3 through 1–5 in part one. This section includes specific information about the reading long-term trend items that were scaled. Both age– and grade–selected students contributed to the reading long-term trend scaling. However, to be consistent with previous long-term trend reports, only students in the "age–only" portion of the reading long-term trend samples contributed to the results presented in Campbell et al.

---

[1]Jo–Lin Liang was the primary person responsible for the planning, specification, and coordination of the reading long-term trend analyses. Data analyses and scaling were performed by Lois Worthington and Ingeborg Novatkoski. Others contributing to the analysis of data were Gerry Kokolis and Duanli Yan. Nancy L. Allen, David Freund, and Bruce A. Kaplan provided consultation.

[2]Although long-term trend writing assessments have also been administered since 1984, the results from these assessments are undergoing evaluation. Therefore, the analysis of the long-term trend writing assessment data is not described in this document.

Table 3–1. NAEP long-term trend reading student samples: 1999

Sample	Book IDs	Mode	Cohort assessed	Time of testing	Age definition	Modal grade	Number assessed
9 [RW–LTTrend]	51–56	Print	Age 9/Grade 4	1/3/99 – 3/8/99 (Winter)	CY	4	5,793
13 [RW–LTTrend]	51–56	Print	Age 13/Grade 8	10/9/98 – 12/22/98 (Fall)	CY	8	5,933
17 [RW–LTTrend]	51–56	Print	Age 17/Grade 11	3/11/99 – 5/10/99 (Spring)	Not CY	11	5,288

LEGEND

RW	Reading and writing
LTTrend	Long-term trend assessment
Print	Print administration
CY	Calendar year: birthdates in 1989 and 1985 for ages 9 and 13, respectively.
Not CY	Age 17 only: birthdates between October 1, 1981, and September 30, 1982

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 3–2 clarifies the relationships between the 1999 long-term trend samples and samples from previous years. For all ages, the 1999 reading long-term trend samples allow direct comparisons with 1996, 1994, 1992, 1990, 1988, and 1984 samples. The long-term trend scale, established in 1984, was linked to the 1971, 1975, and 1980 assessments using a complex equating strategy described in *Implementing the New Design: The NAEP 1983–84 Technical Report* (Beaton, 1987). At each age, several intact booklets were retained from the 1984 assessment, forming the basis of the reading long-term trend assessment in 1988, 1990, 1992, 1994, 1996, and 1999.

Information about the previous reading long-term trend assessments is available in: chapter 9 of *Expanding the New Design: The NAEP 1985–86 Technical Report* (Zwick, 1988), chapter 10 of *Focusing the New Design: The NAEP 1988 Technical Report* (Zwick, 1990); chapter 12 of *The NAEP 1990 Technical Report* (Donoghue, 1992); chapter 12 of *The NAEP 1992 Technical Report* (Donoghue, Isham, Bowker, and Freund, 1994); chapter 15 of *The NAEP 1994 Technical Report* (Chang, Donoghue, and Worthington, 1996); and chapter 14 of *The NAEP 1996 Technical Report* (Liang and Worthington, 1999).

The 1999 reading long-term trend assessment included, at each age level, six of the assessment booklets administered in 1984. These booklets (51–56) contained both reading and writing blocks, as well as background items. Although these long-term trend booklets represented only about one–tenth of the reading booklets administered using the complicated 1984 BIB design,[3] they contained 10 of the 12 reading blocks that were scaled at each age/grade level in 1984.

In the 1999 long-term trend reading assessment, minimum word changes were made to one passage called "nuts!" This policy decision resulted from parental complaints about the word "devil" being scary for their children. The main character in the passage was changed from "the Devil" to "the King;" all "Devil"–related wording was changed to "King"–related wording. This passage is the last passage in block H at each age. The "nuts" items appear in one booklet at each age, and block H is the first of the three cognitive blocks in that booklet. All five items in this passage were treated as new items; the first four are multiple–choice questions and the last is a constructed–response question. At age 9 there are five "nuts" items out of 10 items in the block; at ages 13 and 17 there are five "nuts" items out of 12. Despite this change affecting about 5 percent of the reading items, it was possible to maintain the trend from 1996 to 1999.

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

Table 3–2. NAEP reading samples contributing to the 1999 long-term trend results: 1971–1999

Cohort	Year	Sample	Subjects	Time of testing	Mode of administration	Age definition	Modal grade
Age 9	1971	Main	RL	Winter	Tape	CY	4
	1975	Main	RA	Winter	Tape	CY	4
	1980	Main	RA	Winter	Tape	CY	4
	1984	Main	RW	Winter, Spring	Print	CY	4
	1984	T–84	RW	Winter	Tape	CY	4
	1988[1]	LTTrend[2]	RW	Winter	Print	CY	4
	1990	LTTrend[2]	RW	Winter	Print	CY	4
	1992	LTTrend[2]	RW	Winter	Print	CY	4
	1994	LTTrend[2]	RW	Winter	Print	CY	4
	1996	LTTrend[2]	RW	Winter	Print	CY	4
	1999	LTTrend[2]	RW	Winter	Print	CY	4
Age 13	1971	Main	RL	Fall	Tape	CY	8
	1975	Main	RA	Fall	Tape	CY	8
	1980	Main	RA	Fall	Tape	CY	8
	1984	Main	RW	Winter, Spring	Print	CY	8
	1984	T–84	RW	Fall	Tape	CY	8
	1988[1]	LTTrend[2]	RW	Fall	Print	CY	8
	1990	LTTrend[2]	RW	Fall	Print	CY	8
	1992	LTTrend[2]	RW	Fall	Print	CY	8
	1994	LTTrend[2]	RW	Fall	Print	CY	8
	1996	LTTrend[2]	RW	Fall	Print	CY	8
	1999	LTTrend[2]	RW	Fall	Print	CY	8
Age 17	1971	Main	RL	Spring	Tape	Not CY	11
	1975	Main	RABS	Spring	Tape	Not CY	11
	1980	Main	RA	Spring	Tape	Not CY	11
	1984	Main	RW	Winter, Spring	Print	Not CY	11
	1984	T–84	RW	Spring	Tape	Not CY	11
	1988[1]	LTTrend[2]	RW	Spring	Print	Not CY	11
	1990	LTTrend[2]	RW	Spring	Print	Not CY	11
	1992	LTTrend[2]	RW	Spring	Print	Not CY	11
	1994	LTTrend[2]	RW	Spring	Print	Not CY	11
	1996	LTTrend[2]	RW	Spring	Print	Not CY	11
	1999	LTTrend[2]	RW	Spring	Print	Not CY	11

[1]Data for constructed–response items were omitted from the 1988 reading assessment due to scoring inconsistencies that affected these items (Zwick, 1988).

[2] Within a cohort, these samples received common booklets.

LEGEND

RL	Reading and literature	LTTrend	Long-term trend (these samples received common booklets within an age group)
RA	Reading and art		
RABS	Reading, art, index of basic skills	Print	Print administration
RW	Reading and writing	Tape	Audiotape administration
Main	Main assessment	CY	Calendar year: birthdates (1999 sample) in 1989 and 1985 for ages 9 and 13
T–84	Special sample in the 1984 assessment that was used to establish links to previous assessments (1971–1980) for the purposes of long-term trend	Not CY	Age 17 only (1999 sample): birthdates between October 1 and September 30 of the appropriate years

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The numbers of scaled items in common across different age combinations are presented in table 3–3. As in previous reading long-term trend analyses, each age was scaled separately. The numbers of items scaled in 1999 that were common across assessment years are given in table 3–4. As was the case for previous long-term trend analyses, the long-term trend scale is univariate. Dimensionality analyses conducted following the 1984 assessment showed that the reading items were well summarized by a unidimensional scale (Zwick, 1987).

Table 3–3. Numbers of scaled NAEP reading long-term trend items common across ages: 1999

Age	Number of items
Total	184[1]
9 only	61
13 only	22
17 only	23
9 and 13 only	13
9 and 17 only	2
13 and 17 only	42
9, 13, and 17	21[1]

[1]These figures reflect the deletion of the five new "nuts" items from the reading long-term trend scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 3–4. Numbers of scaled NAEP reading long-term trend items common across assessments: 1984–1999

	Number of items[1]		
Assessment year	Age 9	Age 13	Age 17
1984, 1992, 1994, 1996, 1999	97	98	88
1984, 1990, 1992, 1994, 1996, 1999	96	96	87
1984, 1988, 1990, 1992, 1994, 1996, 1999	93	93	82
1980, 1984, 1988, 1990, 1992, 1994, 1996, 1999	62	66	47
1971, 1975, 1980, 1984, 1988, 1990, 1992, 1994, 1996, 1999	31	40	32

[1]These figures reflect the deletion of the five new "nuts" items from the reading long-term trend scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

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

3.2 Differential Item Functioning (DIF) Analysis

Due to the change of wording in the "nuts!" passage in the 1999 reading long-term trend assessment, a DIF analysis of items was conducted on all five new "nuts" items to identify potentially biased items that were differentially difficult for members of various subgroups with comparable overall scores. The purpose of the analysis was to identify items that should be examined more closely by a committee of trained test developers and subject–matter specialists to determine if any DIF identified during the analysis was actually biased. If NAEP items are identified as being biased, they are excluded from the analysis and reporting. The presence of DIF in an item means that the item is differentially harder for one group of students than another, while controlling for the ability level of the students. DIF analyses were conducted separately at each age using booklet–level matching for criterion on students who received the related booklets. Sample sizes were sufficient enough to compare male and female students, White and Black students, and White and Hispanic students. However, DIF analyses could not be completed to compare results for Black and Hispanic students because the total sample size for the two groups is not large enough.

For dichotomous items, the Mantel–Haenszel procedure as adapted by Holland and Thayer (1988) was used as a test of DIF (this is described in part two, section 2.3.5). The Mantel procedure (Mantel, 1963) was used for detection of DIF in polytomous items and also as described by Zwick, Donoghue, and Grima (1993). This procedure assumes ordered categories. For dichotomous items, the DIF index generated by the Mantel–Haenszel procedure is used to place items into one of three categories: "A," "B," or "C." "A" items exhibit little or no DIF, while "C" items exhibit a strong indication of DIF and should be examined more closely. Positive values of the index indicate items that are differentially easier for the focal group (female, Black, or Hispanic students) than for the reference groups (male or White students). Similarly, negative values indicate items that are differentially harder for the focal group than the reference group. An item that was classified as a "C" item in any analysis was considered to be a "C" item.

As in previous assessments, the constructed–response item associated with the "Nuts" passage was dichotomized according to criteria developed by subject–area experts. Table 3–5 summarizes the results of DIF analyses for the five new "Nuts" items. Two "C" items were identified at age 9, one at age 13, and two at age 17. After reviewing the identified items, the committee decided that these items did not show evidence of bias and they were retained. No item was dropped from the scale as the result of DIF analysis.

Table 3–5. NAEP reading long-term trend DIF analysis on new "nuts" item, DIF C–items: 1999

Age/Cohort	Flagged Item	Block	Favoring	Sample size
Age 9				
Female/Male	1 item (CR)	H	Female	412/430
Black/White	†	†	†	†
Hispanic/White	1 item (MC)	H	White	64/584
Age 13				
Female/Male	1 item (MC)	H	Male	470/500
Black/White	†	†	†	†
Hispanic/White	†	†	†	†
Age 17				
Female/Male	†	†	†	†
Black/White	1 item (MC)	H	Black	141/659
Hispanic/White	1 item (MC)	H	White	121/611

†Not applicable.
NOTE: (CR) = constructed–response item; (MC) = multiple–choice item
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

3.3 Item Analysis for the NAEP 1999 Reading Long-Term Trend Assessment

A preliminary item analysis showed that the overall item statistics for the "King" version of "nuts" items (new nuts items) are similar with the "Devil" version of items (old nuts items), indicating that it was likely that the new items would have little effect on the construct being measured by the original long-term trend scales.

Conventional item analyses did not identify any difficulties with the long-term trend data. The results displayed in table 3–6 contain the number of items, size of the unweighted sample administered the block, average weighted proportion correct, average weighted r–biserial, and average weighted alpha as a measure of reliability for each block. Because the blocks were presented in self–paced, print–administered form, the weighted proportion of students attempting the last item is included in the table to give an indication of the speededness of each block. Common labeling of these blocks across ages does not denote common items. Booklet information is detailed in part one, section 1.3. Student weights were used for all statistics except for the sample sizes. The average values reflect only the items in the block that were scaled. Overall, the 1999 item–level statistics were not very different from those for the 1996 assessment.

Table 3–6. NAEP reading long-term trend descriptive statistics for item blocks as defined after scaling: 1999

Statistics	Blocks H	J	K	L	M	N	O	P[1]	Q	R[2]	V[3]
Age 9											
Number of scaled items	10	8	11	7	11	12	11	†	11	12	9
Number of scaled constructed–response items	1	0	0	1	1	1	0	†	0	0	3
Unweighted sample size	663	722	721	680	659	657	654	†	677	1341	684
Average weighted proportion correct	.61	.52	.44	.53	.43	.56	.50	†	.57	.48	.62
Average weighted r–biserial	.76	.68	.67	.79	.67	.73	.61	†	.72	.67	.77
Weighted alpha reliability	.75	.64	.75	.73	.72	.81	.62	†	.80	.77	.78
Weighted proportion of students attempting last item	.90	.92	.78	.74	.65	.69	.88	†	.88	.83	.96
Age 13											
Number of scaled items	12	9	8	5	11	12	10	9	16	11	†
Number of scaled constructed–response items	1	0	0	0	1	1	1	1	0	0	†
Unweighted sample size	682	666	663	706	662	683	693	663	706	682	†
Average weighted proportion correct	.64	.63	.65	.73	.59	.67	.66	.73	.63	.69	†
Average weighted r–biserial	.69	.61	.77	.87	.66	.68	.63	.79	.57	.76	†
Weighted alpha reliability	.67	.55	.71	.54	.66	.78	.56	.70	.71	.77	†
Weighted proportion of students attempting last item	.96	.88	1.00	.99	.93	.78	.82	.89	.77	.98	†
Age 17											
Number of scaled items	12	4	8	6	11	12	13	10	10	7	†
Number of scaled constructed–response items	1	1	0	1	1	1	1	1	0	0	†
Unweighted sample size	734	684	678	671	678	688	645	683	671	727	†
Average weighted proportion correct	.72	.80	.76	.75	.67	.83	.66	.74	.56	.67	†
Average weighted r–biserial	.76	.92	.79	.89	.73	.80	.57	.74	.65	.81	†
Weighted alpha reliability	.73	.54	.67	.46	.69	.78	.68	.76	.67	.72	†
Weighted proportion of students attempting last item	.96	.98	1.00	.98	.97	.91	.67	.81	.93	.98	†

†Not applicable.
[1]Block P was not administered at age 9.
[2]Unlike the other blocks, block R was administered in two booklets at age 9 (see table 1–3).
[3]Block V was not administered at age 13 or 17.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 3–6a. NAEP reading long-term trend summary response rates by item type: 1999

Statistics	Multiple–choice	Short constructed–response	Extended constructed–response
Age 9			
Number of items	95	3	4
Average percentage–missing[1]	5.29	25.71	30.73
Minimum	0.45	20.30	14.26
Maximum	22.29	35.31	41.71
Average percentage–off–task[2]	†	0	0.95
Minimum	†	†	0
Maximum	†	†	2.24
Average weighted proportion correct	0.51	0.66	0.10
Average r–biserial[3]	0.72	0.81	0.63
Age 13			
Number of items	98	0	5
Average percentage–missing[1]	2.28	†	13.46
Minimum	0	†	3.97
Maximum	23.23	†	22.67
Average percentage–off–task[2]	†	†	0.43
Minimum	†	†	0.14
Maximum	†	†	0.78
Average weighted proportion correct	0.65	†	0.37
Average r–biserial[3]	0.69	†	0.68
Age 17			
Number of items	86	0	7
Average percentage–missing[1]	1.28	†	12.01
Minimum	0	†	2.45
Maximum	17.33	†	35.34
Average percentage–off–task[2]	†	†	0.78
Minimum	†	†	0
Maximum	†	†	1.60
Average weighted proportion correct	0.72	†	0.48
Average r–biserial[3]	0.76	†	0.70

†Not applicable.
[1]Missing includes the categories "omitted" and "not–reached." (Section 2.3 provides detailed information on these categories.)
[2]"Off–task" (constructed–response items only) is a response that is unrelated to the question and considered inappropriate.
[3]R–biserials are computed at the block level.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

3.4 Treatment of Constructed–Response Items

Data for constructed–response items in the long-term trend analysis were used for the 1984, 1990, 1992, 1994, 1996, and 1999 assessments only. Constructed–response items were not included in the original scoring of the 1988 reading assessment because a previous study (Zwick, 1988) had shown that scoring inconsistencies (drops in interrater reliability and/or scorer drift—that is, scorers showed evidence of rating items more strictly or more leniently than did the original 1984 scorers) had affected these items. A similar review was performed on constructed–response items in all subsequent years (1990, 1992, 1994, 1996, and 1999) and scoring did not suffer the same inconsistencies as the 1988 scoring.

Rater reliability within year was computed for the 1999 constructed–response items at each age. Between–year reliability was also studied with the 1996 and the 1984 responses. Results of the rater reliability study conducted in 1999 are provided in part one, table 1–9. In general, the 1999 scoring did not show irregularities.

The items that were excluded from calibration in the previous assessments were deleted in the 1999 calibration and are listed in table 3–7. The remaining constructed–response items were dichotomized according to criteria developed by subject–area experts. The dichotomized versions of the constructed–response items were included in the calibration.

Table 3–7. Items deleted from the NAEP reading long-term trend analysis: 1999

Age	Block	Item	Reason for exclusion
9	J	N001801	Excluded in previous assessments
	M	N003003	Excluded in previous assessments
	J	N008905	Excluded in previous assessments (constructed–response item)
13	J	N001801	Excluded in previous assessments
	J	N001904	Excluded in previous assessments (constructed–response item)
	K	N002302	Excluded in previous assessments
	L	N002804	Excluded in previous assessments (constructed–response item)
	Q	N005001	Excluded in previous assessments
17	J	N001702	Excluded in previous assessments
	K	N002302	Excluded in previous assessments
	Q	N015905	Excluded in previous assessments (constructed–response item)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

3.5 IRT Scaling for the NAEP 1999 Reading Long-Term Trend Assessment

3.5.1 Item Parameter Estimation

The first step in the scaling process was the estimation of item parameters for the long-term trend items. This item calibration was performed using the BILOG/PARSCALE program described in part two, section 2.4. Items were calibrated separately for each of the three age/grade groups. Item parameters were estimated using combined data from the assessment years 1996 and 1999, treating each assessment as a sample from a separate subpopulation. Student weights were used for the calibration. To ensure that each assessment year had a similar influence on the calibration, student weights for the 1996 examinees were multiplied by a constant, to adjust them to have the same sum

as the sum of the weights for the 1999 examinees. Approximately 600–700 examinee responses for each item were present in each assessment year.

Since five new "nuts" items were added to the 1999 assessment, starting values for item parameters were based on the item parameters created by the current item analysis for all items, including the new items, instead of the final item parameter values from the analysis of the 1996 long-term trend assessment. At each age, when scaling both assessment years together for linking, the five old "nuts" items were included in the scale for the 1996 sample and the five new "nuts" items were included in the scale for the 1999 sample.

As described in part two, section 2.4, BILOG/PARSCALE calibrations were completed in two stages. At stage one, the proficiency distribution of each assessment year was constrained to be normal, although the means and variances differed across assessment years. The values of the item parameters from this normal solution were then used as starting values for a second–stage estimation run in which the proficiency distribution (modeled as a separate multinomial distribution for each assessment year) was estimated concurrently with item parameters. Calibration was concluded when changes in item parameters became negligibly small (i.e., less than .005).

3.5.2 Derived Background Variables

In the long-term trend analysis, all derived background variables were used to define subgroups of students for reporting. For this reason, these variables were also used in conditioning. Derived reporting variables are described in part one, section 1.8.

3.5.3 Evaluation of Model Fit

During and subsequent to item parameter estimation, evaluations of the fit of the IRT models were carried out for each of the items. These evaluations were based primarily on graphical analysis. First, model fit was evaluated by examining plots of nonmodel–based estimates of the expected proportion correct (conditional on proficiency) versus the proportion correct predicted by the estimated item response model (see part two, section 2.4, and Mislevy and Sheehan, 1987, p. 302). In making decisions about excluding items from the final scales, a balance was sought between being too stringent, hence deleting too many items and possibly damaging the content representativeness of the pool of scaled items, and being too lenient, hence including items with models that fit poorly enough to endanger the types of model–based inferences made from NAEP results. A certain degree of misfit was tolerated for a number of items included in the final scales.

Most of the items fit the model well. Items excluded from the analysis of the 1999 assessment were the same items that were deleted from the 1996 reading long-term trend analysis. Table 3–7 lists items that were excluded from the analysis of the 1999 long-term trend assessment.

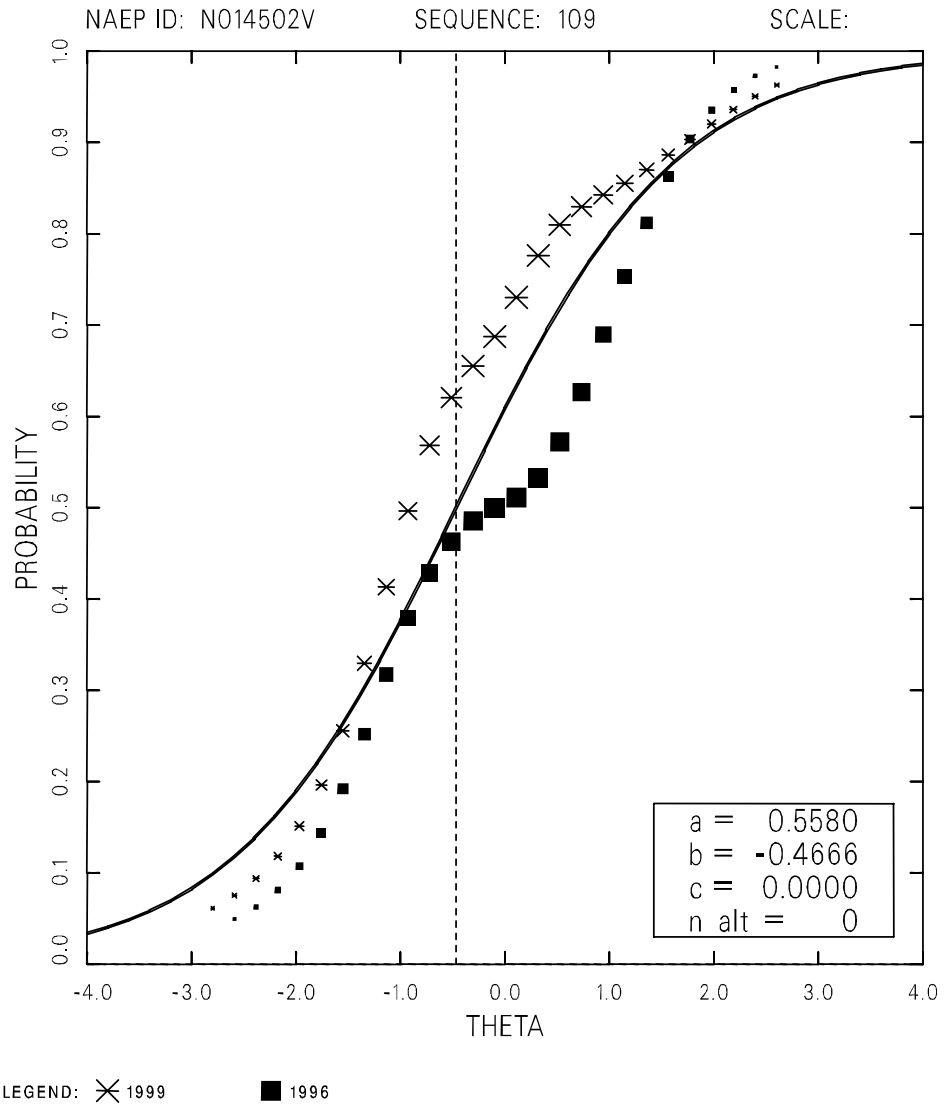
The adequacy of the assumption of a common item response function across assessment years was also evaluated by comparing the nonmodel–based expected proportions for each assessment year to the single, model–based item response function fit by BILOG/PARSCALE. Items that showed clear evidence of functioning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

trend inferences. These separately scaled items will be reexamined in future long-term trend assessments.

At age 9, two long-term trend reading items were calibrated separately by assessment year. Examination of residual plots identified one constructed–response item as functioning differently across assessments. Figure 3–1 shows item N014502 from the analysis for age 9/grade 4. Data are presented for 1996 (squares), and for 1999 (asterisks)[4]. For middle proficiency values, the two sets of symbols diverge and according to expert judgment, the discrepancy of the item characteristic curves of the two years is substantial. The top (1996 data), and the bottom (1999 data) of figure 3–2 shows the plots for the item treated separately by assessment year; the 1996 data showed poorer fit. After the split of N014502, another item, N001101, was also split due to poor fitting. Figure 3–3 shows the two sets of symbols diverge in the middle proficiency area, data are presented for 1996 (squares) and for 1999 (asterisks). Figure 3–4 shows the plots for the item treated separately by assessment year, the 1996 data on the top and 1999 data on the bottom. In order to maintain the link for the trend, item N014502 was kept in the analysis but with the 1999 data calibrated separately and the 1996 data excluded from the final calibration to convergence. Both the 1999 and 1996 versions of N001101 were included in the final calibration because when the data for N014502 from 1996 was excluded from the analysis, both 1999 and 1996 data for N0001101 fit the model well. Parameter estimates from this run served as the final estimates for age 9.

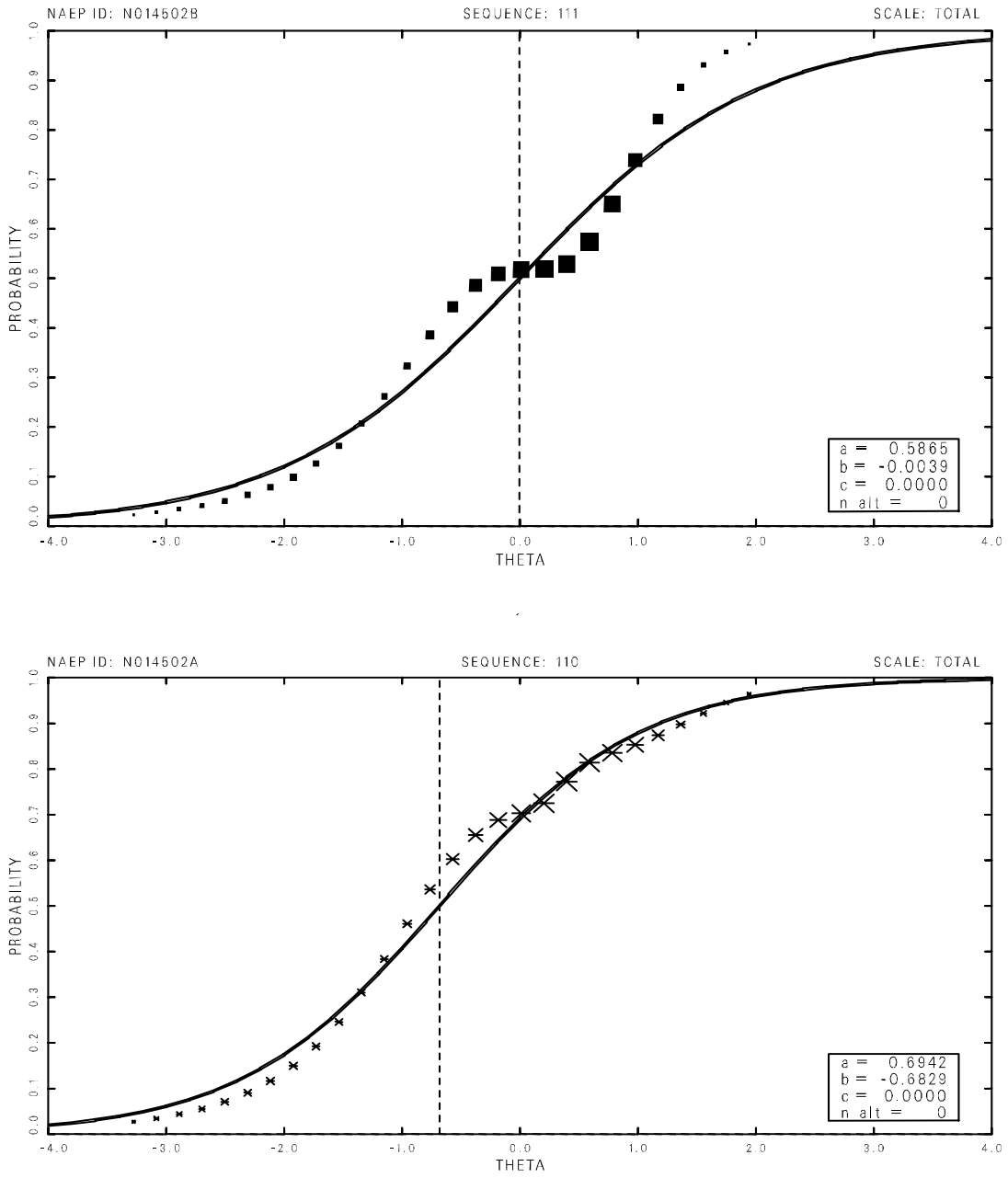
[4]The size of the symbols are proportional to the estimated number of students at a particular scale score level. The symbols are ordinarily larger in the middle of the theta scale, where most students' scale scores fall.

Figure 3–1. Example of NAEP long-term trend item (N014502, age 9) demonstrating DIF across assessment years: 1996 and 1999



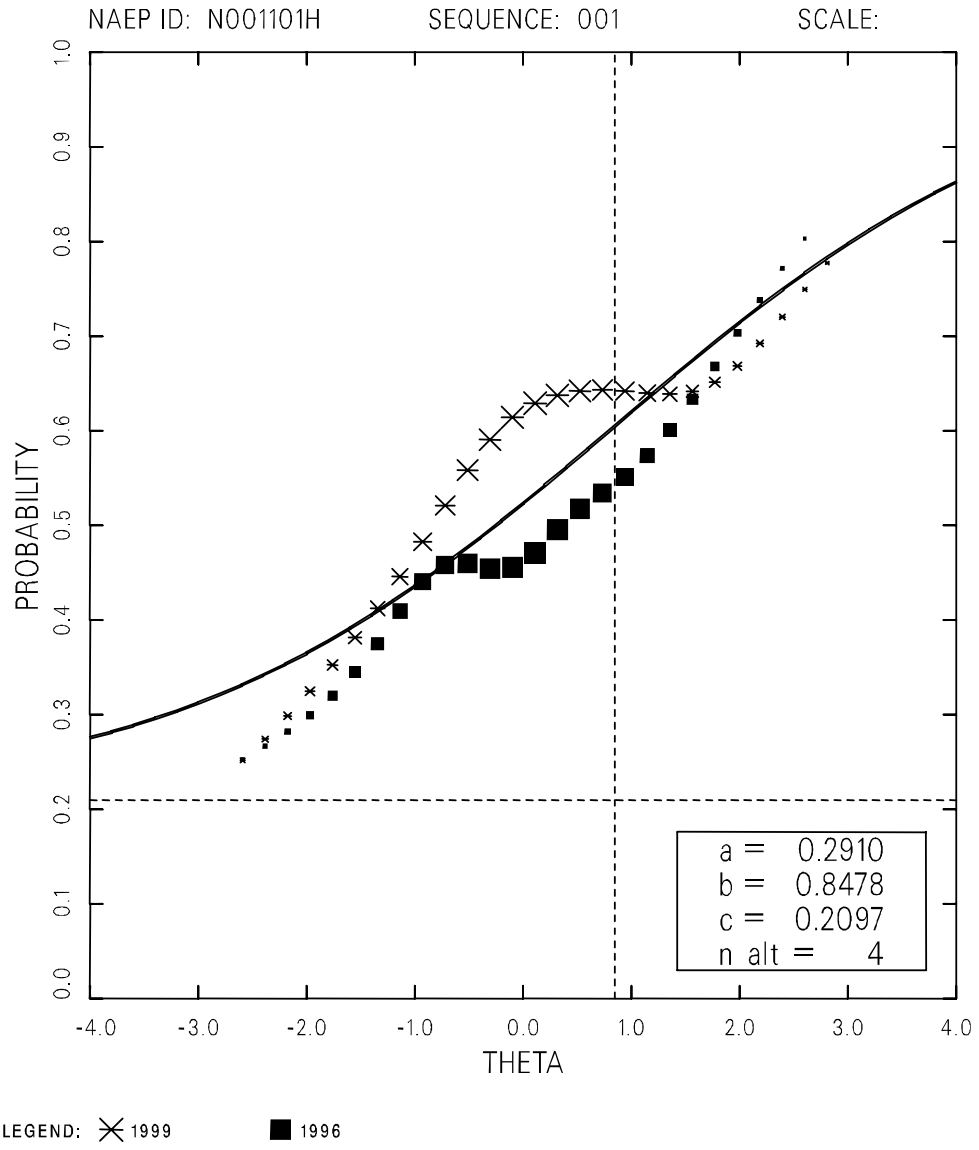
NOTE: This plot compares empirical and model–based estimates of the item response function (IRF). The smooth curve represents the model–based estimate at each provisional proficiency level. The squares represent 1996 data; asterisks represent 1999 data.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Figure 3–2. Example of NAEP long-term trend item (N014502, age 9) fitting separate item response functions for each assessment year: 1996 and 1999



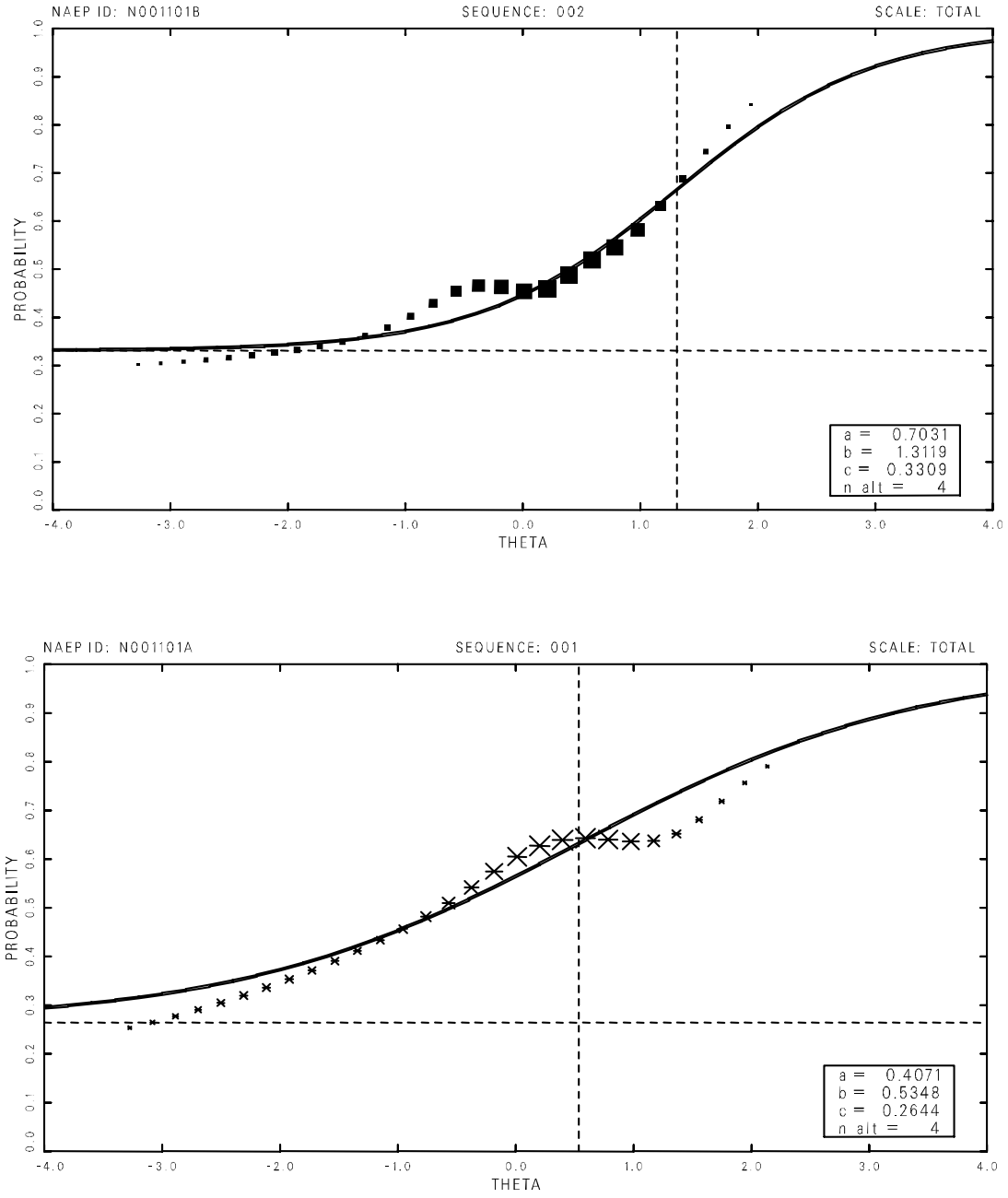
NOTE: The plot compares empirical and model–based estimates of the item response function (IRF). The smooth curve represents the model–based estimate at each provisional proficiency level. The squares represent 1996 data; asterisks represent 1999 data.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Figure 3–3. Example of NAEP long-term trend item (N001101, age 9) demonstrating DIF across assessment years: 1996 and 1999



NOTE: This plot compares empirical and model–based estimates of the item response function (IRF). The smooth curve represents the model–based estimate at each provisional proficiency level. The squares represent 1996 data; asterisks represent 1999 data.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Figure 3–4. Example of NAEP long-term trend item (N001101, age 9) fitting separate item response functions for each assessment year: 1996 and 1999



NOTE: The plot compares empirical and model–based estimates of the item response function (IRF). The smooth curve represents the model–based estimate at each provisional proficiency level. The squares represent 1996 data; asterisks represent 1999 data.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

At age 13, two items (N002201 and N002202) caused difficulty in scaling and both items had large slope parameter values (3.8 and 5.1, respectively) in preliminary calibrations. Further examination of the items indicated that this might be due to local dependence of these two items. The approach of fixing the slope–parameter was taken to obtain stable item parameter estimates. After the convergence of estimation with the proficiency distribution constrained to be normally distributed, the slope–parameter of N002201 was fixed at its converged value. Then the rest of the parameters were calibrated to convergence with the proficiency distribution not constrained to be normally distributed. Parameter estimates from this run served as the final estimates for age 13.

Similar dependence problem also occurred at age 17 for items N002201 and N002202, and their slope parameter values in preliminary calibrations were 3.7 and 4.4, respectively. The same approach used for age 13 was applied. At calibration stage–two, after the estimation of the proficiency distribution was constrained to be normally distributed and calibrated to convergence, the slope–parameter of N002201 was fixed at the value, and all items were calibrated to convergence. Parameter estimates from this run served as the final estimates for age 17.

The remaining misfit is relatively small. All together, six items received treatments during the analysis; table 3–8 lists the two items that were calibrated separately by assessment year. A list of the items scaled for each of the ages, along with their item parameter estimates, appears in appendix B.

Table 3–8. Items calibrated separately by assessment year in the NAEP reading long-term trend analysis

Age	Block	Item	Reason for separate calibration
9	22	N014502	Fit poorly to common item response function across assessments
9	8	N001101	Fit poorly to common item response function across assessments

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

3.6 Generation of Plausible Values

The generation of plausible values was conducted independently for each age/grade level for each of the assessment years. The item parameters from BILOG/PARSCALE, final student weights, item responses, and selected background variables were used with the computer program BGROUP (described in part two, section 2.4) to generate the values for each age. The background variables included student demographic characteristics (e.g., race/ethnicity of the student, highest level of education attained by parents), students' perceptions about reading, and student behavior both in and out of school (e.g., amount of television watched daily, amount of homework done each day). Appendix C gives the codings for the conditioning variables for the three age groups. Table 3–9 contains a list of the number of background contrasts included in conditioning, as well as the proportion of variance accounted for by the conditioning model for each age/grade.

Table 3–9. Proportion of proficiency variance accounted for by the conditioning model for the NAEP reading long-term trend assessment: 1999

Age/grade	Number of conditioning contrasts[1]	Proportion of proficiency variance
9/4	47	0.33
13/8	47	0.35
17/11	45	0.32

[1]Excluding the constant term.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

3.7 The Final NAEP Reading Long-Term Trend Scale

The linear indeterminacy of the long-term trend scale was resolved by linking the 1999 long-term trend scales to previous long-term trend scales. For each age, the item parameters from the joint calibration based on data from both 1996 and 1999 were used with the 1996 data to reestimate plausible values for the 1996 data. The mean and standard deviation of the new 1996 estimates were calculated and matched to the mean and standard deviation of the old 1996 plausible values that were reported previously. The linear constants of this transformation were then applied to transform the 1999 scales to the 1996 proficiency metric. (For score metric transformation, see part two, section 2.4.3). The transformation equations that resulted from this matching of the first two moments for the 1996 data are

$$\text{Age 9: } \theta_{target} = 48.92 \bullet \theta_{calibrated} + 209.64,$$

$$\text{Age 13: } \theta_{target} = 39.51 \bullet \theta_{calibrated} + 257.29, \text{ and}$$

$$\text{Age 17: } \theta_{target} = 43.72 \bullet \theta_{calibrated} + 283.56,$$

where θ_{target} denotes values on the final transformed scale, and $\theta_{calibrated}$ denotes values on the calibration scale. Overall summary statistics for the reading long-term trend samples are given in table 3–10.

Table 3–10. Means and standard deviations on the NAEP reading long-term trend scale: 1984–1999

Age	Assessment year	All five plausible values Mean	Standard deviation
9	1984	211.0	41.1
	1988	211.8	41.2
	1990	209.2	44.7
	1992	210.5	40.4
	1994	211.0	40.5
	1996	212.5[1]	39.0[1]
	1999	211.7	39.1
13	1984	257.1	35.5
	1988	257.5	34.7
	1990	256.8*	36.0
	1992	259.8	39.4
	1994	257.9	39.8
	1996	257.9[1]	39.2[1]
	1999	259.4	38.7
17	1984	288.8	40.3
	1988	290.1	37.1
	1990	290.2	41.3
	1992	289.7	43.0
	1994	288.1	44.4
	1996	287.6[1]	42.2[1]
	1999	287.8	41.8

*Significantly different from 1999, as reported in Campbell, et al. (2000). Note that appropriate standard errors for these statistical tests are provided in table B.1 of that report.
[1]These figures have been updated since the publication in the *1996 NAEP Technical Report* (table 14–9), (Allen et al., 1999).
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

As in the past, interpretation of the long-term trend results was facilitated through the provision of scale anchoring information. In 1984, five NAEP reading scale levels were selected as anchor points. These points described in Campbell et al. (2000) are:

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

Detailed descriptions of the skills required to read at each level were derived and benchmark exercises were selected to exemplify each level. These same anchor points were used in the 1988, 1990, 1992, 1994, 1996, and 1999 reading long-term trend reports. The estimated proportion of students in each reporting category who are at or above each anchor point was examined in Campbell et al.

Part Four

Data Analysis for the NAEP 1999 Long-Term Trend Mathematics Assessment[1]

Catherine A. McClellan and Norma A. Norris
Educational Testing Service

4.1 Introduction

Part four describes the analyses performed on the responses to the cognitive and background items in the 1999 long-term trend assessment of mathematics. The emphasis of part four is on the methods and results of procedures used to develop the IRT–based scale scores. The theoretical underpinnings of the IRT and the plausible values methodology used in this section are described in part two, and therefore are not detailed here.

The objectives of the mathematics analyses were to prepare scale values and perform all analyses necessary to produce a long-term trend report in mathematics. The results obtained from these analyses include the years 1973, 1978, 1982, 1986, 1990, 1992, 1994, 1996 and 1999, and are presented in the *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance* (Campbell et al., 2000).

The student samples that were administered mathematics items in the 1999 long-term trend assessment are shown in table 4–1. (See part one, section 1.2.1 for descriptions of the target populations and the sample design used for the assessment.)

The mathematics long-term trend results reported in Campbell et al. (2000) are based on paced–tape administrations at all three age levels. For ages 9 and 13, the long-term trend booklets administered to the students in the long-term trend mathematics sample contained blocks of reading, mathematics, and science items. The science and mathematics blocks were administered by audiotape to pace the students through blocks and to ensure consistent reading of items (the reading block was presented in print form only). The age 17 long-term trend booklets contained only mathematics and science blocks, both administered by paced tape–recordings as well. All students received a block of common background questions, distinct for each age. Subject–area background questions were presented in the cognitive blocks. The booklets for the age 9 and age 13 samples (Booklets 91–93), and the booklets for the age 17 samples (Booklets 84–85), were the same as those used for mathematics long-term trend assessments in 1986, 1990, 1992, 1994, and 1996. The booklets and the blocks within those booklets are listed in tables 1–3 through 1–5 in part one. This section includes specific information about the mathematics long-term trend items that were scaled.

[1]Catherine A. McClellan was the primary person responsible for the planning, specification, and coordination of the mathematics long-term trend analyses. Computer activities for all long-term trend mathematics scaling and data analyses were performed by Norma A. Norris. Nancy L. Allen, and John R. Donoghue provided consultation.

