NATIONAL CENTER FOR EDUCATION STATISTICS

THE NAEP 1998

TECHNICAL REPORT

THE NATION'S REPORT CARD

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

NCES 2001-509

What is The Nation's Report Card?

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

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

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

The National Assessment Governing Board

NATIONAL CENTER FOR EDUCATION STATISTICS

THE NAEP 1998 TECHNICAL REPORT

Nancy L. Allen
John R. Donoghue
Terry L. Schoeps

in collaboration with
Mary Lyn Bourque, Charles Brungardt, Nancy W. Caldwell, James E. Carlson, Hua-Hua Chang, Patricia L. Donahue, John J. Ferris, David S. Freund, Elissa A. Greenwald, Lucy M. Gray, Steven P. Isham, Frank Jenkins, Eugene G. Johnson, Bruce A. Kaplan, Tom Krenzke, Edward Kulick, Stephen Lazer, Venus Leung, Jo-Lin Liang, Youn-Hee Lim, Robert J. Mislevy, Norma A. Norris, Ingeborg U. Novatkoski, Jiahe Qian, Katharine E. Pashley, Timothy Robinson, Alfred M. Rogers, Keith F. Rust, Connie Smith, Spencer S. Swinton, Mark M. Waksberg, Leslie Wallace, Andrew R. Weiss, Lois H. Worthington, Ibrahim Yansaneh, and Jinming Zhang

June 2001

U.S. Department of Education
Office of Educational Research and Improvement **NCES 2001-509**

U.S. Department of Education
Rod Paige
Secretary

National Center for Education Statistics
Gary W. Phillips
Acting Commissioner

June 2001

SUGGESTED CITATION
U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. *The NAEP 1998 Technical Report,* NCES 2001-509, by Allen, N.L., Donoghue, J.R., & Schoeps, T.L. (2001). Washington, DC: National Center for Education Statistics.

FOR MORE INFORMATION
Content contact:
Arnold A. Goldstein
202-502-7344

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

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

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

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

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

THE NAEP 1998 TECHNICAL REPORT

♦Table of Contents♦

Introduction ***James E. Carlson and Nancy L. Allen, Educational Testing Service*** **1**

Chapter 1 **OVERVIEW OF PART I: THE DESIGN AND IMPLEMENTATION OF THE 1998 NAEP**
Nancy L. Allen, James E. Carlson, and John R. Donoghue, Educational Testing Service **5**

1.1 INTRODUCTION 5

1.2 THE 1998 NAEP DESIGN 6

1.2.1 The 1998 NAEP Samples 10
1.2.2 NAEP Assessments Since 1969 11

1.3 DEVELOPMENT OF ASSESSMENT OBJECTIVES, ITEMS, AND BACKGROUND QUESTIONS 16

1.4 THE 1998 SAMPLE DESIGN 17

1.4.1 Step 1: Primary Sampling Units 17
1.4.2 Step 2: Selection of Schools 18
1.4.3 Step 3: Assigning Assessment Session and Sample Type to Schools 18
1.4.4 Step 4: Sampling Students and Teachers 19

1.5 ASSESSMENT INSTRUMENTS 20

1.6 FIELD OPERATIONS AND DATA COLLECTION 21

1.7 MATERIALS AND DATA PROCESSING 22

1.8 PROFESSIONAL SCORING 22

1.9 CREATION OF THE DATABASE 23

Chapter 2 **DEVELOPING THE NAEP OBJECTIVES, ITEMS, AND BACKGROUND QUESTIONS FOR THE 1998 ASSESSMENTS OF READING, WRITING, AND CIVICS**
Terry L. Schoeps, Educational Testing Service **25**

2.1 INTRODUCTION 25

2.2 OVERVIEW OF THE 1998 ASSESSMENT OBJECTIVES AND FRAMEWORKS 26

2.3 GENERAL OVERVIEW OF PROCEDURES FOR DEVELOPING COGNITIVE ITEMS 28

2.4 DEVELOPING BACKGROUND ITEMS 29

Chapter 3 SAMPLE DESIGN FOR THE NATIONAL ASSESSMENT

Keith F. Rust and Tom Krenzke, Westat
Jiahe Qian and Eugene G. Johnson, Educational Testing Service **31**

3.1 INTRODUCTION 31

3.1.1 Brief Overview of the Sample Design and Sampling Activities 31
3.1.2 Target Population and Sample Size 32
3.1.3 Highlights of Design Changes for the 1998 Assessment 33

3.2 THE SAMPLE OF PRIMARY SAMPLING UNITS AND SCHOOLS 34

3.2.1 The Definition of Primary Sampling Units 34
3.2.2 Definition of PSU Strata 35
3.2.3 Selection of Noncertainty PSUs 37
3.2.4 School Sample 38
3.2.4.1. Frame Construction 38
3.2.4.2. Assigning Size Measures and Selecting School Samples 39
3.2.4.3. Updating the School Frame and Sample 41
3.2.4.4. School Substitution 42
3.2.4.5. School Participation Experience 42

3.3 ASSIGNMENT OF SESSIONS AND SAMPLE TYPES TO SCHOOLS 43

3.3.1 Description of Session Types 43
3.3.2 Allocation of Sessions 43
3.3.2.1 Grade 4 Allocation of Sessions 44
3.3.2.2 Grade 8 Allocation of Sessions 44
3.3.2.3 Grade 12 Allocation of Sessions 45
3.3.3 Assignment of Sample Types 45
3.3.3.1 Grade 4 Assignment of Sample Types 46
3.3.3.2 Grade 8 Assignment of Sample Types 46
3.3.3.3 Grade 12 Assignment of Sample Types 46
3.3.3.4 Schools Selected in Both National and State Assessments 46

3.4 STUDENT SAMPLE 46

3.4.1 Updating Estimates of Grade-Eligible Students 47
3.4.2 Within-School Sampling Rates 47
3.4.3 The Session Assignment Form (SAF) 47
3.4.4 Updating Session Allocation When Generating SAFs 48
3.4.5 Sample Selection 48
3.4.5.1 Oversampling Black and Hispanic Students 49
3.4.5.2 Oversampling SD/LEP Students in Reading 49
3.4.6 Supporting the Field Staff on Sampling Issues 50
3.4.7 Excluded Students 51
3.4.8 Student Participation Results 57
3.4.9 Teacher Survey 59

Chapter 4 SAMPLE DESIGN FOR THE STATE ASSESSMENT

Keith F. Rust and Leslie Wallace, Westat
Jiahe Qian, Educational Testing Service **61**

4.1 INTRODUCTION 61

Chapter 4 SAMPLE DESIGN FOR THE STATE ASSESSMENT (continued)

4.2 TARGET POPULATIONS AND SAMPLING FRAME FOR THE 1998 STATE ASSESSMENT 63
4.2.1 Target Population 63
4.2.2 Sampling Frame 63
4.3 STRATIFICATION OF SCHOOLS IN THE SAMPLING FRAME 66
4.3.1 Stratification Variables 66
4.3.2 Missing Stratification Variables 66
4.3.3 Resources for Stratification Variables 67
4.3.3.1 Urbanization Classification 67
4.3.3.2 Minority Classification 68
4.3.3.3 Median Household Income 68
4.3.3.4 Metro Area Status 68
4.3.3.5 School Type for Nonpublic Schools 69
4.4 SCHOOL SAMPLE SELECTION 69
4.4.1 Measure of Size and Sample Selection 69
4.4.2 Sparse State Sample Option 70
4.4.3 Control of Overlap of School Samples for National Educational Studies 71
4.4.4 Selection of Schools in Small Jurisdictions 72
4.4.5 Selection of New Public Schools 72
4.4.6 Assigning Subject, Sample Type, and Monitor Status 75
4.4.7 School Substitution and Retrofitting 76
4.5 STUDENT SAMPLE SELECTION 77
4.5.1 Student Sampling and Participation 77
4.5.2 The Reduced Sample Option 77

Chapter 5 FIELD OPERATIONS AND DATA COLLECTION
Lucy M. Gray, Mark M. Waksberg, and Nancy W. Caldwell, Westat **79**

5.1 INTRODUCTION 79
5.1.1 Organization of the National Assessment for 1998 79
5.1.1.1 Additional Special Studies 81
5.1.1.2 Exclusions and Accommodations for Students 81
5.1.2 Organization of the State Assessment for 1998 82
5.2 PREPARING FOR THE ASSESSMENTS 85
5.2.1 Gaining the Cooperation of Sampled Schools 85
5.2.2 Supervisor Training 88
5.2.3 Contacting Districts and Nonpublic Schools 89
5.2.4 Recruiting, Hiring, and Training Exercise Administrators 90
5.3 SELECTING THE STUDENT SAMPLES 91
5.3.1 Selecting the National NAEP Student Samples 91
5.3.2 Selecting the Special Studies Samples 93
5.3.3 Selecting the State NAEP Student Samples 94
5.4 CONDUCTING THE ASSESSMENT SESSIONS 94
5.4.1 Conducting the National Assessments 94
5.4.2 Conducting the State Assessments 95

Chapter 5 FIELD OPERATIONS AND DATA COLLECTION (continued)

5.4.3 Participation of DoDEA Schools in State NAEP 96

5.5 RESULTS OF THE NATIONAL NAEP ASSESSMENT 97

5.5.1 School and Student Participation 97

5.5.2 Assessment Questionnaires 97

5.6 RESULTS OF THE STATE NAEP ASSESSMENT 98

5.6.1 School and Student Participation 98

5.6.2 Results of the Observations 102

5.7 FIELD MANAGEMENT 103

Chapter 6 PROCESSING ASSESSMENT MATERIALS

Connie Smith, Charles Brungardt, and Timothy Robinson, National Computer Systems **105**

6.1 INTRODUCTION 105

6.2 PRINTING 105

6.3 PACKAGING AND DISTRIBUTION 107

6.4 PROCESSING 109

6.4.1 Document Receipt and Opening 109

6.4.2 Batching of Booklets 110

6.4.3 Scanning of Documents 110

6.4.4 Data Transcription 110

6.4.4.1 Data Entry 111

6.4.4.2 Data Validation 111

6.5 DATA TRANSMISSION BEFORE SCORING 115

6.6 CLASSROOM-BASED WRITING STUDY 115

Chapter 7 PROFESSIONAL SCORING

Connie Smith, Charles Brungardt, and Timothy Robinson, National Computer Systems **117**

7.1 OVERVIEW 117

7.2 SELECTION OF TRAINING PAPERS 119

7.3 CALIBRATION POLICIES 120

7.4 IMAGE SCORING 120

7.4.1 Reader Qualification 120

7.4.2 Backreading Process 120

7.4.3 Calibration Process 121

7.4.4 Short-Term Trend Rescoring 121

7.4.5 Validity Sets Tool 122

7.4.6 *t*-Tests 122

7.4.7 Procedure for Monitoring Interrater Reliability 122

7.4.8 Process for Monitoring Frequency Distribution of Scores 122

7.4.9 Process for Monitoring the Rate of Scoring 123

7.4.10 Scoring Buttons 123

Chapter 7 PROFESSIONAL SCORING (continued)

7.5 PAPER SCORING ... 123

7.6 LARGE-PRINT BOOKS AND OTHER SPECIAL ACCOMMODATIONS ... 123

7.7 TRAINING ... 124

7.8 SCORING ... 124

7.9 INTERRATER RELIABILITY ... 124

7.9.1 Scoring of Reading ... 125
7.9.2 Scoring of Writing ... 127
7.9.2.1 Selective Rescoring ... 127
7.9.2.2 Prewriting Coding ... 128
7.9.3 Scoring of Civics ... 128

