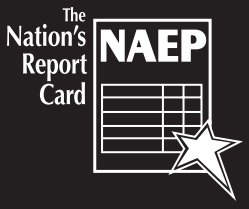




**U.S. Department of Education**
Institute of Education Sciences
NCES 2005–484

# 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

National Center for Education Statistics

The Nation's Report Card NAEP

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 $\Delta$, 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, $\alpha_{MH}$ , 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

$\theta_{target}$ = 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, $\sigma$ is the "true" standard error, and $\varepsilon$ 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 $\alpha$ 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 $T_\alpha$ 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 $\alpha$ be the combined significance level desired, usually 0.05. The procedure will compare $P_q$ with $\alpha$, $P_{q-1}$ 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.

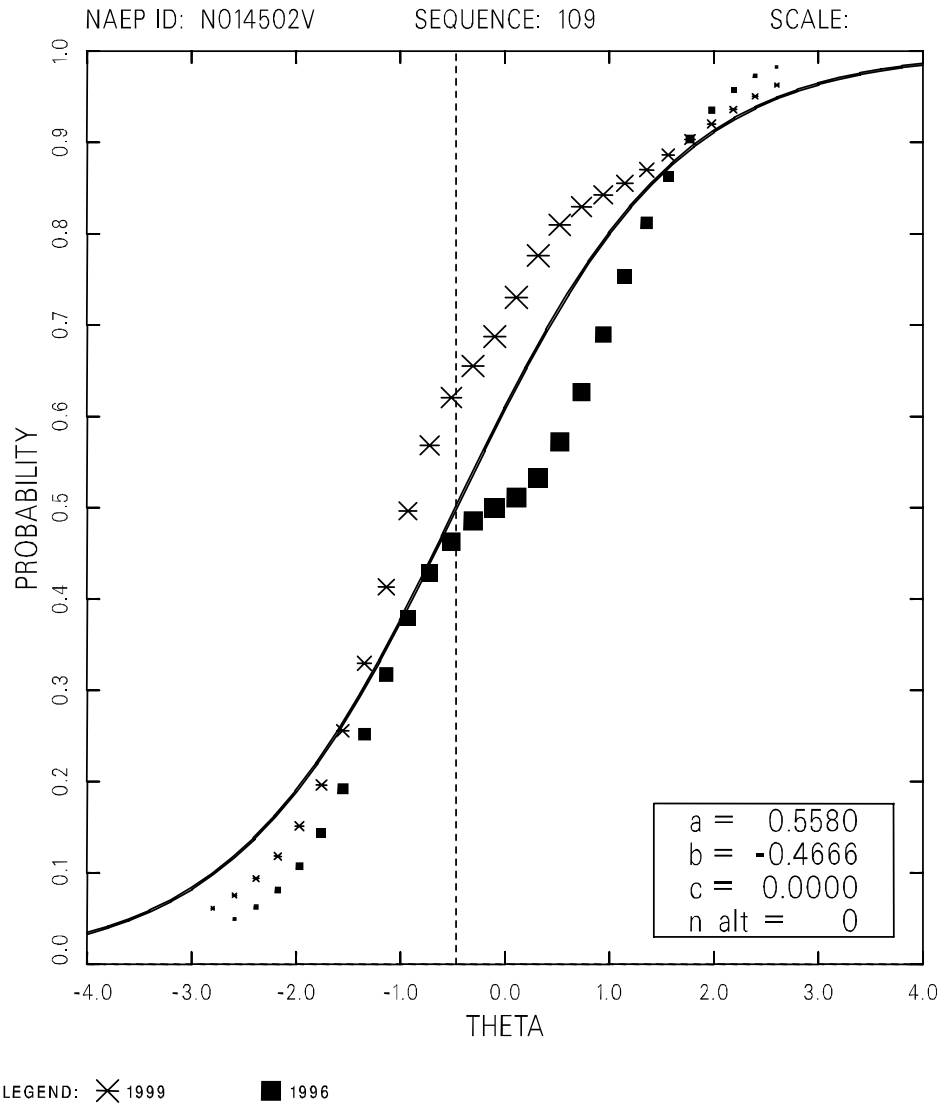
The adequacy of the assumption of a common item response function across assessment years was also evaluated by comparing the nonmodel–based expected proportions for each assessment year to the single, model–based item response function fit by BILOG/PARSCALE. Items that showed clear evidence of functioning 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.