Table 4–1. NAEP mathematics long-term trend student samples: 1999

Sample	Booklet IDs	Mode	Cohort assessed	Time of testing	Age definition	Modal grade	Number assessed
9 [MS–LTTrend]	91–93	Tape	Age 9	1/3/99 – 3/8/99 (Winter)	CY	4	6,032
13 [MS–LTTrend]	91–93	Tape	Age 13	10/9/98 – 12/22/98 (Fall)	CY	8	5,941
17 [MS–LTTrend]	84–85	Tape	Age 17	3/11/99 – 5/10/99 (Spring)	Not CY	11	3,795

LEGEND

MS	Mathematics and science
LTTrend	Long-term trend assessment: booklets are identical to 1986 long-term trend assessments
Tape	Audiotape administration
CY	Calendar year: birthdates in 1989 and 1985 for ages 9 and 13, respectively
Not CY	Age 17 only: birthdates between October 1, 1981, and September 30, 1982

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 4–2 clarifies the relationships among the 1999 mathematics long-term trend samples and samples from previous years. For all ages, the 1999 mathematics long-term trend samples allow direct comparisons with 1986, 1990, 1992, 1994, and 1996 mathematics long-term trend samples because the same booklets were used in these assessments. There was also a tape administration in 1988 at ages 9 and 13 that was comparable to the other years. However, a tape administration was not conducted at age 17 in 1988. Instead, a noncomparable paper–based assessment was conducted. Hence, 1988 is not included as a point in the mathematics long-term trend reporting. In 1986, the mathematics long-term trend items were scaled with common items from the 1977 and 1982 assessments. Because the 1973 assessment had few items in common with the current assessment, data from that assessment was not scaled using the IRT model, but was linked to the mathematics long-term trend line by a linear transformation involving the logit of mean proportion correct for common items (see *Expanding the New Design: The NAEP 1985–86 Technical Report* [Beaton, 1988]). When comparisons were made including the 1973 assessment results, z–tests rather than t–tests were used to test statistical significance (see section 2.5 in part two). Since 1990, successive assessments have been placed on the common scale using data from the preceding assessment.

Information about previous mathematics trend assessment years is available in: chapter 10 of *Expanding the New Design: The NAEP 1985–86 Technical Report* (Johnson, 1988), chapter 13 of *The NAEP 1990 Technical Report* (Yamamoto and Jenkins, 1992), chapter 13 of *The NAEP 1992 Technical Report* (Jenkins and Kulick, 1994), chapter 16 of *The NAEP 1994 Technical Report* (Ip, Jenkins, and Kulick, 1996), and chapter 15 of *The NAEP 1996 Technical Report* (Qian and Norris, 1999).

Table 4–2. NAEP mathematics samples contributing to the 1999 long-term trend results: 1973–1999

Cohort assessed	Year	Sample	Subjects	Time of testing	Mode of administration	Age definition	Modal grade
Age 9	1973	Main	MS	Winter	Tape	CY	4
	1977	Main	M	Winter	Tape	CY	4
	1982	Main	MSC	Winter	Tape	CY	4
	1986	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1990	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1992	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1994	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1996	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1999	LTTrend[1]	MS	Winter	Tape[2]	CY	4
Age 13	1973	Main	MS	Fall	Tape	CY	8
	1977	Main	M	Fall	Tape	CY	8
	1982	Main	MSC	Fall	Tape	CY	8
	1986	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1990	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1992	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1994	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1996	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1999	LTTrend[1]	MS	Fall	Tape[2]	CY	8
Age 17	1973	Main	MS	Spring	Tape	Not CY	11
	1977	Main	M	Spring	Tape	Not CY	11
	1982	Main	MSC	Spring	Tape	Not CY	11
	1986	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11
	1990	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11
	1992	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11
	1994	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11
	1996	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11
	1999	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11

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

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

LEGEND

M	Math	Tape	Audiotape administration
MS	Mathematics and science		
MSC	Mathematics, science, and civics	CY	Calendar year: birthdates in 1989 and 1985 for ages 9 and 13 in the 1999 assessment
Main	Main assessment		
LTTtrend	Long-term trend: booklets are identical to the long-term trend assessment of 1986	Not CY	Age 17 only: birthdates between October 1 and September 30 of the appropriate years

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The numbers of scaled items in common across different age combinations are presented in table 4–3. As in previous mathematics long-term trend analyses, each age was scaled separately. Item parameters were estimated assuming a univariate scale, since the number of items presented to each student was small and there were too few items to estimate several content area scales separately.

The numbers of items scaled in 1999 that were common across assessment years are presented in table 4–4. The 1986, 1990, 1992, 1994, 1996, and 1999 assessments had all items in common. For age 9, the number of items common across assessment years 1978 to 1999 was 35; for age 13, the number was 56; and for age 17, the number was 54.

Table 4–3. Numbers of scaled items in the NAEP mathematics long-term trend assessment common across ages: 1999

Age	Booklet numbers	Number of items
Total		153
9 only	91–93	32
13 only	91–93	30
17 only	84–85	41
9 and 13 only	91–93, 91–93	20
9 and 17 only	91–93, 84–85	0
13 and 17 only	91–93, 84–85	27
9, 13, and 17	91–93, 91–93, 84–85	3

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 4–4. Numbers of scaled NAEP mathematics long-term trend items common across assessments: 1986–1999

	Number of items		
Assessment year	Age 9	Age 13	Age 17
1986, 1990, 1992, 1994, 1996, 1999	55	80	71
1982, 1986, 1990, 1992, 1994, 1996, 1999	53	79	65
1978, 1986, 1990, 1992, 1994, 1996, 1999	35	56	54
1978, 1982, 1986, 1990, 1992, 1994, 1996, 1999	35	56	54

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

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

The second step was to fit an IRT model to the data from the 1999 and 1996 assessments for each age separately. This procedure puts item parameters and ability estimates on the same scale across years. The same item may have different item parameters for different age groups.

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

Finally, the scales of the 1999 mathematics long-term trend assessment were transformed to the proficiency scale used in previous mathematics trend assessments. These proficiency means constitute the last point in the mathematics long-term trend from 1973 to 1999. The only available estimates of the proficiency means for 1973 were linked via extrapolation to the IRT scale, but the data from that year was not scaled using an IRT model (see section 4.6 for further information on the extrapolation).

4.2 Item Analysis for the NAEP 1999 Mathematics Long-Term Trend Assessment

Conventional item analyses did not identify any difficulties with the 1999 mathematics long-term trend data. Table 4–5 contains information about the mathematics long-term trend blocks. The correspondence between blocks, booklets, and samples is given for the mathematics long-term trend assessment in tables 1–3 through 1–5 in part one. Common labeling of these blocks across ages does not denote common items.

Table 4–5 contains the number of scaled items, size of the sample administered to the block, mean weighted proportion correct, mean weighted r–biserial, and mean weighted alpha as a measure of reliability for each block. The average values were calculated using examinee sampling weights and the responses to the items in the block that were scaled. On average, the 1999 item–level statistics were not very different from those for the 1996 assessments. Similar statistics for the 1996 assessment were reported in table 15–5 of *The NAEP 1996 Technical Report* (Allen, Carlson, and Zelenak, 1999). The percent of examinees not reaching items in the mathematics long-term trend blocks was almost always zero because the items were administered with a tape–recording to pace response time.

Table 4–5. NAEP mathematics long-term trend descriptive statistics for item blocks as defined after scaling: 1999

Statistic	Block M1	M2	M3[1]
Age 9			
Number of scaled items	24	26	5
Number of scaled constructed response items	9	9	0
Unweighted sample size	2,032	2,135	1,865
Average weighted proportion correct	.62	.64	.69
Average weighted r–biserial	.62	.65	.80
Weighted alpha reliability	.82	.86	.47
Age 13			
Number of scaled items	36	36	8
Number of scaled constructed response items	9	8	0
Unweighted sample size	2,019	1,962	1,960
Average weighted proportion correct	.69	.63	.66
Average weighted r–biserial	.58	.57	.73
Weighted alpha reliability	.86	.86	.67
Age 17			
Number of scaled items	33	33	5
Number of scaled constructed response items	10	5	1
Unweighted sample size	1,953	1,953	1842
Average weighted proportion correct	.65	.66	.57
Average weighted r–biserial	.70	.64	.75
Weighted alpha reliability	.91	.88	.51

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

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 4–5a. NAEP mathematics long-term trend summary response rates by item type: 1999

Statistics	Multiple–choice	Short constructed–response
Age 9		
Number of items	37	18
Average percentage–missing[1]	1.25	3.13
Minimum	0.04	0.47
Maximum	6.50	6.50
Average weighted proportion correct	0.64	0.65
Average r–biserial[2]	0.65	0.67
Age 13		
Number of items	63	17
Average percentage–missing[1]	1.12	2.53
Minimum	0.13	0.32
Maximum	3.78	6.99
Average weighted proportion correct	0.64	0.72
Average r–biserial[2]	0.59	0.59
Age 17		
Number of items	55	16
Average percentage–missing[1]	0.95	6.02
Minimum	0.23	0.65
Maximum	4.46	9.92
Average weighted proportion correct	0.69	0.52
Average r–biserial[2]	0.67	0.71

[1]Missing includes the categories "omitted" and "not–reached." (Section 2.3 provides detailed information on these categories.)
[2]R–biserials are computed at the block level.
NOTE: The long-term trend mathematics assessments included no extended constructed–response items.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

In the 1999 mathematics long-term trend assessment, 20 percent of the samples of the constructed–response items were used to check the interrater reliability—the score agreement between first and second raters. The percent of exact agreement ranged from 97.1 to 100 percent; and the intraclass correlation ranged from .908 to 1.00. In general, the interrater reliability was very high in the 1999 mathematics long-term trend assessment.

4.3 IRT Scaling for the NAEP 1999 Mathematics Long-Term Trend Assessment

4.3.1 Item Parameter Estimation

The scaling process began with the estimation of item parameters for the long-term trend items. This item calibration was performed using the NAEP version of the BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs as described in part two, section 2.4. Items calibration was performed separately for each of the three age groups, using combined data from the 1996 and 1999 assessment years. The data from the two assessment years were treated as sampling from separate subgroups. Including the 1996 assessment data assures that item parameters will be similar for adjacent assessments so that year–to–year trends will not be distorted by abrupt changes in calibration, and to make it possible to link the current long-term trend assessment to the previous assessments. The calibration was performed on the entire sample of students, resulting in a range of about 1,700 to 1,900 examinee responses to each item in each assessment year. The calibration was

based on student weights that were rescaled for the 1999 data so that the sum of the weights equaled the unweighted sample size. Also, weights for the 1999 data were restandardized to give equal weight to the two assessment years included in the scaling. As with the previous assessment, calculator items were excluded from the analysis. Because calculators have changed greatly since the start of the long-term trend assessment, it was judged that calculator questions are no longer comparable across time. These items were kept in the assessment, since excluding them would have changed the testing context.

Since parameters for items in blocks M1, M2, and M3 were estimated separately for ages 9, 13, and 17, items administered at more than one age have multiple sets of item parameter estimates. Items were examined for lack of fit with the data. Those that exhibited extreme violation of IRT assumptions (i.e., did not have monotonically increasing item characteristic curves) were deleted from the analysis, as they were in previous assessments. Other items were deleted because they were calculator items, which were not considered part of the regular assessment. These excluded items appear in tables 4–6, 4–7, and 4–8. As a result of these deletions, 55 items were scaled for age 9, 80 items were scaled for age 13, and 71 items were scaled for age 17. Of the 153 noncalculator items that were part of the assessment, seven items (5%) were excluded due to poor fit with the data. A list of the items scaled for each of the ages, along with their item parameter estimates, appears in appendix B.

Three items in the 1999 long-term trend mathematics assessment received special treatment. These items are listed in table 4–9. The items were administered in both 1996 and 1999 but showed evidence of having a distinct item response function for each assessment year. It was decided to "split" the item across the assessment years, estimating the item parameters separately for the two years. This resulted in good fit for the items in each year individually.

Table 4–6. Items deleted from the NAEP mathematics long-term trend analysis, age 9: 1999

Booklet IDs	Block	Item	Reason for exclusion
91	M1	N252601	Excluded in previous assessments
		N262502	Excluded in previous assessments
92	M3	N268221	Calculator item
		N276021	Calculator item
		N276022	Calculator item
		N276821	Calculator item
		N276822	Calculator item
		N276823	Calculator item
		N277621	Calculator item
		N277622	Calculator item
		N277623	Calculator item
		N284021	Calculator item
		N284022	Calculator item

NOTE: All calculator items were deleted from the analysis.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 4–7. Items deleted from the NAEP mathematics long-term trend analysis, age 13: 1999

Booklet IDs	Block	Item	Reason for exclusion
91	M1	N262502	Excluded in previous assessments
93	M2	N261601	Excluded in previous assessments
92	M3	N264521	Calculator item
		N259921	Calculator item
		N276821	Calculator item
		N276822	Calculator item
		N276823	Calculator item
		N278921	Calculator item
		N278922	Calculator item
		N278923	Calculator item
		N278924	Calculator item
		N278925	Calculator item
		N280621	Calculator item
		N280622	Calculator item
		N280623	Calculator item
		N280624	Calculator item
		N280625	Calculator item
		N280626	Calculator item

NOTE: All calculator items were deleted from the analysis.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 4–8. Items deleted from the NAEP mathematics long-term trend analysis, age 17: 1999

Booklet IDs	Block	Item	Reason for exclusion
84	M1	N282801	Excluded in previous assessments
		N285701	Excluded in previous assessments
84	M2	N266801	Excluded in previous assessments
		N255301	Excluded in previous assessments
85	M3	N259921	Calculator item
		N264321	Calculator item
		N264521	Calculator item
		N267921	Calculator item
		N276821	Calculator item
		N276822	Calculator item
		N276823	Calculator item
		N278921	Calculator item
		N278922	Calculator item
		N278923	Calculator item
		N278924	Calculator item
		N278925	Calculator item
		N280621	Calculator item
		N280622	Calculator item
		N280623	Calculator item
		N280624	Calculator item
		N280625	Calculator item
		N280626	Calculator item
		N285321	Calculator item

NOTE: All calculator items were deleted from the analysis.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 4–9. Items receiving special treatment in the NAEP mathematics long-term trend analysis: 1999

Booklet ID	Block	Item	Treatment
84	M1	N278501	1996 and 1999 responses split
		N278502	1996 and 1999 responses split
		N278503	1996 and 1999 responses split

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

4.3.2 Derived Background Variables

In the long-term trend analysis, all derived background variables were used to define subgroups of students for reporting. For this reason, these variables were also used in conditioning. Information about the conditioning variables and the respective codings is given in appendix C. A statistical summary of the NAEP 1999 subgroups is displayed in several tables in appendix A.

4.4 Generation of Plausible Values

The generation of plausible values was conducted independently for each age group. The item parameters from NAEP–BILOG/PARSCALE, final student weights, item responses and selected background variables (conditioning variables) were used with the computer program BGROUP (described in part two, section 2.4.3) in order to generate the plausible values for each student. There were 49 contrasts in the conditioning model (See equation 12.8 in chapter 12 of *The NAEP 1998 Technical Report*, [Allen, Carlson, Johnson, and Mislevy, 2001]) at age 9, excluding an overall constant, 52 at age 13, and 58 at age 17. Appendix C gives the codings for the conditioning variables for the three age groups. A check on the distributions of the plausible values for each age was made. The generation of plausible values is described in more detail in part two. Table 4–10 contains a list of the number of background contrasts included in conditioning, as well as the proportion of variance accounted for by the conditioning model for each age. This proportion is the ratio of the difference between the total variance and the BGROUP residual variance, divided by the total variance. The total variance is the mean of the five theta–scale variances obtained by their respective plausible values.

Table 4–10. Proportion of proficiency variance accounted for by the conditioning model for the NAEP mathematics long-term trend assessment: 1999

Age	Number of conditioning contrasts[1]	Proportion of proficiency variance
9	53	.39
13	56	.36
17	63	.52

[1]Excluding the constant term.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

4.5 The Final NAEP Mathematics Long-Term Trend Scale

Since the plausible value (theta) scales have a linear indeterminacy, comparisons with previous assessments will be sensible only if the scale is linearly transformed to a meaningful metric. This indeterminacy was resolved by linking the 1999 scales to previous long-term trend scales. The 1999 data had to be transformed to compensate for linear changes in the scale due to employing newly estimated item parameters and new BGROUP conditioning parameters in 1999. The transformation was accomplished by first reestimating the 1996 student abilities using 1999 item parameters and 1999 BGROUP parameters. (For score metric transformation, see part two, section 2.4.3.) The new 1996 ability estimates were then equated to the old 1996 ability estimates by matching the first two moments (i.e., the mean and standard deviation). The constants for this transformation were then applied to the 1999 data. The transformation equations that resulted are:

$$\text{Age 9:}\quad \theta_{target} = 34.56 \bullet \theta_{calibrated} + 231.15,$$

$$\text{Age 13: } \theta_{target} = 33.07 \bullet \theta_{calibrated} + 274.79, \text{ and}$$

$$\text{Age 17: } \theta_{target} = 30.70 \bullet \theta_{calibrated} + 307.59,$$

where θ_{target} denotes values on the final reporting scale of the 1999 data and $\theta_{calibrated}$ denotes values on the original 1999 calibration (theta) scale. Overall summary statistics for the long-term trend scales are given in table 4–11. The detailed mathematics long-term trend results from the analyses described in this section are reported in Campbell et al. (2000).

Table 4–11. Means and standard deviations on the NAEP mathematics long-term trend scale: 1978–1999

		All five plausible values	
Age	**Assessment**	**Mean**	**Standard deviation**
9	1978	218.6*	36.0
	1982	219.0*	34.8
	1986	221.7*	34.0
	1990	229.6*	32.9
	1992	229.6*	33.1
	1994	231.1	33.2
	1996	231.0	33.8
	1999	232.0	34.1
13	1978	264.1*	39.0
	1982	268.6*	33.4
	1986	269.0*	30.8
	1990	270.4*	31.3
	1992	273.1*	30.9
	1994	274.3	32.4
	1996	274.3	31.6
	1999	275.8	32.6
17	1978	300.4*	34.9
	1982	298.5*	32.4
	1986	302.0*	31.0
	1990	304.6*	31.3
	1992	306.7	30.1
	1994	306.2	30.2
	1996	307.2	30.2
	1999	308.2	30.8

*Significantly different from 1999, as reported in Campbell, et al. (2000). Note that appropriate standard errors for these statistical tests are provided in table B.1 of that report.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

To provide a context for interpreting the overall mathematics long-term trend results, the NAEP mathematics results were "anchored" at five NAEP mathematics scale levels. In 1986, five mathematics scale levels were selected as anchor points, using the process described in *Expanding the New Design: The 1985–86 Technical Report* (Beaton, 1988). These five levels of mathematics proficiency are:

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

These same anchor points were used in 1978, 1982, 1986, 1990, 1992, 1994, 1996, and 1999.

4.6 Extrapolation of the 1973–74 Mean P–Value Results onto the NAEP Mathematics Long-Term Trend Scale

Because of insufficient items in common with the 1986 long-term trend assessment, the 1973–74 mathematics assessment was never included in the scaling of NAEP long-term trend data. However, for the nation and several reporting subgroups (e.g., male, female) at each of the three age levels, an estimate of the 1973–74 mean level of student mathematics proficiency was computed when the data from the 1985–86 assessment were analyzed.

These estimates were obtained by assuming that the relationship within a given age level between the logit of a subgroup's mean p–value (i.e., mean proportion correct) and its respective mathematics proficiency mean was linear and that the same line held for all assessment years and for all subgroups within the age level. Under this assumption, the between–year difference of the mean proficiency values of a subgroup for a pair of assessment years is equal to a constant (B) times the between–year difference of the logits of the mean p–values of that subgroup for the same two years. For each age level, a mean p–value estimate using a common set of items was available for 1973–74, 1977–78, and 1981–82. The constant B was estimated by a regression (through the origin) of the difference between proficiency means in 1977–78 and 1981–82 on the corresponding difference between the logits of the mean p–values for these two years. All subgroups in a given age were included in the regression. The estimate of the 1973–74 proficiency mean for a subgroup was then obtained as the sum of the 1977–78 subgroup mean proficiency and B times the difference between the logits of the 1973–74 and 1977–78 subgroup mean p–values.[2]

The quality of this extrapolation technique was evaluated by comparing its performance in predicting the 1977–78 data. The actual values of the 1977–78 subgroup mean proficiencies were compared with the predicted values formed as the sum of the 1981–82 subgroup mean proficiency and B times the difference between the logits of the 1977–78 and 1981–82 subgroup mean p–values. The predictions were very close to the actual values, the residual means squared error being only .4 percent of the variance of the actual values.

[2]See *Mathematics Data Analysis* (Johnson, 1988).

THIS PAGE INTENTIONALLY LEFT BLANK.

Part Five

Data Analysis for the NAEP 1999 Long-Term Trend Science Assessment[1]

Spencer S. Swinton, Steven P. Isham and Venus Leung
Educational Testing Service

5.1 Introduction

Part five describes the analyses performed on the responses to the cognitive and background items in the 1999 long-term trend assessment of science. The emphasis of part five is on the methods and results of procedures used to develop the IRT–based scale scores. The theoretical underpinnings of the IRT and the plausible values methodology are described in part two, and therefore are not detailed here.

The objectives of the science analyses were to prepare scale values and perform all analyses necessary to produce a long-term trend report in science. The results obtained from these analyses include the years 1969–1970, 1973, 1977, 1982, 1986, 1990, 1992, 1994, 1996 and 1999, and are presented in the *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance* (Campbell et al., 2000).

The student samples that were administered science items in the 1999 long-term trend assessment are shown in table 5–1. (See part one, section 1.2.1 for descriptions of the target populations and the sample design used for the assessment.)

The science long-term trend results reported in Campbell et al. (2000) are based on paced–tape administrations at all three age levels. For ages 9 and 13, the long-term trend booklets administered to the students in the science long-term trend sample contained blocks of reading, mathematics, and science items. The science and mathematics blocks were administered by audiotape to pace the students through blocks and to ensure consistent reading of items (the reading block was presented in print form only). The age 17 long-term trend booklets contained only mathematics and science blocks, both administered by paced tape–recordings as well. All students received a block of common background questions, distinct for each age. Subject–area background questions were presented in the cognitive blocks. The booklets for the age 9 and age 13 samples (Booklets 91–93), and the booklets for the age 17 samples (Booklets 84–85), were the same as those used for science long-term trend assessments in 1986, 1990, 1992, 1994, and 1996. The booklets and the blocks within those booklets are listed in tables 1–3 through 1–5 in part one. This section includes specific information about the science long-term trend items that were scaled.

[1]Spencer Swinton was the primary person responsible for the planning, specification, and coordination of the science long-term trend analyses. Computer activities for all long-term trend science scaling and data analyses were performed by Steven Isham and Venus Leung. Nancy L. Allen provided consultation.

Table 5–1. NAEP science long-term trend student samples: 1999

Sample	Booklet IDs	Mode	Cohort assessed	Time of testing	Age definition	Modal grade	Number assessed
9 [MS–LTTrend]	91–93	Tape	Age 9	1/3/99 – 3/8/99 (Winter)	CY	4	6,032
13 [MS–LTTrend]	91–93	Tape	Age 13	10/9/98 – 12/22/98 (Fall)	CY	8	5,941
17 [MS–LTTrend]	84–85	Tape	Age 17	3/11/99 – 5/10/99 (Spring)	Not CY	11	3,795

LEGEND

MS	Mathematics and science
LTTrend	Long-term trend assessment: booklets are identical to 1986 long-term trend assessments
Tape	Audiotape administration
CY	Calendar year: birthdates in 1989 and 1985 for ages 9 and 13, respectively
Not CY	Age 17 only: birthdates between October 1, 1981, and September 30, 1982

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 5–2 clarifies the relationships among the 1999 science long-term trend samples and samples from previous years. For all ages, the 1999 science long-term trend samples allow direct comparisons with 1986, 1990, 1992, 1994, and 1996 science long-term trend samples because the same booklets were used in these assessments. There was also a tape administration in 1988 at ages 9 and 13 that was comparable to the other years. However, a tape administration was not conducted at age 17 in 1988. Instead, a noncomparable paper–based assessment was conducted. Hence, 1988 is not included as a point in the science long-term trend reporting. In 1986, the science long-term trend items were scaled with common items from the 1977 and 1982 assessments. Because of the small number of items in common with those in the 1969–70 and 1973 assessments, data from those assessments were not scaled using the IRT model, but were linked to the science long-term trend line by a linear transformation involving the logit of mean proportion correct for common items (see *Expanding the New Design: The NAEP 1985–86 Technical Report* [Beaton, 1988]). When comparisons were made including the 1969–70 and 1973 assessment results, z–tests rather than t–tests were used to test statistical significance (see section 2.5 in part two).

Since 1990, successive assessments have been placed on the common scale using data from the preceding assessment. Information about previous assessment years, including 1969–70 and 1973, is available in chapter 11 of *Expanding the New Design: The NAEP 1985–86 Technical Report* (Yamamoto, 1988), chapter 14 of *The NAEP 1990 Technical Report* (Allen, 1992), chapter 14 of *The NAEP 1992 Technical Report* (Allen and Isham, 1994), and chapter 17 of *The NAEP 1994 Technical Report* (Swinton, Allen, Isham and Chen, 1996), and chapter 16 of *The NAEP 1996 Technical Report* (Allen, Carlson, and Zelenak, 1999).

Table 5–2. NAEP science samples contributing to the 1999 long-term trend results: 1970–1999

Cohort assessed	Year	Sample	Subjects	Time of testing	Mode of administration	Age definition	Modal grade
Age 9	1970	Main	SWC	Winter	Tape	CY	4
	1973	Main	MS	Winter	Tape	CY	4
	1977	Main	SCI	Winter	Tape	CY	4
	1982	Main	MSC	Winter	Tape	CY	4
	1986	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1990	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1992	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1994	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1996	LTTrend[1]	MS	Winter	Tape[2]	CY	4
	1999	LTTrend[1]	MS	Winter	Tape[2]	CY	4
Age 13	1970	Main	SWC	Fall	Tape	CY	8
	1973	Main	MS	Fall	Tape	CY	8
	1977	Main	SCI	Fall	Tape	CY	8
	1982	Main	MSC	Fall	Tape	CY	8
	1986	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1990	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1992	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1994	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1996	LTTrend[1]	MS	Fall	Tape[2]	CY	8
	1999	LTTrend[1]	MS	Fall	Tape[2]	CY	8
Age 17	1969	Main	SWC	Spring	Tape	Not CY	11
	1973	Main	MS	Spring	Tape	Not CY	11
	1977	Main	SCI	Spring	Tape	Not CY	11
	1982	Main	MSC	Spring	Tape	Not CY	11
	1986	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11
	1990	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11
	1992	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11
	1994	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11
	1996	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11
	1999	LTTrend[1]	MS	Spring	Tape[2]	Not CY	11

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

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

LEGEND

SCI	Science	LTTrend	Long-term trend: booklets are identical to the long-term trend assessment of 1986
MS	Mathematics and science		
MSC	Mathematics, science, and civics	Tape	Audiotape administration
SWC	Science, writing, and citizenship	CY	Calendar year: birthdates in 1989 and 1985 for ages 9 and 13 in the 1999 assessment
Main	Main assessment	Not CY	Age 17 only: birthdates between October 1 and September 30 of the appropriate years

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The numbers of items scaled in 1999 that were common across different age combinations are presented in table 5–3. As in previous science long-term trend analyses, each age was scaled separately. Item parameters were estimated assuming a univariate scale, since the number of items presented to each student was small and there were too few items to estimate several content area scales separately.

The numbers of items scaled in 1999 that were common across assessment years are presented in table 5–4. The 1986, 1990, 1992, 1994, 1996, and 1999 assessments had all items in common. For age 9, the number of items common across assessment years 1977 to 1999 was 10; for age 13, the number was 58; and for age 17, the number was 45.

Table 5–3. Numbers of scaled items in the NAEP science long-term trend assessments common across ages: 1999

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

[1]One of these items (N406303) was treated as a different item from 1990 in the scaling of the 1992 assessment, but only for age 13. It was treated as an item common to 1992, 1994, 1996, and 1999 for all ages in the 1994, 1996, and 1999 assessments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 5–4. Numbers of scaled items in the NAEP science long-term trend items common across assessments: 1986–1999

	Number of items		
Assessment years	Age 9	Age 13	Age 17
1986, 1990, 1992, 1994,1996, 1999	56	76	78
1982, 1986, 1990, 1992, 1994, 1996, 1999	10[1]	58	47
1977, 1986, 1990, 1992, 1994, 1996, 1999	56	76	76
1977, 1982, 1986, 1990, 1992, 1994, 1996, 1999	10[1]	58[2]	45

[1]Twenty–four items common to years 1977 and 1982, but not later years, were included in the 1986 scaling of these items to stabilize the estimation of the item parameters. See *Expanding the New Design: The NAEP 1985–86 Technical Report* (Beaton, 1988) for more information.
[2]One of these items (N406303) was treated as a different item from 1990 in the scaling of the 1992 assessment, but only for age 13. It was treated as an item common to 1992, 1994, 1996, and 1999 in the 1994,1996, and 1999 assessments for all ages.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

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

The second step was to fit an IRT model to the data from the 1999 and 1996 assessments for each age separately. This procedure puts item parameters and ability estimates on the same scale across years. The same item may have different item parameters for different age groups.

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

Finally, the scales of the 1999 science long-term trend assessment were transformed to the proficiency scale used in previous science trend assessments. These proficiency means constitute the last point in the science long-term trend from 1969–70 to 1999. The only available estimates of the proficiency means for 1969–70 and 1973 were linked via extrapolation to the IRT scale, but the data from those years were not scaled using an IRT model.[2]

5.2 Item Analysis for the NAEP 1999 Science Long-Term Trend Assessment

Conventional item analyses did not identify any difficulties with the 1999 science long-term trend data. Table 5–5 contains information about the science long-term trend blocks. At all ages, the blocks labeled S1, S2, and S3 were presented intact to students in the 1986, 1990, 1992, 1994, 1996 and 1999 long-term trend samples. The age 9 and age 13 blocks appeared in Booklets 91 through 93. For age 17, Block S3 was in Booklet 84, and Blocks S1 and S2 were in Booklet 85. The correspondence between blocks, booklets, and samples is given for the long-term trend assessment in tables 1–3 through 1–5 in part one. Common labeling of these blocks across ages does not denote common items.

Table 5–5 contains the number of scaled items, size of the sample administered the block, mean weighted proportion correct, mean weighted r–biserial, and mean weighted alpha as a measure of reliability for each block. The average values were calculated using examinee sampling weights and the responses to the items in the block that were scaled. On average, the 1999 item–level statistics were not very different from those for the 1996 assessments. Similar statistics for the 1996 assessment were reported in table 16–5 of The *NAEP 1996 Technical Report* (Allen, et al., 1999). The percent of examinees not reaching items in the science long-term trend blocks was almost always zero because the items were administered with a tape–recording to pace response time. The science long-term trend contained no constructed–response items.

[2]See *Science Data Analysis* (Yamamoto, 1988).

Table 5–5. NAEP science long-term trend descriptive statistics for item blocks as defined after scaling: 1999

Statistic	Block S1	S2	S3
Age 9			
Number of scaled items	17	20	19
Number of scaled constructed–response items	0	0	0
Unweighted sample size	2,032	1,865	2,135
Average weighted proportion correct	0.62	0.58	0.70
Average weighted r–biserial	0.56	0.46	0.59
Weighted alpha reliability	0.68	0.60	0.73
Age 13			
Number of scaled items	23	30	23
Number of scaled constructed–response items	0	0	0
Unweighted sample size	2,019	1,960	1,962
Average weighted proportion correct	0.54	0.56	0.60
Average weighted r–biserial	0.52	0.48	0.52
Weighted alpha reliability	0.73	0.76	0.73
Age 17			
Number of scaled items	24	31	23
Number of scaled constructed–response items	0	0	0
Unweighted sample size	1,842	1,842	1,953
Average weighted proportion correct	0.65	0.65	0.61
Average weighted r–biserial	0.48	0.52	0.64
Weighted alpha reliability	0.67	0.77	0.82

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 5–5a. NAEP science long-term trend summary response rates by item type: 1999

Statistics	Multiple–choice
Age 9	
Number of items	56
Average percentage–missing[1]	0.75
Minimum	0.00
Maximum	1.87
Average weighted proportion correct	0.63
Average r–biserial[2]	0.52
Age 13	
Number of items	76
Average percentage–missing[1]	0.63
Minimum	0.05
Maximum	2.86
Average weighted proportion correct	0.57
Average r–biserial[2]	0.49
Age 17	
Number of items	78
Average percentage–missing[1]	0.50
Minimum	0.13
Maximum	1.53
Average weighted proportion correct	0.64
Average r–biserial[2]	0.54

[1]Missing includes the categories "omitted" and "not–reached." (Section 2.3 provides detailed information on these categories.)

[2]R–biserials are computed at the block level.

NOTE: The science long-term trend assessments included no constructed–response items.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

5.3 IRT Scaling for the NAEP 1999 Science Long-Term Trend Assessment

5.3.1 Item Parameter Estimation

The scaling process began with the estimation of item parameters for the long-term trend items. This item calibration was performed using the NAEP version of the BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs described in part two, section 2.4. Item calibration was performed separately for each of the three age groups, using combined data from the 1996 and 1999 assessment years. The data from the two assessment years were treated as sampling from separate subgroups. Including the 1996 assessment data assures that item parameters will be similar for adjacent assessments so that year–to–year trends will not be distorted by abrupt changes in calibration, and to make it possible to link the current long-term trend assessment to the previous assessments. The calibration was performed on the entire sample of students, resulting in a range of about 1,700 to 1,900 examinee responses to each item in each assessment year. The calibration was based on student weights that were rescaled for the 1999 data so that the sum of the weights equaled the unweighted sample size. Also, weights for the 1999 data were restandardized to give equal weight to the two assessment years included in the scaling.

Although other items were examined for irregularities, only items that were deleted from the previous scaling of the paced–tape long-term trend data were excluded in the 1999 analysis. Eight percent of the items (18 items) administered to the long-term trend sample were excluded from analyses of previous assessments. The deleted items appear in tables 4–6, 4–7 and 4–8. As a result of these deletions, 56 items were scaled for age 9, 76 items were scaled for age 13, and 78 items were scaled for age 17. A list of the items scaled for each of the ages, along with their item parameter estimates, appears in appendix B.

Table 5–6. Items deleted from the NAEP science long-term trend analysis, age 9: 1999

Booklet IDs	Block	Item	Reason for Exclusion
91	S1	N400201	Excluded in previous assessments
92	S2	N401701	Excluded in previous assessments
92	S2	N402003	Excluded in previous assessments
92	S2	N402004	Excluded in previous assessments
92	S2	N402601	Excluded in previous assessments
92	S2	N402603	Excluded in previous assessments
93	S3	N403802	Excluded in previous assessments

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 5–7. Items deleted from the NAEP science long-term trend analysis, age 13: 1999

Booklet IDs	Block	Item	Reason for Exclusion
91	S1	N404902	Excluded in previous assessments
91	S1	N404903	Excluded in previous assessments
92	S2	N407501	Excluded in previous assessments
93	S3	N409401	Excluded in previous assessments
93	S3	N409402	Excluded in previous assessments
93	S3	N409403	Excluded in previous assessments
93	S3	N409801	Excluded in previous assessments

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table 5–8. Items deleted from the NAEP science long-term trend analysis, age 17: 1999

Booklet IDs	Block	Item	Reason for Exclusion
85	S1	N410001	Excluded in previous assessments
85	S1	N410002	Excluded in previous assessments
85	S1	N410301	Excluded in previous assessments
85	S2	N407402	Excluded in previous assessments

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

5.3.2 Derived Background Variables

In the long-term trend analysis, all variables derived from background questions were used to define subgroups of students for reporting. For this reason, these variables were also used in conditioning. Information about the conditioning variables and the respective codings is given in appendix C. A statistical summary of the NAEP 1999 subgroups is displayed in several tables in appendix A.

5.4 Generation of Plausible Values

The generation of plausible values was conducted independently for each age group. The item parameters from NAEP–BILOG/PARSCALE, final student weights, item responses and selected background variables (conditioning variables) were used with the computer program BGROUP (described in part two, section 2.4.3) in order to generate the plausible values for each student. There were 49 contrasts in the conditioning model (see equation 12.8 in chapter 12 of the *NAEP 1998 Technical Report* [Allen, Carlson, et al., 2001]) at age 9, excluding an overall constant, 52 at age 13, and 58 at age 17. appendix C gives the codings for the conditioning variables for the three age groups. A check on the distributions of the plausible values for each age was made. The generation of plausible values is described in more detail in part two, section 2.4.2. Table 5–9 contains a list of the number of background contrasts included in conditioning, as well as the proportion of variance accounted for by the conditioning model for each age. This proportion is the ratio of the difference between the total variance and the BGROUP residual variance, divided by the total variance. The total variance is the mean of the five theta–scale variances obtained by their respective plausible values.

Table 5–9. Proportion of proficiency variance accounted for by the conditioning model for the NAEP science long-term trend assessment: 1999

Age	Number of conditioning contrasts[1]	Proportion of proficiency variance
9	49	0.29
13	52	0.34
17	58	0.40

[1]Excluding the constant and intercept terms.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

5.5 The Final NAEP Science Long-Term Trend Scale

Since the plausible value (theta) scales have a linear indeterminacy, comparisons with previous assessments will be sensible only if the scale is linearly transformed to a meaningful metric. This indeterminacy was resolved by linking the 1999 scales to previous long-term trend scales. The 1999 data had to be transformed to compensate for linear changes in the scale due to employing newly estimated item parameters and new BGROUP conditioning parameters in 1999. The transformation was accomplished by first reestimating the 1996 student abilities using 1999 item parameters and 1999 BGROUP parameters. (For score metric transformation, see part two, section 2.4.3.) The new 1996 ability estimates were then equated to the old 1996 ability estimates by matching the first two moments (i.e., the mean and standard deviation). The constants for this transformation were then applied to the 1999 data. The transformation equations that resulted are:

$$\text{Age 9:}\quad \theta_{target} = 41.59 \bullet \theta_{calibrated} + 226.73,$$

Age 13: $\theta_{target} = 39.74 \bullet \theta_{calibrated} + 255.09$, and

Age 17: $\theta_{target} = 46.78 \bullet \theta_{calibrated} + 294.84$,

where θ_{target} denotes values on the final reporting scale of the 1999 data and $\theta_{calibrated}$ denotes values on the original 1999 calibration (theta) scale. Overall summary statistics for the long-term trend scales are given in table 5–10. The detailed science long-term trend results from the analyses described in this section are reported in Campbell et al. (2000).

Table 5–10. Means and standard deviations on the NAEP science long-term trend scale: 1977–1999

		All five plausible values	
Age	**Assessment**	**Mean**	**Standard deviation**
9	1977	219.9*	44.9
	1982	220.8*	40.9
	1986	224.3*	41.6
	1990	228.7	40.2
	1992	230.6	39.9
	1994	231.0	40.9
	1996	229.7	42.2
	1999	229.4	39.8
13	1977	247.4*	43.5
	1982	250.1*	38.6
	1986	251.4*	36.6
	1990	255.2	37.6
	1992	258.0*	36.9
	1994	256.8	37.2
	1996	256.0	38.4
	1999	255.8	36.7
17	1977	289.5*	45.0
	1982	283.3*	46.7
	1986	288.5*	44.4
	1990	290.4*	46.2
	1992	294.1	44.7
	1994	294.0	45.6
	1996	295.7	45.1
	1999	295.3	43.8

*Significantly different from 1999, as reported in Campbell, et al. (2000). Note that appropriate standard errors for these statistical tests are provided in table B.1 of that report.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

To provide a context for interpreting the overall science long-term trend results, the NAEP science results were "anchored" at five NAEP science scale levels. In 1986, five science scale level were selected as anchor points, using the process described in *Expanding the New Design: The 1985–86 Technical Report* (Beaton, 1988). The five levels of science proficiency are:

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

These same anchor points were used in 1977, 1982, 1986, 1990, 1992, 1994, 1996, and 1999.

5.6 Extrapolation of the 1971–72 and 1973–74 Mean P–Value Results onto the NAEP Science Long-Term Trend Scale

Because of insufficient common items between the 1971–72, 1973–74, and 1986 science assessments data from 1971–72 and 1973–74 were never included in the IRT trend analysis. However, for the nation and several reporting subgroups (e.g., gender) at each of the three age levels, an estimate of the 1971–72 and 1973–74 mean level of student science proficiency was computed when the data from the 1985–86 assessment were analyzed.

The method used to derive 1971–72 and 1973–74 science proficiency scores is based on the strong linear relationship between the logit of a subgroup's weighted mean proportion correct and its respective proficiency mean across the assessments of 1976–77, 1981–82, and 1986, given an age level. Assuming this linear relationship would hold for both 1971–72 and 1973–74 data, extrapolation of proficiency scores of subgroups can be obtained from weighted mean correct of corresponding subgroups of those years. For each age, separate linear coefficients between proficiency scores and difference in logits of weighted mean proportion correct were obtained. Common items for each pair of the three assessment years 1976–77, 1981–82, and 1986, as well as common items for all three years, were used to calculate weighted mean proportion correct. These coefficients per age were kept constant to estimate proficiency scores of 1971–72 and 1973–74 from differences in the logits of the weighted mean percent correct of the corresponding year.

All subgroups in a given age were included in the regression. The estimate of the 1973–74 proficiency mean for a subgroup was then obtained as the sum of the 1976–77 mean proficiency of the subgroup and the coefficient times the difference between the logit of the 1973–74 and 1976–77 subgroup mean proportion correct. Insufficient common items between 1971–72 and 1976–77 made it difficult to extrapolate 1971–72 proficiency scores from 1976–77 scores. For that reason, the estimates of 1971–72 proficiency mean were calculated in a fashion similar to that done for 1973–74, except that 1976–77 proficiency scores were replaced by 1973–74 extrapolated proficiency scores.

THIS PAGE INTENTIONALLY LEFT BLANK.

Appendix A

Statistical Summary of the 1999 NAEP Samples[1]

In this appendix, the characteristics of the final reporting NAEP samples are displayed in tables A–1 through A–24. Although the subgroups *Type of Location* and *Region of the Country* were not reported in *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance* (Campbell, et al., 2000), these statistics are provided for informational purposes in this appendix.

Tables A–1, A–2, and A–3 display the distribution of students assessed in the long-term trend reading and writing assessment for several basic categories: gender, racial/ethnic grouping, region of the country (Northeast, Southeast, Central, or West), parental education, type of location (central city, urban fringe/large town, rural/small town), and school type (public, nonpublic, Bureau of Indian Affairs [BIA], or Department of Defense Education Activity [DoDEA]).

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

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

Tables A–4, A–5, and A–6 provide similar information for the long-term trend science and mathematics assessment. Note that since these are age–only samples, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– and grade–eligible. Tables A–7 through A–12 enumerate the excluded students across the various long-term trend samples.

Tables A–13 through A–18 show the sizes of the estimated populations of assessable students and the weighted percentages for the NAEP categories of gender, race/ethnicity, region of the country, parents' education level, type of location, and school type. Tables A–19 through A–24 show the estimated total population of excluded students and the weighted percentages by demographic subgroups. Data about parents' education level is not collected for excluded students.

[1]Bruce A. Kaplan and Yuxin Tang provided the statistical summary data tables.

THIS PAGE INTENTIONALLY LEFT BLANK.