7.10 PREPARATION FOR TAPE CREATION ... 129

7.11 UPLOADING OF SCORES TO THE NAEP DATABASE ... 129

7.12 SD/LEP STUDENT QUESTIONNAIRES ... 129

7.13 SCHOOL QUESTIONNAIRES ... 130

7.14 TEACHER QUESTIONNAIRE MATCH ... 130

7.15 DELIVERY ... 130

7.16 STORAGE OF DOCUMENTS ... 130

7.17 QUALITY CONTROL DOCUMENTS ... 130

Chapter 8 CREATION OF THE DATABASE, QUALITY CONTROL OF DATA ENTRY, AND CREATION OF THE DATABASE PRODUCTS

John J. Ferris, Katharine E. Pashley, David S. Freund, and Alfred M. Rogers, Educational Testing Service ... **131**

8.1 INTRODUCTION ... 131

8.2 CREATION OF THE DATABASE ... 131

8.2.1 Merging Files ... 132
8.2.2 Creating the Master Catalog ... 133

8.3 QUALITY CONTROL OF NAEP DATA ENTRY FOR 1998 ... 133

8.3.1 Student Booklet Data ... 135
8.3.2 SD/LEP Student Questionnaire Data ... 135
8.3.3 Teacher Questionnaire Data ... 135
8.3.4 School Characteristics and Policies Questionnaire Data ... 136

8.4 NAEP DATABASE PRODUCTS ... 136

8.4.1 File Definition ... 136
8.4.2 Definition of the Variables ... 137
8.4.3 Data Definition ... 137
8.4.4 Data File Catalogs ... 138
8.4.5 Data File Layouts ... 139
8.4.6 Data Codebooks ... 139
8.4.7 Control Statement Files for Statistical Packages ... 140
8.4.8 Machine-Readable Catalog Files ... 140
8.4.9 NAEP Data on Disk ... 141

Chapter 9 OVERVIEW OF PART II: THE ANALYSIS OF 1998 NAEP DATA
Nancy L. Allen, James E. Carlson, and John R. Donoghue, Educational Testing Service **143**

9.1 INTRODUCTION 143

9.2 SUMMARY OF THE NAEP DESIGN 144

9.3 ANALYSIS STEPS 145

9.3.1 Preparation of Final Sampling Weights 145
9.3.2 Reliability of Scoring Constructed-Response Items 145
9.3.3 Teacher Questionnaires 146
9.3.4 Analysis of Item Properties: Background and Cognitive Items 146
9.3.4.1 Background Items 146
9.3.4.2 Cognitive Items 147
9.3.4.3 Tables of Item-Level Results 148
9.3.4.4 Tables of Block-Level Results 148
9.3.4.5 Differential Item Functioning Analysis of Cognitive Items 149
9.3.5 Scaling 152
9.3.5.1 Scaling the Cognitive Items 153
9.3.5.2 Generation of Plausible Values for Each Scale 153
9.3.5.3 Transformation to the Reporting Metric 154
9.3.5.4 Definition of Composites for the Multivariate Scales in Reading 155
9.3.5.5 Tables of Scale Score Means and Other Reported Statistics 155
9.3.6 Dimensionality Analysis 155
9.3.6.1 Previous Dimensionality Analyses of NAEP Data 155
9.3.7 Drawing Inferences from the Results 156

9.4 OVERVIEW OF CHAPTERS 10 THROUGH 24 158

Chapter 10 WEIGHTING PROCEDURES AND ESTIMATION OF SAMPLING VARIANCE FOR THE NATIONAL ASSESSMENT
Jiahe Qian, Bruce A. Kaplan, and Eugene G. Johnson, Educational Testing Service
Tom Krenzke and Keith F. Rust, Westat **161**

10.1 INTRODUCTION 161

10.2 WEIGHTING PROCEDURES FOR ASSESSED AND EXCLUDED STUDENTS IN THE NATIONAL SAMPLES 162

10.2.1 Base Weights 163
10.2.2 Adjustment of the Base Weights for Nonresponse 166
10.2.2.1 Session Nonresponse Adjustment (SESNRF) 166
10.2.2.2 Student Nonresponse Adjustment (STUNRADJ) 167
10.2.3 Variation in Weights 168
10.2.3.1 Trimming the Weights for Outliers 169
10.2.4 Reporting Factors 171
10.2.5 Poststratification 171
10.2.5.1 The 50-Minute Writing Session 173
10.2.6 Final Student Reporting Weights 173

10.3 OTHER WEIGHTING PROCEDURES IN THE NATIONAL SAMPLES 174

10.3.1 Modular Weights 174
10.3.2 Linking Weights 175
10.3.3 School Weights 177

Chapter 10 WEIGHTING PROCEDURES AND ESTIMATION OF SAMPLING VARIANCE FOR THE NATIONAL ASSESSMENT (continued)

10.3.4 Reporting Weights with Accommodations 178
10.3.5 Jackknife Replicate Weights 179
10.4 POTENTIAL FOR BIAS DUE TO NONRESPONSE 179
10.5 VARIANCE ESTIMATION 184
10.5.1 Procedure to Estimate Sampling Variability 184
10.5.2 Approximating the Sampling Variance Using Design Effects 187

Chapter 11 STATE WEIGHTING PROCEDURES AND VARIANCE ESTIMATION

Jiahe Qian, Bruce A. Kaplan, and Eugene G. Johnson, Educational Testing Service
Ibrahim S. Yansaneh and Keith F. Rust, Westat **193**

11.1 OVERVIEW 193
11.2 CALCULATION OF BASE WEIGHTS 194
11.2.1 Calculation of School Base Weights 194
11.2.2 Weighting New Schools 195
11.2.3 Trimming School Base Weights for New Schools 195
11.2.4 Treatment of Substitute Schools 195
11.2.5 Calculation of Student Base Weights 196
11.3 ADJUSTMENTS FOR NONRESPONSE 197
11.3.1 Defining Initial School-Level Nonresponse Adjustment Classes 197
11.3.2 Constructing the Final Nonresponse Adjustment Classes 198
11.3.3 School Nonresponse Adjustment Factors 198
11.3.4 Student Nonresponse Adjustment Classes 199
11.3.5 Student Nonresponse Adjustments 200
11.4 CHARACTERISTICS OF NONRESPONDING SCHOOLS AND STUDENTS 207
11.4.1 Weighted Distributions of Schools Before and After School Nonresponse 208
11.4.2 Characteristics of Schools Related to Response 209
11.4.3 Weighted Distributions of Students Before and After Student Absenteeism 215
11.5 VARIATION IN WEIGHTS 216
11.6 CALCULATION OF REPLICATE WEIGHTS 217
11.6.1 Defining Replicate Groups and Forming Replicates for Variance Estimation 217
11.6.1.1 Replicate Group Assignments for Non-SD/LEP Students 218
11.6.1.2 Replicate Group Assignments for SD/LEP Students in Reading 219
11.6.2 School-Level Replicate Weights 220
11.6.3 Student-Level Replicate Weights 222
11.7 RAKING OF WEIGHTS 223
11.7.1 Raking Dimensions for Full Sample Student Weights 223
11.7.2 Raking Student Replicate Weights 224
11.8 APPROXIMATING THE SAMPLING VARIANCE USING DESIGN EFFECTS 224

Chapter 12 SCALING PROCEDURES

Nancy L. Allen, James E. Carlson, Eugene G. Johnson, and Robert J. Mislevy, Educational Testing Service **227**

12.1 INTRODUCTION 227

12.2 BACKGROUND 227

12.3 SCALING METHODOLOGY 229

12.3.1 The Scaling Models 229
12.3.2 An Overview of Plausible Values Methodology 237
12.3.3 Computing Plausible Values in IRT-Based Scales 238

12.4 INFERENCES ABOUT PROFICIENCIES 240

12.4.1 Computational Procedures 240
12.4.2 Statistical Tests 241
12.4.3 Biases in Secondary Analyses 242
12.4.4 A Numerical Example 243

12.5 DESCRIBING STUDENT PERFORMANCE 244

12.5.1 Achievement Levels 244
12.5.2 Item Mapping Procedures 245

12.6 OVERVIEW OF THE 1998 NAEP SCALES 246

Chapter 13 CONVENTIONS USED IN HYPOTHESIS TESTING AND REPORTING NAEP RESULTS

Spencer S. Swinton, David S. Freund, and Nancy L. Allen, Educational Testing Service **247**

13.1 OVERVIEW 247

13.2 MINIMUM SCHOOL AND STUDENT SAMPLE SIZES FOR REPORTING SUBGROUP RESULTS 248

13.3 IDENTIFYING ESTIMATES OF STANDARD ERRORS WITH LARGE MEAN SQUARED ERRORS 249

13.4 TREATMENT OF MISSING DATA FROM THE STUDENT, TEACHER, AND SCHOOL QUESTIONNAIRES 250

13.5 HYPOTHESIS-TESTING CONVENTIONS 251

13.5.1 Comparing Means and Proportions for Different Groups of Students 251
13.5.2 Multiple Comparison Procedures 253
13.5.3 Comparing Proportions Within a Group 254

Chapter 14 ASSESSMENT FRAMEWORKS AND INSTRUMENTS FOR THE 1998 NATIONAL AND STATE READING ASSESSMENTS

Patricia L. Donahue and Terry L. Schoeps, Educational Testing Service **255**

14.1 INTRODUCTION 255

14.2 DEVELOPING THE READING ASSESSMENT FRAMEWORK 255

14.3 READING FRAMEWORK AND ASSESSMENT DESIGN PRINCIPLES 256

14.4 FRAMEWORK FOR THE 1998 READING ASSESSMENT 257

Chapter 14 ASSESSMENT FRAMEWORKS AND INSTRUMENTS FOR THE 1998 NATIONAL AND STATE READING ASSESSMENTS (continued)

14.5 DEVELOPING THE READING COGNITIVE ITEMS ... 260

14.6 DEVELOPING THE READING OPERATIONAL FORMS ... 261

14.7 DISTRIBUTION OF READING ASSESSMENT ITEMS ... 261

14.8 BACKGROUND QUESTIONNAIRES FOR THE 1998 READING ASSESSMENT ... 263

14.8.1 Student Reading Questionnaires ... 263

14.8.2 Language Arts Teacher Questionnaire ... 264

14.9 STUDENT BOOKLETS FOR THE 1998 READING ASSESSMENT ... 265

Chapter 15 INTRODUCTION TO THE DATA ANALYSIS FOR THE NATIONAL AND STATE READING ASSESSMENTS

Jinming Zhang, Jiahe Qian, and Steven P. Isham, Educational Testing Service ... **269**

15.1 INTRODUCTION ... 269

15.2 DESCRIPTION OF STUDENT SAMPLES, ITEMS, ASSESSMENT BOOKLETS, AND ADMINISTRATIVE PROCEDURES ... 269

15.3 SCORING CONSTRUCTED-RESPONSE ITEMS ... 275

15.4 DIF ANALYSIS ... 276

15.5 THE WEIGHT FILES ... 280

Chapter 16 DATA ANALYSIS OF THE NATIONAL READING ASSESSMENT

Jinming Zhang, Steven P. Isham, and Lois H. Worthington, Educational Testing Service ... **283**

16.1 INTRODUCTION ... 283

16.2 NATIONAL ITEM ANALYSES ... 283

16.2.1 Conventional Item and Test Analyses ... 283

16.2.2 Scoring the Constructed-Response Items ... 288

16.3 NATIONAL IRT SCALING ... 288

16.3.1 Overview of Item Parameter Estimation ... 288

16.3.2 Evaluation of Model Fit ... 289

16.3.2.1 Items Deleted from the Final Scale ... 289

16.3.2.2 Recoded Polytomous Items ... 298

16.3.2.3 Item Category Response Functions (ICRF's) Common Across Assessment Years ... 299

16.4 GENERATION OF PLAUSIBLE VALUES ... 300

16.5 THE FINAL READING SCALES ... 302

16.5.1 Purpose-for-Reading Scales ... 302

16.5.2 The Composite Reading Scale ... 303

16.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE ... 304

16.7 READING TEACHER QUESTIONNAIRES ... 305

Chapter 17 DATA ANALYSIS OF THE STATE READING ASSESSMENT
Jiahe Qian, Steven P. Isham, Lois H. Worthington, and Jo-Lin Liang, Educational Testing Service **307**