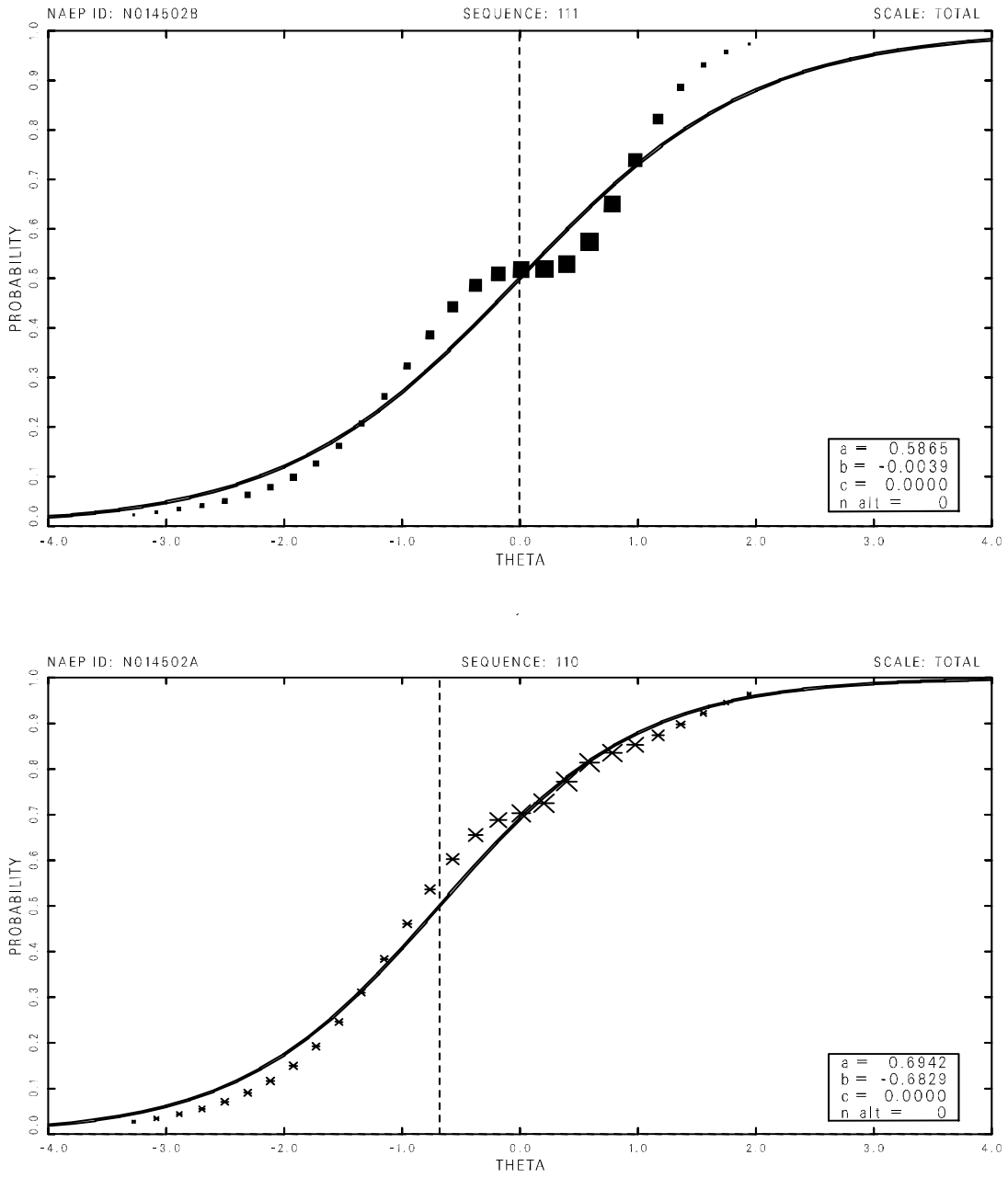
[4]The size of the symbols are proportional to the estimated number of students at a particular scale score level. The symbols are ordinarily larger in the middle of the theta scale, where most students' scale scores fall.