Table A–1. Number of students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 9/grade 4: 1999

	Age	Grade	Age and grade	Age or grade
Total	4,109	4,578	2,894	5,793
Gender				
Male	2,013	2,274	1,352	2,935
Female	2,096	2,304	1,542	2,858
Race/ethnicity				
White	2,271	2,625	1,608	3,288
Black	706	776	470	1,012
Hispanic	870	877	619	1,128
Asian American	141	168	120	189
American Indian	102	113	63	152
Unclassified	19	19	14	24
Region				
Northeast	796	927	625	1,098
Southeast	971	1,041	588	1,424
Central	1,012	1,135	674	1,473
West	1,330	1,475	1,007	1,798
Parents' education				
Less than high school	162	176	113	225
High school	672	734	427	979
Greater Than High School	181	193	129	245
Graduated College	1,614	1,891	1,195	2,310
Unknown	1,480	1,584	1,030	2,034
Type of location				
Central city	1,434	1,532	1,036	1,930
Urban fringe/large town	1,608	1,803	1,162	2,249
Rural/Small Town	1,067	1,243	696	1,614
School type				
Public	3,709	4,091	2,579	5,221
Nonpublic	400	487	315	572
Private	153	219	119	253
Catholic	247	268	196	319
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–2. Number of students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 13/grade 8: 1999

	Age	Grade	Age and grade	Age or grade
Total	4,100	4,531	2,698	5,933
Gender				
Male	2,014	2,227	1,229	3,012
Female	2,086	2,304	1,469	2,921
Race/ethnicity				
White	2,547	2,832	1,655	3,724
Black	704	757	451	1,010
Hispanic	633	682	441	874
Asian American	136	163	104	195
American Indian	70	87	41	116
Unclassified	10	10	6	14
Region				
Northeast	791	871	608	1,054
Southeast	1,037	1,168	642	1,563
Central	918	979	503	1,394
West	1,354	1,513	945	1,922
Parents' education				
Less than high school	264	293	158	399
High school	1,031	1,210	663	1,578
Greater Than High School	422	512	325	609
Graduated College	1,927	2,058	1,277	2,708
Unknown	456	458	275	639
Type of location				
Central city	1,410	1,618	959	2,069
Urban fringe/large town	1,667	1,861	1,152	2,376
Rural/Small Town	1,023	1,052	587	1,488
School type				
Public	3,620	4,016	2,372	5,264
Nonpublic	480	515	326	669
Private	209	242	144	307
Catholic	271	273	182	362
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–3. Number of students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 17/grade 11: 1999

	Age	Grade	Age and grade	Age or grade
Total	4,111	4,400	3,223	5,288
Gender				
Male	2,038	2,203	1,508	2,733
Female	2,073	2,197	1,715	2,555
Race/ethnicity				
White	2,734	2,872	2,173	3,433
Black	671	689	472	888
Hispanic	448	529	367	610
Asian American	205	249	174	280
American Indian	45	51	30	66
Unclassified	8	10	7	11
Region				
Northeast	692	803	541	954
Southeast	1,069	1,076	793	1,352
Central	1,150	1,133	864	1,419
West	1,200	1,388	1,025	1,563
Parents' education				
Less than high school	266	282	183	365
High school	964	990	688	1,266
Greater Than High School	720	798	594	924
Graduated College	1,996	2,139	1,637	2,498
Unknown	165	191	121	235
Type of location				
Central city	1,248	1,395	1,012	1,631
Urban fringe/large town	1,590	1,692	1,247	2,035
Rural/Small Town	1,273	1,313	964	1,622
School type				
Public	3,723	3,971	2,895	4,799
Nonpublic	388	429	328	489
Private	113	139	101	151
Catholic	275	290	227	338
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–4. Number of students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 9: 1999

	Age	Grade	Age and grade	Age or grade
Total	6,032	4,110	4,110	6,032
Gender				
Male	2,948	1,964	1,964	2,948
Female	3,084	2,146	2,146	3,084
Race/ethnicity				
White	3,348	2,274	2,274	3,348
Black	1,123	780	780	1,123
Hispanic	1,228	806	806	1,228
Asian American	175	146	146	175
American Indian	152	100	100	152
Unclassified	6	4	4	6
Region				
Northeast	1,306	1,009	1,009	1,306
Southeast	1,475	871	871	1,475
Central	1,399	866	866	1,399
West	1,852	1,364	1,364	1,852
Parents' education				
Less than high school	262	169	169	262
High school	754	476	476	754
Greater than high school	408	300	300	408
Graduated college	2,650	1,862	1,862	2,650
Unknown	1,958	1,303	1,303	1,958
Type of location				
Central city	2,051	1,412	1,412	2,051
Urban fringe/large town	2,383	1,712	1,712	2,383
Rural/small town	1,598	986	986	1,598
School type				
Public	5,378	3,637	3,637	5,378
Nonpublic	654	473	473	654
Private	194	140	140	194
Catholic	460	333	333	460
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–5. Number of students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 13: 1999

	Age	Grade	Age and grade	Age or grade
Total	5,941	3,797	3,797	5,941
Gender				
Male	2,940	1,788	1,788	2,940
Female	3,001	2,009	2,009	3,001
Race/ethnicity				
White	3,699	2,305	2,305	3,699
Black	1,064	682	682	1,064
Hispanic	859	573	573	859
Asian American	218	180	180	218
American Indian	94	52	52	94
Unclassified	7	5	5	7
Region				
Northeast	1,090	809	809	1,090
Southeast	1,473	859	859	1,473
Central	1,368	759	759	1,368
West	2,010	1,370	1,370	2,010
Parents' education				
Less than high school	389	217	217	389
High school	1,257	762	762	1,257
Greater than high school	982	689	689	982
Graduated college	2,730	1,817	1,817	2,730
Unknown	583	312	312	583
Type of location				
Central city	2,036	1,343	1,343	2,036
Urban fringe/large town	2,495	1,702	1,702	2,495
Rural/small town	1,410	752	752	1,410
School type				
Public	5,328	3,407	3,407	5,328
Nonpublic	613	390	390	613
Private	268	170	170	268
Catholic	345	220	220	345
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–6. Number of students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 17: 1999

	Age	Grade	Age and grade	Age or grade
Total	3,795	2,978	2,978	3,795
Gender				
Male	1,805	1,344	1,344	1,805
Female	1,990	1,634	1,634	1,990
Race/ethnicity				
White	2,475	1,970	1,970	2,475
Black	687	506	506	687
Hispanic	404	306	306	404
Asian American	193	169	169	193
American Indian	33	24	24	33
Unclassified	3	3	3	3
Region				
Northeast	667	543	543	667
Southeast	1,053	786	786	1,053
Central	938	690	690	938
West	1,137	959	959	1,137
Parents' education				
Less than high school	248	162	162	248
High school	794	579	579	794
Greater than high school	882	713	713	882
Graduated college	1,749	1,440	1,440	1,749
Unknown	122	84	84	122
Type of location				
Central city	1,147	947	947	1,147
Urban fringe/large town	1,510	1,190	1,190	1,510
Rural/small town	1,138	841	841	1,138
School type				
Public	3,460	2,695	2,695	3,460
Nonpublic	335	283	283	335
Private	93	72	72	93
Catholic	242	211	211	242
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–7. Number of excluded students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 9/grade 4: 1999

	Age	Grade	Age and grade	Age or grade
Total	343	428	205	566
Gender				
Male	220	268	132	356
Female	123	160	73	210
Race/ethnicity				
White	157	215	91	281
Black	59	72	28	103
Hispanic	117	126	78	165
Asian American	10	12	8	14
American Indian	0	0	0	0
Unclassified	0	3	0	3
Region				
Northeast	72	105	55	122
Southeast	62	81	24	119
Central	57	58	25	90
West	152	184	101	235
Type of location				
Central city	140	152	87	205
Urban fringe/large town	131	166	85	212
Rural/small town	72	110	33	149
School type				
Public	341	427	204	564
Nonpublic	1	0	0	1
Private	1	0	0	1
Catholic	1	1	1	1
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–8. Number of excluded students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 13/grade 8: 1999

	Age	Grade	Age and grade	Age or grade
Total	252	336	121	467
Gender				
Male	174	227	82	319
Female	78	109	39	148
Race/ethnicity				
White	126	180	57	249
Black	52	79	22	109
Hispanic	64	65	34	95
Asian American	8	9	6	11
American Indian	1	2	1	2
Unclassified	1	1	1	1
Region				
Northeast	36	67	28	75
Southeast	78	111	26	163
Central	38	56	15	79
West	100	102	52	150
Type of location				
Central city	94	122	52	164
Urban fringe/large town	97	132	47	182
Rural/small town	61	82	22	121
School type				
Public	251	335	121	465
Nonpublic	0	0	0	0
Private	0	0	0	0
Catholic	1	1	0	2
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–9. Number of excluded students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 17/ grade 11: 1999

	Age	Grade	Age and grade	Age or grade
Total	183	223	77	329
Gender				
Male	114	138	41	211
Female	69	85	36	118
Race/ethnicity				
White	114	136	49	201
Black	38	49	13	74
Hispanic	21	23	10	34
Asian American	10	14	5	19
American Indian	0	0	0	0
Unclassified	0	1	0	1
Region				
Northeast	40	50	17	73
Southeast	62	76	25	113
Central	46	50	15	81
West	35	47	20	62
Type of location				
Central city	48	58	25	81
Urban fringe/large town	91	109	35	165
Rural/small town	44	56	17	83
School type				
Public	180	220	75	325
Nonpublic	1	0	0	1
Private	1	0	0	1
Catholic	2	3	2	3
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–10. Number of excluded students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 9: 1999

	Age	Grade	Age and grade	Age or grade
Total	554	308	308	554
Gender				
Male	357	208	208	357
Female	197	100	100	197
Race/ethnicity				
White	247	127	127	247
Black	103	48	48	103
Hispanic	183	116	116	183
Asian American	18	14	14	18
American Indian	0	0	0	0
Unclassified	3	3	3	3
Region				
Northeast	114	69	69	114
Southeast	110	44	44	110
Central	92	42	42	92
West	238	153	153	238
Type of location				
Central city	252	148	148	252
Urban fringe/large town	207	116	116	207
Rural/small town	95	44	44	95
School type				
Public	550	305	305	550
Nonpublic	0	0	0	0
Private	0	0	0	0
Catholic	4	3	3	4
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–11. Number of excluded students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 13: 1999

	Age	Grade	Age and grade	Age or grade
Total	357	158	158	357
Gender				
Male	235	103	103	235
Female	122	55	55	122
Race/ethnicity				
White	209	85	85	209
Black	82	35	35	82
Hispanic	52	31	31	52
Asian American	10	6	6	10
American Indian	0	0	0	0
Unclassified	4	1	1	4
Region				
Northeast	53	40	40	53
Southeast	118	40	40	118
Central	74	28	28	74
West	112	50	50	112
Type of location				
Central city	143	69	69	143
Urban fringe/large town	138	64	64	138
Rural/small town	76	25	25	76
School type				
Public	355	158	158	355
Nonpublic	0	0	0	0
Private	0	0	0	0
Catholic	2	0	0	2
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–12. Number of excluded students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 17: 1999

	Age	Grade	Age and grade	Age or grade
Total	231	92	92	231
Gender				
Male	146	58	58	146
Female	85	34	34	85
Race/ethnicity				
White	122	57	57	122
Black	70	22	22	70
Hispanic	28	10	10	28
Asian American	10	3	3	10
American Indian	1	0	0	1
Unclassified	0	0	0	0
Region				
Northeast	39	17	17	39
Southeast	81	33	33	81
Central	67	23	23	67
West	44	19	19	44
Type of location				
Central city	63	25	25	63
Urban fringe/large town	99	48	48	99
Rural/small town	69	19	19	69
School type				
Public	224	86	86	224
Nonpublic	2	2	2	2
Private	2	2	2	2
Catholic	5	4	4	5
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–13. Weighted percentage of students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 9/grade 4: 1999

	Age	Grade	Age and grade	Age or grade
Total	3,165,926	3,654,876	2,196,494	4,624,307
Gender				
Male	48.9	49.8	46.4	50.8
Female	51.1	50.2	53.6	49.2
Race/ethnicity				
White	64.8	66.3	66.1	65.4
Black	15.6	15.1	14	16
Hispanic	14.6	14.2	15.2	14
Asian American	2.5	2.6	2.9	2.4
American Indian	2.1	1.7	1.5	2.0
Unclassified	0.3	0.2	0.3	0.3
Region				
Northeast	22.1	21.8	24.4	20.8
Southeast	20.5	21.5	19.2	21.9
Central	26.7	26.8	24.7	27.7
West	30.8	30.0	31.8	29.6
Parents' education				
Less than high school	4.0	4.0	4.1	3.9
High school	15.9	15.8	14.5	16.5
Greater than high school	4.5	4.3	4.6	4.3
Graduated college	40.8	42.9	43.1	41.4
Unknown	34.7	33.1	33.7	33.9
Type of location				
Central city	33.1	31.9	33.3	32.0
Urban fringe/large town	40.9	40.3	41.8	40.0
Rural/small town	26.1	27.8	25.0	28.0
School type				
Public	88.4	86.9	86.7	88.0
Nonpublic	11.5	13.0	13.1	11.9
Private	5.4	7.2	6.3	6.4
Catholic	6.1	5.8	6.8	5.5
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–14. Weighted percentage of students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 13/grade 8: 1999

	Age	Grade	Age and grade	Age or grade
Total	3,196,594	3,466,603	1,981,332	4,681,865
Gender				
Male	49.2	49.7	45.3	51.3
Female	50.8	50.3	54.7	48.7
Race/ethnicity				
White	65.6	66.9	67.8	65.6
Black	15	14.4	13.4	15.2
Hispanic	14.5	13.9	14.2	14.1
Asian American	3.1	3.1	3.1	3.1
American Indian	1.6	1.6	1.3	1.8
Unclassified	0.2	0.2	0.2	0.2
Region				
Northeast	21.9	21.1	24.8	20.1
Southeast	20.4	21.3	18.5	21.9
Central	26.8	27.1	24.8	27.9
West	30.9	30.5	32	30.1
Parents' education				
Less than high school	5.8	6.1	5.2	6.3
High school	25.3	27.1	24.7	26.9
Greater than high school	9.8	10.8	11.7	9.7
Graduated college	48	45.9	48.2	46.4
Unknown	11.1	10.1	10.2	10.8
Type of location				
Central city	32.2	33.4	32.3	33.1
Urban fringe/large town	43	42.6	44.7	42
Rural/small town	24.8	23.9	23.1	24.9
School type				
Public	87.3	87.7	86.8	87.8
Nonpublic	12.6	12.2	13.1	12.1
Private	5.5	5.4	5.7	5.4
Catholic	7.1	6.8	7.4	6.7
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–15. Weighted percentage of students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 17/grade 11: 1999

	Age	Grade	Age and grade	Age or grade
Total	3,458,040	3,402,827	2,181,931	4,678,935
Gender				
Male	52.2	51.8	47.4	54.2
Female	47.8	48.2	52.6	45.8
Race/ethnicity				
White	68.5	67	72.7	65.5
Black	14.0	14.6	12.0	15.4
Hispanic	12.6	13.1	10.2	14.1
Asian American	3.7	4.1	4.2	3.7
American Indian	1.0	0.9	0.8	1.1
Unclassified	0.2	0.2	0.2	0.2
Region				
Northeast	20.6	20.5	20.0	20.9
Southeast	23.6	21.3	20.4	23.4
Central	26.8	26.4	27.5	26.2
West	29.0	31.8	32.1	29.6
Parents' education				
Less than high school	7.2	6.8	5.1	7.9
High school	24.3	22.3	21.0	24.5
Greater than high school	17.0	18.2	18.8	17.0
Graduated college	47.1	48.3	51.5	46.0
Unknown	4.3	4.4	3.7	4.6
Type of location				
Central city	27.9	29.3	28.4	28.7
Urban fringe/large town	42.9	42.6	43.2	42.5
Rural/small town	29.2	28.1	28.4	28.8
School type				
Public	90.1	90.0	89.0	90.6
Nonpublic	9.8	9.9	10.9	9.3
Private	2.7	3.2	3.1	2.8
Catholic	7.1	6.7	7.8	6.5
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–16. Weighted percentage of students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 9: 1999

	Age	Grade	Age and grade	Age or grade
Total	3,444,287	2,222,404	2,222,404	3,444,287
Gender				
Male	49.0	47.9	47.9	49.0
Female	51.0	52.1	52.1	51.0
Race/ethnicity				
White	66.5	66.0	66.0	66.5
Black	15.0	15.5	15.5	15.0
Hispanic	13.5	13.0	13.0	13.5
Asian American	2.7	3.3	3.3	2.7
American Indian	2.2	2.2	2.2	2.2
Unclassified	0.1	0.1	0.1	0.1
Region				
Northeast	21.9	24.5	24.5	21.9
Southeast	21.7	19.9	19.9	21.7
Central	27.3	25.2	25.2	27.3
West	29.1	30.4	30.4	29.1
Parents' education				
Less than high school	3.9	3.7	3.7	3.9
High school	12.1	11.2	11.2	12.1
Greater than high school	6.9	7.5	7.5	6.9
Graduated college	44.8	46.6	46.6	44.8
Unknown	32.3	31.0	31.0	32.3
Type of location				
Central city	32.6	33.0	33.0	32.6
Urban fringe/large town	41.4	43.0	43.0	41.4
Rural/small town	26.0	24.0	24.0	26.0
School type				
Public	88.0	87.6	87.6	88.0
Nonpublic	11.9	12.2	12.2	11.9
Private	4.0	3.9	3.9	4.0
Catholic	7.9	8.3	8.3	7.9
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age–or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–17. Weighted percentage of students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 13: 1999

	Age	Grade	Age and grade	Age or grade
Total	3,396,555	2,083,464	2,083,464	3,396,555
Gender				
Male	50.1	47.4	47.4	50.1
Female	49.9	52.6	52.6	49.9
Race/ethnicity				
White	68.0	66.8	66.8	68.0
Black	14.1	14.3	14.3	14.1
Hispanic	12.9	12.9	12.9	12.9
Asian American	3.5	4.5	4.5	3.5
American Indian	1.5	1.3	1.3	1.5
Unclassified	0.1	0.2	0.2	0.1
Region				
Northeast	20.9	23.5	23.5	20.9
Southeast	21.0	20.1	20.1	21.0
Central	27.7	24.3	24.3	27.7
West	30.3	32.1	32.1	30.3
Parents' education				
Less than high school	6.0	5.2	5.2	6.0
High school	20.7	20.0	20.0	20.7
Greater than high school	16.7	18.1	18.1	16.7
Graduated college	47.1	48.7	48.7	47.1
Unknown	9.5	8.0	8.0	9.5
Type of location				
Central city	31.8	32.7	32.7	31.8
Urban fringe/large town	43.8	46.0	46.0	43.8
Rural/small town	24.4	21.3	21.3	24.4
School type				
Public	88.2	88.4	88.4	88.2
Nonpublic	11.7	11.5	11.5	11.7
Private	5.2	5.0	5.0	5.2
Catholic	6.5	6.5	6.5	6.5
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–18. Weighted percentage of students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 17: 1999

	Age	Grade	Age and grade	Age or grade
Total	3,398,386	2,518,213	2,518,213	3,398,386
Gender				
Male	48.3	45.6	45.6	48.3
Female	51.7	54.4	54.4	51.7
Race/ethnicity				
White	68.5	70.5	70.5	68.5
Black	13.8	12.2	12.2	13.8
Hispanic	12.9	11.9	11.9	12.9
Asian American	4.2	4.8	4.8	4.2
American Indian	0.6	0.5	0.5	0.6
Unclassified	0.1	0.1	0.1	0.1
Region				
Northeast	21.2	22.0	22.0	21.2
Southeast	22.8	21.5	21.5	22.8
Central	26.6	25.0	25.0	26.6
West	29.3	31.6	31.6	29.3
Parents' education				
Less than high school	6.6	5.1	5.1	6.6
High school	20.0	18.3	18.3	20.0
Greater than high school	22.7	23.4	23.4	22.7
Graduated college	47.1	50.0	50.0	47.1
Unknown	3.6	3.2	3.2	3.6
Type of location				
Central city	27.9	28.9	28.9	27.9
Urban fringe/large town	45.5	45.8	45.8	45.5
Rural/small town	26.6	25.3	25.3	26.6
School type				
Public	89.5	89.0	89.0	89.5
Nonpublic	10.5	10.9	10.9	10.5
Private	4.2	3.9	3.9	4.2
Catholic	6.3	7.0	7.0	6.3
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–19. Weighted percentage of excluded students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 9/grade 4: 1999

	Age	Grade	Age and grade	Age or grade
Total	103,181	269,796	62,644	310,333
Gender				
Male	65.9	63.2	65.4	63.7
Female	34.1	36.8	34.6	36.3
Race/ethnicity				
White	53.4	56.7	53.0	56.3
Black	17.9	18.7	13.5	19.5
Hispanic	26.0	21.9	29.9	21.6
Asian American	2.8	1.9	3.6	1.9
American Indian	0	0	0	0
Unclassified	0	0.8	0	0.7
Region				
Northeast	20.0	20.2	25.2	19.2
Southeast	15.4	20.9	11.3	21.0
Central	19.3	17.1	13.8	18.5
West	45.3	41.7	49.7	41.3
Type of location				
Central city	37.6	30.6	39.7	31.1
Urban fringe/large town	40.8	41.2	43.7	40.5
Rural/small town	21.5	28.2	16.6	28.4
School type				
Public	98.6	99.9	99.4	99.5
Nonpublic	1.0	0	0	0.3
Private	1.0	0	0	0.3
Catholic	0.4	0.1	0.6	0.1
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–20. Weighted percentage of excluded students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 13/grade 8: 1999

	Age	Grade	Age and grade	Age or grade
Total	91,527	235,164	39,343	287,348
Gender				
Male	66.1	66.9	67.1	66.6
Female	33.9	33.1	32.9	33.4
Race/ethnicity				
White	46.4	56.7	50.1	54.4
Black	16.5	20.4	14.1	20.1
Hispanic	33.2	20.3	29.8	23.1
Asian American	3.3	2.1	4.8	2.1
American Indian	0.4	0.4	0.9	0.3
Unclassified	0.2	0.1	0.4	0.1
Region				
Northeast	13.9	19.5	22.3	17.3
Southeast	22.6	29.7	16.9	29.2
Central	14.1	20.1	13.1	19.1
West	49.5	30.7	47.7	34.3
Type of location				
Central city	40.9	32.8	40.9	34.3
Urban fringe/large town	39.3	43.9	40.9	42.8
Rural/small town	19.8	23.3	18.3	22.9
School type				
Public	99.7	99.6	100.0	99.5
Nonpublic	0	0	0	0
Private	0	0	0	0
Catholic	0.3	0.4	0	0.5
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–21. Weighted percentage of excluded students in the NAEP reading and writing long-term trend sample by type of eligibility and subgroup classification, age 17/grade 11: 1999

	Age	Grade	Age and grade	Age or grade
Total	97,538	163,303	26,972	233,869
Gender				
Male	64.6	62.4	51.9	64.5
Female	35.4	37.6	48.1	35.5
Race/ethnicity				
White	65.5	60.9	71.1	61.7
Black	16.6	23.6	13.6	21.9
Hispanic	13.1	9.7	9.3	11.1
Asian American	4.8	5.1	6.0	4.8
American Indian	0	0	0	0
Unclassified	0	0.7	0	0.5
Region				
Northeast	25.9	23.5	23.4	24.5
Southeast	27.8	30.1	24.6	29.8
Central	20.8	21.7	16.7	21.9
West	25.5	24.7	35.3	23.9
Type of location				
Central city	25.4	24.0	35.5	23.2
Urban fringe/large town	53.3	52.1	45.3	53.4
Rural/small town	21.2	24.0	19.1	23.4
School type				
Public	98.5	98.6	96.8	98.8
Nonpublic	0.6	0	0	0.2
Private	0.6	0	0	0.2
Catholic	0.9	1.4	3.2	1.0
BIA	0	0	0	0
DoDEA	0	0	0	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–22. Weighted percentage of excluded students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 9: 1999

	Age	Grade	Age and grade	Age or grade
Total	169,991	94,820	94,820	169,991
Gender				
Male	66.0	68.2	68.2	66.0
Female	34.0	31.8	31.8	34.0
Race/ethnicity				
White	53.1	50.9	50.9	53.1
Black	18.8	13.3	13.3	18.8
Hispanic	24.2	30.3	30.3	24.2
Asian American	3.5	4.7	4.7	3.5
American Indian	0	0	0	0
Unclassified	0.4	0.7	0.7	0.4
Region				
Northeast	21.1	21.7	21.7	21.1
Southeast	17.1	13.5	13.5	17.1
Central	19.2	15.1	15.1	19.2
West	42.6	49.7	49.7	42.6
Type of location				
Central city	44.2	46.3	46.3	44.2
Urban fringe/large town	40.3	42.1	42.1	40.3
Rural/small town	15.6	11.6	11.6	15.6
School type				
Public	99.1	98.8	98.8	99.1
Nonpublic	0	0	0	0
Private	0	0	0	0
Catholic	0.9	1.2	1.2	0.9
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–23. Weighted percentage of excluded students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 13: 1999

	Age	Grade	Age and grade	Age or grade
Total	128,785	53,328	53,328	128,785
Gender				
Male	66.1	66.0	66.0	66.1
Female	33.9	34.0	34.0	33.9
Race/ethnicity				
White	62.0	61.9	61.9	62.0
Black	18.7	15.8	15.8	18.7
Hispanic	15.0	18.6	18.6	15.0
Asian American	3.1	3.1	3.1	3.1
American Indian	0	0	0	0
Unclassified	1.2	0.5	0.5	1.2
Region				
Northeast	16.2	28.6	28.6	16.2
Southeast	25.0	17.5	17.5	25.0
Central	22.5	17.7	17.7	22.5
West	36.3	36.2	36.2	36.3
Type of location				
Central city	38.6	39.1	39.1	38.6
Urban fringe/large town	44.8	47.6	47.6	44.8
Rural/small town	16.6	13.3	13.3	16.6
School type				
Public	99.5	100.0	100.0	99.5
Nonpublic	0	0	0	0
Private	0	0	0	0
Catholic	0.5	0	0	0.5
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table A–24. Weighted percentage of excluded students in the NAEP mathematics and science long-term trend sample by type of eligibility and subgroup classification, age 17: 1999

	Age	Grade	Age and grade	Age or grade
Total	124,125	31,277	31,277	124,125
Gender				
Male	63.8	59.9	59.9	63.8
Female	36.2	40.1	40.1	36.2
Race/ethnicity				
White	53.8	69.6	69.6	53.8
Black	23.6	16.8	16.8	23.6
Hispanic	16.9	11.0	11.0	16.9
Asian American	5.0	2.6	2.6	5.0
American Indian	0.7	0	0	0.7
Unclassified	0	0	0	0
Region				
Northeast	18.6	18.7	18.7	18.6
Southeast	27.6	26.6	26.6	27.6
Central	23.2	24.0	24.0	23.2
West	30.6	30.7	30.7	30.6
Type of location				
Central city	29.0	27.6	27.6	29.0
Urban fringe/large town	40.8	55.3	55.3	40.8
Rural/small town	30.2	17.1	17.1	30.2
School type				
Public	97.5	92.5	92.5	97.5
Nonpublic	0.5	1.9	1.9	0.5
Private	0.5	1.9	1.9	0.5
Catholic	2.0	5.5	5.5	2.0
BIA	0	0	0	0
DoDEA	0	0	0	0

NOTE: Since this is an age–only sample, the number of students who are age–eligible only will be the same as the number of students who are age– or grade–eligible. Likewise, the number of students who are grade–eligible only will be the same as the number of students who are both age– *and* grade–eligible.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Appendix B

IRT Parameters

This appendix contains tables of IRT (item response theory) parameters for the 1999 NAEP long-term trend items that were used in the creation of IRT scales.

For each of the items used in scaling, the tables provide estimates of the IRT parameters and the associated standard errors (s.e.) of the estimates. For each of the binary scored items used in scaling (i.e., multiple–choice items and short constructed–response items), the tables provide estimates of the IRT parameters (which correspond to aj, bj, and cj in equation 12.1 in chapter 12 of the *NAEP 1998 Technical Report* (Allen, Carlson et al., 2001).

The tables also show the block in which each item appears for each age class (Block) and the position of each item within its block (Item).

Note that item parameters shown in this section are in the metrics used for the original calibration of the scales.

Table B–1. IRT parameters for the NAEP reading long-term trend items, age 9/grade 4: 1999

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N001101	H	5	0.717	(0.176)	1.304	(0.187)	0.332	(0.039)
N001521[1]	H	17	1.954	(0.253)	–0.629	(0.090)	0.341	(0.041)
N001522[1]	H	18	2.146	(0.242)	0.333	(0.044)	0.220	(0.027)
N001523[1]	H	19	1.296	(0.162)	–0.274	(0.100)	0.310	(0.041)
N001524[1]	H	20	2.231	(0.262)	0.341	(0.045)	0.252	(0.028)
N001527[1]	H	15	1.128	(0.151)	1.949	(0.168)	0.000	0.000
N001601	J	12	0.954	(0.097)	0.231	(0.075)	0.262	(0.029)
N001602	J	13	1.531	(0.144)	0.404	(0.046)	0.286	(0.023)
N001603	J	14	1.163	(0.165)	0.961	(0.067)	0.310	(0.023)
N001604	J	15	1.169	(0.132)	0.823	(0.054)	0.218	(0.021)
N001802	J	20	1.381	(0.301)	2.026	(0.199)	0.225	(0.013)
N002001	K	14	1.842	(0.172)	0.838	(0.035)	0.194	(0.016)
N002002	K	15	1.483	(0.132)	0.557	(0.039)	0.192	(0.020)
N002003	K	16	1.670	(0.159)	0.503	(0.041)	0.268	(0.021)
N002101	K	18	1.146	(0.278)	1.917	(0.202)	0.231	(0.017)
N002102	K	19	1.266	(0.280)	2.011	(0.201)	0.163	(0.014)
N002401	L	22	1.656	(0.145)	0.663	(0.034)	0.149	(0.017)
N002702	L	20	1.493	(0.151)	0.754	(0.041)	0.189	(0.019)
N002801	L	17	2.647	(0.204)	0.194	(0.028)	0.199	(0.020)
N002802	L	18	1.818	(0.143)	–0.024	(0.043)	0.218	(0.024)
N002804	L	26	0.548	(0.059)	1.708	(0.149)	0.000	0.000
N003001	M	10	0.747	(0.164)	2.054	(0.240)	0.172	(0.021)
N003002	M	11	0.492	(0.072)	0.516	(0.157)	0.206	(0.042)
N003101	M	12	1.222	(0.113)	0.080	(0.062)	0.249	(0.027)
N003102	M	13	2.739	(0.206)	0.694	(0.027)	0.220	(0.015)
N003104	M	16	0.820	(0.112)	2.399	(0.247)	0.000	0.000
N003701	N	23	1.324	(0.131)	0.024	(0.069)	0.339	(0.030)
N003702	N	24	1.880	(0.170)	0.353	(0.040)	0.259	(0.022)
N003704	N	25	0.829	(0.067)	1.005	(0.063)	0.000	0.000
N003801	O	12	1.171	(0.270)	1.698	(0.160)	0.323	(0.019)
N003802	O	13	0.529	(0.075)	0.296	(0.160)	0.226	(0.044)
N003803	O	14	0.992	(0.268)	2.260	(0.304)	0.219	(0.016)
N004101	O	19	1.247	(0.115)	–0.150	(0.072)	0.313	(0.031)
N004201	O	18	1.139	(0.164)	1.022	(0.070)	0.265	(0.023)
N004202	O	19	1.011	(0.176)	1.203	(0.096)	0.307	(0.025)
N004701	Q	10	2.087	(0.161)	0.325	(0.031)	0.201	(0.019)
N004702	Q	11	0.988	(0.106)	0.061	(0.089)	0.336	(0.033)
N004703	Q	12	2.119	(0.164)	0.215	(0.033)	0.241	(0.021)
N004801	Q	13	1.190	(0.111)	–0.373	(0.087)	0.341	(0.034)
N004901	Q	14	1.833	(0.171)	0.939	(0.039)	0.226	(0.015)
N005101	Q	15	0.733	(0.063)	–1.847	(0.196)	0.276	(0.059)
N008601	H	15	1.608	(0.130)	–0.273	(0.056)	0.282	(0.028)
N008602	H	16	1.287	(0.105)	0.014	(0.055)	0.229	(0.026)
N008603	H	17	1.145	(0.099)	–0.319	(0.077)	0.269	(0.032)
N008701	H	9	0.615	(0.057)	–2.978	(0.265)	0.274	(0.064)
N008801	J	18	1.493	(0.132)	–0.802	(0.083)	0.314	(0.034)
N008901	L	15	1.692	(0.130)	–0.205	(0.047)	0.221	(0.025)

See notes at end of table.→

Table B–1. IRT parameters for the NAEP reading long-term trend items, age 9/grade 4: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N008902	J	16	1.018	(0.091)	–0.402	(0.089)	0.248	(0.033)
N009001	K	12	1.590	(0.136)	0.430	(0.039)	0.197	(0.021)
N009002	K	13	1.284	(0.131)	0.568	(0.047)	0.202	(0.022)
N009003	K	14	1.510	(0.125)	1.190	(0.062)	0.252	(0.017)
N009004	K	15	1.898	(0.201)	0.393	(0.039)	0.285	(0.022)
N009101	K	16	0.895	(0.195)	–0.641	(0.130)	0.307	(0.042)
N009201	K	17	1.445	(0.100)	–0.751	(0.085)	0.298	(0.034)
N009401	L	13	1.697	(0.068)	–0.366	(0.049)	0.204	(0.026)
N009601	L	21	0.669	(0.073)	–1.593	(0.190)	0.226	(0.056)
N009701	M	5	1.238	(0.149)	0.154	(0.056)	0.247	(0.026)
N009702	M	6	1.790	(0.131)	0.197	(0.041)	0.268	(0.023)
N009703	M	7	1.704	(0.125)	0.588	(0.039)	0.264	(0.020)
N009704	M	8	1.629	(0.201)	0.622	(0.037)	0.208	(0.019)
N009705	M	9	1.752	(0.195)	0.120	(0.041)	0.243	(0.023)
N009801	N	11	1.140	(0.100)	–1.640	(0.144)	0.312	(0.053)
N009901	N	13	1.170	(0.068)	0.045	(0.069)	0.302	(0.029)
N010002	N	18	1.528	(0.073)	0.125	(0.049)	0.276	(0.025)
N010003	N	19	1.525	(0.149)	0.186	(0.045)	0.226	(0.023)
N010102	N	21	2.075	(0.131)	0.727	(0.036)	0.317	(0.018)
N010103	N	22	2.380	(0.125)	0.077	(0.035)	0.273	(0.023)
N010201	O	20	1.112	(0.201)	–1.443	(0.140)	0.302	(0.050)
N010301	O	10	0.708	(0.195)	–1.182	(0.182)	0.273	(0.053)
N010401	O	12	0.648	(0.100)	–0.879	(0.201)	0.294	(0.054)
N010402	O	13	1.241	(0.068)	0.827	(0.055)	0.216	(0.022)
N010403	O	14	1.491	(0.208)	1.383	(0.082)	0.228	(0.016)
N010801	Q	16	1.269	(0.124)	0.333	(0.057)	0.275	(0.026)
N010902	Q	18	2.312	(0.217)	0.459	(0.034)	0.294	(0.020)
N010903	Q	19	2.832	(0.233)	0.261	(0.028)	0.236	(0.020)
N010904	Q	20	2.010	(0.200)	0.585	(0.036)	0.262	(0.020)
N011001	R	5	1.644	(0.104)	0.233	(0.033)	0.300	(0.018)
N011002	R	6	2.488	(0.153)	0.556	(0.020)	0.278	(0.013)
N011003	R	7	2.493	(0.143)	–0.031	(0.025)	0.302	(0.017)
N011004	R	8	2.405	(0.134)	0.290	(0.021)	0.232	(0.014)
N011101	R	9	1.982	(0.110)	0.308	(0.023)	0.207	(0.015)
N011201	R	10	1.209	(0.088)	0.481	(0.039)	0.247	(0.018)
N011301	R	11	1.956	(0.120)	0.271	(0.028)	0.283	(0.016)
N011302	R	12	1.162	(0.103)	0.582	(0.048)	0.327	(0.020)
N011401	R	13	1.798	(0.165)	1.249	(0.051)	0.404	(0.011)
N011402	R	14	1.128	(0.127)	0.997	(0.053)	0.304	(0.018)
N011403	R	15	1.498	(0.148)	1.176	(0.048)	0.285	(0.013)
N011404	R	16	1.331	(0.133)	1.095	(0.044)	0.199	(0.014)
N013201	V	29	1.836	(0.134)	0.109	(0.036)	0.207	(0.021)
N013301	V	12	1.333	(0.126)	–0.517	(0.086)	0.405	(0.032)
N013401	V	31	1.457	(0.124)	0.597	(0.037)	0.153	(0.018)
N013402	V	32	2.465	(0.220)	0.194	(0.036)	0.370	(0.022)
N013403	V	33	2.214	(0.186)	0.507	(0.029)	0.224	(0.017)
N014001	M	13	1.185	(0.106)	–0.028	(0.065)	0.245	(0.028)
N014101	Q	21	0.928	(0.095)	–0.211	(0.100)	0.272	(0.036)
N014201	V	21	1.044	(0.091)	–0.234	(0.077)	0.250	(0.030)
N014301	N	14	2.377	(0.184)	0.220	(0.030)	0.242	(0.020)
N014302	N	15	1.432	(0.127)	0.405	(0.044)	0.235	(0.022)

See notes at end of table.→

Table B–1. IRT parameters for the NAEP reading long-term trend items, age 9/grade 4: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N014303	N	16	2.454	(0.189)	–0.011	(0.033)	0.255	(0.022)
N014501	V	35	0.651	(0.038)	–0.541	(0.057)	0.000	0.000
N014502	V	35	0.684	(0.053)	–0.694	(0.085)	0.000	0.000
N014503	V	35	0.898	(0.046)	–1.136	(0.061)	0.000	0.000

[1]N001521–N001527 are the same questions as those numbered N001501–N001507 in previous assessments. In 1999 these questions refer to the passage in which references to the "Devil" were changed to references to the "King."
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table B–2. IRT parameters for the NAEP reading long-term trend items, age 13/grade 8: 1999

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N001101	H	6	0.230	(0.040)	0.821	(0.377)	0.302	(0.046)
N001201	H	7	0.652	(0.121)	1.410	(0.149)	0.346	(0.031)
N001202	H	8	1.398	(0.131)	0.700	(0.047)	0.212	(0.019)
N001301	H	9	0.635	(0.089)	0.097	(0.190)	0.438	(0.045)
N001302	H	10	0.739	(0.087)	–1.821	(0.281)	0.541	(0.067)
N001303	H	11	0.742	(0.085)	0.607	(0.098)	0.235	(0.032)
N001401	H	12	0.813	(0.075)	–0.292	(0.111)	0.254	(0.040)
N001521[1]	H	25	1.878	(0.237)	–1.859	(0.117)	0.271	(0.058)
N001522[1]	H	26	1.269	(0.129)	–0.674	(0.090)	0.206	(0.041)
N001523[1]	H	27	1.038	(0.118)	–1.094	(0.146)	0.273	(0.056)
N001524[1]	H	28	1.261	(0.130)	–0.668	(0.092)	0.213	(0.042)
N001527[1]	H	18	0.590	(0.071)	2.093	(0.206)	0.000	0.000
N001601	J	11	0.399	(0.048)	–1.091	(0.315)	0.294	(0.065)
N001602	J	12	0.772	(0.064)	–1.680	(0.163)	0.257	(0.057)
N001603	J	13	0.753	(0.087)	–0.119	(0.146)	0.360	(0.044)
N001604	J	14	0.869	(0.081)	–0.438	(0.116)	0.295	(0.042)
N001701	J	17	0.666	(0.068)	–0.811	(0.183)	0.305	(0.055)
N001702	J	18	0.745	(0.201)	2.725	(0.360)	0.262	(0.018)
N001703	J	19	0.642	(0.064)	–0.305	(0.146)	0.243	(0.046)
N001802	J	21	0.714	(0.093)	0.773	(0.109)	0.257	(0.034)
N001901	J	22	0.834	(0.087)	0.079	(0.107)	0.279	(0.038)
N002001	K	22	1.130	(0.086)	–0.094	(0.062)	0.192	(0.028)
N002002	K	23	1.136	(0.092)	–0.149	(0.069)	0.245	(0.031)
N002003	K	24	1.140	(0.098)	–0.545	(0.088)	0.306	(0.038)
N002101	K	12	0.814	(0.133)	1.423	(0.109)	0.272	(0.025)
N002102	K	13	1.263	(0.115)	0.796	(0.046)	0.147	(0.018)
N002201	K	14	1.568	0.000	–0.186	(0.039)	0.237	(0.023)
N002202	K	15	1.827	(0.172)	–0.227	(0.059)	0.432	(0.029)
N002203	K	16	0.531	(0.051)	–1.760	(0.244)	0.279	(0.063)
N002401	L	22	0.888	(0.069)	–0.771	(0.103)	0.192	(0.041)
N002501	L	23	0.489	(0.053)	0.130	(0.159)	0.195	(0.042)
N002701	L	24	0.781	(0.089)	0.521	(0.097)	0.257	(0.033)
N002801	L	20	1.192	(0.091)	–1.306	(0.096)	0.217	(0.043)
N002802	L	21	1.218	(0.098)	–1.502	(0.109)	0.249	(0.049)
N002902	M	6	0.548	(0.051)	–1.451	(0.216)	0.263	(0.059)
N002903	M	7	1.238	(0.100)	–0.812	(0.083)	0.255	(0.038)
N002904	M	8	0.936	(0.078)	–0.281	(0.088)	0.231	(0.035)
N002905	M	9	0.522	(0.068)	0.516	(0.161)	0.237	(0.043)
N002906	M	10	1.417	(0.114)	–0.625	(0.068)	0.275	(0.034)
N003001	M	18	0.650	(0.080)	1.191	(0.100)	0.140	(0.026)
N003002	M	12	0.299	(0.039)	–0.137	(0.287)	0.188	(0.052)
N003003	M	19	1.729	(0.211)	2.374	(0.130)	0.092	(0.007)
N003101	M	29	1.150	(0.097)	–0.986	(0.101)	0.271	(0.044)
N003102	M	30	1.403	(0.118)	–0.367	(0.067)	0.308	(0.032)
N003104	M	16	0.554	(0.046)	1.846	(0.131)	0.000	0.000
N003201	N	12	0.863	(0.073)	–0.857	(0.120)	0.245	(0.045)

See notes at end of table.→

Table B–2. IRT parameters for the NAEP reading long-term trend items, age 13/grade 8: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N003202	N	13	1.048	(0.092)	0.172	(0.070)	0.236	(0.029)
N003203	N	14	1.355	(0.132)	0.282	(0.063)	0.349	(0.026)
N003204	N	15	0.908	(0.092)	0.434	(0.080)	0.243	(0.029)
N003301	N	16	0.879	(0.070)	–0.656	(0.101)	0.210	(0.039)
N003401	N	17	1.066	(0.084)	–0.215	(0.070)	0.200	(0.031)
N003501	N	18	0.915	(0.079)	–0.473	(0.100)	0.250	(0.039)
N003601	N	19	0.936	(0.078)	–1.214	(0.127)	0.253	(0.049)
N003602	N	20	0.935	(0.077)	–0.244	(0.083)	0.202	(0.033)
N003701	N	21	0.713	(0.068)	–0.850	(0.161)	0.273	(0.052)
N003702	N	22	1.293	(0.118)	–0.071	(0.071)	0.323	(0.031)
N003704	N	23	0.648	(0.043)	0.155	(0.052)	0.000	0.000
N003801	O	12	0.484	(0.082)	1.111	(0.177)	0.248	(0.044)
N003802	O	13	0.230	(0.033)	–1.724	(0.461)	0.213	(0.060)
N003803	O	14	0.565	(0.144)	2.584	(0.342)	0.267	(0.025)
N003901	O	16	1.283	(0.130)	–2.473	(0.136)	0.261	(0.060)
N004002	O	15	0.475	(0.048)	–2.329	(0.288)	0.277	(0.064)
N004101	O	17	0.856	(0.070)	–1.597	(0.151)	0.266	(0.056)
N004201	O	18	0.766	(0.071)	–0.150	(0.110)	0.232	(0.039)
N004202	O	19	0.584	(0.068)	0.062	(0.159)	0.257	(0.046)
N004301	O	20	1.297	(0.116)	0.289	(0.059)	0.283	(0.025)
N004303	O	21	1.000	(0.054)	0.135	(0.035)	0.000	0.000
N004401	P	7	1.391	(0.137)	–2.242	(0.122)	0.269	(0.060)
N004402	P	8	0.855	(0.077)	–0.149	(0.097)	0.237	(0.037)
N004403	P	9	1.099	(0.091)	–1.806	(0.131)	0.266	(0.057)
N004501	P	10	0.699	(0.081)	0.188	(0.130)	0.287	(0.041)
N004502	P	11	0.615	(0.056)	–1.098	(0.179)	0.253	(0.054)
N004601	P	16	0.811	(0.080)	0.283	(0.091)	0.227	(0.033)
N004602	P	17	0.980	(0.086)	–0.034	(0.081)	0.254	(0.032)
N004603	P	18	1.318	(0.105)	–0.539	(0.069)	0.261	(0.033)
N004605	P	15	0.735	(0.044)	–1.005	(0.067)	0.000	0.000
N004701	Q	7	1.530	(0.116)	–0.781	(0.062)	0.221	(0.032)
N004702	Q	8	0.697	(0.059)	–1.515	(0.175)	0.261	(0.057)
N004703	Q	9	0.800	(0.060)	–1.159	(0.120)	0.206	(0.044)
N004801	Q	10	1.147	(0.098)	–1.310	(0.118)	0.292	(0.050)
N004901	Q	11	0.883	(0.082)	–0.072	(0.096)	0.273	(0.036)
N005002	Q	16	0.891	(0.182)	1.798	(0.150)	0.344	(0.021)
N005003	Q	17	1.178	(0.171)	1.915	(0.116)	0.182	(0.013)
N005101	Q	12	0.644	(0.065)	–2.935	(0.265)	0.273	(0.064)
N005201	Q	16	0.737	(0.172)	1.354	(0.185)	0.588	(0.027)
N005202	Q	17	0.582	(0.073)	0.434	(0.145)	0.247	(0.042)
N005203	Q	18	0.882	(0.189)	1.938	(0.170)	0.314	(0.020)
N005301	Q	19	0.976	(0.090)	–0.138	(0.088)	0.265	(0.035)
N005302	Q	20	1.530	(0.145)	0.585	(0.046)	0.217	(0.020)
N005303	Q	21	0.655	(0.097)	0.808	(0.129)	0.270	(0.036)
N005304	Q	22	1.620	(0.137)	0.072	(0.048)	0.227	(0.024)
N005305	Q	23	1.168	(0.110)	–0.661	(0.095)	0.297	(0.040)
N005403	R	7	1.210	(0.111)	–0.460	(0.088)	0.369	(0.036)