17.1 INTRODUCTION 307
17.2 STATE ITEM AND TEST ANALYSES 307
17.3 STATE IRT SCALING 316
17.3.1 Item Parameter Estimation 318
17.3.2 Recoded Extended Constructed-Response Items 324
17.4 GENERATION OF PLAUSIBLE VALUES 329
17.5 THE FINAL SCORE SCALES 336
17.5.1 Linking State and National Scales 336
17.5.2 Producing a Reading Composite Scale 341
17.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE 342
17.7 READING TEACHER QUESTIONNAIRES 342

Chapter 18 ASSESSMENT FRAMEWORKS AND INSTRUMENTS FOR THE 1998 NATIONAL AND STATE WRITING ASSESSMENTS
Elissa A. Greenwald and Terry L. Schoeps, Educational Testing Service **345**

18.1 INTRODUCTION 345
18.2 DEVELOPING THE WRITING ASSESSMENT FRAMEWORK 345
18.3 WRITING FRAMEWORK AND ASSESSMENT DESIGN PRINCIPLES 346
18.4 FRAMEWORK FOR THE 1998 WRITING ASSESSMENT 346
18.5 DEVELOPING THE WRITING COGNITIVE ITEMS 348
18.6 DEVELOPING THE WRITING OPERATIONAL FORMS 348
18.7 DISTRIBUTION OF WRITING ASSESSMENT ITEMS 349
18.8 BACKGROUND QUESTIONNAIRES FOR THE 1998 WRITING ASSESSMENT 353
18.8.1 Student Writing Questionnaires 353
18.8.2 Language Arts Teacher Questionnaire 354
18.9 STUDENT BOOKLETS FOR THE 1998 WRITING ASSESSMENT 355
18.10 WRITING CLASSROOM-BASED STUDY IN 1998 355

Chapter 19 INTRODUCTION TO THE DATA ANALYSIS FOR THE NATIONAL AND STATE WRITING SAMPLES
Frank Jenkins, Jiahe Qian, Hua-Hua Chang, and Bruce A. Kaplan, Educational Testing Service **359**

19.1 INTRODUCTION 359
19.2 DESCRIPTION OF STUDENT SAMPLES, ITEMS, ASSESSMENT BOOKLETS, AND ADMINISTRATIVE PROCEDURES 359
19.3 SCORING CONSTRUCTED-RESPONSE ITEMS 367

Chapter 19 INTRODUCTION TO THE DATA ANALYSIS FOR THE NATIONAL AND STATE WRITING SAMPLES (continued)

19.4 DIFFERENTIAL ITEM FUNCTIONING 368

19.5 50-MINUTE WRITING STUDY 369

19.6 THE WEIGHT FILES 369

Chapter 20 DATA ANALYSIS FOR THE NATIONAL WRITING SAMPLES

Frank Jenkins, Bruce A. Kaplan, and Youn-Hee Lim, Educational Testing Service **371**

20.1 INTRODUCTION 371

20.2 NATIONAL ITEM ANALYSIS 371

20.3 ITEM RESPONSE THEORY (IRT) SCALING 375

20.3.1 Item Parameter Estimation 375

20.3.2 Evaluation of Model Fit 376

20.4 GENERATION OF PLAUSIBLE VALUES 377

20.4.1 Principal Components (NSWEEP Program) 377

20.4.2 Conditioning (BGROUP Program) 379

20.5 FINAL REPORTING SCALES 379

20.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE 380

20.7 WRITING TEACHER QUESTIONNAIRES 380

Chapter 21 DATA ANALYSIS OF THE STATE WRITING ASSESSMENT

Jiahe Qian, Hua-Hua Chang, Bruce A. Kaplan, Jo-Lin Liang, and Youn-Hee Lim, Educational Testing Service **381**

21.1 INTRODUCTION 381

21.2 STATE ITEM ANALYSES 382

21.2.1 Conventional Item and Test Analyses 382

21.3 STATE IRT SCALING 385

21.3.1 Samples Used in State IRT Scaling 385

21.3.2 Item Parameter Estimation 387

21.4 GENERATION OF PLAUSIBLE VALUES 388

21.5 FINAL SCORE SCALES 393

21.5.1 Linking State and National Scales 393

21.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE 396

21.7 WRITING TEACHER QUESTIONNAIRES 398

Chapter 22 ASSESSMENT FRAMEWORKS AND INSTRUMENTS FOR THE 1998 CIVICS ASSESSMENT
Andrew R. Weiss and Terry L. Schoeps, Educational Testing Service **399**

22.1 INTRODUCTION 399
22.2 DEVELOPING THE CIVICS ASSESSMENT FRAMEWORK 399
22.3 CIVICS FRAMEWORK AND ASSESSMENT DESIGN PRINCIPLES 400
22.4 FRAMEWORK FOR THE 1998 CIVICS ASSESSMENT 400
22.5 DEVELOPING THE CIVICS COGNITIVE ITEMS 401
22.6 DEVELOPING THE CIVICS OPERATIONAL FORMS 403
22.7 DISTRIBUTION OF CIVICS ASSESSMENT ITEMS 404
22.8 BACKGROUND QUESTIONS FOR THE 1998 CIVICS ASSESSMENT 404
22.8.1 Student Civics Questionnaires 405
22.8.2 Civics Teacher Questionnaire 405
22.9 STUDENT BOOKLETS FOR THE 1998 CIVICS ASSESSMENT 407
22.10 CIVICS SPECIAL TREND STUDY IN 1998 411

Chapter 23 INTRODUCTION TO THE DATA ANALYSIS FOR THE CIVICS ASSESSMENT
Spencer S. Swinton and Edward Kulick, Educational Testing Service **413**

23.1 INTRODUCTION 413
23.2 DESCRIPTION OF STUDENT SAMPLES, ITEMS, ASSESSMENT BOOKLETS, AND ADMINISTRATIVE PROCEDURES 413
23.3 SCORING CONSTRUCTED-RESPONSE ITEMS 417
23.4 DIF ANALYSIS 417
23.5 THE WEIGHT FILES 420

Chapter 24 DATA ANALYSIS FOR THE CIVICS ASSESSMENT
Spencer S. Swinton, Edward Kulick, and Venus Leung, Educational Testing Service **421**

24.1 INTRODUCTION 421
24.2 ITEM ANALYSIS 421
24.2.1 Constructed-Response Items 425
24.3 ITEM RESPONSE THEORY (IRT) SCALING 425
24.3.1 Evaluating the Fit of the IRT Model 426
24.3.2 Derived Background Variables 428
24.4 GENERATION OF PLAUSIBLE VALUES 435
24.5 TRANSFORMATION OF THE CIVICS CALIBRATION SCALE FOR REPORTING 436
24.6 PARTITIONING OF THE ESTIMATION ERROR VARIANCE 437
24.7 CIVICS TEACHER QUESTIONNAIRE 437

Appendix A **STATISTICAL SUMMARY OF THE 1998 NAEP SAMPLES**
Bruce A. Kaplan and Youn-Hee Lim, Educational Testing Service **439**

A.1 INTRODUCTION 439
A.2 MEASUREMENT INSTRUMENTS 439
A.3 SAMPLE CHARACTERISTICS 440
A.4 POPULATION ESTIMATES 440

Appendix B **SUMMARY INFORMATION FOR THE NAEP 1998 STATE SAMPLES AND FOR WEIGHTING THE NAEP 1998 STATE SAMPLES**
Keith F. Rust and Leslie Wallace, Westat **473**

Appendix C **CONSTRUCTED-RESPONSE ITEM SCORE STATISTICS** **563**

Appendix D **DIFFERENTIAL ITEM FUNCTIONING (DIF) RESULTS** **585**

Appendix E **IRT PARAMETERS** **587**

Appendix F **CONDITIONING VARIABLES AND CONTRAST CODINGS** **627**

Appendix G **REPORTING SUBGROUPS AND SPECIAL VARIABLES FOR THE 1998 NAEP ASSESSMENT** **823**

G.1 Major Reporting Subgroups 823
G.2 Writing Derived Variables 827
G.3 Civics Derived Variables 828
G.4 Variables Related to Scaling 829
G.5 Quality Education Data (QED) Variables 830

Appendix H **ESTIMATION ERROR VARIANCE OF THE MEAN BY GENDER AND RACE/ETHNICITY** **831**

Appendix I **SETTING THE ACHIEVEMENT LEVELS FOR THE 1998 NAEP READING ASSESSMENT**
Mary Lyn Bourque, National Assessment Governing Board **841**

I.1 INTRODUCTION 841
I.2 1992 Preparation for the Reading Level Setting Meeting 842
I.3 1992 Reading Level Setting Panel 842
I.4 1992 Process for Developing the Achievements Levels 843
I.5 1992 Process for Selecting Exemplar Items 845

Appendix I **SETTING THE ACHIEVEMENT LEVELS FOR THE 1998 NAEP READING ASSESSMENT (continued)**

I.6 1992 Process for Validating the Levels....847

I.7 Evaluation of the 1992 Levels....848

I.8 1994 Process for Validating the Levels....848

I.9 1994 Exemplars....849

I.10 Mapping the Levels onto the NAEP Scale....849

Appendix J **SETTING THE ACHIEVEMENT LEVELS FOR THE 1998 NAEP CIVICS AND WRITING ASSESSMENTS**
Mary Lyn Bourque, National Assessment Governing Board **869**

J.1 INTRODUCTION....869

J.2 PREPARING THE FINAL DESCRIPTIONS....869

J.3 1998 FIELD TRIALS IN CIVICS AND WRITING....870

J.4 PREPARATION FOR CIVICS AND WRITING LEVEL SETTING MEETINGS....870

J.5 1998 PILOT STUDIES IN CIVICS AND WRITING....871

J.6 RESULTS OF THE 1998 PILOT STUDIES....874

J.7 1998 LEVEL-SETTING PANELS....875

J.8 1998 PROCESS FOR DEVELOPING THE ACHIEVEMENT LEVELS....875

J.9 MAPPING THE LEVELS ONTO THE NAEP SCALE....877

J.10 ADDITIONAL ANALYSIS OF THE 1998 DATA....878

J.11 SELECTING EXEMPLAR ITEMS....885

J.12 1998 RESEARCH STUDIES ON THE ACHIEVEMENT LEVELS....886

Appendix K **PARTICIPANTS IN THE OBJECTIVES AND ITEM DEVELOPMENT PROCESS**....**915**

References **921**

THE NAEP 1998 TECHNICAL REPORT

◆ List of Tables & Figures ◆

Chapter 1 Overview of Part I: The Design and Implementation of the 1998 NAEP

Figure 1-1 Subsamples of the 1998 NAEP Reading Assessment 8
Table 1-1 NAEP 1998 Student Samples 9
Table 1-2 National Assessment of Educational Progress Subject Areas, Grades, and Ages Assessed: 1969–1998 12
Table 1-3 An Example of a BIB Design 20
Table 1-4 An Example of a PBIB Design 21