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



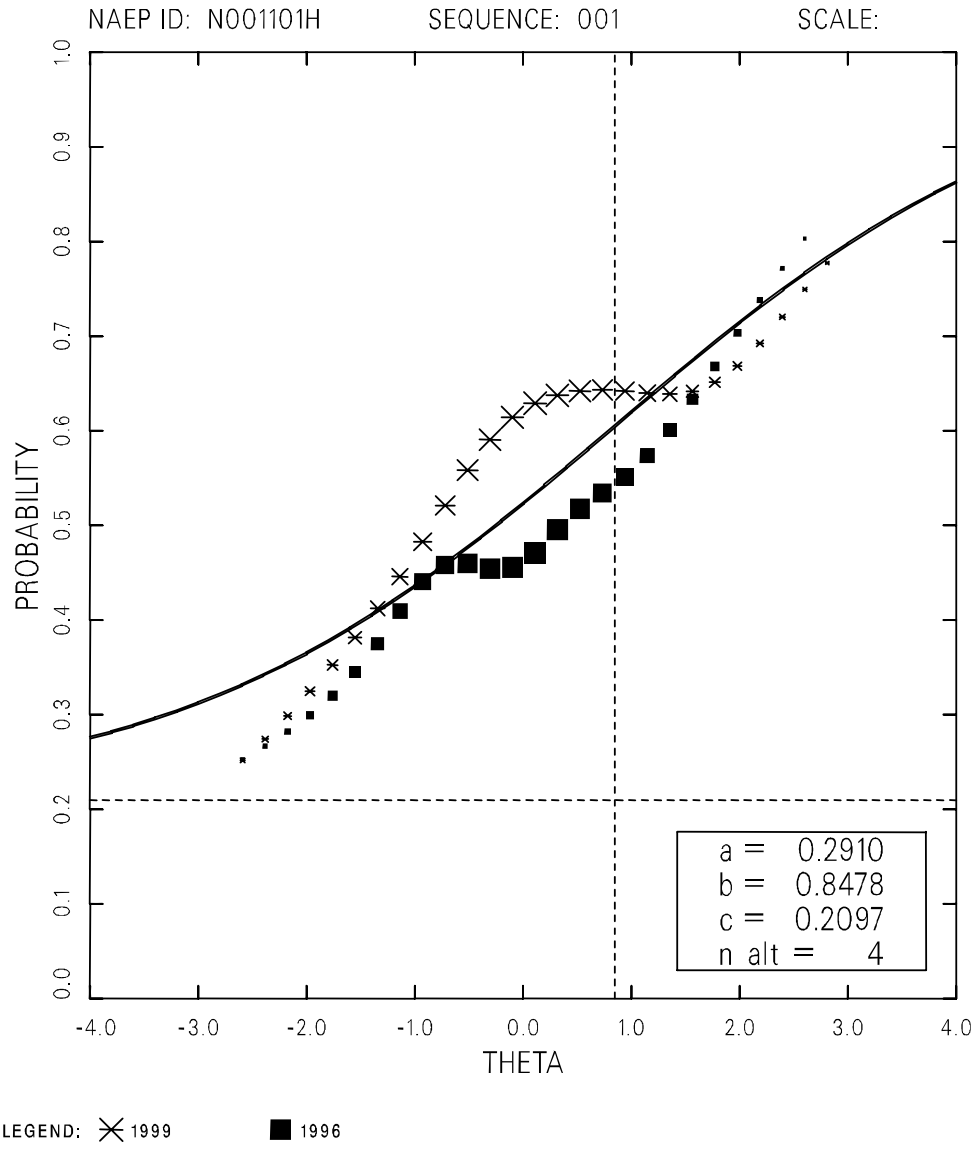

NOTE: This plot compares empirical and model–based estimates of the item response function (IRF). The smooth curve represents the model–based estimate at each provisional proficiency level. The squares represent 1996 data; asterisks represent 1999 data.
SOURCE: U.S. Department of Education, Institute 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**