See notes at end of table.→

Table B–2. IRT parameters for the NAEP reading long-term trend items, age 13/grade 8: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N005404	R	8	1.034	(0.092)	–1.466	(0.142)	0.304	(0.057)
N005405	R	9	1.478	(0.116)	0.037	(0.051)	0.253	(0.025)
N005406	R	10	0.970	(0.083)	–0.322	(0.090)	0.258	(0.036)
N005407	R	11	1.270	(0.109)	–0.523	(0.079)	0.316	(0.035)
N005503	R	14	0.705	(0.083)	0.300	(0.124)	0.284	(0.039)
N005504	R	15	1.387	(0.156)	1.042	(0.051)	0.223	(0.017)
N005505	R	16	0.973	(0.086)	–1.048	(0.128)	0.294	(0.050)
N005601	R	17	1.359	(0.121)	–0.621	(0.081)	0.346	(0.037)
N005602	R	18	1.237	(0.116)	0.551	(0.055)	0.237	(0.022)

[1]N001521–N001527 are the same questions as those numbered N001501–N001507 in previous assessments. In 1999 these questions refer to the passage in which references to the "Devil" were changed to references to the "King."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table B–3. IRT parameters for the NAEP reading long-term trend items, age 17/grade 11: 1999

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N001301	H	10	0.842	(0.123)	–0.096	(0.172)	0.590	(0.037)
N001302	H	11	0.518	(0.073)	–3.081	(0.485)	0.586	(0.070)
N001303	H	12	0.862	(0.084)	–0.237	(0.106)	0.293	(0.038)
N001401	H	13	1.095	(0.114)	–0.635	(0.117)	0.430	(0.041)
N001521[1]	H	25	1.469	(0.186)	–1.915	(0.150)	0.290	(0.057)
N001522[1]	H	26	1.588	(0.183)	–1.016	(0.096)	0.234	(0.040)
N001523[1]	H	27	1.489	(0.191)	–1.422	(0.138)	0.295	(0.050)
N001524[1]	H	28	1.504	(0.171)	–1.051	(0.100)	0.229	(0.040)
N001527[1]	H	19	0.441	(0.059)	2.136	(0.263)	0.000	0.000
N001701	J	12	0.614	(0.065)	–1.494	(0.238)	0.323	(0.063)
N001703	J	14	0.984	(0.097)	–0.525	(0.115)	0.361	(0.042)
N001901	J	15	1.055	(0.099)	–0.684	(0.109)	0.345	(0.043)
N001904	J	17	0.726	(0.045)	–1.325	(0.081)	0.000	0.000
N002001	K	22	1.459	(0.117)	–0.383	(0.060)	0.241	(0.032)
N002002	K	23	0.966	(0.081)	–0.743	(0.098)	0.220	(0.039)
N002003	K	24	1.070	(0.098)	–1.210	(0.122)	0.279	(0.047)
N002101	K	12	0.587	(0.070)	0.287	(0.137)	0.200	(0.041)
N002102	K	13	1.636	(0.131)	0.126	(0.044)	0.198	(0.024)
N002201	K	14	1.493	0.000	–0.786	(0.057)	0.378	(0.035)
N002202	K	15	2.101	(0.239)	–0.684	(0.072)	0.501	(0.036)
N002203	K	16	0.382	(0.050)	–3.417	(0.480)	0.307	(0.066)
N002501	L	27	0.558	(0.069)	–0.480	(0.215)	0.323	(0.055)
N002701	L	28	0.665	(0.064)	–0.422	(0.132)	0.194	(0.043)
N002702	L	29	0.804	(0.072)	–1.051	(0.137)	0.214	(0.048)
N002801	L	20	1.645	(0.169)	–1.832	(0.111)	0.265	(0.049)
N002802	L	21	1.322	(0.126)	–1.948	(0.130)	0.273	(0.052)
N002804	L	32	0.217	(0.032)	2.554	(0.380)	0.000	0.000
N002902	M	6	0.545	(0.056)	–1.815	(0.251)	0.290	(0.061)
N002903	M	7	1.794	(0.175)	–1.105	(0.078)	0.282	(0.038)
N002904	M	8	1.041	(0.096)	–0.917	(0.113)	0.299	(0.044)
N002905	M	9	0.961	(0.110)	0.348	(0.094)	0.327	(0.034)
N002906	M	10	1.897	(0.188)	–0.879	(0.070)	0.346	(0.037)
N003001	M	18	1.126	(0.105)	0.501	(0.060)	0.197	(0.025)
N003002	M	12	0.355	(0.048)	–0.421	(0.285)	0.212	(0.057)
N003003	M	19	1.489	(0.149)	1.240	(0.048)	0.079	(0.011)
N003101	M	29	0.867	(0.083)	–1.682	(0.174)	0.291	(0.057)
N003102	M	30	1.237	(0.115)	–1.111	(0.106)	0.283	(0.044)
N003104	M	16	0.631	(0.046)	0.963	(0.070)	0.000	0.000
N003201	N	21	1.155	(0.119)	–1.435	(0.145)	0.350	(0.052)
N003202	N	22	1.177	(0.105)	–0.734	(0.092)	0.292	(0.039)
N003203	N	23	1.000	(0.085)	–0.595	(0.091)	0.224	(0.037)
N003204	N	24	0.836	(0.072)	–1.158	(0.130)	0.225	(0.045)
N003301	N	25	0.861	(0.080)	–1.304	(0.149)	0.273	(0.050)
N003501	N	27	0.629	(0.064)	–1.142	(0.193)	0.280	(0.054)

See notes at end of table.→

Table B–3. IRT parameters for the NAEP reading long-term trend items, age 17/grade 11: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N003601	N	28	0.915	(0.088)	–1.958	(0.184)	0.305	(0.060)
N003602	N	29	1.023	(0.094)	–0.951	(0.114)	0.274	(0.044)
N003701	N	30	0.802	(0.083)	–1.160	(0.173)	0.319	(0.054)
N003702	N	31	1.622	(0.146)	–0.498	(0.065)	0.323	(0.034)
N003704	N	32	0.756	(0.049)	–0.713	(0.063)	0.000	0.000
N003801	O	12	0.500	(0.081)	0.756	(0.181)	0.248	(0.047)
N003802	O	13	0.237	(0.035)	–1.945	(0.467)	0.206	(0.059)
N003803	O	14	0.642	(0.150)	1.927	(0.221)	0.302	(0.029)
N004201	O	21	0.870	(0.086)	–0.418	(0.116)	0.290	(0.042)
N004202	O	22	0.660	(0.087)	–0.004	(0.168)	0.345	(0.047)
N004301	O	23	1.057	(0.109)	–0.123	(0.093)	0.313	(0.037)
N004303	O	24	0.614	(0.049)	–0.375	(0.072)	0.000	0.000
N004501	P	20	0.665	(0.076)	–0.448	(0.176)	0.337	(0.051)
N004502	P	21	0.482	(0.052)	–1.967	(0.288)	0.298	(0.063)
N004601	P	16	0.888	(0.084)	–0.032	(0.094)	0.259	(0.036)
N004602	P	17	1.443	(0.119)	–0.439	(0.065)	0.280	(0.033)
N004603	P	18	1.433	(0.125)	–0.731	(0.076)	0.304	(0.037)
N004605	P	25	0.593	(0.045)	–1.356	(0.110)	0.000	0.000
N004901	Q	10	1.020	(0.096)	–0.602	(0.106)	0.321	(0.042)
N005001	Q	15	2.315	(0.211)	0.689	(0.033)	0.224	(0.016)
N005002	Q	16	1.032	(0.126)	0.733	(0.078)	0.304	(0.028)
N005003	Q	17	0.745	(0.111)	1.471	(0.109)	0.143	(0.024)
N005201	Q	11	0.833	(0.141)	0.396	(0.167)	0.590	(0.035)
N005202	Q	12	0.526	(0.072)	0.157	(0.193)	0.296	(0.049)
N005203	Q	13	0.618	(0.097)	1.116	(0.135)	0.256	(0.035)
N005503	R	14	0.686	(0.085)	–0.027	(0.152)	0.350	(0.043)
N005504	R	15	1.492	(0.152)	0.526	(0.053)	0.314	(0.023)
N005505	R	16	0.815	(0.080)	–1.786	(0.204)	0.336	(0.063)
N015101	R	17	0.828	(0.095)	0.145	(0.113)	0.349	(0.036)
N015102	R	18	2.653	(0.217)	–0.031	(0.031)	0.252	(0.021)
N015103	R	19	2.548	(0.206)	0.060	(0.031)	0.236	(0.020)
N015104	R	20	2.004	(0.165)	–0.070	(0.042)	0.282	(0.025)
N015201	N	26	0.645	(0.063)	–2.563	(0.248)	0.286	(0.062)
N015502	P	16	1.320	(0.111)	–0.275	(0.068)	0.295	(0.033)
N015503	P	17	1.110	(0.101)	0.141	(0.070)	0.261	(0.030)
N015504	P	18	1.248	(0.101)	–0.378	(0.069)	0.258	(0.033)
N015505	P	19	0.720	(0.071)	–0.690	(0.151)	0.284	(0.048)
N015901	Q	14	1.320	(0.135)	0.097	(0.074)	0.378	(0.031)
N015902	Q	15	1.204	(0.110)	0.188	(0.065)	0.256	(0.029)
N015903	Q	16	1.848	(0.168)	0.558	(0.039)	0.204	(0.019)
N016001	O	15	0.862	(0.085)	–1.062	(0.151)	0.316	(0.051)
N016002	O	16	1.066	(0.149)	0.695	(0.092)	0.421	(0.029)
N016003	O	17	0.886	(0.094)	0.068	(0.102)	0.303	(0.037)
N016004	O	18	1.194	(0.103)	–0.482	(0.079)	0.278	(0.036)

See notes at end of table.→

Table B–3. IRT parameters for the NAEP reading long-term trend items, age 17/ grade 11: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N016005	O	19	1.455	(0.123)	–0.404	(0.065)	0.284	(0.033)
N016006	O	20	0.907	(0.086)	–0.083	(0.093)	0.252	(0.036)
N017001	H	7	1.266	(0.114)	–0.140	(0.071)	0.337	(0.031)
N017002	H	8	1.672	(0.159)	0.507	(0.045)	0.284	(0.021)
N017003	H	9	1.292	(0.175)	1.274	(0.066)	0.224	(0.017)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table B–4. IRT parameters for the NAEP mathematics long-term trend items, age 9: 1999

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N250301	M2	20	0.883	(0.083)	0.872	(0.063)	0.306	(0.020)
N250601	M2	13	1.054	(0.064)	–2.110	(0.099)	0.180	(0.046)
N250602	M2	14	0.535	(0.034)	–2.072	(0.167)	0.181	(0.050)
N250603	M2	15	0.874	(0.046)	–0.297	(0.060)	0.132	(0.026)
N250701	M1	7	0.668	(0.037)	–1.353	(0.109)	0.141	(0.041)
N250702	M1	8	1.206	(0.069)	0.462	(0.033)	0.144	(0.015)
N250703	M1	9	1.037	(0.052)	–0.440	(0.052)	0.125	(0.025)
N250901	M2	17	0.511	(0.032)	–1.750	(0.164)	0.180	(0.049)
N250902	M2	18	1.088	(0.065)	0.416	(0.039)	0.149	(0.017)
N250903	M2	19	1.067	(0.053)	–0.114	(0.043)	0.114	(0.020)
N251401	M2	16	0.783	(0.043)	–0.744	(0.086)	0.172	(0.035)
N252001	M2	25	1.243	(0.110)	1.496	(0.052)	0.233	(0.010)
N252101	M1	25	0.739	(0.085)	1.386	(0.077)	0.242	(0.020)
N257201	M1	11	1.030	(0.060)	–0.685	(0.074)	0.268	(0.033)
N257801	M2	3	0.694	(0.039)	–1.445	(0.114)	0.200	(0.043)
N258501	M3	19	0.570	(0.076)	1.434	(0.106)	0.226	(0.027)
N261401	M2	12	0.450	(0.032)	–0.847	(0.165)	0.197	(0.044)
N262201	M1	10	0.762	(0.057)	–0.608	(0.124)	0.342	(0.041)
N262401	M3	18	0.726	(0.070)	0.500	(0.093)	0.295	(0.029)
N262501	M1	19	0.430	(0.043)	–0.085	(0.199)	0.300	(0.045)
N263401	M2	4	0.775	(0.052)	–1.390	(0.144)	0.316	(0.054)
N263402	M2	5	0.888	(0.063)	–0.495	(0.100)	0.342	(0.037)
N265401	M1	21	0.387	(0.114)	3.954	(0.709)	0.286	(0.022)
N266101	M1	22	0.697	(0.097)	1.665	(0.101)	0.267	(0.020)
N267001	M3	16	0.871	(0.054)	–1.417	(0.111)	0.261	(0.048)
N267601	M1	3	1.341	(0.073)	–0.564	(0.052)	0.253	(0.027)
N267602	M1	18	1.056	(0.053)	–0.073	(0.043)	0.142	(0.019)
N268201	M1	24	1.184	(0.084)	0.699	(0.040)	0.239	(0.016)
N269001	M2	26	0.676	(0.099)	2.641	(0.197)	0.084	(0.011)
N269101	M1	23	0.615	(0.076)	1.371	(0.090)	0.217	(0.024)
N270001	M1	14	0.572	(0.024)	–0.599	(0.042)	0.000	0.000
N270901	M1	1	0.719	(0.040)	–2.695	(0.112)	0.000	0.000
N271101	M2	24	0.746	(0.027)	–0.292	(0.029)	0.000	0.000
N272101	M3	17	0.783	(0.052)	–0.948	(0.119)	0.277	(0.045)
N272102	M1	15	0.840	(0.051)	–0.430	(0.079)	0.196	(0.032)
N272301	M2	1	0.912	(0.059)	–2.208	(0.127)	0.220	(0.054)
N272801	M3	15	0.775	(0.047)	–1.739	(0.124)	0.206	(0.050)
N273501	M2	6	0.663	(0.047)	–1.035	(0.153)	0.301	(0.051)
N275401	M2	7	1.048	(0.036)	–0.904	(0.029)	0.000	0.000
N276001	M2	21	0.959	(0.034)	–0.947	(0.032)	0.000	0.000
N276002	M2	22	0.967	(0.037)	1.012	(0.032)	0.000	0.000
N276101	M1	12	1.019	(0.036)	–1.013	(0.032)	0.000	0.000
N276601	M2	2	1.157	(0.068)	–0.976	(0.078)	0.288	(0.039)
N276801	M1	4	0.648	(0.043)	–3.400	(0.176)	0.000	0.000
N276802	M1	5	0.566	(0.030)	–2.399	(0.109)	0.000	0.000
N276803	M1	6	0.629	(0.025)	–0.070	(0.033)	0.000	0.000

See notes at end of table.→

Table B–4. IRT parameters for the NAEP reading long-term trend items, age 9: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N277401	M1	2	1.007	(0.056)	–1.541	(0.089)	0.210	(0.043)
N277501	M2	8	0.793	(0.029)	–0.700	(0.033)	0.000	0.000
N277601	M2	9	0.864	(0.031)	–0.902	(0.034)	0.000	0.000
N277602	M2	10	0.794	(0.028)	–0.004	(0.026)	0.000	0.000
N277603	M2	11	0.804	(0.028)	–0.232	(0.027)	0.000	0.000
N284001	M1	16	0.794	(0.029)	–0.836	(0.036)	0.000	0.000
N284002	M1	17	0.801	(0.041)	1.836	(0.070)	0.000	0.000
N286101	M1	13	0.893	(0.032)	–0.935	(0.034)	0.000	0.000
N286102	M2	23	0.978	(0.032)	–0.057	(0.023)	0.000	0.000

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table B–5. IRT parameters for the NAEP mathematics long-term trend items, age 13: 1999

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N250201	M2	19	0.582	(0.042)	–1.517	(0.180)	0.285	(0.055)
N250701	M2	14	0.399	(0.037)	–4.266	(0.378)	0.135	(0.048)
N250702	M2	15	0.798	(0.042)	–1.282	(0.083)	0.131	(0.035)
N250703	M2	16	0.648	(0.040)	–2.481	(0.142)	0.107	(0.039)
N250901	M1	25	0.342	(0.031)	–3.446	(0.339)	0.185	(0.053)
N250902	M1	26	0.867	(0.045)	–0.962	(0.074)	0.149	(0.034)
N250903	M1	27	0.795	(0.048)	–2.066	(0.117)	0.134	(0.042)
N252001	M2	40	0.974	(0.075)	0.677	(0.054)	0.257	(0.020)
N252101	M1	41	0.841	(0.080)	0.392	(0.092)	0.386	(0.029)
N252901	M1	32	1.162	(0.056)	–0.156	(0.038)	0.110	(0.019)
N253701	M2	22	0.412	(0.043)	–0.067	(0.212)	0.407	(0.040)
N254001	M3	28	0.946	(0.056)	–0.710	(0.080)	0.223	(0.037)
N254601	M1	16	0.851	(0.067)	–2.016	(0.171)	0.388	(0.059)
N254602	M1	46	0.980	(0.093)	1.256	(0.056)	0.238	(0.016)
N255701	M1	50	0.963	(0.063)	0.653	(0.045)	0.150	(0.017)
N256101	M2	17	0.889	(0.040)	–1.668	(0.056)	0.000	0.000
N256501	M3	30	1.285	(0.088)	0.284	(0.046)	0.300	(0.020)
N256801	M3	32	1.271	(0.087)	0.374	(0.045)	0.286	(0.019)
N257601	M1	35	1.148	(0.040)	–0.633	(0.024)	0.000	0.000
N258801	M1	38	1.455	(0.124)	0.663	(0.046)	0.422	(0.016)
N258802	M2	31	1.407	(0.085)	0.284	(0.036)	0.231	(0.017)
N258803	M2	41	1.123	(0.084)	1.025	(0.042)	0.183	(0.014)
N260101	M1	43	1.417	(0.085)	–0.146	(0.044)	0.267	(0.023)
N261001	M1	47	0.698	(0.056)	0.387	(0.087)	0.226	(0.029)
N261201	M2	38	0.561	(0.106)	2.470	(0.208)	0.235	(0.021)
N261301	M2	37	0.474	(0.040)	0.861	(0.101)	0.115	(0.027)
N261501	M2	34	0.620	(0.045)	–0.785	(0.142)	0.257	(0.046)
N261801	M2	35	0.695	(0.050)	–0.115	(0.098)	0.241	(0.034)
N262201	M2	18	0.579	(0.046)	–1.510	(0.208)	0.349	(0.059)
N262401	M1	28	1.140	(0.072)	–0.521	(0.070)	0.318	(0.032)
N262501	M1	33	0.512	(0.041)	–1.303	(0.208)	0.319	(0.055)
N263101	M1	39	0.662	(0.027)	–0.495	(0.035)	0.000	0.000
N263401	M2	12	0.779	(0.059)	–2.583	(0.177)	0.271	(0.056)
N263402	M2	13	0.734	(0.051)	–1.926	(0.154)	0.295	(0.054)
N263501	M2	30	0.876	(0.042)	–0.073	(0.045)	0.076	(0.019)
N264701	M2	33	1.240	(0.079)	0.347	(0.041)	0.244	(0.018)
N265201	M1	36	0.706	(0.059)	–2.429	(0.218)	0.362	(0.064)
N265202	M1	30	0.689	(0.056)	–0.596	(0.148)	0.357	(0.046)
N265901	M1	40	0.817	(0.072)	1.008	(0.063)	0.221	(0.020)
N265902	M3	31	0.771	(0.100)	1.556	(0.090)	0.299	(0.019)
N266101	M3	27	0.902	(0.065)	–0.417	(0.095)	0.333	(0.037)
N266801	M1	31	0.597	(0.042)	–1.383	(0.161)	0.277	(0.051)
N267201	M1	23	0.898	(0.069)	–0.904	(0.128)	0.434	(0.045)
N269001	M1	44	1.021	(0.059)	–0.078	(0.054)	0.163	(0.025)
N269101	M2	26	0.870	(0.054)	–0.447	(0.082)	0.229	(0.034)
N269201	M2	44	0.897	(0.040)	1.614	(0.053)	0.000	0.000
N269901	M3	29	0.770	(0.054)	–0.567	(0.109)	0.286	(0.040)

See notes at end of table.→

Table B–5. IRT parameters for the NAEP mathematics long-term trend items, age 13: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N270301	M2	20	0.475	(0.038)	–1.805	(0.243)	0.215	(0.066)
N270302	M2	21	1.378	(0.093)	1.618	(0.043)	0.076	(0.006)
N273901	M1	37	1.440	(0.078)	–0.188	(0.039)	0.218	(0.021)
N274801	M1	29	1.342	(0.113)	0.425	(0.054)	0.456	(0.019)
N275001	M1	42	0.863	(0.032)	0.571	(0.028)	0.000	0.000
N275301	M3	25	0.409	(0.032)	–2.050	(0.230)	0.191	(0.053)
N276801	M1	17	0.483	(0.043)	–4.597	(0.354)	0.000	0.000
N276802	M1	18	0.471	(0.039)	–4.251	(0.305)	0.000	0.000
N276803	M1	19	0.396	(0.024)	–2.078	(0.124)	0.000	0.000
N277401	M2	8	0.569	(0.044)	–3.276	(0.236)	0.183	(0.053)
N277601	M1	20	0.737	(0.041)	–2.495	(0.106)	0.000	0.000
N277602	M1	21	0.678	(0.031)	–1.452	(0.059)	0.000	0.000
N277603	M1	22	0.606	(0.030)	–1.766	(0.078)	0.000	0.000
N277901	M2	9	0.703	(0.043)	–2.912	(0.135)	0.000	0.000
N277902	M2	10	0.620	(0.040)	–3.233	(0.169)	0.000	0.000
N277903	M2	11	0.645	(0.036)	–2.525	(0.114)	0.000	0.000
N278901	M2	32	1.277	(0.079)	0.098	(0.045)	0.275	(0.020)
N278902	M2	29	0.982	(0.091)	0.989	(0.059)	0.308	(0.018)
N278903	M2	42	2.011	(0.131)	0.741	(0.026)	0.222	(0.011)
N278904	M1	49	0.619	(0.073)	1.408	(0.093)	0.210	(0.025)
N281401	M2	39	0.775	(0.111)	2.146	(0.125)	0.187	(0.014)
N281901	M1	15	1.180	(0.084)	–2.201	(0.109)	0.201	(0.046)
N282201	M2	28	1.089	(0.077)	0.465	(0.050)	0.269	(0.020)
N282202	M3	26	1.368	(0.093)	–0.326	(0.058)	0.377	(0.027)
N283101	M1	51	1.729	(0.109)	0.952	(0.027)	0.144	(0.009)
N285701	M2	27	0.933	(0.074)	0.085	(0.080)	0.345	(0.029)
N286201	M1	24	0.876	(0.050)	–1.032	(0.086)	0.218	(0.038)
N286301	M1	45	1.332	(0.078)	0.296	(0.036)	0.212	(0.017)
N286501	M1	48	0.870	(0.057)	0.615	(0.050)	0.130	(0.019)
N286502	M2	43	0.955	(0.062)	0.955	(0.041)	0.096	(0.013)
N286601	M2	23	0.932	(0.033)	–0.185	(0.024)	0.000	0.000
N286602	M2	24	0.959	(0.033)	–0.310	(0.024)	0.000	0.000
N286603	M2	25	1.094	(0.038)	0.626	(0.024)	0.000	0.000

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table B–6. IRT parameters for the NAEP mathematics long-term trend items, age 17: 1999

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N251101	M1	49	1.214	(0.043)	0.815	(0.024)	0.000	0.000
N251701	M2	41	0.859	(0.048)	–0.605	(0.074)	0.136	(0.033)
N253901	M1	39	1.240	(0.065)	–0.637	(0.051)	0.196	(0.027)
N253902	M1	40	0.572	(0.068)	0.266	(0.178)	0.378	(0.044)
N253903	M1	41	0.949	(0.071)	0.264	(0.067)	0.304	(0.025)
N253904	M1	42	1.713	(0.126)	0.419	(0.037)	0.386	(0.015)
N254001	M2	21	0.794	(0.050)	–1.355	(0.124)	0.222	(0.052)
N254301	M1	33	1.017	(0.074)	0.127	(0.067)	0.308	(0.026)
N254601	M2	15	0.962	(0.073)	–2.586	(0.153)	0.262	(0.061)
N254602	M1	27	1.338	(0.067)	–0.446	(0.042)	0.165	(0.023)
N255501	M3	33	0.879	(0.071)	0.273	(0.078)	0.305	(0.027)
N255601	M2	45	2.331	(0.124)	1.324	(0.033)	0.340	(0.009)
N255701	M1	32	1.103	(0.059)	–1.220	(0.069)	0.185	(0.036)
N255801	M2	49	0.864	(0.040)	1.659	(0.056)	0.000	0.000
N256001	M3	34	0.971	(0.034)	–0.227	(0.025)	0.000	0.000
N256101	M1	15	0.769	(0.040)	–2.282	(0.088)	0.000	0.000
N256801	M1	36	1.238	(0.080)	–0.431	(0.064)	0.335	(0.029)
N257101	M3	35	0.620	(0.123)	2.305	(0.194)	0.309	(0.020)
N258801	M2	38	1.516	(0.097)	–0.448	(0.052)	0.368	(0.026)
N258802	M1	26	1.750	(0.096)	–0.553	(0.038)	0.234	(0.023)
N258803	M1	37	1.175	(0.065)	–0.211	(0.049)	0.200	(0.024)
N258804	M1	18	0.834	(0.059)	–2.233	(0.159)	0.282	(0.063)
N259001	M2	31	1.015	(0.034)	–0.257	(0.024)	0.000	0.000
N259901	M1	28	0.937	(0.060)	–0.296	(0.074)	0.233	(0.031)
N260101	M2	20	1.240	(0.071)	–1.453	(0.073)	0.213	(0.041)
N260601	M1	16	1.508	(0.072)	–1.745	(0.040)	0.000	0.000
N260801	M2	43	1.406	(0.045)	–0.057	(0.018)	0.000	0.000
N260901	M1	35	1.654	(0.084)	–0.274	(0.033)	0.187	(0.019)
N261001	M2	40	0.788	(0.048)	–0.565	(0.090)	0.221	(0.036)
N261201	M2	26	0.443	(0.041)	0.119	(0.172)	0.210	(0.043)
N261301	M2	28	0.493	(0.039)	0.134	(0.126)	0.148	(0.037)
N261501	M2	24	0.641	(0.041)	–2.166	(0.155)	0.194	(0.053)
N261601	M2	27	0.887	(0.136)	1.768	(0.101)	0.376	(0.016)
N261801	M2	25	0.499	(0.034)	–1.642	(0.177)	0.218	(0.051)
N262301	M2	17	0.582	(0.047)	–1.376	(0.214)	0.325	(0.064)
N262401	M1	17	1.050	(0.066)	–1.497	(0.101)	0.275	(0.050)
N262501	M2	35	0.503	(0.039)	–1.636	(0.226)	0.334	(0.059)
N262502	M2	36	1.150	(0.108)	1.307	(0.051)	0.258	(0.013)
N262601	M1	38	0.700	(0.055)	0.181	(0.095)	0.233	(0.032)
N263001	M1	43	0.679	(0.028)	0.722	(0.037)	0.000	0.000
N263101	M2	37	0.801	(0.031)	–0.899	(0.037)	0.000	0.000
N263201	M2	18	0.819	(0.064)	–1.539	(0.174)	0.417	(0.061)
N263202	M2	19	1.005	(0.076)	–0.635	(0.104)	0.432	(0.038)
N264301	M1	47	0.849	(0.035)	1.234	(0.042)	0.000	0.000

See notes at end of table.→

Table B–6. IRT parameters for the NAEP mathematics long-term trend items, age 17: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N264701	M2	39	1.396	(0.075)	–0.382	(0.043)	0.227	(0.023)
N266501	M3	31	0.815	(0.059)	–0.325	(0.100)	0.289	(0.037)
N268801	M2	48	1.287	(0.080)	1.071	(0.032)	0.091	(0.009)
N268901	M2	47	1.642	(0.092)	0.333	(0.029)	0.204	(0.014)
N269001	M2	22	1.471	(0.083)	–0.436	(0.045)	0.241	(0.025)
N270301	M1	30	0.838	(0.051)	–2.257	(0.120)	0.136	(0.048)
N270302	M1	31	1.235	(0.055)	–0.289	(0.035)	0.086	(0.018)
N271301	M3	32	1.497	(0.094)	–0.006	(0.043)	0.303	(0.021)
N278501	M1	23	0.889	(0.045)	–0.674	(0.042)	0.000	0.000
N278502	M1	24	0.939	(0.045)	–0.216	(0.034)	0.000	0.000
N278503	M1	25	0.760	(0.040)	–0.630	(0.046)	0.000	0.000
N278901	M2	23	1.066	(0.058)	–0.827	(0.067)	0.207	(0.033)
N278902	M2	42	1.030	(0.067)	–0.368	(0.074)	0.296	(0.031)
N278903	M2	44	1.263	(0.067)	–0.341	(0.046)	0.192	(0.024)
N278905	M1	44	0.471	(0.060)	1.114	(0.141)	0.225	(0.037)
N280401	M2	30	0.607	(0.026)	–0.754	(0.043)	0.000	0.000
N281401	M2	29	0.549	(0.063)	1.450	(0.096)	0.150	(0.025)
N286001	M1	19	0.722	(0.038)	–1.317	(0.090)	0.132	(0.036)
N286002	M1	20	0.954	(0.050)	–1.740	(0.083)	0.115	(0.038)
N286301	M2	33	1.027	(0.056)	–1.030	(0.072)	0.201	(0.035)
N286302	M1	22	1.049	(0.066)	–0.892	(0.085)	0.302	(0.039)
N286501	M2	34	1.077	(0.058)	–1.071	(0.070)	0.179	(0.036)
N286502	M1	34	1.359	(0.067)	–0.483	(0.041)	0.156	(0.022)
N287101	M1	29	1.166	(0.070)	–0.598	(0.065)	0.267	(0.032)
N287102	M2	32	0.979	(0.053)	–0.863	(0.073)	0.195	(0.035)
N287301	M1	45	0.808	(0.030)	0.578	(0.030)	0.000	0.000
N287302	M1	46	0.839	(0.031)	0.443	(0.028)	0.000	0.000

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table B–7. IRT parameters for the NAEP science long-term trend items, age 9: 1999

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N400001	S1	6	0.731	(0.066)	–0.975	(0.166)	0.447	(0.040)
N400101	S1	15	1.419	(0.168)	1.245	(0.064)	0.507	(0.013)
N400102	S1	16	1.250	(0.133)	1.163	(0.064)	0.463	(0.015)
N400301	S1	8	0.914	(0.079)	0.053	(0.076)	0.433	(0.023)
N400401	S1	9	1.002	(0.086)	–1.067	(0.131)	0.526	(0.032)
N400402	S1	10	2.077	(0.135)	–0.571	(0.041)	0.357	(0.019)
N400403	S1	11	0.707	(0.068)	–1.416	(0.225)	0.556	(0.045)
N400404	S1	12	1.543	(0.110)	–0.403	(0.053)	0.430	(0.020)
N400405	S1	13	0.858	(0.071)	–0.611	(0.109)	0.431	(0.030)
N400501	S1	14	0.454	(0.055)	0.500	(0.163)	0.324	(0.036)
N400601	S1	17	0.832	(0.070)	0.229	(0.073)	0.359	(0.023)
N400701	S1	18	1.043	(0.074)	0.524	(0.045)	0.270	(0.017)
N400901	S1	19	0.269	(0.044)	2.117	(0.304)	0.342	(0.030)
N401001	S1	20	0.548	(0.055)	0.942	(0.092)	0.218	(0.024)
N401101	S1	21	0.314	(0.051)	1.558	(0.248)	0.378	(0.034)
N401201	S1	22	0.682	(0.169)	2.935	(0.341)	0.271	(0.014)
N401301	S1	23	0.464	(0.056)	0.550	(0.155)	0.313	(0.036)
N401501	S2	1	0.308	(0.058)	1.910	(0.306)	0.433	(0.032)
N401601	S2	2	0.599	(0.053)	–1.062	(0.191)	0.324	(0.049)
N401702	S2	4	0.263	(0.054)	2.105	(0.443)	0.570	(0.027)
N401703	S2	5	0.368	(0.087)	1.816	(0.317)	0.546	(0.034)
N401801	S2	6	1.488	(0.140)	0.078	(0.056)	0.572	(0.019)
N401802	S2	7	1.699	(0.180)	–0.047	(0.062)	0.668	(0.018)
N401803	S2	8	1.312	(0.144)	0.289	(0.069)	0.646	(0.018)
N401804	S2	9	0.839	(0.099)	1.023	(0.092)	0.481	(0.020)
N401901	S2	10	0.316	(0.078)	3.109	(0.464)	0.346	(0.032)
N402001	S2	11	0.774	(0.065)	–1.020	(0.149)	0.421	(0.038)
N402002	S2	12	0.751	(0.066)	–1.051	(0.162)	0.446	(0.040)
N402005	S2	15	0.683	(0.077)	–0.045	(0.140)	0.506	(0.031)
N402101	S2	16	0.554	(0.049)	–0.570	(0.158)	0.256	(0.041)
N402201	S2	17	0.318	(0.039)	0.267	(0.243)	0.316	(0.041)
N402401	S2	18	0.485	(0.155)	3.768	(0.732)	0.360	(0.017)
N402501	S2	19	1.198	(0.107)	1.366	(0.054)	0.229	(0.013)
N402602	S2	21	0.633	0.000	0.578	(0.140)	0.621	(0.020)
N402701	S2	23	0.598	(0.069)	1.534	(0.103)	0.211	(0.021)
N402801	S2	24	1.832	(0.109)	1.758	(0.046)	0.188	(0.007)
N402901	S2	25	0.374	(0.115)	4.771	(1.004)	0.202	(0.018)
N403001	S3	12	0.353	(0.046)	–6.733	(0.802)	0.314	(0.067)
N403101	S3	13	0.470	(0.040)	–4.565	(0.357)	0.303	(0.065)
N403201	S3	14	0.515	(0.032)	–2.546	(0.205)	0.221	(0.054)
N403202	S3	15	0.408	(0.033)	–1.083	(0.214)	0.236	(0.046)
N403301	S3	16	0.690	(0.054)	–0.923	(0.140)	0.327	(0.038)
N403401	S3	17	0.475	(0.055)	0.328	(0.166)	0.355	(0.036)
N403501	S3	18	0.546	(0.059)	0.156	(0.146)	0.385	(0.034)

See notes at end of table.→

Table B–7. IRT parameters for the NAEP science long-term trend items, age 9: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N403502	S3	19	0.592	(0.056)	–1.781	(0.281)	0.534	(0.053)
N403503	S3	20	0.385	(0.055)	0.407	(0.248)	0.459	(0.039)
N403601	S3	21	0.883	(0.069)	0.780	(0.055)	0.257	(0.019)
N403701	S3	22	5.191	0.000	–0.120	(0.014)	0.358	(0.014)
N403702	S3	23	4.971	0.000	–0.145	(0.017)	0.491	(0.015)
N403703	S3	24	5.221	(0.426)	–0.060	(0.020)	0.422	(0.015)
N403801	S3	25	0.691	(0.116)	1.801	(0.141)	0.460	(0.019)
N403803	S3	27	0.599	(0.062)	–0.570	(0.185)	0.471	(0.038)
N403804	S3	28	0.467	(0.059)	0.030	(0.215)	0.448	(0.039)
N403901	S3	29	0.701	(0.055)	–0.160	(0.092)	0.267	(0.028)
N404001	S3	30	0.280	(0.031)	0.730	(0.204)	0.195	(0.035)
N404201	S3	31	0.485	(0.050)	1.152	(0.098)	0.159	(0.025)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table B–8. IRT parameters for the NAEP science long-term trend items, age 13: 1999

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N400001	S1	6	0.537	(0.037)	–1.736	(0.194)	0.264	(0.057)
N400101	S1	15	0.914	(0.070)	0.260	(0.068)	0.299	(0.024)
N400102	S1	16	0.788	(0.052)	–2.558	(0.157)	0.261	(0.061)
N400301	S1	8	0.532	(0.049)	–0.260	(0.171)	0.273	(0.046)
N400401	S1	9	0.700	(0.044)	–1.790	(0.155)	0.271	(0.057)
N400402	S1	10	0.640	(0.054)	0.057	(0.111)	0.244	(0.035)
N400403	S1	11	1.047	(0.091)	–1.543	(0.166)	0.568	(0.051)
N400404	S1	12	1.521	(0.089)	–0.386	(0.046)	0.326	(0.022)
N400405	S1	13	1.348	(0.096)	0.461	(0.042)	0.337	(0.017)
N400501	S1	14	0.771	(0.052)	–0.446	(0.098)	0.254	(0.035)
N400601	S1	17	0.646	(0.082)	0.955	(0.102)	0.344	(0.028)
N400701	S1	18	0.880	(0.073)	0.926	(0.051)	0.188	(0.017)
N400901	S1	19	0.483	(0.072)	0.689	(0.193)	0.378	(0.043)
N401001	S1	20	0.641	(0.086)	1.363	(0.094)	0.280	(0.024)
N401101	S1	21	0.920	(0.085)	1.175	(0.053)	0.200	(0.016)
N401201	S1	22	1.237	(0.094)	0.507	(0.047)	0.348	(0.018)
N401301	S1	23	0.679	(0.141)	1.870	(0.165)	0.452	(0.020)
N401501	S2	1	0.947	(0.064)	0.441	(0.050)	0.198	(0.019)
N401601	S2	2	1.432	(0.109)	1.176	(0.037)	0.241	(0.011)
N401702	S2	4	1.000	(0.103)	1.427	(0.058)	0.190	(0.013)
N401703	S2	5	1.217	(0.203)	2.679	(0.218)	0.136	(0.007)
N401801	S2	6	1.564	(0.197)	2.374	(0.140)	0.186	(0.007)
N401802	S2	7	1.290	(0.145)	2.312	(0.124)	0.111	(0.006)
N401803	S2	8	0.794	(0.228)	2.040	(0.255)	0.660	(0.015)
N401804	S2	9	0.308	(0.035)	–0.914	(0.302)	0.472	(0.040)
N401901	S2	10	1.072	(0.090)	0.224	(0.070)	0.431	(0.023)
N402001	S2	11	0.681	(0.117)	1.295	(0.121)	0.482	(0.024)
N402002	S2	12	0.986	(0.110)	0.267	(0.101)	0.598	(0.023)
N402005	S2	15	1.306	(0.103)	0.243	(0.055)	0.446	(0.020)
N402101	S2	16	1.237	(0.126)	–0.564	(0.117)	0.685	(0.027)
N402201	S2	17	1.367	(0.112)	–0.129	(0.068)	0.543	(0.022)
N402401	S2	18	1.213	(0.101)	–0.190	(0.078)	0.526	(0.025)
N402501	S2	19	0.894	(0.092)	1.018	(0.060)	0.278	(0.019)
N402602	S2	21	0.430	(0.038)	–1.061	(0.239)	0.311	(0.053)
N402701	S2	23	0.978	(0.084)	0.518	(0.063)	0.345	(0.022)
N402801	S2	24	0.734	(0.058)	–1.740	(0.202)	0.463	(0.059)
N402901	S2	25	0.503	(0.093)	0.930	(0.210)	0.528	(0.036)
N403001	S3	12	0.832	(0.062)	–0.765	(0.123)	0.387	(0.040)
N403101	S3	13	0.671	(0.052)	–1.137	(0.171)	0.391	(0.050)
N403201	S3	14	1.706	(0.151)	1.289	(0.053)	0.535	(0.010)
N403202	S3	15	0.413	(0.047)	–0.026	(0.232)	0.423	(0.042)
N403301	S3	16	0.661	(0.066)	0.096	(0.132)	0.349	(0.037)
N403401	S3	17	0.413	(0.078)	1.361	(0.205)	0.350	(0.043)
N403501	S3	18	0.948	(0.123)	1.845	(0.096)	0.157	(0.012)

See notes at end of table.→

Table B–8. IRT parameters for the NAEP science long-term trend items, age 13: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N403502	S3	19	0.815	(0.082)	0.689	(0.076)	0.329	(0.024)
N403503	S3	20	0.465	(0.056)	0.891	(0.139)	0.217	(0.036)
N403601	S3	21	0.846	(0.164)	1.845	(0.140)	0.418	(0.017)
N403701	S3	22	0.706	(0.090)	1.339	(0.083)	0.256	(0.022)
N403702	S3	23	0.846	(0.087)	1.195	(0.061)	0.213	(0.018)
N403703	S3	24	1.193	(0.132)	1.724	(0.080)	0.287	(0.011)
N403801	S3	25	0.568	(0.079)	1.219	(0.110)	0.276	(0.030)
N403803	S3	27	1.519	(0.115)	0.962	(0.032)	0.214	(0.012)
N403804	S3	28	0.901	(0.131)	1.932	(0.114)	0.186	(0.013)
N403901	S3	29	1.188	(0.096)	0.722	(0.045)	0.328	(0.016)
N404001	S3	30	0.729	(0.060)	–1.069	(0.174)	0.452	(0.049)
N404201	S3	31	0.776	(0.078)	–0.719	(0.185)	0.574	(0.042)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table B–9. IRT parameters for the NAEP science long-term trend items, age 17: 1999

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N400201	S1	12	0.475	(0.040)	–3.349	(0.289)	0.222	(0.057)
N401201	S1	30	0.911	(0.070)	–0.253	(0.091)	0.347	(0.033)
N404601	S1	13	0.381	(0.033)	–1.641	(0.263)	0.235	(0.056)
N405001	S1	29	0.349	(0.034)	–0.436	(0.225)	0.255	(0.045)
N405101	S3	14	1.047	(0.068)	0.072	(0.054)	0.246	(0.023)
N405201	S1	31	0.538	(0.064)	0.321	(0.166)	0.307	(0.044)
N405401	S3	19	0.780	(0.057)	0.481	(0.061)	0.170	(0.023)
N405501	S3	21	0.598	(0.045)	–0.641	(0.140)	0.243	(0.043)
N406001	S1	33	0.667	(0.104)	2.045	(0.144)	0.179	(0.018)
N406101	S1	35	1.257	(0.120)	1.729	(0.069)	0.198	(0.010)
N406201	S1	37	0.849	(0.081)	1.652	(0.074)	0.087	(0.011)
N406301	S1	21	1.079	(0.121)	0.520	(0.081)	0.557	(0.021)
N406302	S1	22	0.230	(0.030)	–1.855	(0.427)	0.426	(0.044)
N406303	S1	23	0.783	(0.068)	–0.424	(0.132)	0.427	(0.039)
N406304	S1	24	0.466	(0.057)	–0.098	(0.240)	0.458	(0.045)
N406401	S2	10	0.851	(0.072)	–0.602	(0.130)	0.453	(0.040)
N406402	S2	11	1.068	(0.081)	–0.601	(0.095)	0.439	(0.034)
N406403	S2	12	1.029	(0.081)	–1.536	(0.137)	0.484	(0.049)
N406404	S2	13	1.202	(0.088)	–1.079	(0.099)	0.452	(0.040)
N406405	S2	14	1.049	(0.076)	–1.232	(0.114)	0.439	(0.044)
N406601	S1	28	0.372	(0.031)	–2.340	(0.280)	0.206	(0.056)
N406801	S2	16	0.750	(0.065)	–2.177	(0.211)	0.476	(0.058)
N406802	S2	17	0.358	(0.050)	0.776	(0.235)	0.440	(0.035)
N406803	S2	18	0.638	(0.049)	–1.431	(0.181)	0.392	(0.051)
N406804	S2	19	0.646	(0.050)	–1.891	(0.191)	0.389	(0.054)
N406805	S2	20	0.605	(0.084)	0.631	(0.154)	0.467	(0.034)
N406806	S2	21	0.355	(0.038)	–0.519	(0.245)	0.392	(0.042)
N406901	S2	27	0.522	(0.045)	–0.674	(0.188)	0.257	(0.052)
N407001	S2	33	0.359	(0.033)	–0.558	(0.215)	0.199	(0.046)
N407101	S2	38	1.033	(0.096)	1.172	(0.050)	0.193	(0.015)
N407201	S2	32	0.690	(0.070)	0.361	(0.110)	0.323	(0.033)
N407301	S2	36	0.301	(0.035)	0.701	(0.209)	0.250	(0.036)
N407302	S2	37	0.910	(0.130)	1.263	(0.084)	0.439	(0.019)
N407401	S2	28	0.415	(0.043)	–0.691	(0.249)	0.440	(0.045)
N407403	S2	30	0.589	(0.061)	–0.312	(0.189)	0.420	(0.045)
N407404	S2	31	0.504	(0.045)	–2.523	(0.278)	0.395	(0.060)
N407701	S2	35	0.676	(0.058)	0.757	(0.071)	0.157	(0.024)
N408101	S1	38	0.738	(0.087)	1.408	(0.078)	0.194	(0.020)
N408301	S3	10	1.025	(0.070)	–0.694	(0.090)	0.384	(0.034)
N408302	S3	11	0.797	(0.060)	–1.991	(0.169)	0.424	(0.054)
N408303	S3	12	0.707	(0.056)	–2.144	(0.202)	0.438	(0.057)
N408304	S3	13	1.136	(0.083)	–1.660	(0.118)	0.445	(0.048)
N408601	S1	19	0.335	(0.031)	–2.841	(0.341)	0.209	(0.059)
N408801	S3	24	0.681	(0.053)	–0.558	(0.135)	0.298	(0.043)
N408901	S3	15	0.887	(0.068)	–1.106	(0.137)	0.465	(0.044)
N408902	S3	16	1.404	(0.116)	–1.744	(0.109)	0.486	(0.048)
N408903	S3	17	0.785	(0.074)	0.126	(0.106)	0.420	(0.030)
N408904	S3	18	0.555	(0.058)	–0.063	(0.173)	0.385	(0.041)
N409301	S1	20	0.782	(0.049)	–1.456	(0.122)	0.230	(0.048)
N409501	S1	34	0.690	(0.069)	1.153	(0.071)	0.160	(0.021)
N409901	S1	18	0.762	(0.052)	–0.844	(0.119)	0.266	(0.044)

See notes at end of table.→

Table B–9. IRT parameters for the NAEP 1999 science long-term trend items, age 17: 1999—Continued

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.
N410003	S1	16	0.222	(0.033)	–4.355	(0.770)	0.461	(0.060)
N410004	S1	17	0.323	(0.037)	–1.669	(0.363)	0.485	(0.047)
N410101	S1	25	0.587	(0.059)	–1.261	(0.261)	0.507	(0.057)
N410102	S1	26	0.288	(0.034)	–1.361	(0.346)	0.450	(0.043)
N410103	S1	27	0.301	(0.035)	–1.673	(0.359)	0.450	(0.046)
N410201	S1	32	0.799	(0.099)	1.511	(0.082)	0.229	(0.018)
N410401	S2	15	0.298	(0.035)	0.143	(0.241)	0.340	(0.038)
N410501	S2	22	0.339	(0.027)	–0.956	(0.190)	0.146	(0.040)
N410601	S2	23	1.990	(0.104)	1.130	(0.026)	0.139	(0.008)
N410602	S2	24	0.528	(0.049)	–2.606	(0.295)	0.425	(0.064)
N410603	S2	25	1.475	(0.144)	1.027	(0.045)	0.398	(0.014)
N410604	S2	26	0.494	(0.046)	–2.481	(0.304)	0.425	(0.063)
N410701	S2	34	0.762	(0.072)	0.760	(0.071)	0.233	(0.024)
N410801	S2	39	0.281	(0.038)	1.644	(0.234)	0.224	(0.032)
N410901	S2	40	0.944	(0.068)	1.073	(0.044)	0.098	(0.013)
N411001	S2	41	1.111	(0.101)	1.430	(0.052)	0.131	(0.011)
N411101	S3	22	0.582	(0.043)	–0.253	(0.118)	0.189	(0.037)
N411201	S3	23	0.684	(0.050)	–0.083	(0.093)	0.200	(0.032)
N411301	S3	20	0.612	(0.145)	3.165	(0.410)	0.141	(0.013)
N411401	S3	25	1.941	(0.111)	0.301	(0.025)	0.227	(0.014)
N411501	S3	26	1.102	(0.093)	1.132	(0.045)	0.196	(0.014)
N411502	S3	27	0.878	(0.056)	–1.107	(0.107)	0.292	(0.043)
N411601	S3	28	1.319	(0.094)	0.849	(0.035)	0.203	(0.013)
N411701	S3	29	1.051	(0.080)	0.904	(0.043)	0.183	(0.015)
N411801	S3	30	1.885	(0.103)	0.281	(0.025)	0.196	(0.014)
N411901	S3	31	1.256	(0.117)	1.194	(0.045)	0.252	(0.013)
N412001	S3	32	1.418	(0.118)	1.600	(0.056)	0.251	(0.009)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Appendix C

Conditioning Variables and Contrast Codings

This appendix contains information about the conditioning variables used in scaling/plausible value estimation for the 1999 NAEP assessment. The initial step in construction of conditioning variables involves forming primary student–based vectors of response data from answers to student and school questionnaires, demographic and background data such as supplied by Westat, and other student information known prior to scaling. The initial conditioning vectors concatenate this student background information into a series of identifying "contrasts" comprising:

1. Categorical variables derived by expanding the response options of a questionnaire variable into a binary series of one–degree–of–freedom "dummy" variables or contrasts, (these form the majority of each student conditioning vector);

2. Questionnaire or demographic variables that possess ordinal response options, such as number of hours spent watching television, which are included as linear and/or quadratic multi–degree–of–freedom contrasts;

3. Continuous variables, such as student logit scores based on percent correct values, included as contrasts in their original form or a transformation of their original form, and;

4. Interactions of two or more categorical variables forming a set of orthogonal one–degree–of–freedom dummy variables or contrasts.