Chapter 3 Sample Design for the National Assessment

Table 3-1 1998 NAEP National Samples and Target Sample Sizes 33
Table 3-2 Definition of NAEP Stratification and Reporting Regions 35
Table 3-3 The 22 Largest Primary Sampling Units, by Region 1998 NAEP 36
Table 3-4 The Number of Noncertainty Strata in Each Major Stratum 1998 NAEP 37
Table 3-5 Number of Schools Eligible in QED and PSS Sampling Frame Components by Grade, 1998 Main NAEP 39
Table 3-6 Participation Rates in 1996 National NAEP 40
Table 3-7 Number of Schools in the Original Samples by Major Stratum 41
Table 3-8 Summary of School Participation Experience for 1998 National NAEP, Unweighted 43
Table 3-9 Percentage of Assessed and Absent Students Who Were Specified as SD/LEP National 1998 Reading Samples 50
Table 3-10 Number of Students Per School for Each Subject Type for 1998 National Assessments 53
Table 3-11 Weighted Percentages of Students Excluded (SD and LEP) from 1998 National Reading Assessment 54
Table 3-12 Weighted and Unweighted Distribution of Students Excluded for 1998 National Assessments, by Reason for Exclusion, Subject, and Grade 55
Table 3-13 Student Exclusion Rates for 1998 National Assessment by Grade, School Type, and Sample Type, Weighted 56
Table 3-14 Comparison of Target Assessments to Actual Assessments for 1998 National Samples, by Grade 57
Table 3-15 Unweighted Student Participation Rates for National Assessments, by Grade and School Type 57
Table 3-16 Overall Unweighted Participation Rates (School and Student Combined) for 1998 National Assessments, by Grade 58
Table 3-17 Weighted Participation Rates by Grade and Subject Type for the 1998 National Reporting Samples 58

Chapter 4 Sample Design for the State Assessment

Table 4-1 Distribution of Fourth-Grade Schools and Enrollment in Combined Sampling Frame for 1998 NAEP State Assessments 64
Table 4-2 Distribution of Eighth-Grade Schools and Enrollment in Combined Sampling Frame for 1998 NAEP State Assessments 65
Table 4-3 Estimated Grade Enrollment and Measure of Size, Grade 4 69
Table 4-4 Estimated Grade Enrollment and Measure of Size, Grade 8 69
Table 4-5 The Effect of the Sparse State Option on Sample Sizes, by Grade for Jurisdictions Exercising the Option 71
Table 4-6 Number of Schools Selected for Both State and National NAEP, by Grade and School Type 72
Table 4-7 Jurisdictions Where All Schools Were Selected, by Grade and School Type 72
Table 4-8 NAEP 1998 Distribution of New Schools Coming from Districts Designated as "Medium" or "Large" 74

Chapter 5 Field Operations and Data Collection

Table 5-1 Jurisdictions Participating in the 1998 State Assessment Program 83
Table 5-2 Background Questionnaires Received for Schools, Teachers, and SD/LEP Students in the 1998 National Assessment 98
Table 5-3 School Participation, 1998 State Assessment 100
Table 5-4 Student Participation, 1998 State Assessment 101

Chapter 6 Processing Assessment Materials

Table 6-1 Number of Sessions and Student Booklets Processed for the 1998 National and State Assessments 106
Table 6-2 Questionnaire Totals for the 1998 NAEP Assessment 107

Chapter 7 Professional Scoring

Table 7-1 Processing and Scoring Totals for the 1998 NAEP Assessment 117
Figure 7-1 Image Scoring Flow Chart 118
Figure 7-2 Paper Scoring Flow Chart 119
Table 7-2 Interrater Reliability Ranges for the NAEP 1998 Assessment 125
Table 7-3 Number of Constructed-Response Items by Score-Point Levels for the 1998 NAEP Reading Assessment 126
Table 7-4 Number of Constructed-Response Items by Score-Point Levels for the 1998 NAEP Writing Assessment 127
Table 7-5 Number of Constructed-Response Items by Score-Point Levels for the 1998 NAEP Civics Assessment 129

Chapter 8 Creation of the Database, Quality Control of Data Entry, and Creation of the Database Products

Table 8-1 Summary of Quality Control Error Analysis for NAEP 1998 Data Entry 135

Table 10-1 Reporting Samples for 1998 NAEP National Assessments 163
Table 10-2 Session Allocation Weights Used in the 1998 National Assessment 165
Table 10-3 1998 National Assessment Writing and Civics Sample Allocation 166
Table 10-4 Value of Factor F for Sample Subjects Used in the 1998 National Assessment 169
Table 10-5 Distribution of Populations of Eligible Students Based on Trimmed Weights of Assessed Students in Participating Schools, 1998 National 25-Minute Writing Samples 170
Table 10-6 1998 National Reading Assessment Reporting Factors for Assessed and Excluded Students 171
Table 10-7 Major Subgroups for Poststratification in the 1998 National Assessment 172
Table 10-8 Distributions of Final Student Weights for 1998 National Reporting Samples 174
Table 10-9 Distribution of Modular Weights Used in the 1998 National Assessment 175
Table 10-10 First and Second Categorical Variables Used for Raking 176
Table 10-11 Third Categorical Variable Used for Raking 176
Table 10-12 Percentiles of Raking Adjustments 177
Table 10-13 Reporting Factors for the Reporting Weights with Accommodations for the 1998 National Reading Assessment 178
Table 10-14 Distribution of Accommodated Reporting Weights for the 1998 National Reading Assessment 179
Table 10-15 Distribution of Populations of Eligible Students Based on Full Weighted Sample of Eligible Schools, Before and After School Nonresponse Adjustments, 1998 National 25-Minute Writing Samples 180
Table 10-16 Distribution of Populations of Eligible Students Before and After Student Nonresponse Adjustments, 1998 National 25-Minute Writing Samples 181
Table 10-17 Distribution of Populations of Eligible Students Before School and Student Nonresponse Adjustments, 1998 National 25-Minute Writing Samples 182
Table 10-18 Weighted Distribution of Absent Students by Nature of Absenteeism for All Grades, 1998 National 25-Minute Writing Samples 183
Table 10-19 Design Effects by Demographic Subgroup and Grade for Mean Reading Scale Scores 189
Table 10-20 Design Effects by Demographic Subgroup and Grade for Mean Writing Scale Scores 190
Table 10-21 Design Effects by Demographic Subgroup and Grade for Mean Civics Scale Scores 190
Table 10-22 Within-Grade Mean, Median, and Upper Quartile of the Distribution of Design Effects for 1998 National Assessments by Subject Area and Across Subject Areas 191

Chapter 11 State Weighting Procedures and Variance Estimation

Table 11-1 Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 4 Public Schools, 1998 Reading State Samples 201

Table 11-2 Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 8 Public Schools, 1998 Reading State Samples 202

Table 11-3 Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 8 Public Schools, 1998 Writing State Samples 203

Table 11-4 Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 4 Nonpublic Schools, 1998 Reading State Samples 205

Table 11-5 Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 8 Nonpublic Schools, 1998 Reading State Samples 206

Table 11-6 Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 8 Nonpublic Schools, 1998 Writing State Samples 207

Table 11-7 Jurisdictions Included in Logistic Regression Analysis of the NAEP 1998 State Assessment 209

Table 11-8 Results of Logistic Regression Analysis of School Nonresponse - Grade 4, 1998 Reading State Samples 212

Table 11-9 Results of Logistic Regression Analysis of School Nonresponse - Grade 8, 1998 Reading State Samples 213

Table 11-10 Results of Logistic Regression Analysis of School Nonresponse - Grade 8, 1998 Writing State Samples 214

Table 11-11 Average Design Effects by Demographic Subgroup for 1998 Mean State Reading and Writing Scale Scores Averaged Across State Samples 225

Table 11-12 Mean, Median, and Upper Quartile of the 1998 Across-State Average Design Effects for Mean State Scale Scores (Distribution Across Demographic Subgroups) 225

Chapter 12 Scaling Procedures

Figure 12-1 Dichotomous Item (R016102) Exhibiting Good Model Fit 234

Figure 12-2 Polytomous Item (HC00004) Exhibiting Good Model Fit 234

Figure 12-3 Dichotomous Item (M017901) Exhibiting Good Model Fit Across Assessment Years 235

Figure 12-4 Dichotomous Item (M018901) Exhibiting Different Empirical Item Functions for Different Assessment Years 236

Table 12-1 Estimation Error Variance and Related Coefficients for the 1992 Grade 4 Reading Composite (Based on Five Plausible Values) 244

Chapter 14 Assessment Frameworks and Instruments for the 1998 National and State Reading Assessments

Figure 14-1 Description of Reading Stances 258
Figure 14-2 Description of Purposes for Reading 259
Table 14-1 Percentage Distribution of Items by Reading Purpose as Specified in the NAEP Reading Framework 260
Table 14-2 Percentage Distribution of Items by Reading Stance as Specified in the NAEP Reading Framework 260
Figure 14-3 Distribution of Items for the 1998 Reading Assessment 262
Table 14-3 Percentage Distribution of Assessment Time by Grade and Reading Purpose for the NAEP 1998 Reading Assessment 262
Table 14-4 Percentage Distribution of Assessment Time by Grade and Reading Stance for the NAEP 1998 Reading Assessment 263
Table 14-5 NAEP 1998 Background Sections of Student Reading Booklets 264
Table 14-6 NAEP 1998 Reading Grade 4 Booklet Configuration 266
Table 14-7 NAEP 1998 Reading Grade 8 Booklet Configuration 267
Table 14-8 NAEP 1998 Reading Grade 12 Booklet Configuration 268

Chapter 15 Introduction to the Data Analysis for the National and State Reading Assessments

Table 15-1 NAEP 1998 Reading Student Samples 270
Table 15-2 1998 Reading Blocks and Items Common to the 1992 and 1994 Assessments 271
Table 15-3 Number of Items in Subscales in the Reading Main Assessment, by Reading Purposes 271
Table 15-4 1998 NAEP Reading Block Composition by Purpose for Reading and Item Type As Defined Before Scaling, Grade 4 272
Table 15-5 1998 NAEP Reading Block Composition by Purpose for Reading and Item Type As Defined After Scaling, Grade 4 272
Table 15-6 1998 NAEP Reading Block Composition by Purpose for Reading and Item Type As Defined Before Scaling, Grade 8 273
Table 15-7 1998 NAEP Reading Block Composition by Purpose for Reading and Item Type As Defined After Scaling, Grade 8 273
Table 15-8 1998 NAEP Reading Block Composition by Purpose for Reading and Item Type As Defined Before Scaling, Grade 12 274
Table 15-9 1998 NAEP Reading Block Composition by Purpose for Reading and Item Type As Defined After Scaling, Grade 12 274
Table 15-10 DIF Category for National Samples by Grade for Dichotomous Items 277
Table 15-11 DIF Category for National Samples by Grade for Polytomous Items 278
Table 15-12 The Category of DIF between Public and Nonpublic Schools for State Samples, by Grade for Dichotomous Items 279
Table 15-13 The Category of DIF between Public and Nonpublic Schools for State Samples, by Grade for Polytomous Items 280

Chapter 16 Data Analysis of the National Reading Assessment

Table 16-1 Descriptive Statistics for Item Blocks by Position Within Test Booklet and Overall Occurrences for the National Main Reading Sample, Grade 4, As Defined After Scaling 285