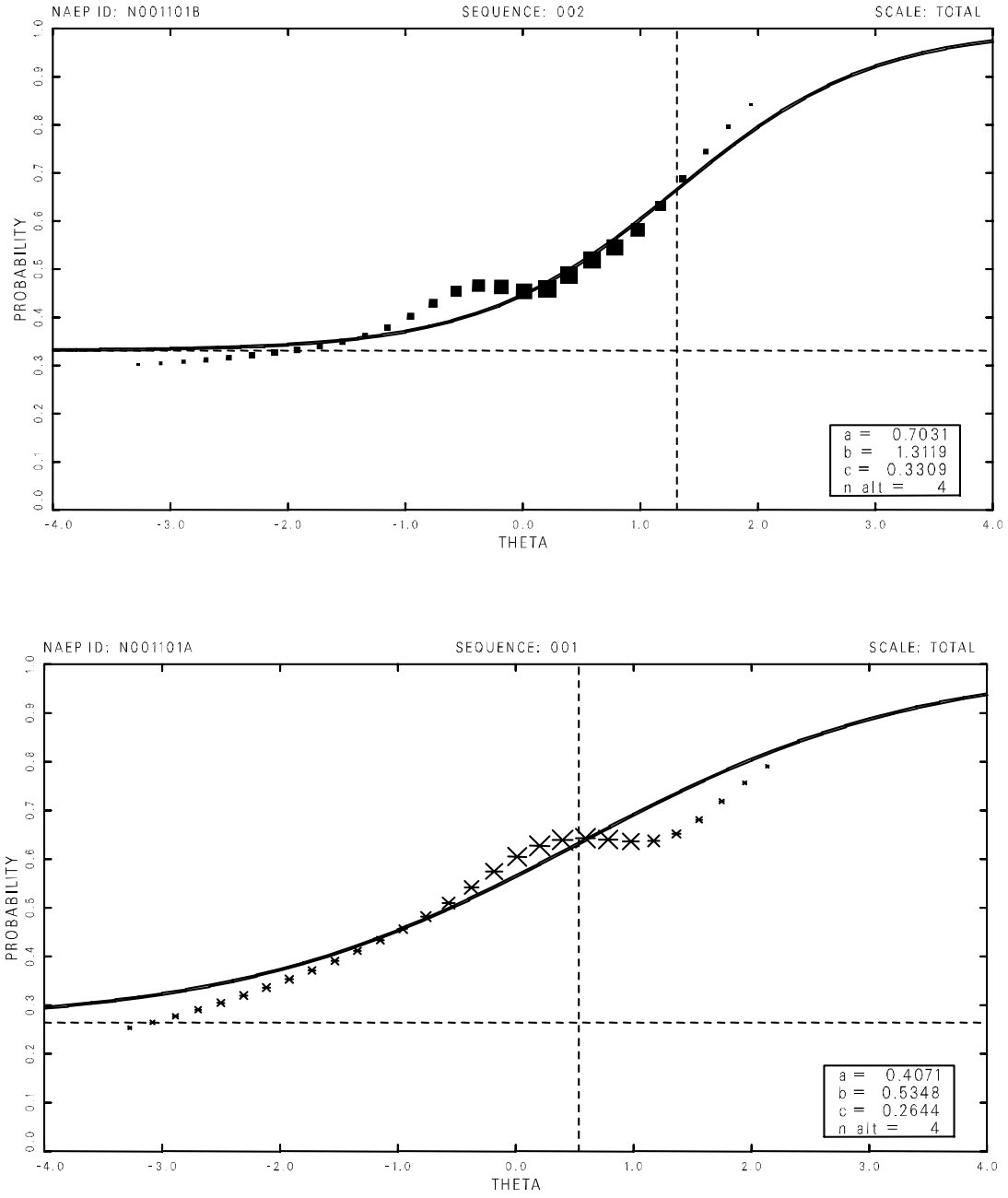

NOTE: The plot compares empirical and model–based estimates of the item response function (IRF). The smooth curve represents the model–based estimate at each provisional proficiency level. The squares represent 1996 data; asterisks represent 1999 data.
SOURCE: U.S. Department of Education, Institute 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 $\theta_{target}$ 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 $\theta_{target}$ 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 $\theta_{target}$ 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.