This appendix gives the specifications used for constructing the conditioning variables. Table C–1 provides a description of the specifications provided for each of the conditioning variables. Table C–2 provides a summary of the conditioning variables specific to reading, and tables C–3 and C–4 provide the variables for mathematics and science respectively.

The conditioning variables differ by subjects and age classes due to different questions being included on questionnaires for each subject and age class. They also differ because the current conditioning variables and contrast codings were selected to match those used in analyses of data from previous assessment years. In the past, computational limitations determined the number of contrasts that could be included in the conditioning models. Therefore, the conditioning variables and contrast codings specified in this appendix reflect earlier limitations in technology.

Table C–1. Description of specifications provided for each conditioning variable in the NAEP long-term trend assessment: 1999

Title	Description
Conditioning variable	A short description of the conditioning variable.
Age classes	Specifies student age cohort(s) (9=9 years old, 13=13 years old, 17=17 years old, and All=all ages) in which the conditioning variable was used.
Variable name	The seven–character NAEP database identification for the conditioning variable.
Variable coding	Short description of the variable coding.
Contrast coding	The codes that correspond to each set of contrasts for a given conditioning variable.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table C–2. Conditioning variables for the NAEP long-term trend reading assessment: 1999

Conditioning variable	Age classes	Variable name(s)	Variable coding	Contrast coding
Overall	All		—	1
Gender	All	DSEX	Male	0
			Female	1
Region	All	REGION	Northeast	000
			Southeast	100
			Central	010
			West	001
Parental education	All	PARED	Less than high school	0000
			High school graduate	1000
			Post–high school	0100
			College graduate	0010
			Missing and I don't know	0001
Items in the home	All	B000901	None of the six items	00
		B000902	One of the six items	10
		B000903	Two of the six items	20
		B000904	Three of the six items	30
		B000905	Four of the six items	40
		B000906	Five of the six items	50
			Six of the six items	60
			Missing	01
Television watching	All	B001801	None	00
			One hour or less	10
			Two hours	20
			Three hours	30
			Four hours	40
			Five hours	50
			Six or more hours	60
			Missing	01
Homework	All	B001701	Don't have any	00
			Don't do any	00
			Less than 1 hour	10
			1–2 hours	20
			More than 2 hours	30
			Missing	01

See notes at end of table→

Table C–2. Conditioning variables for the NAEP long-term trend reading assessment: 1999—Continued

Conditioning Variable	Age classes	Variable name(s)	Variable coding	Contrast coding
Language spoken at home	All	B000401	English	00
			Spanish	10
			Other	10
			Missing	01
Language spoken in the home (other than English)	All	LANGHOM	Never	00
			Sometimes	10
			Always	01
			Missing	00
Pages read	All	B001101	More than 20	10
			16–20	10
			11–15	10
			6–10	10
			5 or fewer	00
			Missing	01
Percent in school lunch program	All	C032001	None	00000000
			1–5%	10000000
			6–10%	01000000
			11–25%	00100000
			26–50%	00010000
			51–75%	00001000
			76–90%	00000100
			over 90%	00000010
			Missing	00000001
Percent White	All	PCTWHTQ	0–49%	100
			50–79%	010
			80–100%	001
			Missing	000
Derived race/ethnicity	All	DRACE	White	000
			Black	100
			Hispanic	010
			Asian American	001
			American Indian	000
			Unclassified	000
			Missing	000
Age by grade	All	MODGRAG	< age, = grade	0000
			= age, < grade	1000
			= age, = grade	0100
			= age, > grade	0010
			> age, = grade	0001

See notes at end of table→

Table C–2. Conditioning variables for the NAEP long-term trend reading assessment: 1999—Continued

Conditioning Variable	Age classes	Variable name(s)	Variable coding	Contrast coding
School type	All	SCHTYPE	Public	1
			Private, catholic, bureau of indian affairs, department of defense	0
Type of location (94, 96 and 99 only)	All	TOL8	Big city	00000000
			Medium city	10000000
			Fringe of big city	01000000
			Fringe of medium city	00100000
			Large town	00010000
			Small place	00001000
			Rural – MSA	00000100
			Rural – non MSA	00000010
			Missing	00000001
Courses taken	9, 13	B001001	None	00
		B001002	1	10
		B001003	2	20
		B001004	3	30
		B001005	4	40
		B001006	5	50
		B001007	6	60
			7	70
			Missing	01

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table C–3. Conditioning variables for the NAEP long-term trend mathematics assessment: 1999

Conditioning variable	Age classes	Variable name(s)	Variable coding	Contrast coding
Overall	All		—	1
Gender	All	GENDER	Male	0
			Female	1
Observed race/ethnicity	All	ETHNIC	White	000
			Black	100
			Hispanic	010
			Asian American	001
			American Indian	000
			Other	000
			Missing	000
Items in the home	All	HOMEEN2	0–2 Items	00
			3 Items	10
			4 Items	01
			Missing	00
Region	All	REGION	Northeast	000
			Southeast	100
			Central	010
			West	001
Parents' education	All	PARED	Less than high school	0000
			High school graduate	1000
			Post–high school	0100
			College graduate	0010
			Missing and I Don't Know	0001
Modal grade	All	MODGRD	< modal grade	10
			= modal grade, missing	00
			> modal grade	01
Observed race/ethnicity by gender (White includes American Indian and other)	All	RACE x GENDER	White, male	000
			Black, male	000
			Hispanic, male	000
			Asian American, male	000
			White, female	000
			Black, female	100
			Hispanic, female	010
			Asian American, female	001

See notes at end of table→

Table C–3. Conditioning variables for the NAEP long-term trend mathematics assessment: 1999—Continued

Conditioning variable	Age classes	Variable name(s)	Variable coding	Contrast coding
Observed race/ethnicity by parents' education (White includes American Indian and other) coded differently for each age class	9	RACE x PARED	White, < HS	0000 0000 0000
			White, HS graduate	0000 0000 0000
			White, post–HS	0000 0000 0000
			White, college grad.	0000 0000 0000
			White, missing	0000 0000 0000
			Black, < HS	0000 0000 0000
			Black, HS grad & post–HS	1000 0000 0000
			Black, college grad.	0010 0000 0000
			Black, missing	0001 0000 0000
			Hispanic, < HS	0000 0000 0000
			Hispanic, HS grad & post–HS	0000 1000 0000
			Hispanic, coll. grad.	0000 0010 0000
			Hispanic, missing	0000 0001 0000
			Asian Amer., < HS	0000 0000 0000
			Asian Amer., HS grad & post–HS	0000 0000 1000
			Asian Amer., coll. grad.	0000 0000 0010
			Asian Amer., missing	0000 0000 0001
Observed race/ethnicity by parents' education (White includes Americans Indian and other) coded differently for each age class	13	RACE x PARED	White, < HS	0000 0000 0000
			White, HS graduate	0000 0000 0000
			White, post–HS	0000 0000 0000
			White, college grad.	0000 0000 0000
			White, missing	0000 0000 0000
			Black, < HS	0000 0000 0000
			Black, HS graduate	1000 0000 0000
			Black, post–HS	0100 0000 0000
			Black, college grad.	0010 0000 0000
			Black, missing	0001 0000 0000
			Hispanic, < HS	0000 0000 0000
			Hispanic, HS grad.	0000 1000 0000
			Hispanic, post–HS	0000 0100 0000
			Hispanic, coll. grad.	0000 0010 0000
			Hispanic, missing	0000 0001 0000
			Asian Amer., < HS	0000 0000 0000
			Asian Amer., HS grad.	0000 0000 1000
			Asian Amer., post–HS	0000 0000 0100
			Asian Amer., coll. grad.	0000 0000 0010
			Asian Amer., missing	0000 0000 0001

See notes at end of table→

Table C–3. Conditioning variables for the NAEP long-term trend mathematics assessment: 1999—Continued

Conditioning variable	Age classes	Variable name(s)	Variable coding	Contrast coding
Observed race/ethnicity by parents' education (White includes American Indian and other) coded differently for each age class	17	RACE x PARED	White, < HS	0000 0000 0000
			White, HS graduate	0000 0000 0000
			White, post–HS	0000 0000 0000
			White, college grad.	0000 0000 0000
			White, missing	0000 0000 0000
			Black, < HS	0000 0000 0000
			Black, HS graduate	1000 0000 0000
			Black, post–HS	0100 0000 0000
			Black, college grad.	0010 0000 0000
			Black, missing	0001 0000 0000
			Hispanic, < HS	0000 0000 0000
			Hispanic, HS grad.	0000 1000 0000
			Hispanic, post–HS	0000 0100 0000
			Hispanic, coll. grad.	0000 0010 0000
			Hispanic, missing	0000 0001 0000
			Asian Amer., < HS	0000 0000 0000
			Asian Amer., HS grad.	0000 0000 1000
			Asian Amer., post–HS, coll. grad.	0000 0000 0100
			Asian Amer., missing	0000 0000 0001
Language in the home	All	LANGHOM	Never	00
			Sometimes	10
			Always	01
Observed race by language at home		RACE x LANGHOM	White, often	00 00 00
			White, sometimes	00 00 00
			White, never	00 00 00
			Black, often and sometimes	10 00 00
	All		Black, often	10 00 00
			Black, sometimes	01 00 00
			Black, never	00 00 00
			Hispanic, often and sometimes	00 10 00
			Hispanic, often	00 10 00
			Hispanic, sometimes	00 01 00
			Hispanic, never	00 00 00
			Asian American, often and sometimes	00 00 10
			Asian American, often	00 00 10
			Asian American, sometimes	00 00 01
			Asian American, never	00 00 00
Derived race/ethnicity	All	DRACE	White	000
			Black	100
			Hispanic	010
			Asian American	001
			Other	000
			Missing	000

See notes at end of table→

Table C–3. Conditioning variables for the NAEP long-term trend mathematics assessment: 1999—Continued

Conditioning variable	Age classes	Variable name(s)	Variable coding	Contrast coding
Homework	13, 17	HW	None assigned	100
			Didn't do	010
			1/2 hour or less	012
			1 hour	013
			2 hours	014
			More than 2 hours	000
			Missing	000
Highest level of mathematics class	17	NMATH	Pre–algebra	10000
			Algebra	01000
			Geometry	00100
			Algebra 2	00010
			Calculus	00001
			Something else	00000
High school program	17	HS_PGM	General	00
			College preparatory	10
			Vocational/technical	01
			Missing	00
School type	All	SCHTY98	Public	0
			Religious	1
			Other private	1
			Catholic	1
			Bureau of Indian affairs	1
			Dept. Of defense	1
			Charter school	0
Type of location (94, 96 and 99 only)	All	TOL8	Big city	00000000
			Medium city	10000000
			Fringe of big city	01000000
			Fringe of medium city	00100000
			Large town	00010000
			Small place	00001000
			Rural – MSA	00000100
			Rural – non MSA	00000010
			Missing	00000001

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Table C–4. Conditioning variables for the NAEP long-term trend science assessment: 1999

Conditioning Variable	Age classes	Variable name(s)	Variable Coding	Contrast coding
Overall	All		—	1
Gender	All	DSEX	Male	0
			Female	1
Observed race	All	RACE	White	000
			Black	100
			Hispanic	010
			Asian American	001
			American Indian, pacific islander	000
			Other, blank, missing	000
Size and type of community (92 only)	All	STOC	Low metro	10
			High metro	01
			All others, missing	00
Type of location (94 96, and 99 only)	All	TOL8	Big city	00000000
			Medium city	10000000
			Fringe of big city	01000000
			Fringe of medium city	00100000
			Large town	00010000
			Small place	00001000
			Rural – MSA	00000100
			Rural – non MSA	00000010
			Missing	00000001
Region	All	REGION	Northeast	000
			Southeast	100
			Central	010
			West	001
			Missing	000
Parents' education	All	PARED	Less than high school	0000
			High school graduate	1000
			Post–high school	0100
			College graduate	0010
			Missing and "I don't know"	0001
Modal grade	All	MODGRD	< modal grade	10
			= modal grade	00
			> modal grade	01
			Missing	00

See notes at end of table→

Table C–4. Conditioning variables for the NAEP long-term trend science assessment: 1999—Continued

Conditioning variable	Age classes	Variable name(s)	Variable Coding	Contrast coding
Observed race by gender	All	RACE x DSEX	White, male	000
			Black, male	000
			Hispanic, male	000
			Asian American, male	000
			White, female	000
			Black, female	100
			Hispanic, female	010
			Asian American, female	001
			Other combinations, missing	000
Observed race by parents' education	All	RACE x PARED	White, < high school	000000000000
			White, = high school	000000000000
			White, > high school	000000000000
			White, graduated college	000000000000
			White, missing or unknown	000000000000
			Black, < high school	000000000000
			Black, = high school	100000000000
			Black, > high school	010000000000
			Black, graduated college	001000000000
			Black, missing or unknown	000100000000
			Hispanic, < high school	000000000000
			Hispanic, = high school	000010000000
			Hispanic, > high school	000001000000
			Hispanic, graduated college	000000100000
			Hispanic, missing or unknown	000000010000
			Asian American, < high school	000000000000
			Asian American, = high school	000000001000
			Asian American, > high school	000000000100
			Asian American, graduated college	000000000010
			Asian American, missing or unknown	000000000001
School type	All	SCHTYPE	Public	0
			Private, catholic, BIA, DoDEA	1
			Missing	0

See notes at end of table→

Table C–4. Conditioning variables for the NAEP long-term trend science assessment: 1999—Continued

Conditioning variable	Age classes	Variable name(s)	Variable Coding	Contrast coding
Derived race	All	DRACE	White	000
			Black	100
			Hispanic	010
			Asian American	001
			American Indian, pacific islander	000
			Other, missing	000
Observed race by language in the home	All	RACE x LANGHOM	White, always	000000
			White, sometimes	000000
			White, never	000000
			Black, always	100000
			Black, sometimes	010000
			Black, never	000000
			Hispanic, always	001000
			Hispanic, sometimes	000100
			Hispanic, never	000000
			Asian American, always	000010
			Asian American, sometimes	000001
			Asian american, never	000000
			One or both missing	000000
Homework	13, 17	B001701	None assigned	100
			Didn't do	010
			1/2 hour or less	012
			One hour	013
			Two hours	014
			More than two hours	000
			Missing	000
Highest level of science class	17	B005308	General science	1000
		B005309	Biology	0100
		B005310	Chemistry	0010
		B005311	Physics	0001
			Nothing, something else	0000
			Missing	0000
High school program	17	B005001	General	00
			College preparatory	10
			Vocational, technical	01
			Missing	00

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

Appendix D

NAEP 1999 Long-Term Trend [1]
Data Collection, Sampling and Weighting Report

Westat

Nancy W. Caldwell
Jean A. Fowler
Andrea R. Piesse
Mark M. Waksberg
Leslie S. Wallace

[1]This report was submitted to NCES by Westat, contractor for the sampling, administration, and weighting for the NAEP 1999 Long-Term Trend Assessment.

THIS PAGE INTENTIONALLY LEFT BLANK.

D.1. Data Collection Activities

D.1.1 Pre-Assessment Activities

During the fall period (mid-September through mid-December), a number of activities were conducted for the National Assessment of Educational Progress (NAEP) 1999 Long-Term Trend (LTT) Assessment. These included:

- Initiate telephone contacts to district superintendents and private school principals to gain their participation;
- Conduct introductory meetings with school principals to explain NAEP; and
- Conduct the fall assessments in about 290 schools beginning in early October.

D.1.2 Supervisor Training

The assessment supervisors attended a five-day training session in early September. Also in attendance were representatives from Educational Testing Service (ETS), National Computer Systems (NCS), and the National Center for Education Statistics (NCES). The training was conducted by the Westat project director and field director assisted by the field managers. ETS Princeton office staff also made presentations and provided explanatory notes throughout the session.

The topics that were covered included an overview of NAEP and the supervisors' responsibilities; a discussion of various reports from recent assessments; procedures for contacting districts and conducting introductory meetings; scheduling assessments, recruiting and training Exercise Administrators (EAs); procedures for drawing the sample of students, conducting assessments, preparation and distribution of questionnaires, administrative forms, and procedures. Also featured were practice exercises in sampling and filling out the various administrative forms.

In addition, a mock assessment session was held with the supervisors acting as "students." This included reading verbatim from one of the actual assessment scripts (to be used during an assessment); and following prescribed procedures for distributing materials, reading directions, and recording the results of the assessment.

D.1.3 Gaining Cooperation of Sampled Schools

The process of gaining cooperation of the schools selected for NAEP began in the summer with a series of letters and contacts with state and district level officials.

The National Center for Education Statistics (NCES) contacted the Chief State School Officers (CSSO) in each state notifying them of the districts and schools in their states that were in the sample. In August, Westat sent a set of recent NAEP reports, a letter, and listings of sampled schools to the district superintendents and heads of private schools inviting their participation.

Once the supervisors had been trained, they began working on obtaining cooperation. As the supervisors contacted superintendents and private school officials to establish cooperation and to set up the introductory meetings, they completed two forms. The *Introductory Meeting Form* was used to record the names of the school representatives expected to attend each meeting. A *Results of Contact Form* was completed documenting the discussion the supervisor had with each administrator concerning the district's willingness to participate and any special circumstances regarding the introductory meeting or assessments.

Copies of these forms were sent to the field manager and to the home office. Once received in the home office, the forms were used as the basis for mailing packages of materials to the persons scheduled to attend the meetings.

D.1.4 Introductory Meetings

During the period from late September through the middle of December, supervisors conducted introductory meetings with superintendents and principals of selected schools. The supervisors had a number of tasks to perform during the introductory meetings, including the following:

- Collecting and checking completed Principal Questionnaires.
- Presenting an overview of NAEP, using the slide presentation.
- Answering questions.
- Explaining the tasks that were required of each school.
- Setting preliminary sampling and assessment dates for each school.
- Verifying information on and completing the *School Control Form*.
- Distributing appropriate *Student Listing Forms* (SLFs) and explaining the method of completion.
- Identifying a School Coordinator (if not already identified).
- Inquiring about possible Exercise Administrator candidates.

In general, introductory meetings lasted about one hour. They ranged in size from small meetings between the supervisor and one school coordinator to formal meetings attended by 20-30 school officials (superintendents, curriculum specialists, testing personnel, principals, and coordinators). The introductory meetings often were the first opportunity for principals and other officials at the school level to discuss National Assessment with NAEP staff. Thus, the meetings were particularly important for establishing rapport with the schools, assuring school cooperation, and explaining the details of the schools' tasks to the individuals responsible for them.

D.1.5 Making Arrangements for the Assessments

During the introductory meetings, the supervisor discussed arrangements for the assessments with representatives from each school. Within the weeks scheduled for each primary sampling unit (PSU) (see section D.3), 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 the information from the schools, the supervisors set up the assessment schedule for the PSU.

The *School Control Form* was used by the supervisors to record information about the school's assessment plan. The form gave estimates of the number of students to be assessed in the school as well as the type of sessions to be held. Using this information, the supervisor and school staff could discuss the approximate number of sessions to be held in the school and the space required.

The supervisor usually learned during the introductory meeting whether a school required some form of parental notification or permission. In preparation for this, the supervisor had copies of three versions of standard NAEP letters to parents. These letters were made available to schools requesting them. The first version informed parents about the assessment. The second version assumed parental consent unless parents sent the form back stating they did not want their child to participate in the assessment. The third version required that parents sign and return the form before students could

be assessed. Schools were offered their choice of the letters, although when the issue of parental permission came up in discussions, supervisors offered the least restrictive version first. Of course, schools could send out their own letters and notices if they preferred not to use the ones prepared by NAEP.

D.1.6 Recruiting, Hiring, and Training Exercise Administrators

During the fall, while the supervisors were conducting introductory meetings and scheduling assessments, they also were to recruit and hire Exercise Administrators (EAs). The EA's primary job was to administer the assessment sessions. EAs were recruited from many sources. Each supervisor was given a PSU-by-PSU computerized list of interviewers and EAs who had worked for Westat on NAEP and other studies. During introductory meetings, the supervisors asked the school principals and other staff to recommend potential EAs. Where necessary, ads were placed in local newspapers and the job service was notified.

Supervisors were told that, in general, two EAs 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 particular times of the year were all factors taken into consideration by supervisors in developing their plan for EAs.

The assessment supervisors had complete responsibility for recruiting, hiring, training, and supervising their EAs. The supervisors' first task upon arriving in a PSU for the assessments was to train the EAs. The *Supervisor's Manual* discussed the training and use of EAs in conducting assessments. In addition, one session of the supervisors' training included a discussion for EA training and a thorough review of the *Exercise Administrator's Manual*. The supervisors gave a copy of the EA's manual to each EA before the training session was held.

Exercise Administrators were required to study the manual before being trained and then to attend a half-day training session conducted by the supervisor. During the training, the supervisor reviewed, in detail, all aspects of the EA's job including preparing materials, booklets, and Administration Schedules for assessments; the actual conduct of the session; post-assessment collection of booklets, pencils, and other assessment materials; coding booklet covers; record keeping; and administrative matters.

D.2 Assessment Activities

D.2.1 Overview

To provide continuity and comparability with the past, the long-term trend assessments replicated what had been done in prior years. Tape sessions were conducted with samples of age-eligible students, as had been done in all previous years. Additional samples of age- and grade-eligible students were assessed with spiral (self-administered) booklets, following procedures initiated in NAEP 1984. The three age/grade groups were assessed during the same time periods as in the past: 13-year old/8th graders were assessed during October to December; 9-year old/4th graders during January to mid-March and 17-year-old/11th graders were assessed from mid-March to early May.

D.2.2 Selecting the Student Sample

Two weeks prior to a school's assessment date, the assessment supervisor contacted the School Coordinator to make sure that the lists of eligible students were prepared and that all arrangements were set as agreed. The supervisor then visited the school (or district office) a few days to a week or more before the assessment date to select the sample of students. The time interval between the selection of the sample and the assessment varied depending on several factors, most notably the size of the school. The average elapsed time was about a week.

The supervisor's first task upon arriving at the school to select the sample was to review the *Student Listing Forms* or comparable list of students in an effort to be sure that they had been completed correctly. The supervisor made certain checks to help assure that all age- and grade-eligible students had been listed. The supervisor also checked that the students to be excluded from the assessment were listed so that they could be included in the sample.

For each school, the Westat home office produced a *Session Assignment Form* (SAF) that told the supervisor how to select the sample in that school.

Following the sampling instructions, the supervisor was instructed to fill out administration schedules for each session listing the sampled students. Before listing the students on the administration schedules, the supervisor reviewed the plans for the assessment with the school coordinator. If, for example, a large number of students were sampled for a spiral session, the supervisor discussed our preference for this group to be divided into sessions of about 30 each. Also discussed were procedures that might be helpful to the school such as listing students on the administration schedules alphabetically or by homeroom. Sometimes the coordinator had very specific ideas about the organization of the assessment.

After the excluded students were identified, the supervisors were instructed to prepare and distribute the *Excluded Student Questionnaires*. If the 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 an estimated number of questionnaires needed.

D.2.3 Conduct of the Assessment

The primary responsibility for conducting assessment sessions was with the EAs. Supervisors were required to observe the first session an EA conducted to ensure that he/she followed the procedures properly. Supervisors were also required to be present in all schools during the assessments if at all possible, especially in large schools with several sessions. Previous experience has shown that the supervisor can play an important role acting as the liaison between the National Assessment and

school staff and ensuring that the assessments go smoothly. If, for example, the supervisor is present, he/she can help direct students to the correct rooms when more than one session is being conducted at the same time.

To ensure that sessions were administered in a uniform way, the EA was provided with scripts for each session type from which he/she was to read verbatim. The scripts began with a brief introduction to the study followed by directions to the EA to distribute the booklets, being careful to give each student the correct preassigned booklet.

Following the distribution of booklets, the scripts differed depending on whether the session was a spiral or tape session. In spiral sessions, the EA read from the script and followed its directions as he/she continued the session administration and timed the sections of the booklets. In tape sessions, the EA was instructed to turn the tape recorder on after distributing the booklets and the tape did most of the administration and timing of the sections.

During the sessions, the EAs walked around the room monitoring the students, being sure that they were working in the correct section of the booklet and discouraging them from looking at a neighbor's booklet. During the background (first) section, EAs were allowed to assist students in understanding questions and responding to them. After the students began working on the other sections of the booklets, the EA was not allowed to answer any students' questions.

At the end of an assessment session, booklets were collected and the students dismissed according the school's policy. The EA was then responsible for entering information about the results of the assessment on the booklets. EAs then packed completed materials and paperwork and sent it to NCS for scoring.

D.2.4 Results of the Assessment

Information regarding the numbers of schools and students that participated in the NAEP 1999 long-term trend assessment are provided in section D.3. As in the past, response rates were highest at the elementary grades.

D.3. Sample Design

This section describes sampling activities for the NAEP 1999 Long-Term Trend Assessment. Section D.3.1 provides an overview of the sample design; section D.3.2 summarizes the selection of primary sampling units (PSUs) and schools within PSUs; and section D.3.3 discusses allocating sessions to schools, and section D.3.4 discusses sampling students within schools.

D.3.1 Overview of the Sample Design

The sample for the NAEP 1999 long-term trend was a multistage probability sample. Counties or groups of counties were the first-stage sampling units, and elementary and secondary schools were second-stage units. Assignment of sessions by type to selected schools was the third sampling stage. The fourth stage was selection of students within schools and their assignment to session types.

Fifty-two primary sampling units (PSUs) were included in the 1999 long-term trend sample. A school sample was drawn for each of three age classes, where age class refers to student eligibility: age 9 or in grade 4; age 13 or in grade 8; age 17 or in grade 11. Because of these student eligibility requirements, schools having any of several grades were eligible for selection. The number of schools participating for each age class 9, 13, and 17 was 258, 238, and 194, respectively. According to a partial balanced incomplete block design used for previous long-term trend assessments, exercises in reading, writing, mathematics, and science were administered in these schools to 11,825 age class 9 students, to 11,874 age class 13 students, and to 9,038 age class 17 students. Assessments were divided into three time periods, with age class 13 assessments conducted in the fall, age class 9 in the winter; and age class 17 in the spring. Target sample sizes, eligibility criteria, and assessment periods are shown in table D-1.

The school base weight, i.e., the reciprocal of the probability that a school was selected for a particular age-class sample, was calculated and adjusted for nonresponse in the same manner as for previous long-term trend samples.[2] Because of the increasing rate of refusal among participating schools to assess age-eligible students not in one of grades 4, 8, or 11, an additional nonresponse adjustment was calculated in 1999. This adjustment was incorporated into the student base weight, the reciprocal of the overall probability that a student was invited to a particular type of session. The student base weight was then adjusted for nonresponse and further adjusted by a post-stratification procedure as in previous years. School and student participation rates are discussed in sections D.3.2.2.4 and D.3.4.5.

D.3.1.1 Target Population and Sample Size

The target population for the NAEP 1999 long-term trend assessment was the same as for previous assessments. Target sample sizes were increased slightly for 1999, based on examination of 1994 and 1996 yields. These targets were intended to yield approximately 11,200 assessed students in age classes 9 and 13, and 9,200 in age class 17.

[2]See *Sample Design* (Wallace and Rust, 1999).

Table D-1. NAEP long-term trend target sample sizes, eligibility criteria, and assessment periods: 1999

Sample	Sample target size	Eligibility criteria	Assessment period
Age class 9	14,600	born 1/89–12/89 or in grade 4	Winter 1998-1999
Age class 13	15,200	born 1/85–12/85 or in grade 8	Fall 1998
Age class 17	14,000	born 10/81–9/82 or in grade 11	Spring 1999

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.3.2 The Sample of Primary Sampling Units and Schools

The sample for the NAEP 1999 long-term trend assessment was selected using a complex multistage sample design involving the sampling of students from selected schools within 52 selected geographic areas, called primary sampling units (PSUs), across the United States. The sample design included a four-step selection process:

1. Selection of geographic PSUs (counties or groups of counties),
2. Selection of schools within PSUs,
3. Assignment of session types and sample types to schools, and
4. Selection of students for session types within schools.

D.3.2.1 Definition and Selection of Primary Sampling Units

PSU samples for NAEP 1999 are stratified probability samples. PSUs were selected with probability proportional to the population for the long-term trend assessments. A PSU consists of a Consolidated Metropolitan Statistical Area (CMSA), a Metropolitan Statistical Area (MSA), a New England County Metropolitan Area (NECMA), a county, or group of contiguous counties in the U.S.

Construction of the NAEP PSU sampling frame and selection of PSUs are described in chapter 3 of *The NAEP 1998 Technical Report* (Rust, Krenske, Qian, and Johnson, 2001). For the long-term trend assessments, 52 PSUs were drawn for each sample, selecting with certainty only the 10 largest of the 22 main sample certainty PSUs (specified for the 1998 main assessment). Six additional PSUs were selected with probability proportional to population from the 12 remaining main sample certainties. The 72 main sample noncertainty strata (specified for the 1998 main assessment) were then paired and one PSU per pair was selected for the 1999 long-term trend samples. Overlap was minimized from one assessment to the next.

D.3.2.2 School Sample

D.3.2.2.1 Frame Construction

The second stage of sampling was to select schools within each selected PSU. The school sampling frame was a list of schools developed from two sources. Public, BIA, and DoDEA schools were obtained from the 1996 list of schools maintained by Quality Education Data, Inc. (QED), which included information from the 1994-95 NCES Common Core of Data (CCD). 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 the Private School Survey (PSS) developed for the National Center for Education Statistics' 1995-1996 Schools and Staffing Survey. The majority of the PSS list comes from complete enumeration of schools, but a small portion of the PSS list was obtained from a sample of counties selected for the PSS. A weight component was computed from this PSS list for main 1998 sample schools; similar to previous long-term trend assessments, this weight component was used for the 1999 long-term trend sample.

The population of eligible schools for each age class was restricted to the 52 selected PSUs. Because students' ages vary within each grade level, schools having any of several grades were eligible at each age class. As required, the following practice replicates that of previous long-term trend assessments:

Sample:	**Grades defining school eligibility:**
Age class 9	grades 2 to 5
Age class 13	grades 6 to 9
Age class 16	grades 9 to 12

Any school having one or more of the eligible grades was included in the sampling frame. Schools were included in the frame for a particular age class without regard to eligibility for either of the other two age classes. As a result there was considerable overlap among the three frames. An independent sample was selected for each age class. Thus, some schools were selected for assessment of more than one age class.

D.3.2.2.2 Assigning Size Measures and Selecting School Samples

For each age class schools were selected without replacement across all PSUs, with probabilities proportional to measures of size. The measure of size assigned to each school was based on the estimated number of age-eligible students. To increase cost efficiency, lower measures of size, and thus lower probabilities of selection, were assigned to schools having fewer than 20 estimated age-eligible students.

Let
Ai = The estimated number of age-eligible students in school i;
Gi = The estimated number of grade-eligible students in school i.

The maximum sample size in terms of age-eligibles was 60; the maximum sample size of all eligible students, i.e.: age-eligible, grade-eligible or both, was (Gi /Ai)*60.

The measure of size was:

.25, if Ai < 6
Ai/20, if 6 <= Ai <= 19
1, if 20 <= Ai <= 60
Ai/60, if Ai > 60

The measure of size was based on the estimated number of age-eligibles because in the large majority of schools selected, students will be assigned to at least one session for which eligibility is determined by age only. This was the case for approximately 95 percent of the 1999 long-term trend sample schools. In most schools having the target grade, some additional students will be selected who are in the target grade but are not age-eligible. Among schools participating in the 1999 long-term trend assessment, the maximum sample size of all eligible students was almost always less than 90. In 11 schools having much smaller than expected proportions of age eligible students, sample sizes were greater than 90; 4 schools had sample sizes greater than 100.

The total number of schools selected for each age class was such that the predesignated student sample sizes would be achieved by selecting the maximum sample size of all eligible students in each school. 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 exception of the concession to cost mentioned above, had the goal of yielding a sample of students in a given age class or grade with approximately uniform probabilities of selection. The distributions of selection probabilities of the selected students, as reflected by their sampling weights, are shown in section D.5.

D.3.2.2.3 Identifying Substitute Schools

Potential substitute schools were identified for all sample schools in the 1999 long-term trend assessment when a close match could be made on several attributes. Substitute candidates were those schools in the frame not already selected for any 1999 long-term trend sample. An attempt was made to select, before field processes began, a maximum of two substitute schools for each sampled public school (one in-district and one out-of-district) and each sampled Catholic school, and one for each sampled BIA, DoDEA, state, or non-Catholic private school. Within a given age class, a sample school was replaced by a substitute when it was determined to be a final refusal for that age class. To minimize bias, a substitute school resembled the original selection as closely as possible.

Substitutes were assigned by matching on the following attributes:

- Affiliation;
- Grade span;
- Estimated grade enrollment; and
- Minority composition.

A substitute was always selected from the same PSU as the refusing school. Out-of-district substitutes were pre-identified so that replacements would be available in cases of school non-participation due to district refusal. When nonparticipation was due to principal refusal, however, preference was given to the pre-identified in-district substitute. The identity of the substitute school was unknown to the field staff until after the corresponding original selection was designated as a final refusal. This was to protect against any temptation to move on to an "easier" substitute school.

The net numbers of substitutes added to the sample by the above procedure are shown in table D-2. The number of substitutes participating in the 1999 long-term trend assessment was substantially higher than in 1996, probably due to the increased efficiency of pre-selection; more refusing schools had substitutes identified.

D.3.2.2.4 School Participation

Overall, the school participation rate was lower for the 1999 assessment than for the 1996 and 1994 assessments. Most of this decline can be accounted for by decreased participation of originally selected schools in the age class 17 sample. Table D-2 shows the numbers of in-scope schools selected, cooperating, and replaced by substitutes; participation rates are based on the original sample of schools, excluding substitutes.

Note that there was a considerable number of schools in the age class 13 sample that had no eligible students enrolled. The grade structure of the age class 13 sample was such that a school could have one of grades 6, 7, or 9, but no grade 8. There was a reasonable chance that some age 13 students would be enrolled; this was often the case, but sometimes there were none.

Table D-2. School sample sizes, refusals, and substitutes for the NAEP long-term trend samples: 1999

Status	Age class 9	Age class 13	Age class 17	Total
Selected, in scope	286	299	243	828
Refusals	43	54	58	155
Participation rate of originally selected schools	85%	82%	76%	81%
1996 participation rate	85%	84%	81%	84%
Participating, no eligible students enrolled	2	18	3	23
Substitutes participating	17	11	12	40
Final assessed sample	258	238	194	690

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.3.3 Assignment of Sessions to Schools

D.3.3.1 Initial Session Assignments

There were six session types, identical to those conducted in previous long-term trend assessments: reading and writing (spiral) sessions and five types of mathematics and science (tape) sessions. Sessions conducted for each age class are listed in table D-3.

Sessions were allocated among the sampled schools according to estimated age-eligible enrollment:

Estimated number of age-eligible students:	Number of sessions allocated:
1-20	1
21-40	2
41 or more	3

Sorting the list of selected schools in sampling order and randomly choosing the first session, session types were assigned according to the following sequence:

Age class 9, 13: T1, SP, T2, SP, T3, T2, SP, T3, SP, T1, T3, SP, T1, SP, T2
Age class 17: T4, SP, T5, SP,

where T1, T2, T3, T4, and T5 represent the five tape session types and SP represents spiral sessions.

Thus the sessions assigned were about 60 percent tape and 40 percent spiral for age classes 9 and 13; for age class 17 about half were tape, half spiral. The approximate distributions of session type combinations by number of sessions assigned is included in table D-3:

Table D-3. Distributions of session type combination by number of sessions assigned: 1999

Estimated age-eligible enrollment	Number of sessions allocated	Session type combination	Distribution by age class 9 and 13	17
41 or more	3	2 spiral, 1 tape	20%	50%
	3	1 spiral, 2 tape	80%	50%
21-40	2	1 spiral, 1 tape	80%	100%
	2	2 tape	20%	0%
1-20	1	1 spiral	40%	50%
	1	1 tape	60%	50%

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.3.3.2 Revised Session Assignments

Up-to-date student enrollments were obtained for sampled schools in the field. Given its initial session allocation, if a school's current age-eligible enrollment was within a specified interval, the initial session allocation and session type assignment were revised. Field staff used laptop computers to accomplish this task for the NAEP 1999 Long-Term Trend Assessment.

For reasons of cost and operational efficiency, one or two sessions were dropped in schools whose updated age-eligible enrollment was smaller than expected; no sessions were added in schools whose expected age-eligible enrollment was larger than expected. Criteria for dropping sessions were based on the number of sessions initially allocated and the updated age-eligible enrollment and are outlined in table D-4.

Table D-4. NAEP criteria for dropping sessions: 1999

Number of sessions initially allocated	Updated number of age-eligible students	Number of sessions to drop
3	35 or more	0
3	17-34	1
3	1-16	2
2	17 or more	0
2	1-16	1
1	1 or more	0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Given the initial number and type of sessions assigned, sessions were randomly dropped with probabilities that preserved the approximate distributions of session type combinations shown above.

D.3.4 Student Sample

The sample of students within selected schools was drawn by systematic sampling from school-prepared lists of eligible students. *Student Listing Forms* (SLF) were prepared for each participating school; all grade-eligible and age-eligible students were to be entered on the SLFs. Field staff obtained current enrollment figures when scheduling assessment dates; they updated the estimated enrollments and used laptop computers to calculate sampling rates and assign students to sessions.

D.3.4.1 Within-School Sampling Rates

Let

N_i= number of students on SLF in school i;
A_i= age-eligible enrollment in school i, updated using the SLF

The student sample size within school i was:

S_i= N_i, if A_i<=60;
$(N_i/A_i)*60$, otherwise.

The sampling rate applied to the list of eligible students was then:

R_i= 1, if A_i<=60;
$A_i/60$, otherwise.

Students were assigned systematically to sessions in proportion to the number and types of sessions allocated to the school. Note that only the age-eligible students assigned to tape sessions were included in the sample; those who were in the target grade but not age-eligible were dropped. Since the sample size was defined in terms of age-eligibles, the actual sample size of all eligible students depended on the type of sessions assigned and the proportion of age-eligible students. In all but 15 of the 886 participating schools, the number of students selected was less than 90.

D.3.4.2 The Session Assignment Form (SAF)

To control the student sampling operations as closely as possible, Westat generated a *Session Assignment Form (SAF)* for each school where sampling was to be carried out. This computer-generated form listed:

- Updated enrollments of age-eligibles and all (grade/age) eligibles;
- The revised session assignment;
- The line numbers (from the SLF) specifying the students to be selected for spiral and for tape sessions; and
- Instructions for the sampling process.

D.3.4.3 Sample Selection

District supervisors implemented student sampling procedures in the field, usually a week before the assessment, and *Student Listing Forms* (SLFs) were prepared for each participating school. All students in the target grade and all age-eligible students in other grades were entered on the SLF, or the school produced a computer-generated list. Before carrying out the sampling, the district supervisor reviewed the form and made comparisons with other information in an effort to make sure that the list included all eligible students.

The sampling was carried out according to specific instructions described in the supervisor's manual. Sampling statisticians and systems analysts were available by phone and email to help resolve sampling problems.

Briefly, student sampling procedures involved the following:

- Numbering sequentially the students listed on the SLF or computer-generated list.
- Selecting students from the SLF whose line numbers corresponded to the line numbers generated for each session type on the SAF. Two sets of line numbers were generated on the SAF, one for spiral sessions and one for tape sessions. Line numbers of students who were in the target grade but were not age-eligible were eliminated from the set of tape session line numbers. If more than one type of tape session was assigned, students selected for tape sessions were then systematically selected for the different types.
- Identifying excluded students and preparing an *Excluded Student Questionnaire* for each excluded student.

After student sampling was completed, the updated enrollment figures, revised session assignments, and other sampling data stored in the laptop computers were transmitted to Westat's central office for use in sample weighting.

Table D-5 shows the number of students assessed for each session type and the number of students per school for each session type for the three age classes.

Table D-5. Number of students assessed and number of students per school for each session type: 1999

Sample	Session type	Number of assessed students	Number of schools	Mean number of students per assessment per school
Age class 9	Print booklets 51-56	5,793	234	24.8
	Tape booklet 91	2,032	133	15.3
	Tape booklet 92	1,865	125	14.9
	Tape booklet 93	2,135	135	15.8
Age class 13	Print booklets 51-56	5,933	217	27.3
	Tape booklet 91	2,019	125	16.2
	Tape booklet 92	1,960	123	15.9
	Tape booklet 93	1,962	121	16.2
Age class 17	Print booklets 51-56	5,288	187	28.3
	Tape booklet 84	1,953	134	14.6
	Tape booklet 85	1,842	131	14.1

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.3.4.4 Excluded Students

Using the same criteria as in long-term trend assessments dating to the early 1980s, a distinct sample of excluded students was identified for each age class. Operationally, students were first assigned to sessions and then the excluded were identified. Thus, the only excluded students who were not age-eligible had been selected for spiral sessions. Students whose SLF line numbers were selected for tape sessions and who were not age-eligible were dropped; any among them who would have met the criteria for exclusion were not identified. Since the exclusion criteria were not session-specific, the excluded student sample was weighted to account for this procedure (see section D.5).

Table D-6 shows weighted exclusion rates for each age class by session type and school type, calculated using the student base weights as in previous long-term trend assessments. These weights reflect the number of age-eligible excluded students selected from the SLF, but not the numbers in the entire age class cohort. Table D-7 shows the weighted exclusion rates calculated while accounting for assignment of age-eligible excluded students to sessions. As in previous assessments, exclusion rates were generally higher in the lower grades, and much higher in public schools than in private schools.