Table 16-2 Descriptive Statistics for Item Blocks by Position Within Test Booklet and Overall Occurrences for the National Main Reading Sample, Grade 8, As Defined After Scaling 286
Table 16-3 Descriptive Statistics for Item Blocks by Position Within Test Booklet and Overall Occurrences for the National Main Reading Sample, Grade 12, As Defined After Scaling 287
Figure 16-1 Dichotomous Item (R017002) Exhibiting Good Model Fit 290
Figure 16-2 Polytomous Item (R017104) Exhibiting Good Model Fit 291
Figure 16-3 Polytomous Item (R016603) Exhibiting Unacceptably Poor Model Fit 292
Figure 16-4 Polytomous Item (R017110) Exhibiting Poor Model Fit 293
Figure 16-5 Dichotomous Item (R017110) After Collapsing Categories 1 and 2 294
Figure 16-6 Short-Term Trend Polytomous Item (R016210) Demonstrating Differential Item Functioning Across Assessment Years 1994 and 1998 295
Figure 16-7a Short-Term Trend Polytomous Item (R016210) Fitting Separate Item Response Functions for Each Assessment Year 296
Figure 16-7b Short-Term Trend Polytomous Item (R016210) Fitting Separate Item Response Functions for Each Assessment Year 297
Table 16-4 Items Deleted from the Final Scaling 298
Table 16-5 Recoding of Polytomous Items for Scaling 298
Table 16-6 Grade 4 Items Scaled Separately by Assessment Years 299
Table 16-7 Grade 8 Items Scaled Separately by Assessment Years 300
Table 16-8 Grade 12 Items Scaled Separately by Assessment Years 300
Table 16-9 Proportion of Scale Score Variance Accounted for by the Conditioning Model for the National Main Reading Assessment 301
Table 16-10 Conditional Correlations and Variances from Conditioning (CGROUP) 302
Table 16-11 Marginal Correlations of Reading Scales 302
Table 16-12 Coefficients of Linear Transformations of the Purpose-for-Reading Scales from the Calibrating Scale Units to the Units of the Reporting Scale 303
Table 16-13 Weighting of the Purpose-for-Reading Scales on the Reading Composite Scale 303
Table 16-14 Means and Standard Deviations on the Reading Composite Scale 304
Table 16-15 Estimation Error Variance and Related Coefficients for the National Main Reading Assessment 304

Chapter 17 Data Analysis of the State Reading Assessment

Table 17-1 Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall – Public Schools, Grade 4 308
Table 17-2 Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall – Nonpublic Schools, Grade 4 309
Table 17-3 Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall – Public Schools, Grade 8 310
Table 17-4 Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall – Nonpublic Schools, Grade 8 311
Table 17-5 Block-Level Descriptive Statistics for Monitored and Unmonitored Public-School Sessions, Grade 4 312
Table 17-6 Block-Level Descriptive Statistics for Monitored and Unmonitored Nonpublic-School Sessions, Grade 4 312

Table 17-7 Block-Level Descriptive Statistics for Monitored and Unmonitored Public-School Sessions, Grade 8 313
Table 17-8 Block-Level Descriptive Statistics for Monitored and Unmonitored Nonpublic-School Sessions, Grade 8 313
Table 17-9 Effect of Monitoring Sessions by Jurisdiction: Average Jurisdiction Item Scores for Monitored and Unmonitored Sessions, Grade 4 315
Table 17-10 Effect of Monitoring Sessions by Jurisdiction: Average Jurisdiction Item Scores for Monitored and Unmonitored Sessions, Grade 8 316
Table 17-11 Block-Level Descriptive Statistics for Overall Public- and Nonpublic-School Sessions, Grade 4 317
Table 17-12 Block-Level Descriptive Statistics for Overall Public- and Nonpublic-School Sessions, Grade 8 317
Table 17-13 Extended Constructed-Response Items, 1998 State Assessment in Reading 319
Figure 17-1 Dichotomous Items (R012106, R012711, and R013405) Exhibiting Good Model Fit 321
Figure 17-2 Polytomous Item (R013201) Exhibiting Good Model Fit 323
Figure 17-3 Polytomous Item (R012111) Before Collapsing Unsatisfactory and Partial-Response Categories 324
Figure 17-4 Polytomous Item (R012111) After Collapsing Unsatisfactory and Partial-Response Categories 326
Figure 17-5 Polytomous Item (R017110) After Collapsing Unsatisfactory and Partial-Response Categories 327
Figure 17-6 Polytomous Item (R016212) After Collapsing Unsatisfactory and Partial-Response Categories 328
Table 17-14 Summary Statistics for State Assessment Conditioning Models, Grade 4 330
Table 17-15 Summary Statistics for State Assessment Conditioning Models, Grade 8 331
Table 17-16 Average Correlations and Ranges of Scale Correlations Among the Reading Scales for 40 Jurisdictions for Grade 8 333
Figure 17-7 Plot of Mean Scale Score Versus Mean Item Score by Jurisdiction, Grade 4 334
Figure 17-8 Plot of Mean Scale Score Versus Mean Item Score by Jurisdiction, Grade 8 335
Table 17-17 Coefficients of Linear Transformations for the 1998 State Reading Assessment 338
Figure 17-9 Rootogram Comparing Scale Score Distributions for the State Assessment Aggregate Sample and the National Linking Sample for the Reading for Literary Experience Scale, Grade 4 339
Figure 17-10 Rootogram Comparing Scale Score Distributions for the State Assessment Aggregate Sample and the National Linking Sample for the Reading to Gain Information Scale, Grade 4 339
Figure 17-11 Rootogram Comparing Scale Score Distributions for the State Assessment Aggregate Sample and the National Linking Sample for the Reading for Literary Experience Scale, Grade 8 340
Figure 17-12 Rootogram Comparing Scale Score Distributions for the State Assessment Aggregate Sample and the National Linking Sample for the Reading to Gain Information Scale, Grade 8 340
Figure 17-13 Rootogram Comparing Scale Score Distributions for the State Assessment Aggregate Sample and the National Linking Sample for the Reading to Perform a Task Scale, Grade 8 340
Figure 17-14 Rootogram Comparing Scale Score Distributions for the State Assessment Aggregate Sample and the National Linking Sample for the Reading Composite Scale, Grade 4 341

Figure 17-15 Rootogram Comparing Scale Score Distributions for the State Assessment Aggregate Sample and the National Linking Sample for the Reading Composite Scale, Grade 8 342
Table 17-18 Estimation Error Variance and Related Coefficients for the Reading State Assessment, Grade 4 343
Table 17-19 Estimation Error Variance and Related Coefficients for the Reading State Assessment, Grade 8 344

Chapter 18 Assessment Frameworks and Instruments for the 1998 National and State Writing Assessments

Figure 18-1 Description of NAEP 1998 Writing Purposes 347
Table 18-1 Percentage Distribution of Items by Purpose for Writing as Specified in the NAEP Writing Framework 348
Figure 18-2 NAEP 1998 Forms of Writing 348
Table 18-2 NAEP 1998 Writing Grade 4 Blocks by Title and Purpose 350
Table 18-3 NAEP 1998 Writing Grade 8 Blocks by Title and Purpose 351
Table 18-4 NAEP 1998 Writing Grade 12 Blocks by Title and Purpose 352
Table 18-5 Percentage Distribution of Assessment Time by Grade and Purpose for Writing for the NAEP 1998 Writing Assessment 353
Table 18-6 NAEP 1998 Background Sections of Student Writing Booklets 354
Table 18-7 NAEP 1998 National and State Writing Grade 4 Booklet Configuration 356
Table 18-8 NAEP 1998 National and State Writing Grade 8 Booklet Configuration 357
Table 18-9 NAEP 1998 National and State Writing Grade 12 Booklet Configuration 358

Chapter 19 Introduction to the Data Analysis for the National and State Writing Samples

Table 19-1 NAEP 1998 Writing Student Samples 360
Table 19-2 Number of 25-Minute Items in the National Main Writing Assessment Within the Three Purposes of Writing 361
Table 19-3 Number of 50-Minute Items in the National Writing Assessment Within the Three Purposes of Writing 361
Table 19-4 Grade 4: Prompt, Block, and Purpose Correspondence 362
Table 19-5 Grade 8: Prompt, Block, and Purpose Correspondence 363
Table 19-6 Grade 12: Prompt, Block, and Purpose Correspondence 364
Table 19-7 Correspondence of Prompts, Blocks, and Books: Grade 4 365
Table 19-8 Correspondence of Prompts, Blocks, and Books: Grade 8 366
Table 19-9 Correspondence of Prompts, Blocks, and Books: Grade 12 367
Table 19-10 Items with Absolute SMD (Standardized Mean DIF) >.10 369

Chapter 20 Data Analysis for the National Writing Samples

Table 20-1 Descriptive Statistics for 25-Minute Writing Prompts: Grade 4 372
Table 20-2 Descriptive Statistics for 25-Minute Writing Prompts: Grade 8 373
Table 20-3 Descriptive Statistics for 25-Minute Writing Prompts: Grade 12 374
Figure 20-1 Polytomous Item (W010002) Exhibiting Good Model Fit 377
Figure 20-2 Polytomous Item (W008402) Exhibiting Less Than Optimal Model Fit 378
Table 20-4 Proportion of Scale Score Variance Accounted for by the Conditioning Model for the 1998 National Main Writing Assessment 378
Table 20-5 Coefficients of Linear Transformations of the Writing Scales from the Scaling Metric to the Reporting Metric 379
Table 20-6 Estimation Error Variance and Related Coefficients for the National Main Writing Assessment 380

Chapter 21 Data Analysis of the State Writing Assessment

Table 21-1 Descriptive Statistics Writing Prompts, Writing 25-Minute State Samples, Grade 8 383
Table 21-2 Descriptive Statistics for Each Item of the Writing State Assessment Using Senate Weights (Scaled from 0 to 5), Grade 8 384
Table 21-3 Effect of Monitoring Sessions by Jurisdiction: Average Jurisdiction Item Scores for Monitored and Unmonitored Sessions, Grade 8 386
Figure 21-1 Polytomous Item (W006502) Exhibiting Good Model Fit 389
Figure 21-2 Polytomous Item (W007602) Exhibiting Good Model Fit 390
Table 21-4 Proportion of Scale Score Variance Accounted by Conditioning Model for the Writing State Assessment, Grade 8 391
Figure 21-3 Plot of Mean Scale Score Versus Mean Item Score by Jurisdiction, Grade 8 393
Table 21-5 Coefficients of Linear Transformations for the 1998 State Writing Assessment 395
Figure 21-4 Rootogram Comparing Scale Score Distributions for the State Assessment Aggregate Sample and the National Linking Sample for the Composite Scale, Grade 8 396
Table 21-6 Estimation Error Variance and Related Coeffecients for the Writing State Assessment, Grade 8 397

Chapter 22 Assessment Frameworks and Instruments for the 1998 Civics Assessment

Table 22-1 Percentage Distribution of Questions by Intellectual Skill as Recommended in the NAEP Civics Framework 401
Table 22-2 Actual Percentage Distribution of Questions by Intellectual Skill 401
Figure 22-1 Description of the NAEP 1998 Civics Framework Components 402
Table 22-3 NAEP 1998 Civics Assessment Percentage of Student Assessment Time by Question Format 403
Figure 22-2 Distribution of Items for the 1998 Civics Assessment 404
Table 22-4 NAEP 1998 Background Sections of Student Civics Booklets 405
Table 22-5 NAEP 1998 Civics Grade 4 Booklet Configuration 408
Table 22-6 NAEP 1998 Civics Grade 8 Booklet Configuration 409
Table 22-7 NAEP 1998 Civics Grade 12 Booklet Configuration 410

Chapter 23 Introduction to the Data Analysis for the Civics Assessment

Table 23-1 NAEP 1998 National Main Civics Assessment Student Samples 414
Table 23-2 Number of Items in the National Main Civics Assessment by Content Area 415
Table 23-3 1998 NAEP Civics Block Composition As Defined Before Scaling, Grade 4 415
Table 23-4 1998 NAEP Civics Block Composition After Scaling, Grade 4 415
Table 23-5 1998 NAEP Civics Block Composition As Defined Before Scaling, Grade 8 416
Table 23-6 1998 NAEP Civics Block Composition After Scaling, Grade 8 416
Table 23-7 1998 NAEP Civics Block Composition As Defined Before Scaling, Grade 12 416
Table 23-8 1998 NAEP Civics Block Composition After Scaling, Grade 12 417
Table 23-9 DIF Category by Grade for Dichotomous Civics Items 419
Table 23-10 DIF Category by Grade for Polytomous Civics Items 420

Chapter 24 Data Analysis for the Civics Assessment

Table 24-1 Descriptive Statistics for Item Blocks by Position Within Test Booklet and Overall Occurrences for the National Main Civics Sample, Grade 4, As Defined After Scaling 422
Table 24-2 Descriptive Statistics for Item Blocks by Position Within Test Booklet and Overall Occurrences for the National Main Civics Sample, Grade 8, As Defined After Scaling 423
Table 24-3 Descriptive Statistics for Item Blocks by Position Within Test Booklet and Overall Occurrences for the National Main Civics Sample, Grade 12, As Defined After Scaling 424
Table 24-4 1998 Civics Items Receiving Special Treatment 426
Figure 24-1 Dichotomous Item (P040719) Exhibiting Good Model Fit 428
Figure 24-2 Polytomous Item (P042008) Exhibiting Good Model Fit 429
Figure 24-3 Dichotomous Item (P041209) Exhibiting Moderate Model Misfit 430
Figure 24-4 Polytomous Item (P041902) Exhibiting Moderate Model Misfit 431
Figure 24-5 Dichotomous Item (P040506) Exhibiting Poor Model Fit 432
Figure 24-6 Polytomous Item (P040402) Exhibiting Poor Model Fit in the Lower Two Categories 433
Figure 24-7 Same Polytomous Item (P040402) with the Lower Two Categories Collapsed, Now Exhibiting Improved Model Fit 434
Table 24-5 Proportion of Scale Score Variance Accounted for by the Conditioning Model for the National Main Civics Assessment 435
Table 24-6 Means and Standard Deviations for the Civics Scale 436
Table 24-7 Transformation Constants for the National Main Civics Assessment 437
Table 24-8 Estimation Error Variance and Related Coefficients for the National Main Civics Assessment 437

Appendix A Statistical Summary of the 1998 NAEP Samples

Table A-1 Measurement Instruments Used in 1998 NAEP 441
Table A-2 Number of Items Administered, by Sample and Age Class 442
Table A-3 School Participation in NAEP 1998 Main Samples (All Subsamples) 443
Table A-4 School Characteristics in NAEP 1998 Main Samples 444

Table A-5 Numbers of Responses to Teacher Questionnaires and Students Matched with Teacher Data 445
Table A-6 Number of Students Assessed, Accommodated, and Excluded, by Reporting Sample and Grade 446
Table A-7 Number of Students in the Reading Reporting Samples, by Subgroup Classification, National Grades 4, 8, and 12, & State Grades 4 and 8 447
Table A-8 Number of Students in the Writing 25-Minute and 50-Minute Samples by Subgroup Classification, Grades 4, 18, and 12, & State Grade 8 449
Table A-9 Number of Students in the Civics Main Samples by Subgroup Classification, Grades 4, 8, and 12 451
Table A-10 Number of Excluded Students in the Reading Reporting Samples by Subgroup Classification, Grades 4, 8, and 12, & State Grades 4 and 8 452
Table A-11 Number of Excluded Students in the Writing Samples by Subgroup Classification, Grades 4, 8, and 12, & State Grade 8 454
Table A-12 Number of Excluded Students in the Civics Main Samples by Subgroup Classification, Grades 4, 8, and 12 455
Table A-13 Number of Accommodated Students in the Writing Samples by Subgroup Classification, Grades 4, 8, and 12, & State Grade 8 456
Table A-14 Number of Accommodated Students in the Civics Main Samples by Subgroup Classification, Grades 4, 8, and 12 458
Table A-15 Weighted Percentages of Students in the Reading Reporting Samples by Subgroup Classification, National Grades 4, 8, and 12, & State Grades 4 and 8 460
Table A-16 Weighted Percentages of Students in the Writing 25-Minute and 50-Minute Samples by Subgroup Classification, Grades 4, 8, and 12, & State Grade 8 462
Table A-17 Weighted Percentages of Students in the Civics Main Samples by Subgroup Classification, Grades 4, 8, and 12 464
Table A-18 Weighted Percentages of Excluded Students in the Reading Reporting Samples by Subgroup Classification, National Grades 4, 8, and 12, & State Grades 4 and 8 466
Table A-19 Weighted Percentages of Excluded Students in the Writing Samples by Subgroup Classification, Grades 4, 8, and 12, & State Grade 8 467
Table A-20 Weighted Percentages of Excluded Students in the Civics Main Samples by Subgroup Classification, Grades 4, 8, and 12 468
Table A-21 Weighted Percentages of Accommodated Students in the Writing Samples by Subgroup Classification, Grades 4, 8, and 12, & State Grade 8 469
Table A-22 Weighted Percentages of Accommodated Students in the Civics Main Samples by Subgroup Classification, Grades 4, 8, and 12 471

Appendix B Summary Information for NAEP 1998 State Samples and Weighting NAEP 1998 State Samples

Table B-1 Weighted Mean Values Derived from Sampled Public Schools – Grade 4, Reading 475
Table B-2 Weighted Mean Values Derived from Sampled Public Schools – Grade 8, Reading 477
Table B-3 Weighted Mean Values Derived from Sampled Public Schools – Grade 8, Writing 479
Table B-4 Weighted Mean Values Derived from Sampled Nonpublic Schools – Grade 4, Reading 481

Table B-5 Weighted Mean Values Derived from Sampled Nonpublic Schools – Grade 8, Reading 482
Table B-6 Weighted Mean Values Derived from Sampled Nonpublic Schools – Grade 8, Writing 483
Table B-7 Weighted Student Percentages Derived from Sampled Public Schools – Grade 4, Reading 484
Table B-8 Weighted Student Percentages Derived from Sampled Public Schools – Grade 8, Reading 486
Table B-9 Weighted Student Percentages Derived from Sampled Public Schools – Grade 8, Writing 488
Table B-10 Weighted Student Percentages Derived from All Schools Sampled – Grade 4, Reading 490
Table B-11 Weighted Student Percentages Derived from All Schools Sampled – Grade 8, Reading 492
Table B-12 Weighted Student Percentages Derived from All Schools Sampled – Grade 8, Writing 493
Table B-13 Final Collapsed Levels Used for Raking Dimensions for All Jurisdictions – Grade 4, Reading 494
Table B-14 Final Collapsed Levels Used for Raking Dimensions for All Jurisdictions – Grade 8, Reading 496
Table B-15 Distribution of Selected Public Schools by Sampling Strata, Fourth Grade 498
Table B-16 Distribution of Selected Public Schools by Sampling Strata, Eighth Grade 517
Table B-17 Distribution of Selected Nonpublic Schools by Sampling Strata, Fourth Grade 536
Table B-18 Distribution of Selected Nonpublic Schools by Sampling Strata, Eighth Grade 541
Table B-19 Weighted School Participation Rates and Sample Counts - Grade 4, Reading for Public Schools 545
Table B-20 Weighted School Participation Rates and Sample Counts - Grade 4, Reading for Nonpublic Schools 547
Table B-21 Weighted School Participation Rates and Sample Counts - Grade 8, Reading and Writing for Public Schools 548
Table B-22 Weighted School Participation Rates and Sample Counts - Grade 8, Reading and Writing for Nonpublic Schools 550
Table B-23 Weighted Student Participation Rates, Exclusion Rates, and Sample Counts for the Reporting Samples - Grade 4 Reading for Public Schools 551
Table B-24 Weighted Student Participation Rates, Exclusion Rates, and Sample Counts for the Reporting Samples - Grade 4 Reading for Nonpublic Schools 553
Table B-25 Weighted Student Participation Rates, Exclusion Rates, and Sample Counts for the Reporting Samples - Grade 8 Reading for Public Schools 554
Table B-26 Weighted Student Participation Rates, Exclusion Rates, and Sample Counts for the Reporting Samples - Grade 8 Reading for Nonpublic Schools 556
Table B-27 Weighted Student Participation Rates, Exclusion Rates, and Sample Counts for the Reporting Samples - Grade 8 Writing for Public Schools 557
Table B-28 Weighted Student Participation Rates, Exclusion Rates, and Sample Counts for the Reporting Samples - Grade 8 Writing for Nonpublic Schools 559
Table B-29 Results of Logistic Regression Analysis of School Nonresponse - Grade 4 Reading 560
Table B-30 Results of Logistic Regression Analysis of School Nonresponse - Grade 8 Reading 561
Table B-31 Results of Logistic Regression Analysis of School Nonresponse - Grade 8 Writing 562

Appendix C Constructed-Response Item Score Statistics

Table C-1 Score Range, Percent Agreement, and Cohen's Kappa for the Dichotomously Scored Constructed-Response Reading Items Used in 1998 National Main Assessment Scaling, Grade 4 565

Table C-2 Score Range, Percent Agreement, and Cohen's Kappa for the Dichotomously Scored Constructed-Response Reading Items Used in 1998 National Main Assessment Scaling, Grade 8 566

Table C-3 Score Range, Percent Agreement, and Cohen's Kappa for the Dichotomously Scored Constructed-Response Reading Items Used in 1998 National Main Assessment Scaling, Grade12 567

Table C-4 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Reading Items Used in 1998 National Main Assessment Scaling, Grade 4 568

Table C-5 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Reading Items Used in 1998 National Main Assessment Scaling, Grade 8 569

Table C-6 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Reading Items Used in 1998 National Main Assessment Scaling, Grade 12 570

Table C-7 Score Range, Percent Agreement, and Cohen's Kappa for the Dichotomously Scored Constructed-Response Reading Items from 1994 That Were Rescored in 1998, Grade 4 572

Table C-8 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Reading Items from 1994 That Were Rescored in 1998, Grade 4 573

Table C-9 Score Range, Percent Agreement, and Cohen's Kappa for the Dichotomously Scored Constructed-Response Reading Items from 1994 That Were Rescored in 1998, Grade 8 574

Table C-10 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Reading Items from 1994 That Were Rescored in 1998, Grade 8 575

Table C-11 Score Range, Percent Agreement, and Cohen's Kappa for the Dichotomously Scored Constructed-Response Reading Items from 1994 That Were Rescored in 1998, Grade 12 576

Table C-12 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Reading Items from 1994 That Were Rescored in 1998, Grade 12 577

Table C-13 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Writing Items Used in 1998 National Main Assessment Scaling, Grade 4 578

Table C-14 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Writing Items Used in 1998 National Main Assessment Scaling, Grade 8 579

Table C-15 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Writing Items Used in 1998 National Main Assessment Scaling, Grade 12 580

Table C-16 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Civics Items Used in 1998 National Main Assessment Scaling, Grade 4 581

Table C-17 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Civics Items Used in 1998 National Main Assessment Scaling, Grade 8 582

Table C-18 Score Range, Percent Agreement, and Intraclass Correlation for the Polytomously Scored Constructed-Response Civics Items Used in 1998 National Main Assessment Scaling, Grade 12 583

Appendix D Differential Item Functioning (DIF) Results

Table D-1 1998 Reading Items Identified as "C" or "CC" Items in at Least One Comparison 585

Table D-2 1998 Civics Items Identified as "C" or "CC" Items in at Least One Comparison 585

Appendix E IRT Parameters

Table E-1 IRT Parameters for the 1998 Reading Items Reading for Literary Experience Scale, Grade 4 589

Table E-2 IRT Parameters for the 1998 Reading Items Reading to Gain Information Scale, Grade 4 591

Table E-3 IRT Parameters for the 1998 Reading Items Reading for Literary Experience Scale, Grade 8 593

Table E-4 IRT Parameters for the 1998 Reading Items Reading to Gain Information Scale, Grade 8 594

Table E-5 IRT Parameters for the 1998 Reading Items Reading to Perform a Task Scale, Grade 8 596

Table E-6 IRT Parameters for the 1998 Reading Items Reading for Literary Experience Scale, Grade 12 597

Table E-7 IRT Parameters for the 1998 Reading Items Reading to Gain Information Scale, Grade 12 598