Table D-6. NAEP long-term trend student exclusion rates by age class and school type and subject, weighted (calculated as in previous assessments): 1999

Subject	Age class 9			Age class 13			Age class 17		
	Public	Non-public	Total	Public	Non-public	Total	Public	Non-public	Total
Reading/writing print	7.0	0.4	6.3	6.0	0.3	5.3	4.7	0.6	4.4
Mathematics/science tape	6.4	0.5	5.8	4.5	0.2	4.0	2.9	0.8	2.7

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-7. NAEP long-term trend student exclusion rates by age class and school type and subject, weighted (calculated to account for assignment of age-eligible excluded students to sessions): 1999

Subject	Age class 9			Age class 13			Age class 17		
	Public	Non-public	Total	Public	Non-public	Total	Public	Non-public	Total
Reading/writing print	10.9	0.9	9.9	8.9	0.5	8.0	6.6	1.1	6.1
Mathematics/science tape	10.1	0.9	9.1	7.4	0.4	6.7	5.5	1.3	5.2

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.3.4.5 Student Participation Rates

The NAEP long-term trend sample was designed to produce target yields for the reading/writing (spiral) and mathematics/science (tape) assessment components. Tape session yields for the two previous long-term trend assessments were short (4.7 to 11.5 percent) of the target

numbers in all but the 1994 age 13 sample; age 17 spiral session yields were 6.9 and 10.2 percent low in 1994 and 1996, respectively. Based on these results, and taking into account the response rates for these assessments, target numbers were increased for the 1999 long-term trend assessment. Table D-8 compares target numbers to actual assessments for the three age classes. Tape session targets were quite closely met in the age 9 and age 13 samples; age 17 tape samples were 5.1 percent below the target. The spiral session target was closely met in the age 17 sample, but considerably exceeded in the age 9 and age 13 samples. Achieving sampling goals precisely is dependent on many factors, including the reliability of frame enrollment data, and the actual response and exclusion rates encountered. Additional complicating factors for long-term trend assessments are the proportions of age-eligibles in participating schools, and the increasing refusal among participating schools to assess age-eligible students who are not in the modal grades.

Table D-8. NAEP long-term trend target yields and number assessed by age class: 1999

Sample	Target yield	Number assessed
Age class 9		
Spiral	5,200	5,793
Tape	6,000	6,032
Age class 13		
Spiral	5,200	5,933
Tape	6,000	5,491
Age class 17		
Spiral	5,200	5,288
Tape	4,000	3,795

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-9 shows the unweighted student participation rates of invited students. Invited students are the set of selected students, after removing the excluded students and those selected for tape sessions who are 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 were absent from their originally scheduled session. 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 public and nonpublic schools separately. Overall participation rates are also shown for comparable samples from the 1996 long-term trend assessment. Student participation rates have remained fairly steady since 1994 in the age 9 and age 13 long-term trend samples; they dropped 3.6 percent in public schools in the 1999 age 17 sample. The participation rate of nonpublic-school students continued to exceed that of public-school students in 1999 for all age classes, with the difference, both relative and absolute, increasing with age class.

Table D-9. Student participation rates by age class and school type, unweighted: 1999

	1999 Public		1999 Nonpublic		1999 Combined		1996
Sample	Number invited	Participation rate	Number invited	Participation rate	Number invited	Participation rate	Participation rate
Age class 9	11,276	94.0%	1282	95.6%	12,558	94.2%	95.5%
Age class 13	11,534	91.8%	1338	95.8%	12,872	92.2%	91.9%
Age class 17	10,347	79.8%	901	91.5%	11,248	80.8%	84.0%

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The combined impact of school nonparticipation and student absenteeism from sessions within participating schools is summarized in table D-10. The table shows the unweighted 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. Consistent with previous long-term trend assessments, the overall level of participation decreases as age class increases. Overall participation rates in all age classes were lower in 1999 than in 1996, considerably lower in the age 17 sample.

Table D-10. Overall participation rates (school and student combined) by age class, unweighted: 1999

1999 long-term trend samples	Age class 9	Age class 13	Age class 17	Overall
School participation	85.0%	81.9%	76.1%	81.3%
Student participation	94.2%	92.2%	80.8%	89.4%
Overall student participation	80.0%	75.6%	61.5%	72.6%
Number of participating students	11,825	11,874	9,083	32,782

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-11 shows weighted participation rates by age class and session type. Within each age class, the weighted rates for spiral sessions are similar to those for tape sessions. They are also similar to the unweighted rates.

Table D-11. Weighted participation rates by age class and session type, long-term trend samples: 1999

Participation	Reading/writing print	Mathematics/science tape
Age class 9		
School participation	84.9%	83.5%
Student participation	94.4%	93.7%
Overall participation	80.2%	78.3%
Age class 13		
School participation	80.8%	79.3%
Student participation	92.1%	92.5%
Overall participation	74.4%	73.4%
Age class 17		
School participation	74.0%	72.1%
Student participation	80.2%	81.3%
Overall participation	59.4%	58.6%

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.4 Age 17 Nonresponse Bias Analysis

D.4.1 Introduction

Response rates at the school and student level for ages 9 and 13 were considered acceptable, however response rates for the age 17 group were low enough to warrant an investigation into possible nonresponse bias. The methodology and results of that investigation follow.

D.4.2 Methodology

Nonresponse bias was analyzed at both the school and student level. Although substitutes were used for nonresponding schools, the school level analysis presented here is based on the original sample of schools.

For both schools and students, nonresponse is considered separately for the reading (or spiral) assessments, and the mathematics/science (or tape) assessments. Note that the writing assessments are not considered here.

In order to compare respondents and nonrespondents it is necessary to use frame characteristics that are available for both groups. Comparing frame characteristics is not always a good measure of nonresponse bias if the characteristics are unrelated or weakly related to more substantive items in the survey, however this is often the only approach available. For categorical variables, response rates by characteristic were calculated. The hypothesis of independence between the characteristic and response status was tested using a Rao-Scott modified Chi-square statistic. For continuous variables, summary means were calculated. The 95% confidence interval for the difference between the mean for respondents and the mean for nonrespondents was tested to see whether or not it included zero. In addition to these tests, logistic regression models were set up to identify whether any of the frame characteristics were significant in predicting response status. All analyses were performed using WesVar and replicate weights to properly account for the complex sample design. The base weights used did not include nonresponse adjustment factors at either the school or student level. Note that for the school level analysis, the weights used included a measure of the size of school, namely the number of age eligible students. The paired jackknife variance replication method was used, as with all other NAEP analyses.

D.4.3 Results

D.4.3.1 School Level Analysis - Reading

The following nonresponse bias analysis is based on the original sample of 236 schools selected for reading assessment. All schools that were substituted by a replacement were treated as nonrespondents, as were any nonresponding original schools that were not substituted. The unweighted response rate was 76.27%, with 180 out of 236 schools responding. The weighted response rate was 73.18%. Standard errors are given throughout in parentheses.

D.4.3.1.1 Categorical Variables

The following characteristics were available for analysis:

- Metropolitan area
- NAEP region
- Supervisor region
- Community type
- School type

- Number of sessions
- Number of reading (spiral) sessions

Table D-12 shows school response rates by metropolitan area status. The test of independence gives RS3 = 3.23, with a p-value of 0.072. There is some evidence that non-metropolitan schools were more likely to respond than others, though it is not significant at the 5% level.

Table D-12. School reading response rate by metropolitan area, weighted: 1999

Area	Response rate	
Non-Metropolitan Area	84.49%	(5.941%)
Metropolitan Area	70.11%	(5.368%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-13 shows school response rates by NAEP region. The test of independence gives RS3 = 2.47, with a p-value of 0.466. This indicates that there is no significant relationship between response status and NAEP region at the 5% level.

Table D-13. School reading response rate by NAEP region, weighted: 1999

NAEP region	Response rate	
Northeast	67.52%	(8.359%)
Southeast	84.71%	(8.227%)
Central	72.30%	(8.490%)
West	68.79%	(8.651%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-14 shows school response rates by NAEP supervisor region. The test of independence gives RS3 = 5.24, with a p-value of 0.391. This must be interpreted with caution due to the presence of a cell with less than five observations, however it would suggest that there is no significant relationship between response status and supervisor region at the 5% level.

Table D-14. School reading response rate by NAEP supervisor region, weighted: 1999

Supervisor region	Response rate	
1	69.75%	(10.692%)
2	61.00%	(11.480%)
3	75.30%	(13.087%)
4	94.49%	(6.174%)
5	95.10%	(3.161%)
6	57.40%	(15.513%)
7	80.71%	(7.823%)
8	67.33%	(9.860%)
9	66.73%	(6.341%)
10	67.41%	(25.153%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-15 shows school response rates by community type. The test of independence gives RS3 = 3.78, with a p-value of 0.146. This indicates that there is no significant relationship between response status and community type at the 5% level.

Table D-15. School reading response rate by community type, weighted: 1999

Community type	Response rate	
Central city	73.98%	(8.041%)
Urban fringe or large town	66.26%	(6.288%)
Rural or small town	83.85%	(6.035%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-16 shows school response rates by school type. The test of independence gives RS3 = 2.82, with a p-value of 0.391. This must be interpreted with caution due to the presence of a cell with less than five observations, however it would suggest that there is no significant relationship between response status and school type at the 5% level.

Table D-16. School reading response rate by school type, weighted: 1999

School type	Response rate	
Catholic	79.22%	(13.909%)
Other religious	41.39%	(17.521%)
Other private	69.26%	(19.142%)
Public	73.85%	(4.474%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-17 shows school response rates by the total number of sessions the school was asked to conduct. The test of independence gives RS3 = 0.05, with a p-value of 0.973. This must be interpreted with caution due to the presence of a cell with less than five observations, however it would suggest that there is no significant relationship between response status and number of sessions at the 5% level.

Table D-17. School reading response rate by number of sessions, weighted: 1999

Number of sessions	Response rate	
1 session	77.21%	(24.766%)
2 sessions	73.81%	(15.178%)
3 sessions	73.10%	(4.473%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-18 shows school response rates by the number of reading (spiral) sessions the school was asked to conduct. The test of independence gives RS3 = 3.09, with a p-value of 0.079. There is some evidence that schools asked to conduct two reading sessions were more likely to respond than others, though it is not significant at the 5% level.

Table D-18. School reading response rate by number of reading sessions, weighted: 1999

Number of sessions	Response rate	
1 reading session	68.53%	(5.376%)
2 reading sessions	77.88%	(4.886%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.4.3.1.2 Continuous Variables

The following characteristics were available for analysis.

- Number of age eligible students
- Percent Asian or Pacific Islander students
- Percent Black, non-Hispanic students
- Percent Hispanic students
- Percent American Indian or Alaskan Native students
- Percent White, non-Hispanic students

Table D-19 shows the mean number of age eligible students for responding and nonresponding schools. The difference in the mean number of age eligible students is –54.9, with a 95% confidence interval of (-116.4, 6.7). The confidence interval just includes zero. Therefore there is some evidence that the mean number of age eligible students is lower for responding schools, though it is not significant at the 5% level.

Table D-19. Mean number of age eligible students by school reading response status, weighted: 1999

	Responding		Nonresponding	
Number of age eligible students	280.1	(18.97)	335.0	(20.44)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-20 shows the mean race/ethnicity percentages for responding schools and nonresponding schools.

Table D-20. Mean race/ethnicity percentages by school reading response status, weighted: 1999

Race/ethnicity	Responding		Nonresponding	
Percent of Asian or Pacific Islander students	3.78%	(0.577%)	4.14%	(1.033%)
Percent of Black, Non-Hispanic students	18.38%	(1.533%)	13.45%	(2.862%)
Percent of Hispanic students	7.22%	(0.893%)	8.74%	(2.487%)
Percent of American Indian or Alaskan Native students	0.63%	(0.274%)	0.71%	(0.260%)
Percent of White, Non-Hispanic students	70.00%	(1.592%)	72.96%	(3.353%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The difference in the mean percentage of Asian or Pacific Islander students is –0.36%, with a 95% confidence interval of (-2.93%, 2.21%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of Black, non-Hispanic students is 4.93%, with a 95% confidence interval of (-1.20%, 11.06%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of Hispanic students is –1.53%, with a 95% confidence interval of (-5.71%, 2.65%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of American Indian or Alaskan Native students is –0.08%, with a 95% confidence interval of (-0.78%, 0.62%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of White, non-Hispanic students is –2.96%, with a 95% confidence interval of (-9.99%, 4.07%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

D.4.3.1.3 Logistic Regression Model

A logistic regression model was set up treating response status as the binary dependent variable, and frame characteristics as the predictor variables. Response was treated as "success" and nonresponse as "failure". The following variables were used as predictors:

- Metropolitan area
- NAEP region
- Supervisor region
- Community type
- School type
- Number of sessions
- Number of reading (spiral) sessions
- Number of age eligible students
- Percent Asian or Pacific Islander students
- Percent Black, non-Hispanic students
- Percent Hispanic students
- Percent American Indian or Alaskan Native students
- Percent White, non-Hispanic students

The final model, estimated using WesVar to take proper account of the complex sample design, contained number of reading sessions, school type and number of age eligible students, as follows.

$$\log\left(\frac{P(\text{Response})}{P(\text{Non - response})}\right) = 1.984 - 0.485 * \text{One Reading Session} + 0.034 * \text{Catholic}$$
$$- 0.798 * \text{Other Private} - 1.826 * \text{Other Religious} - 0.002 * \text{Age Eligibles}$$

In the above equation, "One Reading Session" is an indicator variable equal to 1 if the school was asked to conduct only one reading session, and equal to 0 otherwise. "Catholic", "Other Private" and "Other Religious" are mutually exclusive indicator variables of the implied school characteristics. "Age Eligibles" is the number of age eligible students at the school. Because number of reading

sessions and school type are categorical variables, the solution to the model provides coefficients for all but the last level of each of these variables. Hence the intercept term incorporates the coefficients relevant to the characteristics: two reading sessions and public school.

The negative "One Reading Session" coefficient indicates that schools asked to conduct one reading session were less likely to respond than schools asked to conduct two reading sessions. The positive "Catholic" coefficient indicates that Catholic schools were more likely to respond than public schools. Other private and other religious schools were less likely to respond than public schools. The negative "Age Eligibles" coefficient indicates that schools with more age eligible students were less likely to respond. Standard errors and tests of hypotheses for the model parameter estimates are presented in table D-21.

Table D-21. Final model parameters for school reading response: 1999

Parameter	Estimate	Standard error	Test for H0: parameter = 0	P-value
Intercept	1.984	0.4330	4.5821	0.0001
One reading session	-0.485	0.2858	-1.6987	0.0980
Catholic	0.034	0.9088	0.0374	0.9704
Other private	-0.798	0.9078	-0.8787	0.3854
Other religious	-1.826	0.7920	-2.3056	0.0270
Age eligibles	-0.002	0.0011	-1.9493	0.0591

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The p-values above indicate that the effect of the number of reading sessions is moderately significant. There is no significant difference between the effect of public schools and Catholic or other private schools, however public schools are significantly different from other religious schools in their response propensity. The effect of the number of age eligible students is also moderately significant.

D.4.3.2 School Level Analysis – Mathematics/Science

The following nonresponse bias analysis is based on the original sample of 236 schools selected for mathematics/science assessment. All schools that were substituted by a replacement were treated as nonrespondents, as were any nonresponding original schools that were not substituted. The unweighted response rate was 75.00%, with 177 out of 236 schools responding. The weighted response rate was 72.08%. Standard errors are given throughout in parentheses.

D.4.3.2.1 Categorical Variables

The following characteristics were available for analysis.

- Metropolitan area
- NAEP region
- Supervisor region
- Community type
- School type
- Number of sessions
- Number of mathematics/science (tape) sessions

Table D-22 shows school response rates by metropolitan area status. The test of independence gives RS3 = 3.51, with a p-value of 0.061. There is some evidence that non-metropolitan schools were more likely to respond than others, though it is not significant at the 5% level.

Table D-22. School mathematics/science response rate by metropolitan area, weighted: 1999

Area	Response rate	
Non-Metropolitan Area	83.23%	(5.769%)
Metropolitan Area	68.96%	(5.273%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-23 shows school response rates by NAEP region. The test of independence gives RS3 = 2.54, with a p-value of 0.456. This indicates that there is no significant relationship between response status and NAEP region at the 5% level.

Table D-23. School mathematics/science response rate by NAEP region, weighted: 1999

NAEP region	Response rate	
Northeast	66.29%	(7.904%)
Southeast	84.15%	(8.508%)
Central	69.54%	(8.350%)
West	69.50%	(8.529%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-24 shows school response rates by NAEP supervisor region. The test of independence gives RS3 = 5.91, with a p-value of 0.312. This must be interpreted with caution due to the presence of a cell with less than five observations, however it would suggest that there is no significant relationship between response status and supervisor region at the 5% level.

Table D-24. School mathematics/science response rate by NAEP supervisor region, weighted: 1999

Supervisor region	Response rate	
1	65.21%	(9.205%)
2	61.00%	(11.480%)
3	76.83%	(11.651%)
4	94.33%	(6.375%)
5	94.77%	(3.275%)
6	53.23%	(14.015%)
7	78.20%	(7.309%)
8	68.75%	(9.561%)
9	65.97%	(6.078%)
10	68.77%	(24.386%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-25 shows school response rates by community type. The test of independence gives RS3 = 4.75, with a p-value of 0.090. There is some evidence that schools from rural or small towns were more likely to respond than others, though it is not significant at the 5% level.

Table D-25. School mathematics/science response rate by community type, weighted: 1999

Community type	Response rate	
Central city	73.03%	(8.215%)
Urban Fringe or Large Town	64.54%	(6.025%)
Rural or Small Town	83.88%	(5.938%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-26 shows school response rates by school type. The test of independence gives RS3 = 5.63, with a p-value of 0.123. This must be interpreted with caution due to the presence of a cell with less than five observations, however it would suggest that there is no significant relationship between response status and school type at the 5% level.

Table D-26. School mathematics/science response rate by school type, weighted: 1999

School type	Response rate	
Catholic	68.54%	(14.423%)
Other religious	28.49%	(11.716%)
Other private	63.23%	(21.402%)
Public	73.87%	(4.387%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-27 shows school response rates by the total number of sessions the school was asked to conduct. The test of independence gives RS3 = 0.72, with a p-value of 0.671. This must be interpreted with caution due to the presence of a cell with less than five observations, however it would suggest that there is no significant relationship between response status and number of sessions at the 5% level.

Table D-27. School mathematics/science response rate by number of sessions, weighted: 1999

Number of sessions	Response rate	
1 session	62.96%	(30.729%)
2 sessions	57.69%	(17.955%)
3 sessions	72.57%	(4.453%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-28 shows school response rates by the number of mathematics/science (tape) sessions the school was asked to conduct. The test of independence gives RS3 = 3.42, with a p-value of 0.064. There is some evidence that schools asked to conduct one mathematics/science session were more likely to respond than others, though it is not significant at the 5% level.

Table D-28. School mathematics/science response rate by number of tape sessions, weighted: 1999

Number of tape sessions	Response rate	
1 mathematics/science session	76.57%	(4.808%)
2 mathematics/science sessions	66.92%	(5.342%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.4.3.2.2 Continuous Variables

The following characteristics were available for analysis.

- Number of age eligible students
- Percent Asian or Pacific Islander students
- Percent Black, non-Hispanic students
- Percent Hispanic students
- Percent American Indian or Alaskan Native students
- Percent White, non-Hispanic students

Table D-29 shows the mean number of age eligible students for responding and nonresponding schools. The difference in the mean number of age eligible students is –41.0, with a 95% confidence interval of (-104.2, 22.2). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

Table D-29. Mean number of age eligible students by school mathematics/science response status, weighted: 1999

	Responding		Nonresponding	
Number of age eligible students	282.2	(20.07)	323.2	(20.22)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-30 shows the mean race/ethnicity percentages for responding schools and nonresponding schools.

Table D-30. Mean race/ethnicity percentages by school mathematics/science response status, weighted: 1999

Race/ethnicity	Responding		Nonresponding	
Percent of Asian or Pacific Islander students	3.90%	(0.576%)	4.07%	(0.987%)
Percent of Black, Non-Hispanic students	17.91%	(1.614%)	12.91%	(2.773%)
Percent of Hispanic Students	7.32%	(0.905%)	8.01%	(2.267%)
Percent of American Indian or Alaskan Native students	0.64%	(0.276%)	0.68%	(0.248%)
Percent of White, Non-Hispanic students	70.23%	(1.659%)	74.34%	(3.244%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The difference in the mean percentage of Asian or Pacific Islander students is –0.17%, with a 95% confidence interval of (-2.66%, 2.32%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of Black, non-Hispanic students is 5.00%, with a 95% confidence interval of (-0.96%, 10.96%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of Hispanic students is –0.69%, with a 95% confidence interval of (-4.57%, 3.20%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of American Indian or Alaskan Native students is –0.03%, with a 95% confidence interval of (-0.74%, 0.67%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of White, non-Hispanic students is –4.11%, with a 95% confidence interval of (-11.01%, 2.79%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

D.4.3.2.3 Logistic Regression Model

A logistic regression model was set up treating response status as the binary dependent variable, and frame characteristics as the predictor variables. Response was treated as "success" and nonresponse as "failure". The following variables were used as predictors:

- Metropolitan area
- NAEP region
- Supervisor region
- Community type
- School type
- Number of sessions
- Number of reading (spiral) sessions
- Number of age eligible students
- Percent Asian or Pacific Islander students
- Percent Black, non-Hispanic students
- Percent Hispanic students
- Percent American Indian or Alaskan Native students
- Percent White, non-Hispanic students

The final model, estimated using WesVar to take proper account of the complex sample design, contained number of mathematics/science sessions, school type and number of age eligible students, as follows.

$$\log\left(\frac{P(\text{Response})}{P(\text{Non - response})}\right) = 1.385 + 0.598 * \text{One Tape Session} - 0.501 * \text{Catholic}$$
$$- 1.288 * \text{Other Private} - 2.691 * \text{Other Religious} - 0.002 * \text{Age Eligibles}$$

In the above equation, "One Tape Session" is an indicator variable equal to 1 if the school was asked to conduct only one mathematics/science session, and equal to 0 otherwise. "Catholic", "Other Private" and "Other Religious" are mutually exclusive indicator variables of the implied school characteristics. "Age Eligibles" is the number of age eligible students at the school. Because number

of mathematics/science sessions and school type are categorical variables, the solution to the model provides coefficients for all but the last level of each of these variables. Hence the intercept term incorporates the coefficients relevant to the characteristics: two mathematics/science sessions and public school.

The positive "One Tape Session" coefficient indicates that schools asked to conduct one mathematics/science session were more likely to respond than schools asked to conduct two such sessions. Catholic, other private and other religious schools were all less likely to respond than public schools. The negative "Age Eligibles" coefficient indicates that schools with more age eligible students were less likely to respond. Standard errors and tests of hypotheses for the model parameter estimates are presented in table D-31.

Table D-31. Final model parameters for school mathematics/science response: 1999

Parameter	Estimate	Standard error	Test for H0: parameter = 0	P-value
Intercept	1.385	0.3755	3.6884	0.0007
One tape session	0.598	0.2781	2.1486	0.0385
Catholic	-0.501	0.6356	-0.7882	0.4357
Other private	-1.288	0.8412	-1.5314	0.1344
Other religious	-2.691	0.6701	-4.0161	0.0003
Age eligibles	-0.002	0.0011	-1.8581	0.0713

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The p-values above indicate that the effect of the number of mathematics/science sessions is significant at the 5% level. There is no significant difference between the effect of public schools and Catholic or other private schools, however public schools are highly significantly different from other religious schools in their response propensity. The effect of the number of age eligible students is moderately significant. The F-value measuring the overall fit of the model is 3.3316, with a p-value of 0.0156. This indicates that this set of independent variables as a group is significantly related to school response rate.

D.4.3.3 Student Level Analysis – Reading

The following nonresponse bias analysis is based on the original sample of 6517 students selected for reading assessment. The unweighted response rate was 81.14%, with 5288 out of 6517 students responding. The weighted response rate was 80.52%. Standard errors are given throughout in parentheses.

D.4.3.3.1 Categorical Variables

The following characteristics were available for analysis.

- Metropolitan area
- NAEP region
- Community type
- School type
- Student grade
- Student achievement level

Table D-32 shows student response rates by metropolitan area status. The test of independence gives RS3 = 8.84, with a p-value of 0.003. The data indicate that students in non-

metropolitan areas were significantly more likely to respond than other students, at the 5% level. This points to a potential source of bias in the student reading assessment results, however the student nonresponse adjustments that were made directly addressed this imbalance.

Table D-32. Student reading response rate by metropolitan area, weighted: 1999

Area	Response rate	
Non-metropolitan area	85.43%	(1.658%)
Metropolitan area	79.14%	(1.247%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-33 shows student response rates by NAEP region. The test of independence gives RS3 = 2.03, with a p-value of 0.462. This indicates that there is no significant relationship between response status and NAEP region at the 5% level.

Table D-33. Student reading response rate by NAEP region, weighted: 1999

NAEP region	Response rate	
Northeast	79.47%	(2.132%)
Southeast	83.06%	(1.267%)
Central	80.57%	(1.720%)
West	79.23%	(2.362%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-34 shows school response rates by community type. The test of independence gives RS3 = 11.14, with a p-value of 0.003. The data indicate that students in rural or small towns were significantly more likely to respond than other students, at the 5% level. This points to a potential source of bias in the student reading assessment results.

Table D-34. Student reading response rate by community type, weighted: 1999

Community type	Response rate	
Central city	75.79%	(1.991%)
Urban fringe or large town	80.34%	(2.048%)
Rural or small town	85.79%	(1.506%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-35 shows student response rates by school type. The test of independence gives RS3 = 12.01, with a p-value of 0.003. The data indicate that students in Catholic or other religious schools were significantly more likely to respond than other students, at the 5% level. This points to a potential source of bias in the student reading assessment results.

Table D-35. Student reading response rate by school type, weighted: 1999

School type	Response rate	
Catholic	93.09%	(1.297%)
Other religious	92.99%	(7.523%)
Other private	76.07%	(9.216%)
Public	79.67%	(1.008%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-36 shows student response rates by grade. The test of independence gives RS3 = 19.10, with a p-value < 0.001. The data indicate that students in the modal grade, grade 11, were significantly more likely to respond than other students, at the 5% level. This points to a potential source of bias in the student reading assessment results, however student nonresponse adjustments addressed this imbalance to some extent.

Table D-36. Student reading response rate by grade, weighted: 1999

Grade	Response rate	
Grade 9 or below	61.78%	(5.573%)
Grade 10	79.34%	(1.875%)
Grade 11	81.40%	(1.116%)
Grade 12	74.16%	(3.791%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-37 shows student response rates by achievement level. The test of independence gives RS3 = 15.12, with a p-value < 0.001. The data indicate that students at or above the modal grade for their age who more likely to respond than other students, at the 5% level. This points to a potential source of bias in the student reading assessment results, however student nonresponse adjustments directly addressed this imbalance.

Table D-37. Student reading response rate by achievement level, weighted: 1999

Achievement level	Response rate	
Below modal grade for age	77.24%	(1.414%)
At or above modal grade for age	82.35%	(1.048%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.4.3.3.2 Continuous Variables

The following characteristics were available for analysis.

- Number of age eligible students
- Percent Asian or Pacific Islander students
- Percent Black, non-Hispanic students
- Percent Hispanic students
- Percent American Indian or Alaskan Native students
- Percent White, non-Hispanic students
- Student date of birth

Table D-38 shows the mean number of age eligible students for schools attended by responding and nonresponding students. The difference in the mean number of age eligible students is –33.9, with a 95% confidence interval of (-58.61, -9.18). The confidence interval does not include zero, therefore there is evidence that the mean number of age eligible students is lower for schools attended by responding students, at the 5% level of significance. This indicates a potential source of bias in the student reading assessment results.

Table D-38. Mean number of age eligible students by student reading response status, weighted: 1999

	Responding		Nonresponding	
Number of age eligible students	280.1	(17.35)	313.9	(16.49)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-39 shows the mean race/ethnicity percentages for schools attended by responding and nonresponding students.

Table D-39. Mean race/ethnicity percentages by student reading response status, weighted: 1999

Race/ethnicity	Responding		Nonresponding	
Percent of Asian or Pacific Islander students	4.14%	(0.581%)	5.23%	(0.909%)
Percent of Black, Non-Hispanic students	16.35%	(1.499%)	21.47%	(2.076%)
Percent of Hispanic students	7.59%	(0.938%)	9.04%	(1.064%)
Percent of American Indian or Alaskan Native students	0.71%	(0.249%)	0.51%	(0.122%)
Percent of White, Non-Hispanic students	71.20%	(1.557%)	63.75%	(2.132%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The difference in the mean percentage of Asian or Pacific Islander students is –1.09%, with a 95% confidence interval of (-2.50%, 0.32%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of Black, non-Hispanic students is –5.12%, with a 95% confidence interval of (-8.86%, -1.37%). The confidence interval does not include zero, therefore there is evidence that the mean percentage of Black, non-Hispanic students is lower for schools attended by responding students, at the 5% level of significance. This indicates a potential source of bias in the student reading assessment results.

The difference in the mean percentage of Hispanic students is –1.44%, with a 95% confidence interval of (-3.21%, 0.32%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of American Indian or Alaskan Native students is 0.20%, with a 95% confidence interval of (-0.17%, 0.56%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of White, non-Hispanic students is 7.45%, with a 95% confidence interval of (3.28%, 11.62%). The confidence interval does not include zero, therefore there

is evidence that the mean percentage of White, non-Hispanic students is higher for schools attended by responding students, at the 5% level of significance. This indicates a potential source of bias in the student reading assessment results.

Table D-40 shows the mean month of birth for responding and nonresponding students. The variable being analyzed is coded such that 0 corresponds to April 1978, 1 corresponds to May 1978, etc. The difference in the mean month of birth is 0.72, with a 95% confidence interval of (0.21, 1.22). The confidence interval does not include zero, therefore there is evidence that responding students tended to be older than nonresponding students, at the 5% significance level. This indicates a potential source of bias in the student reading assessment results.

Table D-40. Mean month of birth by student reading response status, weighted: 1999

	Responding	Nonresponding
Month of birth	46.15 (0.118)	45.43 (0.221)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.4.3.3.3 Logistic Regression Model

A logistic regression model was set up treating response status as the binary dependent variable, and frame characteristics as the predictor variables. Response was treated as "success" and nonresponse as "failure". The following variables were used as predictors:

- Metropolitan area
- NAEP region
- Community type
- School type
- Student grade
- Student achievement level
- Number of age eligible students
- Percent Asian or Pacific Islander students
- Percent Black, non-Hispanic students
- Percent Hispanic students
- Percent American Indian or Alaskan Native students
- Percent White, non-Hispanic students
- Student date of birth

The final model, estimated using WesVar to take proper account of the complex sample design, contained NAEP region, community type, school type, student grade, student achievement level and all of the percent variables related to school racial/ethnic composition, as follows.

$$\log\left(\frac{P(\text{Response})}{P(\text{Non - response})}\right) = 7.797 + 0.114 * \text{Northeast} + 0.568 * \text{Southeast} + 0.170 * \text{Central}$$
$$- 0.570 * \text{Central City} - 0.337 * \text{Urban Fringe/Large Town}$$
$$+ 1.285 * \text{Catholic/Other Religious} - 0.103 * \text{Grade 9 or Below}$$
$$+ 0.572 * \text{Grade 10} + 0.432 * \text{Grade 11} - 0.369 * \text{Below Modal Grade for Age}$$
$$- 0.070 * \text{Percent Asian/PI} - 0.072 * \text{Percent Black} - 0.061 * \text{Percent Hispanic}$$
$$- 0.065 * \text{Percent White}$$

The meaning of each of the variables in the model should be obvious from the naming convention. Because NAEP region, community type, school type, student grade and student achievement level are categorical variables, the solution to the model provides coefficients for all but the last level of each of these variables. Hence the intercept term incorporates the coefficients relevant to the characteristics: West, rural or small town, public or other private school, grade 12, and at or above modal grade for age.

The positive "Northeast", "Southeast" and "Central" coefficients indicate that students from these NAEP regions were more likely to respond than students from the West. The negative "Central City" and "Urban Fringe/Large Town" coefficients indicate that students from these community types were less likely to respond than students from rural or small towns. The positive "Catholic/Other Religious" coefficient indicates that students in Catholic or other religious schools were more likely to respond than students in public or other private schools. Students in grade 9 or below were less likely to respond than students in grade 12, while students in grade 10 or 11 were more likely to respond. Students below the modal grade for their age were less likely to respond than those at or above the modal grade for their age. The interpretation of the coefficients related to school racial/ethnic composition is not straightforward due to the relationship between these percent variables. For instance, if the percentage Hispanic increases, then one or more of the other percent variables will decrease. Standard errors and tests of hypotheses for the model parameter estimates are presented in table D-41.

Table D-41. Final model parameters for student reading response: 1999

Parameter	Estimate	Standard error	Test for H0: parameter = 0	P-value
Intercept	7.797	2.9985	2.6005	0.0134
Northeast	0.114	0.2542	0.4488	0.6563
Southeast	0.568	0.2892	1.9624	0.0575
Central	0.170	0.2720	0.6236	0.5368
Central city	-0.570	0.1553	-3.6693	0.0008
Urban fringe/large town	-0.337	0.2005	-1.6832	0.1010
Catholic/other religious	1.285	0.2172	5.9148	< 0.0001
Grade 9 or below	-0.103	0.3400	-0.3017	0.7646
Grade 10	0.572	0.2790	2.0495	0.0477
Grade 11	0.432	0.2212	1.9549	0.0584
Below modal grade for age	-0.369	0.1001	-3.6876	0.0007
Percent Asian/PI	-0.070	0.0300	-2.3491	0.0244
Percent Black	-0.072	0.0314	-2.2839	0.0284
Percent Hispanic	-0.061	0.0298	-2.0489	0.0478
Percent white	-0.065	0.0318	-2.0488	0.0478

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The p-values above indicate that most of the effects are significant at the 5% level. The F-value measuring the overall fit of the model is 5.6288, with a p-value of 0.0001. This indicates that this set of independent variables as a group is highly significantly related to student response rate.

D.4.3.4 Student Level Analysis – Mathematics/Science

The following nonresponse bias analysis is based on the original sample of 4731 students selected for mathematics/science assessment. The unweighted response rate was 80.22%, with 3795 out of 4731 students responding. The weighted response rate was 81.71%. Standard errors are given throughout in parentheses.

D.4.3.4.1 Categorical Variables

The following characteristics were available for analysis.

- Metropolitan area
- NAEP region
- Community type
- School type
- Student grade
- Student achievement level

Table D-42 shows student response rates by metropolitan area status. The test of independence gives RS3 = 8.52, with a p-value of 0.004. The data indicate that students in non-metropolitan areas were significantly more likely to respond than other students, at the 5% level. This points to a potential source of bias in the student reading assessment results, however the student nonresponse adjustments that were made directly addressed this imbalance.

Table D-42. Student mathematics/science response rate by metropolitan area, weighted: 1999

Area	Response rate	
Non-metropolitan area	86.55%	(1.440%)
Metropolitan area	80.48%	(1.452%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-43 shows student response rates by NAEP region. The test of independence gives RS3 = 5.27, with a p-value of 0.077. There is some evidence that students in the Southeast were more likely to respond than other students, though it is not significant at the 5% level.

Table D-43. Student mathematics/science response rate by NAEP region, weighted: 1999

NAEP region	Response rate	
Northeast	78.97%	(2.337%)
Southeast	86.50%	(0.583%)
Central	82.51%	(1.456%)
West	79.37%	(3.080%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-44 shows school response rates by community type. The test of independence gives RS3 = 9.73, with a p-value of 0.007. The data indicate that students in rural or small towns were significantly more likely to respond than other students, at the 5% level. This points to a potential source of bias in the student reading assessment results.

Table D-44. Student mathematics/science response rate by community type, weighted: 1999

Community type	Response rate	
Central city	77.09%	(2.257%)
Urban fringe or large town	82.14%	(1.877%)
Rural or small town	85.93%	(1.576%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-45 shows student response rates by school type. The test of independence gives RS3 = 9.15, with a p-value of 0.011. The data indicate that students in Catholic or other religious schools were significantly more likely to respond than other students, at the 5% level. This points to a potential source of bias in the student reading assessment results.

Table D-45. Student mathematics/science response rate by school type, weighted: 1999

School type	Response rate	
Catholic	96.39%	(0.658%)
Other religious	97.61%	(3.502%)
Other private	77.95%	(8.382%)
Public	80.44%	(1.189%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-46 shows student response rates by grade. The test of independence gives RS3 = 19.70, with a p-value < 0.001. The data indicate that students in the modal grade, grade 11, were significantly more likely to respond than other students, at the 5% level. This points to a potential source of bias in the student reading assessment results, however student nonresponse adjustments addressed this imbalance to some extent.

Table D-46. Student mathematics/science response rate by grade, weighted: 1999

Grade	Response rate	
Grade 9 or below	62.73%	(4.614%)
Grade 10	79.97%	(1.706%)
Grade 11	83.18%	(1.349%)
Grade 12	71.74%	(4.511%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-47 shows student response rates by achievement level. The test of independence gives RS3 = 4.61, with a p-value of 0.032. The data indicate that students at or above the modal grade for their age were significantly more likely to respond than other students, at the 5% level. This points to a potential source of bias in the student reading assessment results, however student nonresponse adjustments directly addressed this imbalance.

Table D-47. Student mathematics/science response rate by achievement level, weighted: 1999

Achievement level	Response rate	
Below modal grade for age	78.48%	(1.547%)
At or above modal grade for age	82.67%	(1.368%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.4.3.4.2 Continuous Variables

The following characteristics were available for analysis.

- Number of age eligible students
- Percent Asian or Pacific Islander students
- Percent Black, non-Hispanic students
- Percent Hispanic students
- Percent American Indian or Alaskan Native students
- Percent White, non-Hispanic students
- Student date of birth

Table D-48 shows the mean number of age eligible students for schools attended by responding and nonresponding students. The difference in the mean number of age eligible students is –36.4, with a 95% confidence interval of (-67.88, -4.86). The confidence interval does not include zero, therefore there is evidence that the mean number of age eligible students is lower for schools attended by responding students, at the 5% level of significance. This indicates a potential source of bias in the student mathematics/science assessment results.

Table D-48. Mean number of age eligible students by student mathematics/science response status, weighted: 1999

	Responding		Nonresponding	
Number of age eligible students	278.9	(17.95)	315.3	(17.99)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-49 shows the mean race/ethnicity percentages for schools attended by responding and nonresponding students.

Table D-49. Mean race/ethnicity percentages by student mathematics/science response status, weighted: 1999

Race/ethnicity	Responding		Nonresponding	
Percent of Asian or Pacific Islander students	4.16%	(0.552%)	5.76%	(1.080%)
Percent of Black, Non-Hispanic students	15.87%	(1.634%)	19.41%	(2.135%)
Percent of Hispanic students	7.03%	(0.850%)	9.52%	(0.981%)
Percent of American Indian or Alaskan Native students	0.58%	(0.203%)	0.49%	(0.120%)
Percent of White, Non-Hispanic students	72.36%	(1.702%)	64.82%	(2.250%)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The difference in the mean percentage of Asian or Pacific Islander students is –1.60%, with a 95% confidence interval of (-3.34%, 0.13%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of Black, non-Hispanic students is –3.54%, with a 95% confidence interval of (-7.65%, 0.57%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of Hispanic students is –2.49%, with a 95% confidence interval of (-4.01%, -0.98%). The confidence interval does not include zero, therefore there is evidence that the mean percentage of Hispanic students is lower for schools attended by responding students, at the 5% level of significance. This indicates a potential source of bias in the student reading assessment results.

The difference in the mean percentage of American Indian or Alaskan Native students is 0.09%, with a 95% confidence interval of (-0.17%, 0.35%). The confidence interval includes zero, therefore the difference is not significant at the 5% level.

The difference in the mean percentage of White, non-Hispanic students is 7.54%, with a 95% confidence interval of (3.11%, 11.98%). The confidence interval does not include zero, therefore there is evidence that the mean percentage of White, non-Hispanic students is higher for schools attended by responding students, at the 5% level of significance. This indicates a potential source of bias in the student reading assessment results.

Table D-50 shows the mean month of birth for responding and nonresponding students. The variable being analyzed is coded such that 0 corresponds to April 1978, 1 corresponds to May 1978, etc. The difference in the mean month of birth is 0.31, with a 95% confidence interval of (-0.00, 0.63). The confidence interval barely includes zero. Therefore there is some evidence that responding students tended to be older than nonresponding students, though it is not significant at the 5% level. This indicates a potential source of bias in the student reading assessment results.

Table D-50. Mean month of birth by student mathematics/science response status, weighted: 1999

	Responding		Nonresponding	
Month of birth	47.73	(0.057)	47.42	(0.144)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

D.4.3.4.3 Logistic Regression Model

A logistic regression model was set up treating response status as the binary dependent variable, and frame characteristics as the predictor variables. Response was treated as "success" and nonresponse as "failure". The following variables were used as predictors:

- Metropolitan area
- NAEP region
- Community type
- School type
- Student grade
- Student achievement level
- Number of age eligible students
- Percent Asian or Pacific Islander students
- Percent Black, non-Hispanic students
- Percent Hispanic students

- Percent American Indian or Alaskan Native students
- Percent White, non-Hispanic students
- Student date of birth

The final model, estimated using WesVar to take proper account of the complex sample design, contained NAEP region, community type, school type, student achievement level and all of the percent variables related to school racial/ethnic composition, as follows.

$$\log\left(\frac{P(\text{Response})}{P(\text{Non - response})}\right) = 6.066 - 0.088 * \text{Northeast} + 0.683 * \text{Southeast} + 0.170 * \text{Central}$$
$$- 0.447 * \text{Central City} - 0.198 * \text{Urban Fringe/Large Town}$$
$$+ 2.027 * \text{Catholic/Other Religious} - 0.307 * \text{Below Modal Grade for Age}$$
$$- 0.052 * \text{Percent Asian/PI} - 0.051 * \text{Percent Black} - 0.044 * \text{Percent Hispanic}$$
$$- 0.044 * \text{Percent White}$$

The meaning of each of the variables in the model should be obvious from the naming convention. Because NAEP region, community type, school type and student achievement level are categorical variables, the solution to the model provides coefficients for all but the last level of each of these variables. Hence the intercept term incorporates the coefficients relevant to the characteristics: West, rural or small town, public or other private school, and at or above modal grade for age.

The negative "Northeast" coefficient indicates that students from this NAEP region were less likely to respond than students from the West. Students from the Southeast and Central regions were more likely to respond. The negative "Central City" and "Urban Fringe/Large Town" coefficients indicate that students from these community types were less likely to respond than students from rural or small towns. The positive "Catholic/Other Religious" coefficient indicates that students in Catholic or other religious schools were more likely to respond than students in public or other private schools. Students below the modal grade for their age were less likely to respond than those at or above the modal grade for their age. The interpretation of the coefficients related to school racial/ethnic composition is not straightforward due to the relationship between these percent variables. For instance, if the percentage Hispanic increases, then one or more of the other percent variables will decrease. Standard errors and tests of hypotheses for the model parameter estimates are presented in table D-51.

Table D-51. Final model parameters for student mathematics/science response: 1999

Parameter	Estimate	Standard error	Test for H0: parameter = 0	P-value
Intercept	6.066	2.4329	2.4936	0.0174
Northeast	-0.088	0.2811	-0.3117	0.7570
Southeast	0.683	0.3099	2.2039	0.0340
Central	0.170	0.2976	0.5721	0.5708
Central city	-0.447	0.1757	-2.5442	0.0154
Urban fringe/large town	-0.198	0.1906	-1.0375	0.3064
Catholic/other religious	2.027	0.3962	5.1173	< 0.0001
Below modal grade for age	-0.307	0.1213	-2.5346	0.0157
Percent Asian/PI	-0.052	0.0250	-2.0883	0.0439
Percent Black	-0.051	0.0263	-1.9223	0.0625
Percent Hispanic	-0.044	0.0255	-1.7103	0.0958
Percent White	-0.044	0.0262	-1.6696	0.1037

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The p-values above indicate that many of the effects are significant at the 5% level. Several others are moderately significant. The F-value measuring the overall fit of the model is 9.6665, with a p-value < 0.0001. This indicates that this set of independent variables as a group is highly significantly related to student response rate.

D.4.4 Conclusions

The investigation into nonresponse bias at the school and student levels for the age 17 group of the 1999 NAEP Long-Term Trend Assessment has revealed some possible areas of concern.

At the school level for reading, other religious schools were significantly less likely to respond than public schools. This was also true at the school level for mathematics/science. In addition for mathematics/science, those schools asked to conduct one tape session were significantly more likely to respond than those asked to conduct two sessions. Schools assigned two sessions were, on average, slightly larger than those assigned one session, but because session types were randomly assigned to schools, this relationship is fairly weak. All schools with more than 40 age eligible students were randomly assigned either one or two sessions. Smaller schools were assigned only one mathematics/science session.

At the student level for reading, students from central cities were significantly less likely to respond than students from rural or small towns. Students from Catholic or other religious schools were significantly more likely to respond than students from public or other private schools. Students at or below the modal grade for their age were significantly less likely to respond than others. There was also a complicated effect due to the racial/ethnic composition of the school the student attended. In terms of ramifications for the survey results, potentially the most serious of these effects is the one relating to student grade level. Fortunately, nonresponse adjustments were directly targeted in this area.