Table E-8 IRT Parameters for the 1998 Reading Items Reading to Perform a Task Scale, Grade 12 600

Table E-9 IRT Parameters for the 1998 Writing Items, Grade 4 602

Table E-10 IRT Parameters for the 1998 Writing Items, Grade 8 603

Table E-11 IRT Parameters for the 1998 Writing Items, Grade 12 604

Table E-12 IRT Parameters for the 1998 Civics Items, Grade 4 605

Table E-13 IRT Parameters for the 1998 Civics Items, Grade 8 608

Table E-14 IRT Parameters for the 1998 Civics Items, Grade 12 613

Table E-15 IRT Parameters for the 1998 State Reading Items Reading for Literary Experience Scale, Grade 4 618

Table E-16 IRT Parameters for the 1998 State Reading Items Reading to Gain Information Scale, Grade 4 620

Table E-17 IRT Parameters for the 1998 State Reading Items Reading for Literary Experience Scale, Grade 8 622

Table E-18 IRT Parameters for the 1998 State Reading Items Reading to Gain Information Scale, Grade 8 623

Table E-19 IRT Parameters for the 1998 State Reading Items Reading to Perform a Task Scale, Grade 8 625

Table E-20 IRT Parameters for the 1998 State Writing Items, Grade 8 626

Appendix F Conditioning Variables and Contrast Codings

Table F-1 Description of Specifications Provided for Each Conditioning Variable 628
Table F-2 Summary Table of the 1998 Reading Conditioning Variable Specifications..... 629
Table F-3 Summary Table of the 1998 Writing Conditioning Variable Specifications 636
Table F-4 Summary Table of the 1998 Civics Conditioning Variable Specifications..... 643
Table F-5 1998 Reading Conditioning Variable Specifications..... 649
Table F-6 1998 Writing Conditioning Variable Specifications..... 690
Table F-7 1998 Civics Conditioning Variable Specifications..... 731
Table F-8 Proportion of Variance of the Conditioning Variable Contrasts Accounted for by the Principal Components Used in the Conditioning Model for National Reading Conditioning Variables, Grade 4 768
Table F-9 Proportion of Variance of the Conditioning Variable Contrasts Accounted for by the Principal Components Used in the Conditioning Model for National Reading Conditioning Variables, Grade 8 775
Table F-10 Proportion of Variance of the Conditioning Variable Contrasts Accounted for by the Principal Components Used in the Conditioning Model for National Reading Conditioning Variables, Grade 12 782
Table F-11 Proportion of Variance of the Conditioning Variable Contrasts Accounted for by the Principal Components Used in the Conditioning Model for National Writing Conditioning Variables, Grade 4 786
Table F-12 Proportion of Variance of the Conditioning Variable Contrasts Accounted for by the Principal Components Used in the Conditioning Model for National Writing Conditioning Variables, Grade 8 793
Table F-13 Proportion of Variance of the Conditioning Variable Contrasts Accounted for by the Principal Components Used in the Conditioning Model for National Writing Conditioning Variables, Grade 12 801
Table F-14 Proportion of Variance of the Conditioning Variable Contrasts Accounted for by the Principal Components Used in the Conditioning Model for National Civics Conditioning Variables, Grade 4..... 805
Table F-15 Proportion of Variance of the Conditioning Variable Contrasts Accounted for by the Principal Components Used in the Conditioning Model for National Civics Conditioning Variables, Grade 8..... 811
Table F-16 Proportion of Variance of the Conditioning Variable Contrasts Accounted for by the Principal Components Used in the Conditioning Model for National Civics Conditioning Variables, Grade 12..... 817

Appendix G Reporting Subgroups and Special Variables for the 1998 NAEP Assessment

Table G-1 NAEP Geographic Regions 825
Table G-2 Scaling Variables for the 1998 National and State Assessment Samples 830

Appendix H Estimation Error Variance of the Mean by Gender and Race/Ethnicity

Table H-1 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 4 Literacy Scale 831
Table H-2 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 4 Information Scale 831
Table H-3 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 4 Composite Scale 832
Table H-4 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 8 Literacy Scale 832
Table H-5 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 8 Information Scale 833
Table H-6 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 8 Perform a Task Scale 833
Table H-7 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 8 Composite Scale 834
Table H-8 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 12 Literacy Scale 834
Table H-9 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 12 Information Scale 835
Table H-10 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 12 Perform a Task Scale 835
Table H-11 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Reading Grade 12 Composite Scale 836
Table H-12 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Writing Grade 4 836
Table H-13 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Writing Grade 8 837
Table H-14 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Writing Grade 12 837
Table H-15 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Civics Grade 4 838
Table H-16 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Civics Grade 8 838
Table H-17 Estimation Error Variance of the Mean for the 1998 NAEP Assessment National Main Civics Grade 12 839

Appendix I Setting the Achievement Levels for the 1998 NAEP Reading Assessment

Table I-1 Results of First Review for Achievement-Level Exemplars 846
Table I-2 Results of Review of Additional Items for Achievement-Level Exemplars 847
Table I-3 Cut Points for Achievement Levels–Grade 4 850
Figure I-1 Final Descriptions of 1992 Reading Achievement Levels 851
Figure I-2 Draft Descriptions of the Achievement Levels Prepared by the Original Level-Setting Panel 857
Figure I-3 Revised Draft Descriptions of the Achievement Levels Recommended by the Follow-Up Validation Panel 860
Figure I-4 Meeting Participants, NAEP Reading Achievement Level Setting Original Meeting, St. Louis, Missouri, August 21–25, 1992 863

Figure I-5 Meeting Participants, NAEP Reading Achievement Level Setting Follow-Up Validation Meeting, San Diego, California, October 9–11, 1992 865
Figure I-6 Meeting Participants, NAEP Reading Revisit Validation Meeting, St. Louis, Missouri, October 14–16, 1994 866

Appendix J Setting the Achievement Levels for the 1998 NAEP Civics and Writing Assessments

Figure J-1 Sample Reckase Chart Portion 872
Figure J-2 Sample Reckase Chart - Complete 873
Table J-1 Pilot Study Cut Scores (Standard Deviations) on the 1998 Civics NAEP 874
Table J-2 Pilot Study Cut Scores (Standard Deviations) on the 1998 Writing NAEP 874
Table J-3 Civics Achievement-Level Cut Scores and Standard Deviations, by Rounds and Percent Correct Data 877
Table J-4 Writing Achievement-Level Cut Scores and Standard Deviations, by Rounds and Percent Correct Data 878
Table J-5 Mean Cut Scores and Standard Deviations in Writing, by Panelist Type 879
Table J-6 Mean Cut Scores and Standard Deviations in Writing, by Ethnicity 881
Table J-7 Mean Cut Scores and Standard Deviations in Civics, by Ethnicity 882
Table J-8 Mean Cut Scores and Standard Deviations in Civics, by Panelist Type 884
Table J-9 Mean Differences Between Polytomous and Dichotomous Cut Scores for Civics 885
Figure J-3 Achievement-Level Descriptions for Civics 888
Figure J-4 Achievement-Level Descriptions for Writing 893
Figure J-5 Meeting Participants, NAEP Civics Achievement Level Setting Pilot Study, St. Louis, Missouri, August 13–17, 1998 897
Figure J-6 Meeting Participants, NAEP Writing Achievement Level Setting Pilot Study, St. Louis, Missouri, October 1–5, 1998 900
Figure J-7 Meeting Participants, NAEP Civics Achievement Level Setting, St. Louis, Missouri, November 12–16, 1998 903
Figure J-8 Meeting Participants, NAEP Writing Achievement Level Setting Study, St. Louis, Missouri, December 9–13, 1998 908
Figure J-9 Meeting Participants, Civics NAEP Similarities Classification Study Validation Meeting, St. Louis, Missouri, July 9–11, 1999 913

Appendix K Participants in the Objectives and Item–Development Process

Figure K-1 1998 NAEP Reading Item Development Committee 915
Figure K-2 1998 NAEP Reading/Writing Standing Committee 916
Figure K-3 1998 NAEP Writing Item Development Committee 917
Figure K-4 1998 NAEP Civics Item Development Committee 918
Figure K-5 1998 NAEP Civics Standing Committee 920

ACKNOWLEDGMENTS

The design, development, administration, analysis, and reporting of the 1998 National Assessment of Educational Progress (NAEP) program was a collaborative effort among staff from the National Center for Education Statistics (NCES), the National Assessment Governing Board (NAGB), the Council of Chief State School Officers (CCSSO), Educational Testing Service (ETS), Westat, and National Computer Systems (NCS). This report documents NAEP design, administration, and data analysis procedures, indicating what technical decisions were made and the rationale behind those decisions. The development of this report and of the national assessment program is the result of the considerable knowledge, experience, creativity, and dedication of many individuals. I would like to acknowledge these individuals for their contribution to NAEP.

The 1998 NAEP assessment was funded through NCES, in the Office of Educational Research and Improvement of the U.S. Department of Education. The NCES staff played a crucial role in all aspects of the program. We are grateful for the reviews of this report contributed by Marilyn Binkley, Janis Brown, Peggy Carr, Chris Chapman, Michael Cohen, Arnold Goldstein, Steven Gorman, Elvira Hauskens, Janet Johnson, Daniel Kaspryzk, Steve Kaufman, Vonda Kiplinger, Andrew Kolstad, Ralph Lee, Larry Ogle, Andrew Malizio, Marilyn McMillen, Gary Skaggs, Suzanne Triplett, Holly Spurlock, Michael Ross, Bruce Taylor, Sheida White, and Shi-Chang Wu. Additional reviews were provided by Luz Bay and Eugene Johnson.

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

ETS Management has encouraged high quality work on all NAEP activities. Thanks go to several members of ETS Management: Nancy Cole, former President of ETS; Paul Ramsey, Vice-President for the School and College Services Division; Henry Braun, former Vice-President for Research; Charles Davis, former Director of the Psychometrics and Statistics Research Division; John Barone, Executive Director of the Division of Data Analysis, Statistics, and Technology Research; and James Carlson, former Group Leader of the Large-Scale Assessment Research Group.

ETS management and the NAEP program development and reporting areas within ETS's School and College Services Division have been very supportive of NAEP's technical work. Special thanks go to the following staff members in the NAEP program area within ETS's School and College Services Division who provided direct leadership for the NAEP project: Steve Lazer, Executive Director for NAEP; John Mazzeo, Center Director, Large-Scale Assessment Research; Beth Durkin, Lauren Fried, and Kim Whittington. Significant contributions to the project were also received from Nada Ballator, Jay Campbell, Patricia Donahue, Elissa Greenwald, Christine O'Sullivan, Hilary Persky, Shari Santapau, and Andrew Weiss.

The guidance of the NAEP Design and Analysis Committee on technical aspects of NAEP has been outstanding. During the period of analysis of the 1998 data, the members were: Anthony Nitko (chair), Sylvia Johnson (former chair), Albert Beaton, Johnny Blair, Jeremy Finn, Paul Holland, Huynh Huynh, Edward Kifer, David Lohman, Serge Madhere, Ingram Olkin, Tej Pandey, and Hariharan Swaminathan.

The design and data analysis of the 1998 National Assessment was primarily the responsibility of the NAEP research and data analysis staff at ETS with significant contributions from NAEP management, Westat, and NCS staffs. In addition to managing day-to-day data analytic operations, NAEP Large-Scale Assessment Research staff members have made many innovative statistical and psychometric contributions. Major contributions were made by Hua-Hua Chang, John Donoghue, Frank Jenkins, Jiahe Qian, Spencer

Swinton, and Jinming Zhang under the leadership of James Carlson. Jo-lin Liang served as research associate. Eugene Johnson and Robert Mislevy provided valuable statistical and psychometric advice.