At the student level for mathematics/science, students from the Southeast were significantly more likely to respond than students from the West. Students from central cities were significantly less likely to respond than students from rural or small towns. Students from Catholic or other religious schools were significantly more likely to respond than students from public or other private schools. Students at or below the modal grade for their age were significantly less likely to respond than others. There was also a complicated effect due to the racial/ethnic composition of the school the student attended. In terms of ramifications for the survey results, potentially the most serious of these effects is the one relating to student grade level. Fortunately, nonresponse adjustments were directly targeted in this area.

D.5. Weighting Procedures and Estimation of Sampling Variance

D.5.1 Introduction

As in previous assessments, the 1999 long-term trend assessment used a complex sample design with the goal of securing a sample from which estimates of population and subpopulation characteristics could be obtained with reasonably high precision (as measured by low sampling variability). At the same time, it was necessary that the sample be economically and operationally feasible to obtain. The resulting sample had certain properties that had to be taken into account to ensure valid analyses of the data.

The sampling design for the 1999 long-term trend assessment was the same as used for previous long-term trend assessments. This was a multistage probability sampling design which provided that schools with small enrollments of eligible students be assigned lower probabilities of selection. To account for the differential selection probabilities, and to allow for adjustments for nonresponse, each student was assigned a sampling weight. Nonresponse adjustments of the sampling weights for the 1999 assessment included a factor to account for refusal among participating schools to assess age-eligible students not in the target grades. Section D.5.2 discusses the procedures used to derive these sampling weights.

Another consequence of the long-term trend sample design is its effect on the estimation of sampling variability. Because of the effects of cluster selection (cluster of elements: students within schools, schools within primary sampling units) and because of the effects of certain adjustments to the sampling weights (nonresponse adjustment and poststratification), observations made on different students cannot be assumed to be independent of one another. In particular, as a result of clustering, ordinary formulas for the estimation of the variance of sample statistics, based on assumptions of independence, will tend to underestimate the true sampling variability. Section D.5.3 discusses the jackknife technique used by NAEP to estimate sampling variability.

D.5.2 Weighting Procedures for Assessed and Excluded Students

Since the sample design determines the derivation of the sampling weights and the estimation of sampling variability, it will be helpful to note the key features of the NAEP 1999 long-term trend sample design. A description of the design is given in section D.3.

The 1999 sample was a multistage probability sample consisting of four stages. The first stage of selection, the primary sampling units (PSUs), consisted of counties or groups of counties. The second stage of selection consisted of elementary and secondary schools. The assignment of sessions to sampled schools comprised the third stage of sampling, and the fourth stage involved the selection of students within schools and their assignment to sessions.

The probabilities of selection of the first-stage sampling units were proportional to measures of their size, while the probabilities for subsequent stages of selection were such that the overall probabilities of selection of students were approximately uniform. Students from schools with smaller numbers of eligibles received lower probabilities of selection, as a means of enhancing the cost efficiency of the sample.

The 1999 long-term trend samples are intended to provide statistical linkage from 1999 data to data from previous assessments. These samples used the age definitions, times of testing, and modes of administration used in previous assessments. They represent two overlapping student populations,

the first of specified grades (of any age) and the second of specified ages (in any grade). Students were age-eligible if they were born in the appropriate year (1989, 1985, or October 1981 to September 1982. The corresponding grades were 4, 8, and 11. Each student cohort is called an "age class". The samples and their target populations are as follows:

Reading and Writing. These consist of samples comparable to the 1984 main assessment and address the subject areas of reading and writing. The samples were collected by grade and age for age 9/grade 4, age 13/grade 8, and age 17/grade 11, using the age definitions and time of testing from 1984. As in that assessment, print administration was used. Six assessment booklets were administered at each age class. The respondents to the combined set of assigned booklets at a given age class constitute a representative sample of the population of students who were in the specified grade *or* of the specified age. The respondents to any one of the booklets also constitute a representative sample.

Mathematics and Science. These consist of samples comparable to those used for the measurement of trends in 1986. The samples were collected by age only and using the same age definitions and time of testing as in the long-term trend assessment in 1986. As in that assessment, the administration of mathematics and science questions was paced with an audiotape. For ages 9 and 13, three assessment booklets were administered to each age group while two booklets were administered at age 17. The respondents to any one of the booklets assigned to a given age constitute a representative sample of the population of all students of that age. Each booklet was administered in a separate assessment session, but the booklets were combined for weighting and reporting.

For purposes of sampling and weighting, the assessment samples are categorized as "tape-administered" or "print-administered," according to mode of administration:

Tape-administered samples are samples that required audiotape pacing in the assessment (mathematics and science). For these samples all students within a particular assessment session received the same booklet and were paced through at least part of the booklet with an audiotape.

Print-administered samples are the assessments of reading and writing. For these samples, no audiotape pacing was employed and the assessment booklets were spiraled through each assessment session (that is, the different booklets that were part of a given session type were systematically interspersed and assigned for testing in that order).

Each age class was weighted separately. The tape- and print-administered samples were weighted separately; excluded students were weighted together, apart from assessed students.

D.5.2.1 Derivation of the Sample Weights

As indicated earlier, lower sampling rates were introduced for very small schools, those schools having only one to 19 age-eligible students. This reduced level of sampling from small schools was undertaken in a near optimal manner as a means of reducing variances per unit of cost, since it is relatively costly to administer assessments in these small schools. Appropriate estimation of population characteristics must take disproportionate representation into account. This is accomplished by assigning a weight to each respondent, where the weights approximately account for the sample design and reflect the appropriate proportional representation of the various types of individuals in the population.

The weighting procedures for 1999 included computing the student's base weight, the reciprocal of the probability that the student was selected for a particular session type. These base weights were adjusted for nonresponse and then a trimming algorithm was applied to reduce a few excessively large weights. The weights were further adjusted by a student-level poststratification

procedure to reduce the sampling error. This poststratification was accomplished by adjusting the weights of the sampled students so that the resulting estimates of the total number of students in a set of specified subgroups of the population corresponded to population totals based on information from the Current Population Survey and U.S. Census Bureau estimates of the population. The subpopulations were defined in terms of race, ethnicity, geographic region, grade, and age relative to grade.

D.5.2.1.1 Student Base Weight

The base weight assigned to a student is the reciprocal of the probability that the student was selected for a particular assessment. That probability is the product of five factors:

1. The probability that the PSU was selected;
2. The conditional probability, given the PSU, that the school was selected;
3. The conditional probability, given the sample of schools in a PSU, that the school was allocated the specified session type;
4. The conditional probability, given the school, that the student was selected
5. The conditional probability, given the school, that the selected student was assigned to the specified session type.

Thus, the base weight for a student may be expressed as the product

$$W_B = PSUWT \times SCHWT \times SESSWT \times STUSELWT \times STUSCHW$$

where PSUWT, SCHWT, SESSWT, STUSELWT, and STUSCHW are, respectively the reciprocals of the preceding probabilities. SESSWT and STUSCHW are not factors of the student base weight for age-eligible excluded students.

D.5.2.1.2 Session Nonresponse Adjustment (SESNRF)

Sessions were assigned to schools before cooperation status was final. The session nonresponse adjustment was intended to compensate for session nonresponse due to school nonparticipation or refusal to conduct a particular session type. Nonresponse cells, called "subuniverse", were formed by grouping PSUs according to socioeconomic characteristics. The adjustment factors were calculated separately for each age class for the spiral assessment, the tape assessment, and excluded students, within subuniverse. Occasionally, collapsing of cells was necessary to improve the stability of the adjustment factors. Most cells needing collapsing contained small numbers of cooperating schools; occasionally, cells with low response rates were collapsed.

In subuniverse *s* in session type *h*, the session nonresponse adjustment factor $SESNRF_{hs}$ is given by

$$SESNRF_{hs} = \frac{\sum_{i \in B_{hs}} PSUWT_i \times SCHWT_i \times SESSWT_{hi} \times G_i}{\sum_{i \in C_{hs}} PSUWT_i \times SCHWT_i \times SESSWT_{hi} \times G_i}$$

where

$PSUWT_i$ = the PSU weight for the PSU containing school *i;*

$SCHWT_i$	=	the school weight for school *i;*
$SESSWT_{hi}$	=	the session allocation weight for session type *h* in school *i* (spiral or tape, not excluded);
G_i	=	the estimated number of age-plus grade-eligible students in school *i* for the spiral assessment and excluded students; the estimated number of age-eligible students for the tape assessment;
set B_{hs}	=	all in-scope originally sampled schools in subuniverse *s*, excluding substitutes
set C_{hs}	=	all schools in subuniverse *s* that ultimately participated, including substitutes.

D.5.2.1.3 Age-Only Eligible Nonresponse Adjustment (AOENRF)

Historically, schools have occasionally refused to assess age-eligible students who are not in the modal grade, one of grades 4, 8, or 11. The distribution of age-eligibles is such that most of the students missed have been 3rd-, 7th-, and 10th- graders. This practice appears to have increased recently. There was a considerable increase for the 1999 age 9 and age 17 samples. See table D-52 for a comparison of the 1996 and 1999 long-term trend samples.

Table D-52. Long-term trend participating schools refusing to assess age-eligible students not in the modal grade: 1996 and 1999

Sample	Participating schools	Having modal grade and grade below	Refusing to assess age-only eligibles	
1999				
Age class 9	258	250	47	18.8%
Age class 13	238	218	22	10.1%
Age class 17	194	192	34	17.7%
1996				
Age class 9	248	232	30	12.9%
Age class 13	242	226	23	10.2%
Age class 17	191	187	17	9.1%

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

For the 1999 long-term trend samples, 97% of age class 9 and age class 13 students and 95% of age class 17 students were in the modal grade and the grade below. An age-only-eligible adjustment factor was calculated for spiral sessions, tape sessions, and excluded students in participating schools having both grades. This factor was set to 1 for students not in one of grades 3, 7, or 10 and for all students in schools not having both the modal grade and the grade below. The adjustment cells were the collapsed subuniverse classes described in section D.5.2.1.2.

In subuniverse *u* in session type *h*, the age-only-eligible nonresponse adjustment factor $AOENRF_{hu}$ is given by

$$AOENRF_{hu} = \frac{\sum_{i \in M_{hu}} PSUWT_i \times SCHWT_i \times SESSWT_{hi} \times SESNRF_{hi} \times g_i}{\sum_{i \in N_{hu}} PSUWT_i \times SCHWT_i \times SESSWT_{hi} \times SESNRF_{hi} \times g_i}$$

where

$PSUWT_i$ = the PSU weight for the PSU containing school *i;*

$SCHWT_i$ = the school weight for school *i;*

$SESSWT_{hi}$ = the session allocation weight for session type *h* in school *i* (spiral or tape, not excluded);

$SESNRF_{hi}$ = the session nonresponse adjustment factor for session type *h* in school *i*;

g_i = the estimated enrollment in the grade below the modal grade in school *i*;

Set M_{hu} = all participating session-type *h* (spiral, tape, excluded) schools in subuniverse *u* having both grades.

Set N_h = participating session-type *h* (spiral, tape, excluded) schools in subuniverse *u* having both grades, assessing all eligibles

D.5.2.1.4 Student Nonresponse Adjustment (STUNRF)

Student nonresponse adjustments were calculated separately at each age class for the spiral assessment and the tape assessment within classes formed by subuniverse and modal grade status (at or above modal grade, below modal grade). For excluded students at each age class, the adjustments were calculated within classes formed by subuniverse. Distributions of the student nonresponse adjustment factors are shown in table D-54a.

For each class *c* in session type *h*, the student nonresponse adjustment factor $STUNRF_{hc}$ is given by

$$STUNRF_{hc} = \frac{\sum_{j \in A_{hc}} PSUWT_j \times SCHWT_j \times STUSELWT_{hj} \times SESSWT_{hj} \times SESNRF_{hj} \times AOENRF_{hj} \times STUSCHW_{hj}}{\sum_{j \in B_{hc}} PSUWT_j \times SCHWT_j \times STUSELWT_{hj} \times SESSWT_{hj} \times SESNRF_{hj} \times AOENRF_{hj} \times STUSCHW_{hj}}$$

where

$PSUWT_j$ = the PSU weight for the PSU containing student *j;*

$SCHWT_j$ = the school weight for school containing student *j;*

$STUSELWT_{hj}$ = the within school weight for student *j*;

$SESSWT_{hj}$ = the session allocation weight for the school containing student *j* in session type *h* (spiral or tape, not age-eligible excluded);

$SESNRF_{hj}$	=	the session nonresponse adjustment factor for the school containing student *j* in session type *h*;
$AOENFR_{hj}$	=	the age-only-eligible nonresponse factor for the school containing student j in session type *h*;
$STUSCHW_{hj}$	=	the within-school weight for student *j* in session type *h* (spiral or tape, not age-eligible excluded);
Set A_{hc}	=	students in class *c* who were sampled for session type *h* and not excluded; or all excluded students in class *c*
Set B_{hc}	=	students in class *c* who were assessed in session type *h*; or excluded students in class *c* for whom an *Excluded Student Questionnaire* was completed.

D.5.2.1.5 Trimming of Weights

In a number of cases, students were assigned relatively large weights. One cause of large weights is underestimation of the number of eligible students in some schools, leading to inappropriately low probabilities of selection for those schools. A second major cause is the presence of large schools (high schools in particular) in PSUs with small selection probabilities. In such cases, the maximum permissible within-school sampling rate (determined by the maximum sample size allowed per school—see section D.3) could be smaller than the desired overall within-PSU sampling rate for students. Large weights arise also because very small schools were, by design, sampled with low probabilities. Other large weights arise as the result of high levels of nonresponse coupled with low to moderate probabilities of selection and the compounding of nonresponse adjustments at various levels.

Students with notable large weights have an unusually large impact on estimates such as weighted means. Since, under some simplifying assumptions, the variability in weights contributes to the variance of an overall estimate by an approximate factor of $1+V^2$, where V^2 is the relative variance of the weights, an occasional unusually large weight is likely to produce large sampling variances of the statistics of interest, especially when the large weights are associated with students with atypical performance characteristics.

To reduce this problem, a procedure of trimming a few of the more extreme weights to values somewhat closer to the mean weight was applied. This trimming can increase the accuracy of the resulting survey estimates, substantially reducing V^2 and hence the sampling variance, while introducing a small bias. The trimming algorithm was identical to that used since 1984 and had the effect of trimming the weights of students from any school that contributed more than a specified proportion, ξ, to the estimated variance of the estimated number of students eligible for assessment. The trimming was done separately for the spiral assessment, the tape assessment, and excluded students at each age class. Weights for students from five age class 9 schools were trimmed; two age class 13 schools, and one age class 17 school. Distributions of the student weight trimming factors are shown in table D-54b.

D.5.2.1.6 Poststratification

As in most sample surveys the respondent weights are random variables that are subject to sampling variability. Even if there were no nonresponse, the respondent weights would at best provide unbiased estimates of the various subgroup proportions. However, since unbiasedness refers to

average performance over a conceptually infinite number of replications of the sampling, it is unlikely that any given estimate, based on the achieved sample, will exactly equal the population value. Furthermore, the respondent weights have been adjusted for nonresponse and a few extreme weights have been reduced in size.

To reduce the mean squared error of estimates using the sampling weights, these weights were further adjusted so that estimated population totals for a number of specified subgroups of the population, based on the sum of weights of students of the specified type, were the same as presumably better estimates based on composites of estimates from the 1995 and 1996 Current Population Survey and 1999 population projections made by the U.S. Census Bureau. This adjustment, called poststratification, is intended especially to reduce the mean squared error of estimates relating to student populations that span several subgroups of the population, and thus also to reduce the variance of measures of changes over time for such student populations.

Poststratification adjustments were calculated separately at each age class for the spiral assessment, the tape assessment, and excluded students. Adjustment cells were formed by race/region and eligibility class (eligible by grade and of modal age; eligible by age only; and eligible by grade but not of modal age).

Race/Region	**Eligibility Class**
White/Northeast	Grade and age
White/North Central	Age only
White/South	Grade only
White/West	
Black	
Hispanic	
Other	

Thus 21 cells were used for the spiral assessment and excluded students at each age class. Seven cells (by race/region only) were used for the tape assessment at each age class. For each cell the poststratification factor is a ratio whose denominator is the sum of weights (after adjustments for nonresponse and trimming) of assessed and excluded students, and whose numerator is an adjusted estimate of the total number of students in the population who are members of the cell. The poststratification factor for student *j* in session type *h* and poststratification adjustment class c is given by

$$PSFCTR_{hc} = \frac{TOTAL_c}{\sum_{j \in C_{hc}} W_{Bj} \times SESNRF_j \times AOENRF_j \times STUNRF_j \times TRIMFCTR_j}$$

where

W_{Bj} = the base weight for student *j* (see section D.5.2.1)*;*

$TOTAL_c$ = the total number of eligible students in class c, described above, from the October 1995 and 1996 Current Population surveys and 1999 population projections;

$SESNRF_j$ = the session nonresponse adjustment factor for the school containing student *j* in session type *h* (spiral or tape, not excluded);

$STUNRF_j$ = the student nonresponse adjustment for student *j*;

$AOENRF_j$ = the age-only nonresponse adjustment for the school containing student j in session type h;

$TRMFCTR_j$ = the trimming factor for student j;

$Set\ C_{hc}$ = students in class c who were assessed in session type h.

D.5.2.1.7 The Final Student Weights

The final weight assigned to a student is the student's full-sample base weight after the application of the various adjustments described above. The distributions of the NAEP 1999 long-term trend final student weights are given in table D-53.

D.5.2.1.8 School Weights

School weights for the 1999 Long-Term Trend were computed separately by age class. The weight for school i in session type h is given by

$$W_{hi} = PSUWT_i \times SCHWT_i \times SESSWT_{hi} \times SESNRF_{hi}$$

where $PSUWT_i$, $SCHWT_{hi}$, $SESSWT_{hi}$, and $SESNRF_{hi}$ are defined in section D.5.2.1.1. The school nonresponse adjustment factors used for excluded students ($SESNRF_{hi}$) are not subject-specific.

D.5.2.1.9 Jackknife Replicate Weights

In addition to the weights that were used to derive all estimates of population and subpopulation characteristics, other sets of weights, called jackknife replicate weights, were derived to facilitate the estimation of sampling variability by the jackknife variance estimation techniques. These weights are discussed in the next section.

D.5.3 Procedures Used to Estimate Sampling Variability

A major source of uncertainty in the estimation of the population value of a variable of interest exists because information about the variable is obtained on only a sample from the population. To reflect this fact, it is important to attach to any statistic (e.g., a mean) an estimate of the sampling variability to be expected for that statistic. Estimates of sampling variability provide information about how much the value of a given statistic would be likely to change if the statistic had been based on another equivalent sample of individuals drawn in exactly the same manner as the achieved sample.

The estimation of the sampling variability of any statistic must take into account the sample design. In particular, because of the effects of cluster selection (students within schools, schools within PSUs) and because of effects of nonresponse and poststratification adjustments, observations made on different students cannot be assumed to be independent of each other (and are, in fact generally positively correlated). Furthermore, to account for the differential probabilities of selection (and the various adjustments), each student has an associated sampling weight, which should be used in the computation of any statistic and which is itself subject to sampling variability. Ignoring the special characteristics of the sample design and treating the data as if the observations were independent and

identically distributed will generally produce underestimates of the true sampling variability, due to the clustering and unequal sampling weights.

Through the creation of student replicate weights, the jackknife procedure allows the measurement of variability attributable to the use of poststratification and other weight adjustment factors that are dependent upon the observed sample data. Once these replicate weights are derived, it is a straightforward matter to obtain the jackknife variance estimate of any statistic.

The jackknife procedure (as applied to the Long-Term Trend samples) is based on the development of a set of jackknife replicate weights for each assessed student (or excluded student, or school, depending upon the file involved). The replicate weights are developed in such a way that approximately unbiased estimates of the sampling variance of an estimate result, with an adequate number of degrees of freedom to be useful for purposes of making inferences about the parameter of interest.

The estimated sampling variance of a parameter estimator t is the sum of M squared differences (where M is the number of replicate weights developed):

$$\hat{V}ar(t) = \sum_{i=1}^{M} (t_i - t)^2 ,$$

where t_i denotes the estimator of the parameter of interest, obtained using the *ith* set of replicate weights in place of the original sample of full-sample estimates. Essentially, the jackknife method requires repeatedly dividing the full sample into subsamples, or replicates, and calculating the statistic of interest for each replicate. Replicates are created by randomly deleting first-stage sampling units from the full sample. In the case of the Long-Term Trend samples, these are noncertainty PSUs, or groups of schools in certainty PSUs, described below.

D.5.3.1 Replicate Weights

Replicate weights were developed for the 1999 Long-Term Trend samples according to the procedure used in previous assessment years. It is analogous to the procedure used for developing replicate weights for the 1998 main NAEP samples; see chapter 10 of *The NAEP 1998 Technical Report* (Allen, et al., 2001)

Thirty-six replicate weights were developed at each age class for each session type. For age class 9 and age class 13, 22 replicates reflect the amount of sampling variance contributed by the noncertainty PSUs, with the remaining 14 replicates reflecting the variance contribution of the certainty PSUs. For the age class 17 sample, 23 replicates represent the noncertainty PSUs, and 13 represent the certainty PSUs. The derivation of the replicates reflecting the variance of the noncertainty PSUs involves defining pairs of PSUs in a manner that models a design in which two PSUs are drawn with replacement per stratum. This definition of pairs is undertaken in a manner closely reflective of the actual design, in that PSUs are pairs that are drawn from strata within the same subuniverse, with similar stratum characteristics. In the case of the certainty PSUs, strata were defined by grouping schools within school type (public, private)/urbanicity classes. Within each class, replicates were defined by pairs of school groups.

Replicate base weights were calculated for each set of sampled schools. All nonresponse, trimming, and poststratification adjustments described above were then applied to produce final replicate weights.

Table D-53. Distribution of final student weights, NAEP long-term trend samples: 1999

Sample	Number of cases	Mean	Standard deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Age class 9								
Reading/writing	5,793	798.23	400.21	41.35	495.33	766.17	1,024.46	4,260.76
Mathematics/science	6,032	571.00	244.49	83.36	399.37	530.38	689.24	2,288.14
Excluded students	1,120	428.86	320.94	47.46	229.98	334.22	464.12	2,526.04
Age class 13								
Reading/writing	5,933	789.12	387.07	112.88	518.08	751.67	992.69	5,317.46
Mathematics/science	5,941	571.71	253.03	148.72	419.43	510.73	651.52	3,719.37
Excluded students	824	505.02	357.06	98.93	285.80	369.87	584.67	2,550.58
Age class 17								
Reading/writing	5,288	884.82	554.75	99.25	541.10	755.38	1,063.51	7,560.85
Mathematics/science	3,795	895.49	489.32	150.22	175.46	547.25	736.22	4,848.16
Excluded students	560	639.28	371.19	99.25	374.79	555.72	769.46	2,064.74

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-54a. Distribution of student nonresponse adjustment factors, NAEP long-term trend samples: 1999

Sample	Number of cases	Mean	Standard deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Age class 9								
Reading/writing	5,793	1.05	0.03	1.0000	1.0436	1.0584	1.0613	1.2360
Mathematics/science	6,032	1.06	0.03	1.0269	1.0508	1.0648	1.0744	1.2364
Excluded students	1,120	1.14	0.13	1.0000	1.0266	1.0631	1.2578	1.3710
Age class 13								
Reading/writing	5,933	1.08	0.03	1.0000	1.0678	1.0882	1.1080	1.1807
Mathematics/science	5,941	1.07	0.03	1.0000	1.0561	1.0784	1.0962	1.2215
Excluded students	824	1.20	0.26	1.0000	1.0122	1.0574	1.5001	1.7228
Age class 17								
Reading/writing	5,288	1.23	0.09	1.1018	1.1811	1.1948	1.3064	1.4307
Mathematics/science	3,795	1.21	0.08	1.0886	1.1487	1.2203	1.3053	1.4674
Excluded students	560	1.23	0.30	1.0000	1.0617	1.0730	1.2341	1.8411

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Table D-54b. Distribution of student weight trimming factors, NAEP long-term trend samples: 1999

Sample	Number of cases	Mean	Standard deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Age class 9								
Reading/writing	5,793	0.99	0.01	0.8169	1.0000	1.0000	1.0000	1.0000
Mathematics/science	6,032	0.99	0.00	0.8867	1.0000	1.0000	1.0000	1.0000
Excluded students	1,120	0.95	0.12	0.3984	1.0000	1.0000	1.0000	1.0000
Age class 13								
Reading/writing	5,933	1.00	0.00	1.0000	1.0000	1.0000	1.0000	1.0000
Mathematics/science	5,941	0.99	0.00	0.9809	1.0000	1.0000	1.0000	1.0000
Excluded students	824	0.98	0.06	0.6252	1.0000	1.0000	1.0000	1.0000
Age class 17								
Reading/writing	5,288	1.00	0.00	1.0000	1.0000	1.0000	1.0000	1.0000
Mathematics/science	3,795	1.00	0.00	1.0000	1.0000	1.0000	1.0000	1.0000
Excluded students	560	0.99	0.04	0.7530	1.0000	1.0000	1.0000	1.0000

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

THIS PAGE INTENTIONALLY LEFT BLANK.

Appendix E

NAEP Report of Processing and Professional Scoring Activities[1]

Long-Term Trend
1998–99 Math/Science and Reading/Writing

National Computer Systems
March 2000

[1] This report was submitted to ETS by National Computer Systems, subcontractor for the processing and professional scoring for the NAEP's 1999 Long–Term Trend Assessment. A copy of the full report can be obtained by contacting Connie Smith, National Computer Systems, 2510 North Dodge Street, Iowa City, IA 52240 (crsmith@ncspearson.com).

THIS PAGE INTENTIONALLY LEFT BLANK.

E.1. Introduction

In 1998/9, the national component of the National Assessment of Educational Progress (NAEP) included the Long–term Trend (L–TT) Assessment. The L–TT can date itself back to 1969. As a result, the outcome of the assessment is compared to that of other years. These assessments include mathematics, science, reading, and writing in grades 4, 8, and 11 (ages 9, 13, and 17). The 4th grade (age 9) assessment is given in the winter (January through early March, 1999); the 8th grade (age 13) assessment is given in the fall (October through December, 1998); while the 11th grade (age 17) assessment is given in the spring (March through May, 1999).

There were 18 reading/writing and 8 math/science booklets. There were more than 40,000 reading/writing forms printed. The demographic and multiple–choice information was captured using the Falcon key–entry system. All scoring for these booklets was completed using the PSC paper–based system. There were just less than 40,000 math/science forms printed. These were OMR, ICR, and Image scannable forms but scoring was accomplished using the PSC paper–based system. The decision to use the paper–based scoring system was made to hold the trend line consistent with other years' scoring process.

Scoring for the L–TT assessment occurred after all materials were received: fall trend scoring occurred in mid–December, 1998; winter trend scoring occurred in March, 1999; while spring trend scoring occurred in May, 1999. Approximately 381,000 open–ended responses were scored during the three scoring windows in reading/writing and math/science. Holistics and Mechanics were not done with this year's Long–term Trend Writing assessment.

In addition, 25,000 copies of various tracking and/or questionnaires were printed. These forms include the Administration Schedule, the Roster of Questionnaires, the Excluded Student Questionnaire (ESQ), the Grade 4 School Characteristics and Policies Questionnaire (SCPQ), the Grade 8 SCPQ, and the Grade 11 SCPQ.

Figure E–1. NAEP long–term trend math/science and reading/writing schedule: 1998–99

Task Name	Start Date	Finish Date	Actual Start	Actual Finish
Long-Term Trend	7/1/98	5/10/99	7/1/98	6/8/99
Task 17 – Print/Pack/Ship	7/1/98	2/23/99	7/27/98	2/23/99
Printing	7/1/98	9/30/98	7/27/98	9/29/98
Develop/Modify Covers/Rosters/Schedules etc.	7/20/98	7/23/98	7/15/98	8/12/98
Approval Received for Covers/Rosters/Admin Schedules	7/24/98	7/24/98	8/4/98	8/12/98
Print Rosters	7/24/98	8/28/98	8/5/98	8/26/98
Print Administration Schedule	7/24/98	8/28/98	8/4/98	8/26/98
R/W Fall Trend Books sent/received – Vendor	7/28/98	8/17/98	8/20/98	8/20/98
M/S Fall Trend Books sent/received – Columbia	7/28/98	8/25/98	8/20/98	8/26/98
Questionnaires sent/received – Columbia	7/28/98	8/28/98	8/21/98	8/26/98
R/W Winter Trend Books sent and received from printer	8/15/98	9/30/98	8/18/98	9/14/98
M/S Winter Trend Books sent/received – Columbia	8/15/98	9/30/98	8/18/98	9/29/98
School Ques–WT sent/received – Columbia	8/15/98	9/30/98	8/18/98	9/23/98
R/W Spring Trend Books sent/received – Vendor	8/15/98	9/30/98	8/20/98	9/14/98
M/S Spring Trend Books sent/received – Columbia	8/15/98	9/30/98	8/20/98	9/29/98
School Ques–ST sent/received – Columbia	8/15/98	9/30/98	8/20/98	9/29/98
Package/Distribute	8/3/98	2/23/99	8/17/98	2/23/99
Fall Trend	8/11/98	9/18/98	8/17/98	9/17/98
Packaging Kick–off Meeting	8/17/98	8/17/98	8/17/98	8/17/98
Materials Lists Delivered– Fall, Winter, Spring	8/17/98	8/17/98	8/17/98	8/17/98
Packaging Specs	8/17/98	8/17/98	8/17/98	9/14/98
Blue Dot –	9/1/98	9/17/98	8/28/98	9/17/98
Pre–Packaging	9/1/98	9/1/98	9/14/98	9/14/98
Barcoding	8/24/98	8/24/98	8/28/98	8/28/98
Spiraling	8/27/98	8/27/98	8/31/98	8/31/98
Final Packaging	9/17/98	9/17/98	9/17/98	9/17/98
Pre–Packaging	9/1/98	9/4/98	9/1/98	9/14/98
Barcoding	8/24/98	8/25/98	8/28/98	9/1/98
Spiraling	8/24/98	9/1/98	8/31/98	9/1/98
Final Packaging	8/16/98	8/18/98	9/17/98	9/17/98
Session Data File from Westat	8/25/98	8/25/98	8/25/98	8/25/98
Supervisor Address File from Westat	8/27/98	8/27/98	8/26/98	8/26/98
Bulk Shipment Address file from Westat	8/27/98	8/27/98	8/26/98	8/26/98
Ship Admin Sched/Rosters/Ques.	9/10/98	9/10/98	9/9/98	9/9/98
Ship Bulk Material	9/18/98	9/18/98	9/16/98	9/16/98
Ship Session Material	9/18/98	9/18/98	9/17/98	9/17/98
Winter Trend	9/7/98	12/8/98	9/25/98	12/9/98
Packaging Specs	9/25/98	9/25/98	9/25/98	10/2/98
Blue Dot –	10/1/98	12/7/98	10/1/98	10/1/98
Pre–Packaging	11/23/98	11/23/98	12/9/98	12/9/98
Barcoding	10/1/98	10/1/98	10/1/98	10/1/98
Spiraling	10/3/98	10/3/98	10/5/98	10/5/98
Final Packaging	12/7/98	12/7/98	12/9/98	12/9/98
Pre–Packaging	11/24/98	11/30/98	12/9/98	12/9/98
Barcoding	10/1/98	10/5/98	10/1/98	10/2/98
Spiraling	10/2/98	10/6/98	10/5/98	10/7/98
Final Packaging	12/7/98	12/8/98	12/9/98	12/9/98
Final Session Data File from Westat	11/23/98	11/23/98	11/23/98	11/23/98
Supervisor Address File from Westat	11/30/98	11/30/98	11/30/98	11/30/98

See notes at end of figure→

Figure E–1. NAEP long–term trend math/science and reading/writing schedule: 1998–99— Continued

Task Name	Start Date	Finish Date	Actual Start	Actual Finish
Admin. Schedule Address file from Westat	11/30/98	11/30/98	11/30/98	11/30/98
Bulk Shipment Address File from Westat	11/30/98	11/30/98	11/30/98	11/30/98
Ship Admin. Schedules/Rosters/Questionnaires	12/2/98	12/2/98	12/2/98	12/2/98
Ship Bulk Material	12/7/98	12/7/98	12/9/98	12/9/98
Ship Session Material	12/7/98	12/7/98	12/9/98	12/9/98
Spring Trend	9/7/99	2/23/99	10/1/98	2/23/99
Packaging Specs	2/1/98	2/1/98	1/24/99	1/24/99
Blue Dot –	10/1/98	2/23/99	10/1/98	10/1/98
Pre–Packaging	2/1/99	2/1/99	2/1/99	2/1/99
Barcoding	10/1/98	10/1/98	10/1/98	10/1/98
Spiraling	10/3/98	10/3/98	10/5/98	10/5/98
Final Packaging	2/23/99	2/23/99	2/23/99	2/23/99
Pre–Packaging	2/1/99	2/2/99	2/1/99	2/2/99
Barcoding	10/1/98	10/5/98	10/2/98	10/6/98
Spiraling	10/3/98	10/6/98	10/6/98	10/7/98
Final Packaging	2/22/99	2/24/99	2/23/99	2/23/99
Final Session Data File from Westat	2/5/99	2/5/99	2/5/99	2/5/99
Supervisor Address File from Westat	2/8/99	2/8/99	2/5/99	2/5/99
Admin. Schedule Address File from Westat	2/8/99	2/8/99	2/5/99	2/5/99
Ship Admin. Schedules/Rosters/Questionnaires	2/15/99	2/15/99	2/15/99	2/15/99
Ship Bulk Material	2/23/99	2/23/99	2/22/99	2/22/99
Ship Session Material	2/23/99	2/23/99	2/22/99	2/22/99
Task 18 – Receipt Control and Tracking	7/1/98	5/10/99		
Develop/Modify Receipt Control System	8/25/98	9/30/98	8/25/98	10/1/98
Fall Trend	7/1/98	12/21/98	7/1/98	12/22/98
Processing Specs Complete	8/17/98	9/25/98	8/17/98	9/30/98
Test Administration	10/12/98	12/18/98	10/8/98	12/18/99
Cut–off Dates for Questionnaires	1/13/99	1/13/99	1/13/99	1/13/99
Document Receipt(Rcpt/Dock Sort/Open/Log)	10/14/98	12/21/98	10/12/98	12/22/98
Winter Trend	7/1/98	3/22/99	7/1/98	3/22/99
Processing Specs Complete	12/15/98	12/15/98	8/17/98	9/30/98
Test Administration	1/4/99	3/12/99	1/14/99	3/12/99
Cut–off Dates for Questionnaires	3/26/99	3/26/99	3/26/99	3/26/99
Document Receipt(Rcpt/Dock Sort/Open/Log)	1/6/99	3/16/99	1/6/99	3/31/99
Spring Trend	10/1/98	5/10/99	10/1/98	5/28/99
Processing Specs Complete	3/1/98	3/1/98	8/17/98	9/30/98
Test Administration	3/15/99	5/14/99	3/15/99	5/17/99
Cut–off Dates for Questionnaires	5/21/99	5/21/99	5/24/99	5/24/99
Document Receipt(Rcpt/Dock Sort/Open/Log)	3/17/99	5/18/99	3/17/99	5/19/99
Task 19 – Professional Scoring	12/7/98	5/21/99	12/7/98	5/28/99
Fall Trend				
Rescore Pulls	9/18/98	11/1/98	9/18/98	11/1/98
Training Read/Writing	12/7/98	12/9/98	12/7/98	12/10/98
Scoring	12/10/98	12/31/98	12/11/98	12/31/98
Math Training and Scoring	12/21/98	12/23/98	12/21/98	12/23/98

See notes at end of figure→

Figure E–1. NAEP long–term trend math/science and reading/writing schedule: 1998–99—Continued

Task Name	Start Date	Finish Date	Actual Start	Actual Finish
Winter Trend				
Rescore Pulls	9/25/98	12/2/98	9/28/98	12/2/98
Training Read/Writing	3/22/99	3/23/99	3/22/99	3/23/99
Scoring	3/24/99	4/2/99	3/24/99	4/2/99
Math Training and Scoring	3/29/99	4/2/99	3/25/99	4/2/99
Spring Trend				
Rescore Pulls	9/25/98	1/5/99	9/28/98	3/26/99
Training Read/Writing	5/3/99	5/4/99	5/3/99	5/4/99
Scoring	5/5/99	5/28/99	5/5/99	5/28/99
Math Training and Scoring	5/17/99	5/21/99	5/17/99	5/28/99
Task 20 – Processing	6/12/98	3/12/99	10/23/98	5/19/99
Planning/Development	6/12/98	3/12/99	7/1/98	5/19/99
NCS Receives File Format from ETS	8/1/98	8/1/98	7/6/98	7/31/98
NCS Receives List of Data Elements to Deliver from ETS	8/1/98	8/1/98	7/15/98	7/31/98
NCS Receives List of Data Elements to Deliver from Westat	8/1/98	8/1/98	7/15/98	7/15/98
Scanning	10/20/98	3/12/99	10/23/98	5/19/99
Blue Dot– Fall Trend	10/19/98	11/5/98	10/23/98	10/29/98
Math/Science	10/19/98	10/23/98	10/23/98	10/29/98
Gr8 School Ques.	10/23/98	10/30/98	11/12/98	11/16/98
Excluded Student Ques.	11/5/98	11/12/98	11/3/98	12/23/98
Rosters	10/23/98	10/30/98	10/28/98	10/30/98
Administration Schedules	10/19/98	10/23/98	10/23/98	10/30/98
Blue Dot – Winter Trend	1/11/99	1/15/99	1/11/99	1/19/99
Math/Science	1/11/99	1/15/99	1/11/99	1/15/99
Gr4 School Ques.	1/15/99	1/19/99	1/15/99	1/19/99
Blue Dot – Spring Trend	3/22/99	4/2/99	3/22/99	4/2/99
Math/Science	3/22/99	3/26/99	3/2/99	3/26/99
Gr11 School Ques.	3/26/99	4/2/99	3/26/99	4/2/99
* Scanning/Processing	10/23/98	5/18/99	10/23/99	5/19/99
Fall Trend Math/Science	10/23/98	12/21/98	10/29/98	12/21/98
– Through Clean Post	12/23/98	12/23/98	12/23/98	12/23/98
Fall Trend Gr8 School Questionnaires	10/30/98	12/21/98	11/16/98	1/15/99
– Through Clean Post	1/15/99	1/15/99	1/15/99	1/15/99
Winter Trend Math/Science	1/15/99	3/24/99	1/15/99	3/24/99
– Through Clean Post	3/24/99	3/24/99	4/16/99	4/16/99
Winter Trend Gr4 School Questionnaires	1/19/99	3/24/99	1/19/99	3/24/99
– Through Clean Post	3/24/99	3/24/99	3/24/99	3/24/99
Spring Trend Math/Science	3/26/99	5/18/99	3/26/99	5/19/99
– Through Clean Post	5/19/98	5/19/98	5/27/99	5/27/99
Spring Trend Gr11 School Questionnaires	4/2/99	5/18/99	4/2/99	5/19/99
– Through Clean Post	5/19/99	5/19/99	6/9/99	6/9/99
ESQ's	11/12/98	5/18/99	11/4/98	5/25/99
– Through Clean Post – Fall	12/31/98	12/31/98	12/31/98	12/31/98
– Through Clean Post – Winter	3/25/99	3/25/99	3/25/99	3/25/99
– Through Clean Post – Spring	5/25/99	5/25/99	5/25/99	5/25/99

See notes at end of figure→

Figure E–1. NAEP Long–term Trend Math/Science and Reading/Writing Schedule: 1998–99—Continued

Task Name	Start Date	Finish Date	Actual Start	Actual Finish
Rosters	10/30/98	5/18/99	10/30/98	5/27/99
– Through Clean Post – Fall	12/29/98	12/29/98	12/29/98	12/29/98
– Through Clean Post – Winter	3/24/99	3/24/99	3/24/99	3/24/99
– Through Clean Post – Spring	5/27/99	5/27/99	5/27/99	5/27/99
Administration Schedules	10/23/98	5/14/99	10/30/98	5/28/99
– Through Clean Post – Fall	1/18/98	1/18/98	1/18/98	1/18/98
– Through Clean Post – Winter	3/24/99	3/24/99	3/24/99	3/24/99
– Through Clean Post – Spring	5/28/99	5/28/99	5/28/99	5/28/99
Key Entry	10/16/98	5/12/99	10/16/98	5/28/99
Key Entry Screen Setup to Data Input–Fall Trend	9/28/98	9/28/98	9/28/98	9/28/98
Blue Dot– Fall Trend R/W	10/19/98	10/23/98	10/23/98	10/30/98
Fall Trend Reading/Writing Processing	10/23/98	12/21/98	10/30/98	12/31/98
–Through Clean Post	12/28/98	12/28/98	1/6/99	1/6/99
Key Entry Screen Setup to Data Input–Winter Trend	12/28/98	12/28/98	11/4/98	11/11/98
Blue Dot – Winter Trend R/W	1/11/99	1/15/99	1/11/99	1/15/99
Winter Trend Reading/Writing Processing	1/15/99	3/17/99	1/15/99	3/17/99
–Through Clean Post	3/18/99	3/18/99	3/18/99	3/18/99
Key Entry Screen Setup to Data Input–Spring Trend	3/8/99	3/8/99	3/8/99	3/8/99
Blue Dot – Spring Trend R/W	3/22/99	3/26/99	3/22/99	3/26/99
Spring Trend Reading/Writing Processing	3/26/99	5/19/99	3/26/99	5/19/99
–Through Clean Post	5/19/99	5/19/99	5/19/99	5/19/99
Ship Score Data Tape to ETS				
Fall Trend Data	1/18/99	1/18/99	1/22/99	1/27/99
Winter Trend Data	4/12/99	4/12/99	4/12/99	4/12/99
Spring Trend Data	5/28/99	5/28/99	6/8/99	6/8/99
Ship Weights Data Tape to Westat				
Fall Trend Weights Data	1/15/99	1/15/99	1/15/99	1/15/99
Winter Trend Weights Data	4/5/99	4/5/99	4/2/99	4/2/99
Spring Trend Weights Data	5/24/99	5/24/99	5/28/99	5/28/99
Ship QC Books To ETS				
Fall Trend	7/1/99	7/1/99	7/1/99	7/1/99
Winter Trend	7/1/99	7/1/99	7/1/99	7/1/99
Spring Trend	7/1/99	7/1/99	7/1/99	7/1/99

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

E.2. Printing

Printing preparations began with the design of the booklet covers in June 1998. 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.

The Long-Term Trend assessments included 26 assessment booklets, the Administration Schedule, the Excluded Student Questionnaire (ESQ), three School Characteristics and Policies Questionnaires, and the Roster of Questionnaires. All materials for the Long-Term Trend assessments were printed by September 28, 1998.

The 26 booklets used for the three Long-Term Trend assessments were direct reprints of booklets used in previous years' assessments. Only the front covers were redesigned for the 1998–99 assessments. Eighteen of these 26 booklets were non–scannable; the other eight were scannable.

Figure E–2. NAEP long–term trend math/science and reading/writing printed documents: 1998–99

N C S Inventory Number	N C S Document Code	Grade /Age	Document Description	Type	Sample per Book	Est. No. Pages	ACTUAL No. Pages	Printing Method (PrinTech™, offset, etc.)	Type of Document	Total Print Quantity	Book to Printer	Proof from Printer	Approval to Print	Doc. Ship/ Receipt Date	Pntd Samples Distributed*: W=2,MS=2,CB=2, PR=2, LH=2
			Long–Term Trend												
NA9000	163426–001	all	Admin Schedule–Trend	Long-Term Trend	—	2	2	offset	I C R/Image	10,425	7/26/94	8/2/94	8/3/94	8/25/94	8/25/94
NA9001	163427–001	all	Roster of Quest–Trend	Long-Term Trend	—	2	2	offset	I C R	5,255	7/26/94	8/3/94	8/4/94	8/23/94	8/25/94
NA9002	36760–405	all	ESQ–Trend	Long-Term Trend	—	4	4	offset	I C R/Image	4,200	7/27/94	8/2/94	8/4/94	8/25/94	8/25/94
NA9003	153876–203	4	Trend SCPQ–Gr 4	Long-Term Trend	—	12	12	offset	I C R/Image	1,577	8/17/94	9/2/94	9/7/94	9/21/94	9/22/94
NA9004	153593–203	8	Trend SCPQ–Gr 8	Long-Term Trend	—	12	12	offset	I C R/Image	1,570	7/27/94	8/2/94	8/4/94	8/23/94	8/25/94
NA9005	153875–203	11	Trend SCPQ–Gr 11	Long-Term Trend	—	12	12	offset	I C R/Image	1,577	8/19/94	8/30/94	9/1/94	9/27/94	9/28/94
NA9006	—	115	R/W Gr8 Bk 51W	Long-Term Trend	867	32	32	offset	Key	2,595	7/29/94	8/4/94	8/10/94	8/19/94	8/23/94
NA9007	—	115	R/W Gr8 Bk 52W	Long-Term Trend	867	32	32	offset	Key	2,595	7/29/94	8/4/94	8/10/94	8/19/94	8/23/94
NA9008	—	115	R/W Gr8 Bk 53W	Long-Term Trend	867	32	32	offset	Key	2,594	7/29/94	8/4/94	8/11/94	8/19/94	8/23/94
NA9009	—	115	R/W Gr8 Bk 54W	Long-Term Trend	867	32	32	offset	Key	2,595	7/29/94	8/4/94	8/10/94	8/19/94	8/23/94
NA9010	—	115	R/W Gr8 Bk 55W	Long-Term Trend	867	32	32	offset	Key	2,595	7/29/94	8/4/94	8/10/94	8/19/94	8/23/94

See notes at end of figure→

Figure E–2. NAEP long–term trend math/science and reading/writing printed documents: 1998–99—Continued

N C S Inventory Number	N C S Document Code	Grade /Age	Document Description	Type	Sample per Book	Est. No. Pages	ACTUAL No. Pages	Printing Method (PrinTech™, offset, etc.)	Type of Document	Total Print Quantity	Book to Printer	Proof from Printer	Approval to Print	Doc. Ship/ Receipt Date	Pntd Samples Distributed*: W=2,MS=2,CB=2, PR=2, LH=2
			Long-Term Trend												
NA9011	—	115	R/W Gr8 Bk 56W	Long-Term Trend	867	32	32	offset	Key	2,595	7/29/94	8/4/94	8/10/94	8/19/94	8/23/94
NA9012	36684–405	Age 13	M/S Ag13 Bk 91T	Long-Term Trend	2000	44	44	offset	I C R/Image	5,000	7/26/94	8/2/94	8/3/94	8/25/94	8/25/94
NA9013	36685–405	Age 13	M/S Ag13 Bk 92TC	Long-Term Trend	2000	36	36	offset	I C R/Image	5,037	7/26/94	8/2/94	8/3/94	8/23/94	8/24/94
NA9014	36683–405	Age 13	M/S Ag13 Bk 93T	Long-Term Trend	2000	48	48	offset	I C R/Image	5,252	7/26/94	8/2/94	8/3/94	8/23/94	8/24/94
NA9015	—	4	R/W Gr4 Bk 51W	Long-Term Trend	867	28	28	offset	Key	2,048	8/17/94	8/25/94	8/30/94	9/16/94	9/17/94
NA9016	—	4	R/W Gr4 Bk 52W	Long-Term Trend	867	28	28	offset	Key	2,044	8/17/94	8/26/94	8/30/94	9/10/94	9/13/94
NA9017	—	4	R/W Gr4 Bk 53W	Long-Term Trend	867	28	28	offset	Key	2,090	8/17/94	8/25/94	8/30/94	9/10/94	9/13/94
NA9018	—	4	R/W Gr4 Bk 54W	Long-Term Trend	867	28	28	offset	Key	2,089	8/17/94	8/26/94	8/30/94	9/9/94	9/10/94
NA9019	—	4	R/W Gr4 Bk 55W	Long-Term Trend	867	32	32	offset	Key	2,094	8/17/94	8/26/94	8/30/94	9/10/94	9/13/94
NA9020	—	4	R/W Gr4 Bk 56W	Long-Term Trend	867	28	28	offset	Key	2,030	8/17/94	8/25/94	8/30/94	9/9/94	9/10/94
NA9021	37401–405	Age 9	M/S Ag9 Bk 91T	Long-Term Trend	2000	32	32	offset	I C R/Image	4,622	8/17/94	8/25/94	9/2/94	9/27/94	9/28/94
NA9022	37040–405	Age 9	M/S Ag9 Bk 92TC	Long-Term Trend	2000	32	32	offset	I C R/Image	4,622	8/17/94	8/25/94	8/27/94	9/27/94	9/28/94

See notes at end of figure→

Figure E–2. NAEP long–term trend math/science and reading/writing printed documents: 1998–99—Continued

NA9023	37038–405	Age 9	M/S Ag9 Bk 93T	Long-Term Trend	2000	36	36	offset	I C R/Image	4,522	8/17/94	8/25/94	8/27/94	9/27/94	9/28/94
NA9024	—	147	R/W Gr11 Bk 51W	Long-Term Trend	867	32	32	offset	Key	2,084	8/19/94	8/26/94	8/30/94	9/10/94	9/13/94
NA9025	—	147	R/W Gr11 Bk 52W	Long-Term Trend	867	32	32	offset	Key	2,033	8/19/94	8/26/94	8/30/94	9/9/94	9/10/94
NA9026	—	147	R/W Gr11 Bk 53W	Long-Term Trend	867	28	28	offset	Key	2,029	8/19/94	8/26/94	8/30/94	9/9/94	9/10/94
NA9027	—	147	R/W Gr11 Bk 54W	Long-Term Trend	867	36	36	offset	Key	2,063	8/19/94	8/25/94	8/30/94	9/9/94	9/10/94
NA9028	—	147	R/W Gr11 Bk 55W	Long-Term Trend	867	28	28	offset	Key	2,070	8/19/94	8/25/94	8/30/94	9/10/94	9/13/94
NA9029	—	147	R/W Gr11 Bk 56W	Long-Term Trend	867	32	32	offset	Key	2,064	8/19/94	8/25/94	8/30/94	9/9/94	9/10/94
NA9030	377224–05	Age 17	M/S Ag17 Bk 84T	Long-Term Trend	2000	48	48	offset	I C R/Image	4,475	8/19/94	8/30/94	9/1/94	9/27/94	9/27/94
NA9031	377354–05	Age 17	M/S Ag17 Bk 85TC	Long-Term Trend	2000	40	40	offset	I C R/Image	4,622	8/19/94	8/30/94	9/1/94	9/27/94	9/28/94

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

THIS PAGE INTENTIONALLY LEFT BLANK.

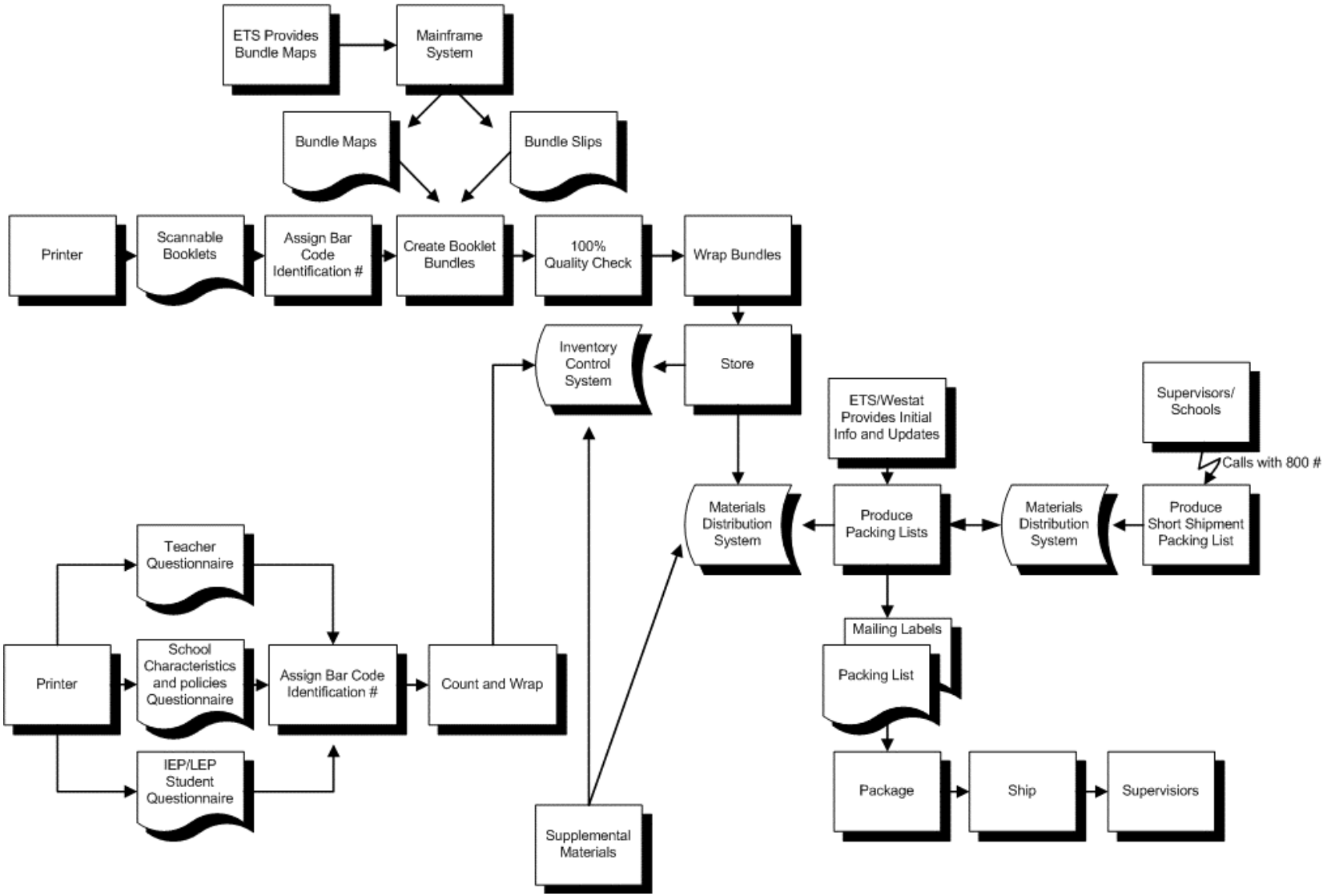
E.3. Packaging, Distribution and Short Shipments

E.3.1 Packaging and Distribution

The distribution effort for the 1999 NAEP Long-Term Trend assessment involved packaging and shipping documents and associated forms to the Westat supervisors. The NCS NAEP Materials Distribution System (MDS), initially developed by NCS in 1990 to control shipments to the schools and supervisors, was utilized again in 1998/99. 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. Figure E–3 illustrates the Packaging and Distribution flow for the 1998–99 Long–term Trend.

THIS PAGE INTENTIONALLY LEFT BLANK.

Figure E–3. NAEP long–term trend packaging/distribution process flow: 1998–99



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

THIS PAGE INTENTIONALLY LEFT BLANK.

Bar code technology continued to be utilized in document control. To identify each document, NCS imprinted a unique ten–digit booklet number or form type consisting of a three digit prefix/book type identifier, a six digit sequential number, and a check digit. Each form was assigned a range of ID numbers. Bar codes reflecting all ID numbers 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 eleven documents. 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 that the booklets were evenly distributed across the bundles. This assured that sample sizes of individual book types would not be jeopardized if entire bundles were not used. Each bundle of documents contained a bundle slip/header sheet that indicated the following:

- Subject area
- Bundle type
- Bundle number
- Unique bar code number
- First three digits of each booklet type in the bundle

All booklets were arranged in the exact order listed on the bundle header sheet. To ensure the accuracy of each bundle and the security of the NAEP assessment, a quality control plan was utilized to verify the document order of each bundle and to account for all booklets. All bundles that contained a bundle slip were taken to a bar code reader/document transport machine where they were scanned to interpret each bundle's bar code. The file of scanned bar codes was then transferred from the personal computer connected to the scanner to a mainframe data set. The unique bundle number on the header sheet informed the system program what type of bundle should follow. A computer job was run to compare the bundle type expected to the sequence of booklets that was scanned after the header. This job also verified that the appropriate number of booklets was included in each bundle. Any discrepancies were printed on an error listing. The NCS packaging department corrected the error and the bundle was again read into the system. This process was repeated until no discrepancies existed. By using this quality–control plan, NCS could verify the document order of each bundle and account for all booklets. See Figure E–4 for 1998/99 NAEP Long-Term Trend Bundle Types and Distribution by Session.

Once all bundles for a subject area passed the bundle QC process, information from the bundle QC file was uploaded to the mainframe computer system and used in the creation of administration schedules. All administration schedules for each scheduled session were pre–printed with the booklet ID's designated for that session. Three bundles of booklets were pre–assigned to each Reading/Writing session, giving each session 33 booklets. Two bundles were pre–assigned to each Math/Science session, giving each session 22 booklets. These numbers most closely approximated the average projected session size plus an additional supply of booklets for any extra students.

Using sampling files provided by Westat, NCS assigned bundles to schools and customized the packing lists. File data from Westat was coupled with the file of bundle numbers and the corresponding booklet numbers. This file was then used to pre–print all booklet identification numbers, school name, school number and session type, directly on to the scannable administration schedule. As a result, every pre–scheduled session had specific bundles assigned to it in advance. This increased the quality level of the booklet accountability system by enabling NCS to identify where any booklet should be at any time during the assessment. It also eliminated the possibility of transcription errors by field staff and assessment administrators for booklet ID numbers. Lastly, by pre–printing booklet ID numbers, the burden on the Westat field staff for transcription of data was notably reduced. NCS distributed the pre–printed administration schedules to Westat supervisors before their session material arrived. This assisted them with sampling in the schools.

NCS was also responsible for the packaging and distribution of bulk materials for use by the Westat supervisors for the Long-Term Trend assessment. Bulk shipments included materials that could be reused by supervisors from one session to another, such as audio tapes, tape recorders and additional booklets to accommodate any students added to a session or to replace defective booklets or materials. As with session shipments, NCS packaging staff pre–assembled materials into the appropriate–sized grouping for distribution prior to final packaging. Distribution of materials for the Long-Term Trend assessments was accomplished in three phases. Initial distribution included a bulk shipment and session materials for all schools tested in the Fall session. Winter sessions were sent out in mid December and Spring sessions were sent in mid March. Figure E–4 illustrates the Bulk Materials shipped by NCS. Figure E–5 illustrates the amount of materials shipped to each session.

Figure E–4. NAEP long–term trend bulk materials: 1998–99

Item Description	Quantity Shipped in Bulk to each Supervisor
General Bulk	
Calculators – Simple TI–108	75
GE Tape Recorder	2
"AA" batteries	4
Digital Timers	5
Express Mail Labels	10
Fed Ex Labels	50
Plastic Sleeve/Fed Ex Labels	50
Laminated "Do Not Disturb" Signs	10
Rubberbands	100
Sealing Tape – Rolls	3
Tape Dispensers	2
Administration Schedules	10
Roster of Questionnaires	10
#2 Pencils	1,440
Fall Trend Bulk	
Gr. 8 R/W Spiral Bundle	5
Age 13 Bk 91T M/S Bundle	3
Age 13 Bk 92TC M/S Bundle	3
Age 13 Bk 93T M/S Bundle	3
Stimulus Tape Book 91T	2
Stimulus Tape Book 92TC	2
Stimulus Tape Book 93T	2
Winter Trend Bulk	
Gr. 4 R/W Sprial Bundle	5
Age 9 Bk 91T M/S Bundle	3
Age 9 Bk 92TC M/S Bundle	3
Age 9 Bk 93T M/S Bundle	3
Stimulus Tape Book 91T	2
Stimulus Tape Book 92TC	2
Stimulus Tape Book 93T	2
Fed Ex Labels	30
Fed Ex Plastic Sleeves	30
Spring Trend Bulk	
Gr. 11 R/W Spiral Bundle	5
Age 17 84T M/S Bundle	3
Age 17 85TC M/S Bundle	3
Stimulus Tape Bk 84T	2
Stimulus Tape Bk 85TC	2
Fed Ex Labels	30
Fed Ex Plastic Sleeves	30
#2 Pencils	720

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

Figure E–5. NAEP long–term trend materials shipped by session: 1998–99

Long-Term Trend	Item Discription	Quantity Distributed per Session
Fall Trend	Gr. 8 Bks 51–56 R/W Spiral Bundle	3 bundles
	Age 13 Bk 91T M/S Bundle	2 bundles
	Age 13 Bk 92TC M/S Bundle	2 bundles
	Age 13 Bk 93T M/S Bundle	2 bundles
Winter Trend	Gr. 4 Bks 51–56 R/W Spiral Bundle	3 bundles
	Age 9 Bk 91T M/S Bundle	2 bundles
	Age 9 Bk 92TC M/S Bundle	2 bundles
	Age 9 Bk 93T M/S Bundle	2 bundles
Spring Trend	Gr. 11 Bk 51–56 R/W Spiral Bundle	3 bundles
	Age 17 Bk 84T M/S Bundle	2 bundles
	Age 17 Bk 85TC M/S Bundle	2 bundles

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

A total of 2,187 sessions were shipped to 806 schools for the Fall, Winter and Spring assessments. An additional 210 short shipments were sent during the Trend assessments.

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.

E.3.2 Toll–Free Line, E–mail and Short Shipments

A toll–free telephone line was maintained for Westat staff to request additional materials for the Trend Assessment. NCS also set up an e–mail address for additional material requests. A total of 163 short shipments were sent during the assessment. To process a shipment, NCS phone staff asked the caller for information such as PSU, school ID, assessment type, city, state, and zip code. This information was then entered into the on–line short shipment system and the 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, which 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. Figure E–6 lists the total number of inventory items sent out for short shipments during 1998/99 Long-Term Trend. Phone staff also took phone calls concerning shipment delivery dates, tracing of shipments and any questions concerning NAEP.

Figure E–6. NAEP long–term trend short shipment inventory items: 1998–99

Inventory Item	Quantity
"AA" Batteries	123
Rubber bands	605
Supplemental Shipping Envelopes	20
Simple Calculator TI–108	192
Sealing Tape Rolls	12
Tape Dispenser	1
Digital Timers	31
Tape Recorder	45
#2 Pencils	23,813
Stimulus Tape 91T Gr.8	9
Stimulus Tape 92TC Gr.8	8
Stimulus Tape 93T Gr.8	7
Stimulus Tape 91T Gr.4	7
Stimulus Tape 92TC Gr.4	7
Stimulus Tape 93T Gr.4	5
Stimulus Tape 84T Gr.11	8
Stimulus Tape 85TC Gr.11	7
Laminated "Do Not Disturb" Signs	19
Admin. Schedule	246
Roster of Questionnaires	136
Excluded Student Questionnaires	1,561
Trend SCPQ – Gr.4	86
Trend SCPQ – Gr.8	74
Trend SCPQ – Gr.11	44
R/W Gr.8 Spiral 51–56 Bundle	51
M/S Age 13 Bk 91T Bundle	14
M/S Age 13 Bk 92TC Bundle	9
M/S Age 13 Bk 93T Bundle	15
R/W Gr.4 Spiral 51–56 Bundle	91
M/S Age 9 Bk 91T Bundle	22
M/S Age 9 Bk 92T Bundle	32
M/S Age 9 Bk 93T Bundle	37
R/W Gr.11 Spiral 51–56 Bundle	21
M/S Age 17 Bk 84T Bundle	2
M/S Age 17 Bk 85TC Bundle	4
Fed Ex Return Labels	449
Fed Ex Plastic Sleeves	419

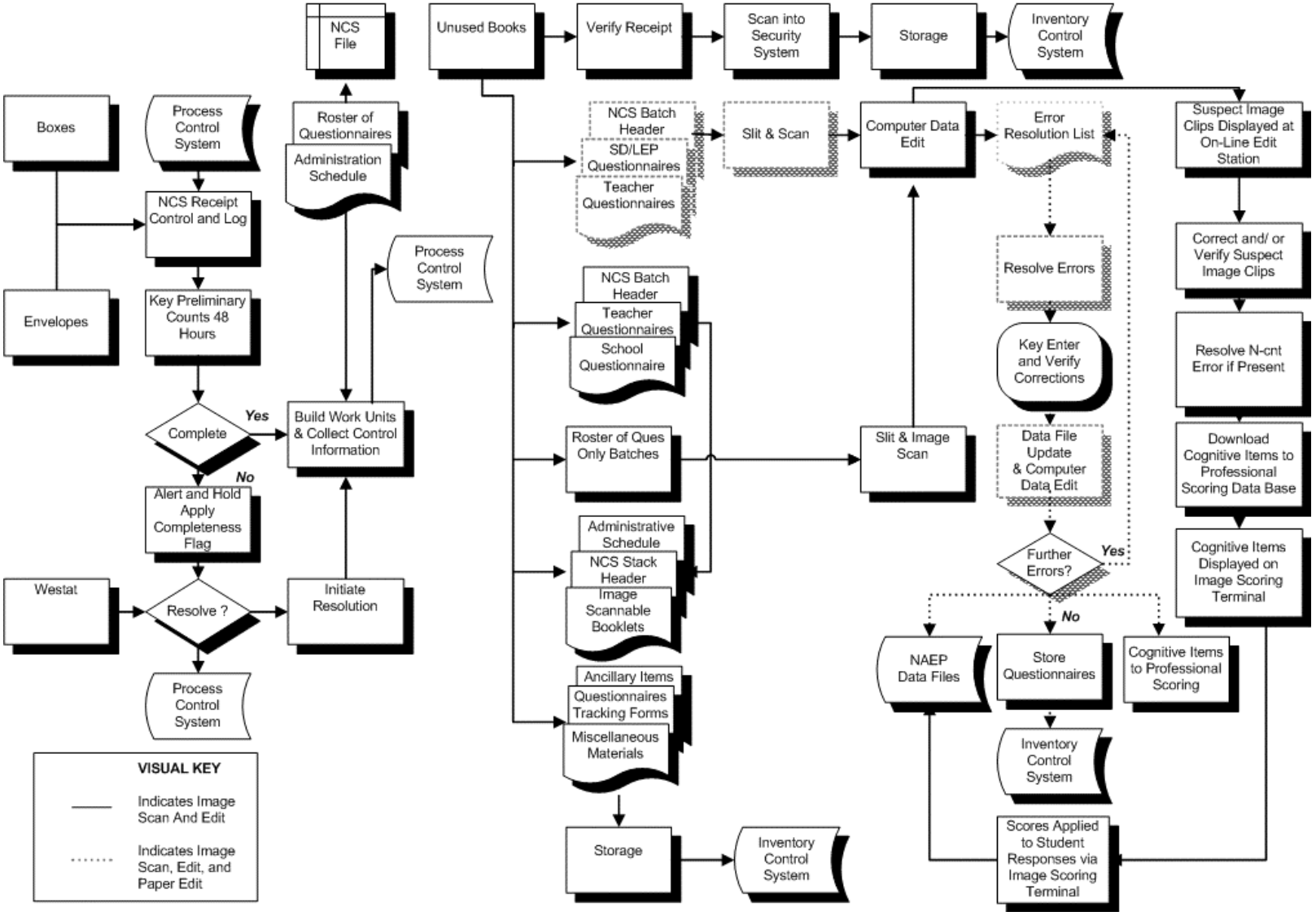
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

E.4. Processing

E.4.1 Overview

This chapter describes the various stages of work involved in receiving and processing the documents used in the 1999 NAEP Long-Term Trend 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. The processing flow is illustrated in figure E–7.

Figure E–7. NAEP long–term trend math/science and reading/writing processing flow chart: 1998–99



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long–Term Trend Assessment.

THIS PAGE INTENTIONALLY LEFT BLANK.

An Alerts 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 by 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" which were forwarded to project personnel for resolution. Once resolved, the problems and resolutions were recorded on–line in the PCS system.

NCS's Work Flow Management System (WFM) 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 workload, by project, across all workstations. By routinely monitoring this 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.

E.4.2 Document Receipt

Shipments were returned to NCS packaged in their original boxes. As mentioned in the earlier section on distribution, 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 to a personal computer (PC) file, after which the shipment was forwarded to the receiving area. The PC 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 Receiving department and supplied to Westat and ETS via electronic file transfer. The PCS 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.

Preliminary information, such as Number Assessed, Absent, Excluded, etc., was entered from the Administration Schedule into the PCS. This information was used to provide Westat with timely student response rates, it 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. The completeness flags used to identify problem sessions and their definitions are listed in figure E–8.

Figure E–8. NAEP long–term trend completeness flags: 1998–99

Completeness Flag	Definition
I = Incomplete	Entire session missing from school shipment Booklets listed on the administration schedule missing from an individual session
M = Held for Makeup	Booklets listed on the administration schedule with absent administration codes missing from a shipment (only used when documentation provided by Westat staff indicated that a make–up session was being held)
A = Alerted	Session held for an alert situation (not used for info–alert situations resolved by opening staff)
N = Not Administered	Schools with multiple sessions choosing not to do one of the sessions (not used if a school refused to do any of their scheduled sessions)

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

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 pre–printed or hand–written 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 sesssion documents arrived. If no make–up session was indicated, Westat was contacted for the disposition 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.

E.4.3 Batching and Scanning of Booklets

Once all the Math and Science Tape Session booklets listed on the administration schedule was verified as present, the entire session (both the administration schedule and booklets) was batched by grade level and session type. The administration schedule for these document types was used as a session header within a batch. Each batch was assigned a unique batch number. This number, created on the Image Capture Environment (ICE) system for all image–scannable documents. Since the Reading and Writing booklets were key–entry, these sessions were created on the Work Flow Management (WFM) system as were the OMR–Scannable documents. This facilitated the internal tracking of the batches and allowed departmental resource planning. All other scannable documents (School Characteristics and Policies Questionnaires, Excluded Student Questionnaires, and Rosters) were batched by document type in the same manner.

The administration schedules from Trend reading/writing sessions were processed in an Administration–Schedule–Only batch through the Image Scanning system. A computerized match occurred with Trend reading/writing materials once the Administration–Schedule–Only batch that contained a session's administration schedule passed through processing.

E.4.4 Batching and Scanning of Questionnaires

The 1998–99 NAEP Long-Term Trend assessments used one roster to account for all questionnaires. The Roster of Questionnaires for the Long-Term Trend assessments recorded the distribution and return of the School Characteristics and Policies Questionnaires (SCPQ) and the Excluded Student Questionnaires (ESQ).

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 ICE system. The batch was then forwarded to scanning where all information on the rosters was scanned into the system.

E.4.5 Booklet Accountability

In 1998–99, 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. Each unique booklet was batched and the booklet ID bar code was read into a file by the bar code scanner or manually key–entered. 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 from the Long-Term Trend assessment returned all unused materials. These booklets' IDs were also read into a file by the bar code scanner. Westat was notified of major discrepancies for follow–up. All unused materials received were then inventoried and sent to the NCS warehouse for storage while awaiting authorization from ETS to salvage them.

E.4.6 Data Transcription

The transcription of the student response data into machine–readable form was achieved through the use of the following three separate systems:

- Data Entry (which included OMR and image scanning, Intelligent Character Recognition [ICR], and key entry)
- Data Validation (edit)
- Data Resolution

These systems are described in the subsections that follow.

E.4.6.1 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. Non–scannable materials were keyed through an interactive on–line system. In both of these cases, the data were edited and suspect cases were resolved before further processing.

E.4.6.1.1 OMR Scanning/Image Scanning

The Math and Science 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 (PAS) number consisting of the scan batch number, the sequential number within the batch, and the bar code ID of the booklet. The number was assigned to and printed on one side of 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 PAS 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 processing 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 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 which 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.

E.4.6.1.2 Intelligent Character Recognition

NCS again used the Intelligent Character Recognition (ICR) engine to read various hand and machine printing on the front cover of the assessment and supervisor documents. 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 on–line key–entry operator. In all, the ICR engine read over 2,000,000 handwritten and machine–printed characters.

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 able 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 not displayed to on–line editing staff for verification. This eliminated a significant amount of on–line editing time needed to process the NAEP assessments. If the scanning programs could not decode the 2–out–of–5 code, booklet IDs were clipped and routed to edit stations for on–line verification.

E.4.6.1.3 Key Entry

A process of key entry and verification was used to make corrections to the non–scannable Trend reading/writing documents and large print booklets. Excluded Student 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.

E.4.6.2 Data Validation (editing) and Resolution

Each dataset produced by the scanning system contains 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 on–line 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 on–line 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 on–line edit file.

Document definition files described each document as a series of blocks which 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 data set 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 on–line edit file and processing continued with the next field. No action was taken on blank fields for multiple–choice items since 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 on–line 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 requiring corrections were displayed at an on–line editing terminal.

E.4.6.2.1 Image–Processed Documents

The paper edit log for key–entered documents is replaced by on–line 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 which 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

verifier 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 for the 1998–99 processing cycle. 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). Any 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, the paper edit log was forwarded to another editor. When this edit was satisfied, the PCS and WFM tracking systems were updated. Since 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.

E.4.6.2.2 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.

E.4.7 Processing Reports

The NCS NAEP (PCS) produced various status reports, one of which was the Receipt Control Status Report. This report displayed the current status of all schools. It could be sorted by school number or by scheduled administration date. As the receipt status of a school was updated through the receiving, opening, and batching processes, the data collected was added to this report. Data represented on this report included participation status, shipment receipt date, and receipt of the Roster of Questionnaires. The comment field in this report showed any school for which a shipment had not been received within two days of the completion of the scheduled assessment administration. NCS transmitted an electronic file to Westat weekly for any shipments not received within two days of the assessment administration date.

E.5. Professional Scoring

E.5.1 Long-Term Trend Assessments

The 1998–99 National Assessment of Educational Progress Long–term Trend Assessments included Mathematics and Reading/Writing scoring at three grades. Grade 8 was administered in the fall; Grade 4 in the winter; and Grade 11 in the spring. The volumes were comparable to previous cycles. The Performance Scoring Center (PSC) scored these assessments using teams of highly experienced and knowledgeable scorers at three different times throughout the year – December, March, and May.

NCS provided trainers for the mathematics and writing scoring. The writing trainer worked with ETS' staff member, who trained the reading items. These assessments were scored from the student booklets, with scores recorded on scannable sheets and captured with the PSC's scanning system. See figure E–9 for Long–term Trend Processing and Scoring Totals.

Figure E–9. NAEP long–term trend processing and scoring totals: 1998–99

	Books Processed	Constructed Responses Scored 1st & 2nd	Discrete Constructed Response Items	Number of Scorers and Scoring Supervisors	Length of Training and Scoring
Fall Trend Reading/Writing	5,946	38,288	22	10/1	12/7/98 – 12/22/98
Fall Trend Math	5,597	74,198	28	33/3	12/21/98 – 12/22/98
Winter Trend Reading/Writing	5,799	28,130	20	12/1	3/22/99 – 4/2/99
Winter Trend Math	6,045	77,877	29	18/1	3/9/99 – 3/26/99
Spring Trend Reading/Writing	5,316	40,470	25	12/1	5/3/99 – 5/18/99
Spring Trend Math	3,828	74,083	29	18/1	5/17/99 – 5/18/99

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

E.5.1.1 Long-Term Trend Mathematics

The Trend mathematics items were scored on a right/wrong basis. 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. Any reading items that appeared in the Mathematics booklets were scored only for attemptedness. 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 attempted. The scoring here was 0 for not attempting the item (blank) or 1 for any writing in the space provided.

Scoring of the Long-Term Trend mathematics items was identical to previous years. Preparation for scoring included copying the scoring guides from previous cycles of the assessment, pulling from the

warehouse the books listed in the samples, and printing and matching scoring sheets for those books. The scoring guides are a listing of the correct answers for each item.

Because the mathematics items were scored as right, wrong, or omitted, lengthy training for scoring these items was unnecessary. For each component (Fall, Winter, Spring), teams of scorers were trained to follow the procedures for scoring the mathematics items. They also became familiar with the scoring standards, which listed general guidelines and also the correct answer for the items in each of the blocks.

A different team scored each grade level at the appropriate time of year. The number of teams and scorers varied for each component, depending on the number of days in which scoring was to be completed. In December, 33 scorers were supervised by three scoring supervisors; in March and May the scoring teams each included 18 scorers and one scoring supervisor. For each component, the entire scoring was completed in one or two weeks at the end of the administration period. All scorers had at least a bachelor's degree. Many of them had previous experience scoring NAEP and mathematics assessments.

In order to establish the consistency of scoring across years, the readers rescored a subset of the responses from previous assessment cycles. Samples of 350 responses to each item from the 1990 assessment and 250 from the 1996 assessment were drawn. The scanning system produced 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 scoring supervisors monitored daily inter–reader agreement reports and t–tests to verify consistency of scores within the current year and across years. Summaries of the inter–reader agreements can be found in figure E–10.

Figure E–10. NAEP long–term trend inter–reader reliability: 1998–99

Assessment	Number of Unique Items	Number of Items in Percentage Agreement Range			
	Total	60–69%	70–79%	80–89%	90% and Above
Fall Trend Reading/Writing	22	1	4	7	10
Fall Trend Math	28	0	0	0	28
Winter Trend Reading/Writing	19	0	2	2	15
Winter Trend Math	29	0	0	0	29
Spring Trend Reading/Writing	25	0	1	6	18
Spring Trend Math	29	0	0	0	29
Note: Not all Long-Term Trend items received second scoring. Figures are included here only for those which were second scored. Figures for writing holistic include adjacent scores.					

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

E.5.1.2 Long-Term Trend Reading and Writing (Primary Trait)

All of the writing items for the three Long-Term Trend assessments (Fall, Winter, 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: not rated, unsatisfactory, minimal, adequate, and 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 writing items focused on students' abilities to write in an informative, persuasive, and narrative manners. 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 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 secondary scoring 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 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."

As with mathematics, the scorers used the same reading and writing training materials as were used for previous cycles for reading and writing. Thus, there was no need to select any new training material from current year responses. Preparation for the Long-Term Trend reading and writing scoring included identifying samples from previous years. Scores assigned in assessment booklets from 1984 (reading) and 1988 (writing) had been masked in previous years to ensure that scoring for training, and subsequent trend rescoring, would be done without knowledge of the original scores assigned. The 1996 books 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.

For the fall trend, a team of 10 scorers and one scoring supervisor was trained to score the reading and writing items. For the winter and spring components, the team was increased to 12 scorers. To the extent possible, the same scorers returned to score each component. All readers for this project were experienced scorers with a minimum of a bachelor's degree. Figure E–11 illustrates the number of readers and scoring supervisors needed to accomplish Long–term Trend Scoring.

Figure E–11. NAEP long–term trend readers and dates: 1998–99

Assessment	Number of Scoring Supervisors	Number of Scorers	Dates
Fall Trend Reading/Writing	1	10	12/7/98 – 12/22/98
Fall Trend Math	3	33	12/21/98 – 12/22/98
Winter Trend Reading/Writing	1	12	3/22/99 – 4/2/99
Winter Trend Math	1	18	3/9/99 – 3/26/99
Spring Trend Reading/Writing	1	12	5/3/99 – 5/18/99
Spring Trend Math	1	18	5/17/99 – 5/18/99

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

The formal training for the Trend assessments was divided into two parts to accommodate the reading and writing items. 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. Following the formal training sessions, the readers scored a sample of the 1984 assessment materials for reading and the 1988 assessment materials for writing. They recorded their scores on scannable scoring sheets that were produced for specific book types with the appropriate trend items pre–printed on the scoring sheets. These sheets were then scanned under a special job number to ensure that this material was designated as training scoring only.

A report that listed the individual and group percent agreement by item was then produced. For that report, the system automatically compared the new scores with the scores assigned in the 1984 or 1988 scoring. Therefore, the report showed the reliability of scores across the years, allowing the scoring director to determine if training in the current year was consistent with that in previous years. T–tests were also generated for each item to verify comparability of scoring across years. Prior to scoring any

1998–99 reading and writing Trend materials, the NCS scoring director carefully reviewed the training reliability agreement report before proceeding with scoring.

Reliability studies were conducted for the scoring of the Trend reading and writing items. For the 1998–99 material, 33 percent of the constructed–response items were scored by a second reader to produce inter–reader reliability statistics. In addition, a Trend reliability study was conducted to ensure that the scoring procedures were consistent with those used in 1984, 1988 and 1996. Three hundred fifty of the 1984 reading responses and 350 of the 1988 writing responses were sampled. Two hundred fifty of both reading and writing responses from 1996 were selected for rescore.

The scoring of these Trend samples was intermixed with the scoring of the current reading and writing Trend material. The readers selected a bundle 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 Trend reliability reporting.

The scoring supervisor monitored consistency within the current year as well as across years on a daily basis. 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 trend sample.

References

Allen, N. L. (1992). Data analysis for the science assessment. In E. G. Johnson and N. L. Allen, *The NAEP 1990 technical report*. (pp. 243-274), (No. 21-TR-20). Washington, DC: National Center for Education Statistics.

Allen, N. L., Carlson, J. E., Johnson, E. G., and Mislevy, R. J. (2001). Scaling procedures. In N. L. Allen, J. R. Donoghue, and T. L. Schoeps, *The NAEP 1998 technical report* (pp. 227-246), (NCES 2001–509). Washington, DC: National Center for Education Statistics.

Allen, N. L., Kline, D. L., and Zelenak, C. A. (1996). *The NAEP 1994 technical report*, NCES 1997-897. Washington, DC: National Center for Education Statistics.

Allen, N. L., Carlson, J. E., and Zelenak, C. A. (1999). *The NAEP 1996 technical report*, NCES 1999-452. Washington, DC: National Center for Education Statistics.

Allen, N. L., and Donoghue, J. R. (1994, April). *Differential item functioning based on complex samples of dichotomous and polytomous items.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Allen, N. L., and Donoghue, J. R. (1996). Applying the Mantel-Haenzel procedure to complex sample items. *Journal of Educational Measurement, 33*, 231-251.

Allen, N. L., Donoghue, J. R., and Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES 2001–509). Washington, DC: National Center for Education Statistics.

Allen, N. L. and Isham S. P. (1994). Data analysis for the science long-term trend assessment. In E. G. Johnson and J. E. Carlson, *The NAEP 1992 technical report* (pp. 343-354), (No. 23-TR-20). Washington, DC: National Center for Education Statistics.

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-84 technical report* (No.15-TR-20) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Beaton, A. E. (1988). *Expanding the new design: The NAEP 1985-86 technical report*, (No.17-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*(1*)*, 289–300.

Bickel, P. J., and Doksum, K. A. (1977). Linear models—regression and analysis of variance. In E. L. Lehmann, (Ed.). *Mathematical statistics: Basic ideas and selected topics*. Oakland, CA.

Caldwell, N. W., Fowler, J., A., Waksberg, M. M., Wallace, L. (2002). *NAEP 1999 long-term trend data collection, sampling, and weighting report*. Rockville, MD: Westat.

Campbell, J. R., Hombo, C. M., and Mazzeo, J. (2000). *NAEP 1999 trends in academic progress: Three decades of student performance.* (NCES 2000-469). Washington, DC: National Center for Education Statistics.

Chang, H., Donoghue, J. R., Wang, M., Worthington, L. H., and Freund. D. S. (1996). Data analysis for the long-term trend reading assessment. In N. L. Allen, D. L. Kline, and C. A. Zelenak, *The NAEP 1994 technical report*. (pp. 357-372). Washington, DC: National Center for Education Statistics.

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley and Sons.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika 16, 297-334.

Donoghue, J. R. (1992). Data analysis for the reading assessment. In E. G. Johnson and N. L. Allen, *The NAEP 1990 technical report* (pp. 215-242), (No. 21-TR-20). Washington, DC: National Center for Education Statistics.

Donoghue, J. R. (1988). *NAEP SIBTEST* (computer program). Princeton, NJ: Educational Testing Service.

Donoghue, J. R., Isham, S. P., Bowker, D. W., and Freund, D. S. (1994). Data analysis for the reading long-term trend assessment. In E. G. Johnson and J. E. Carlson, *The NAEP 1992 technical report* (pp. 257-298), (No. 23-TR-20). Washington, DC: National Center for Education Statistics.

Ferris, J. J., Pashley, K. E., Freund, D. S., and Rogers, A. M. (1999). Creation of the database, quality control of the data entry,and creation of the database products. In N. L. Allen, J. E. Carlson, and C. A. Zelenak, *The NAEP 1996 Technical Report*, (NCES 1999-452). Washington, DC: National Center for Education Statistics.

Holland, P. W., and Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity*, Hillsdale, NJ: Erlbaum.

Ip, E., Jenkins, F., and Kulick, E. (1996). Data analysis for the mathematics long-term trend assessment. In N. L. Allen, D. L. Kline, and C. A. Zelenak, *The NAEP 1994 technical report* (pp. 373-386), NCES 97-897. Washington, DC: National Center for Education Statistics.

Jenkins, F., and Kulick, E. M. (1994). Data analysis for the mathematics assessment. In E. G. Johnson and J. E. Carlson, *The NAEP 1992 technical report* (pp. 299-342), (No. 23-TR-20). Washington, DC: National Center for Education Statistics.

Johnson, E. G. (1988). Mathematics data analysis. In A. E. Beaton, *Expanding the new design: The NAEP 1985-86 technical report* (pp. 215-242), (No.17-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics*, *14* (4).

Johnson, E. G. and Allen, N. L. (1992). *The NAEP 1990 technical report*. (Report No. 21-TR-20). Washington, DC: National Center for Education Statistics.

Johnson, E. G. and Carlson, J. E. (1994). *The NAEP 1992 technical report*. (Report No. 23-TR20). Washington, DC: National Center for Education Statistics.

Johnson, E. G. and Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175-190.

Johnson, E. G. and Rust, K. F. (1993). Effective degrees of freedom for variance estimates from a complex sample survey. *American Statistical Association 1993 Proceedings: Survey research methods section*, 863–866.

Johnson, E. G. and Zwick, R. (1990). *Focusing the new design: The NAEP 1988 technical report*. (Report No. 19-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Kaplan, B. A., Beaton, A. E., Johnson, E. G., and Johnson, J. R. (1988). *National Assessment of Educational Progress: 1986 bridge studies*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Liang, J., and Worthington, L. H. (1999). Data analysis for the long-term trend reading assessment. In N. L. Allen, J. E. Carlson, and C. E. Zelenak, *The NAEP 1996 technical report* (pp. 323-334), NCES 1999-452. Washington, DC: National Center for Education Statistics.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley and Sons.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690–700.

Mantel, N., and Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institution*, *22*, 719-748.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.

Mislevy, R. J., and Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., and Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report*, [No. 15–TR–20] (pp. 293–360). Princeton, NJ: National Assessment of Educational Progress.

Mislevy, R. J., and Wu, P. K. (1988). *Inferring examinee ability when some item responses are missing* (ETS Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Muraki, E., and Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data*. Chicago, IL: Scientific Software, Inc.

National Assessment of Educational Progress. (1984). *Reading Objectives: 1983-84 Assessment*. Princeton, NJ: Educational Testing Service.

National Computer Systems. (2000). *NAEP report of processing and professional scoring activities*. Iowa City, IA: Author.

Petersen, N. (1988). *DIF procedures for use in statistical analysis*. Internal memorandum. Princeton, NJ: Educational Testing Service.

Qian, J., Kaplan, B. A., and Johnson, E., Krenzke, T., and Rust, K. (2001). Weighting procedures and estimation of sampling variance for the national assessment. In N. L. Allen, J. R. Donoghue, and T. L. Schoeps, *The NAEP 1998 technical report* (pp. 161-191). (NCES 2001–509). Washington, DC: National Center for Education Statistics.

Qian, J and Norris, N. (1999). Data analysis for the long-term trend mathematics assessment. In N. L. Allen, J. E. Carlson, and C. E. Zelenak, *The NAEP 1996 technical report* (pp. 335-344), NCES 1999-452. Washington, DC: National Center for Education Statistics.

Rust, K. F., Krenzke, T., Qian, J., and Johnson, E. G. Sample design for the national assessment. In N. L. Allen, J. R. Donoghue, and T. L. Schoeps, *The NAEP 1998 technical report* (pp. 31-59). NCES 2001-509. Washington, DC: National Center for Education Statistics.

Swinton, S. S., Allen, N. L., Isham, S. P., and Chen, C. J. (1996). Data analysis for the science long-term trend assessment. In N. L. Allen, D. L. Kline, and C. A. Zelenak, *The NAEP 1994 technical report* (pp. 386-400), NCES 97-897. Washington, DC: National Center for Education Statistics.

Thomas, N. (1994). *CGROUP and BGROUP: Modifications of the MGROUP program to estimate group effects in multivariate models* [Computer programs]. Princeton, NJ: Educational Testing Service.

Wallace, L., and Rust, K. F. (1999). Sample design. In N. L. Allen, J. E. Carlson, and C. A. Zelenak, *The NAEP 1996 technical report*, NCES 1999-452. Washington, DC: National Center for Education Statistics.

Williams, V. S. L., Jones, L. V., and Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*(1), 42–69.

Wolter, K. M. (1985*). Introduction to variance estimation*. New York, NY: John Wiley and Sons.

Yamamoto, K. (1988). Science data analysis. In A. E. Beaton, *Expanding the new design: The NAEP 1985-86 technical report* (pp. 243-255), (No.17-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Yamamoto, K. (1990). Data analysis for mathematics and science. In R. Zwick, *Focusing the new design: The NAEP 1988 technical report* (pp. 251-265), (No. 19-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Yamamoto, K., and Jenkins, F. (1992). Data analysis for the mathematics assessment. In E. G. Johnson and N. L. Allen, *The NAEP 1990 technical report*. (pp. 243-274), (No. 21-TR-20). Washington, DC: National Center for Education Statistics.

Zhang, J and Norris, N. (1999). Data analysis for the long-term trend science assessment. In N. L. Allen, J. E. Carlson, and C. E. Zelenak, *The NAEP 1996 technical report* (pp. 345-354), NCES 1999-452. Washington, DC: National Center for Education Statistics.

Zieky, M. (1993). Practical questions in the use of DIF statistics. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Zwick, R. (1987). Assessing the dimensionality of NAEP year 15 reading data. *Journal of Educational Measurement, 24*(4), 293–308.

Zwick, R. (1988). Reading data analysis. In A. E. Beaton, *Expanding the new design: The NAEP 1985-86 technical report* (pp. 207-214), (No.17-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Zwick, R. (1990). Data analysis for the reading assessment. In R. Zwick, *Focusing the new design: The NAEP 1988 technical report* (pp. 251-265), (No. 19-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Zwick, R., Donoghue, J. R., and Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251.

Zwick, R., and Grima, A. (1991). *Policy for differential item functioning (DIF) analysis in NAEP.* Technical memorandum. Princeton, NJ: Educational Testing Service.

U.S. Department of Education
ED Pubs
8242-B Sandy Court
Jessup, MD 20794-1398

Official Business
Penalty for Private Use, $300

U.S. POSTAGE PAID
U.S. DEPARTMENT
OF EDUCATION
PERMIT NO. G-17