The Division of Data Analysis, Statistics, and Technology Research at ETS, under the leadership of John Barone, was responsible for developing the operating systems and carrying out the data analyses. David Freund and Alfred Rogers developed and maintained the large and complex NAEP data management systems, and Katharine Pashley managed database activities. Alfred Rogers developed the production versions of key analysis and scaling systems. Thanks also go to David Freund, Steven Isham, Bruce Kaplan, Debbie Kline, Edward Kulick, and Alfred Rogers for their continuing roles as leaders and developers of innovative solutions to NAEP data analysis challenges. Many other members of this division made important contributions of their time and talent to NAEP data analyses and analysis software and data products, including John Ferris, Gerry Kokolis, Laura Jerry, Venus Leung, Youn-Hee Lim, Ting Lu, Mike Narcowich, Norma Norris, Ingeborg Novatkoski, Tanya Petrovicheva, Steve Szyszkiewicz, Lois Worthington, and Fred Yan. Special recognition goes to Phillip Leung for his web technology expertise.

The NAEP web site is managed at ETS by Madeline Goodman, Jeffrey Jenkins, and Richard Bohlander. Pat O'Reilly manages the creation of Internet editions of NAEP reports, including the *NAEP 1998 Technical Report.* She is ably assisted by Rick Hasney.

The staff at Westat contributed their talents and efforts in all areas of the sample design and data collection. Particular recognition is due to Nancy Caldwell, Rob Dymowski, Lucy Gray, Brice Hart, Tom Krenzke, Keith Rust, Renee Slobasky, Debby Vivari, Mark Waksberg, Leslie Wallace, and Dianne Walsh for directing the sampling and data collection activities. Thanks are also due to Carlos Arieira, John Burke, Fran Cohen, Karen Dennis, Sharon Hirabayashi, Drew Kistler, Maida Montes, Tom Mule, Lana Ryaboy, Rick Valliant, Ngoan Vo, and Ibrahim Yansaneh.

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

Cindy Hammel, Joan Stoeckel, Martha Thompson, and Karen Damiano of ETS are acknowledged for their editorial and administrative assistance during the preparation of this report.

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

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

Nancy L. Allen
Center for Large-Scale Assessment Research
Division of Psychometrics and Statistics Research, ETS

INTRODUCTION[1]

James E. Carlson and Nancy L. Allen
Educational Testing Service

The 1998 National Assessment of Educational Progress (NAEP) monitored the performance of students in United States schools in the subject areas of reading, writing, and civics. The national main sample involved public- and nonpublic-school students who were in grades 4, 8, or 12. State assessments were also conducted at grades 4 and 8 in reading and at grade 8 in writing. Nearly 448,000 students were assessed in the national and state samples. Although a special study was done comparing 1998 civics results with those for 1988, no NAEP long-term trend (LTT) assessments of reading, writing, math, or science national samples were conducted in 1998.

For previous assessments in which there were both national (main and/or long-term trend) and state components, separate technical reports were produced for the national assessment and each state component (subject area). For 1998, this publication contains technical information about both the state and national components. Information common to both national and state components is presented in the first two parts, while later chapters contain detailed information for each subject area and for the national and state components.

The purpose of this technical report is to provide details on the instrument development, sample design, data collection, and data analysis procedures for the 1998 assessment. This document provides information necessary to show adherence to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2000) and to the Educational Testing Service (ETS) *Standards for Quality and Fairness* (Educational Testing Service [ETS], 1987). Detailed substantive results are not presented here but can be found in a series of NAEP reports covering the status of and trends in student performance; several additional reports provide information on how the assessment was designed and implemented. The reader is directed to the following reports for 1998 results:

- *NAEP 1998 Civics Report Card for the Nation* (Lutkus, Weiss, Campbell, Mazzeo, & Lazer, 1999)
- *NAEP 1998 Reading Report Card for the Nation and the States* (Donahue, Voelkl, Campbell, & Mazzeo, 1999)
- *NAEP 1998 Reading Report for {each state}* (Ballator & Jerry, 1999a)
- *NAEP 1998 Writing Report Card for the Nation and the States* (Greenwald, Persky, Campbell, & Mazzeo, 1999)
- *NAEP 1998 Writing Report for {each state}* (Ballator & Jerry, 1999b)

[1] James E. Carlson , Nancy L. Allen, and John R. Donoghue were responsible for psychometric and statistical analyses of NAEP for the 1998 assessment.

The *Report Card* publications highlight results for the nation, states, and selected subgroups. The frameworks for the 1998 assessment content areas are in:

- *Civics Framework for the 1998 National Assessment of Educational Progress* (National Assessment Governing Board [NAGB], 1996a)
- *Reading Framework for the National Assessment of Educational Progress: 1992-1998* (NAGB, 1990)
- *Writing Framework and Specifications for the 1998 National Assessment of Educational Progress* (NAGB, 1996b)

Other technical information is in:

- *Sampling Activities and Field Operations for 1998 NAEP* (Gray, Krenzke, & Wallace, 2000)
- *Report on Data Collection Activities for All States* (Westat, 1998)
- *1998 NAEP Assessment Report of Processing and Professional Scoring Activities* (National Computer Systems, 1998)

The *NAEP 1998 Reading Data Companion* (Rogers, Kokolis, Stoeckel, & Kline, 2000), the *NAEP 1998 Writing Data Companion* (Rogers, Kokolis, Stoeckel, & Kline, 2000), and the *NAEP 1998 Civics Data Companion* (Rogers, Kokolis, Stoeckel, & Kline, 2000) provide information needed to analyze the 1998 NAEP results, and *The NAEP Guide: A Description of the Content and Methods of the 1997 and 1998 Assessments* (Calderone, King, & Horkay, 1997) contains a description of the content and methods used in both the main and state components of the 1998 assessments.

Many of the NAEP reports, including summary data tables, are available on the Internet at *http://nces.ed.gov/nationsreportcard*. For information about ordering printed copies of these reports, go to the Department of Education web page *http://www.ed.gov/pubs/edpubs.html*, call toll free 1–877– 4ED PUBS (877–433–7827), or write to:

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

The *Frameworks* are descriptions and plans for subject-area assessment content. For ordering information on these reports, write to:

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

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

AN OVERVIEW OF NAEP ANALYSIS CHANGES OVER TIME

NAEP strives to maintain its links to the past and still implement innovations in measurement technology. To that end, long-term trend samples use the same methodology and population definitions as in previous assessments. Main assessment samples incorporate innovations associated with new NAEP technology and address current educational issues. Both long-term trend samples and main assessment samples are nationally represented. The main assessment sample data are used primarily for analyses involving the current student population, but also to estimate short-term trends for a small number of recent assessments. Some of the assessment materials administered to the main assessment samples are periodically administered to state as well as national samples. In continuing to use this two-tiered approach, NAEP reaffirms its commitment to continuing to study trends while at the same time implementing the latest in measurement technology and educational advances.

In succeeding assessments, many of the innovations that were implemented for the first time in 1988 were continued and enhanced. For example, a focused balanced incomplete block (focused BIB) booklet design was used in 1988. Since that time, either focused BIB or focused partially balanced incomplete block (focused PBIB) designs have been used. Variants of the focused PBIB were used with the 1998 main national and state assessment samples in reading and writing, and a focused BIB was used in the 1998 main national civics assessment. Both the BIB and PBIB designs provide for booklets of interlocking blocks of items, so that no student receives too many items, but all receive groups of items that are also presented to other students. The booklet design is focused, because each student receives blocks of cognitive questions in the same subject area. The focused BIB or PBIB design allows for improved estimation within a particular subject area, and estimation continues to be optimized for groups rather than individuals.

Since 1984, NAEP has applied the plausible values approach to estimating means for demographic as well as curriculum-related subgroups. Scale score estimates were drawn from a posterior distribution that was based on an optimum weighting of two sets of information: the student's responses to cognitive questions, and his or her demographic and associated educational process variables. This Bayesian procedure was developed by Mislevy (1991). An improvement that was implemented first in 1988 and refined for the 1994 assessment continues to be used. This is a multivariate procedure that uses information from all scales within a given subject area in the estimation of the scale score distribution on any one scale in that subject area.

To shorten the timetable for reporting results, the period for national main assessment data collection was shortened in 1992, 1994, 1996, and 1998 from the five-month period (January through May) used in 1990 and earlier assessments to a three-month period in the winter (January through March, corresponding to the period used for the winter half-sample of the 1990 national main assessment).

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

One important innovation in reporting the assessment data that has been continued since 1990 is the use of simultaneous comparison procedures in carrying out significance tests for the differences across assessment years. Methods such as the Bonferroni procedure allow one to control for the type I error rate for a fixed number of comparisons. Beginning with the 1996 assessment, a procedure providing more powerful statistical tests that control for the false discovery rate (FDR) as applied by Benjamini and Hochberg (1994) was used for comparisons involving a large number of groups (e.g., state comparisons). In 1998 the FDR procedure was used for all comparisons in NAEP. While the Bonferroni procedure controls the probability of making even one false rejection, the FDR procedure used in NAEP controls

the expected proportion of falsely rejected hypotheses. The Bonferroni procedure is more conservative than the Benjamini procedure for large families of comparison.

ORGANIZATION OF THE TECHNICAL REPORT

This report begins with the details of the design of the 1998 main and state assessments, summarized in Chapter 1. Chapters 2 through 8 provide an overview of the objectives and frameworks for items used in the assessment, the sample selection procedures, the administration of the assessment in the field, the processing of the data from the assessment instruments into computer-readable form, the professional scoring of constructed-response items, and the methods used to create a complete NAEP database.

The 1998 NAEP data analysis procedures are described in Chapters 9 through 13. Chapter 9 provides a summary of the analysis steps. Subsequent chapters provide a general discussion of the weighting and variance estimation procedures used in NAEP, an overview of NAEP scaling methodology, and information about the conventions used in significance testing and reporting NAEP results.

Details of the reading assessment data analysis are provided in Chapters 14 through 17. These chapters describe assessment frameworks and instruments, student samples, items, booklets, scoring, DIF analysis, weights, and item analyses of the main and state assessments. Similar details are provided for the writing assessment (Chapters 18 through 21) and the civics assessment (Chapters 22 through 24).

The appendices provide detailed information on a variety of procedural and statistical topics. Appendices I and J explain how achievement levels for the subject areas were set by the National Assessment Governing Board (NAGB). The last appendix (Appendix K) provides lists of committee members who contributed to the development of objectives and items.

Chapter 1

OVERVIEW OF PART I: THE DESIGN AND IMPLEMENTATION OF THE 1998 NAEP[1]

Nancy L. Allen, James E. Carlson, and John R. Donoghue
Educational Testing Service

1.1 INTRODUCTION

The 1998 National Assessment of Education Progress (NAEP) collected information on the knowledge and skills of American students in reading, writing, and civics. The 1998 NAEP assessment included three components: the national main assessments of reading, writing, and civics; the state assessments of reading and writing; and national special assessments of aspects of writing and civics. The main assessments were administered to national samples of students. No long-term trend (LTT) assessment was included in 1998. The basis for the information collected for the national main assessments was a complex sample survey involving nearly 448,000 students, consisting of national samples of public- and nonpublic-school students who were in grades 4, 8, and 12. Additional NAEP data came from the state assessment program, which in 1998 assessed about 300,000 students in reading at grades 4 and 8 and in writing at grade 8. Grade 4 state samples included public-school students from 40 states, the District of Columbia, the Department of Defense Dependent Elementary and Secondary Schools (DoDEA/DDESS[2]), the Department of Defense Dependents Schools (DoDEA/DoDDS[2]), and Virgin Islands, as well as nonpublic-school students from 29 states and Virgin Islands. Grade 8 state samples for reading included public-school students from 37 states, the District of Columbia, DoDEA/DDESS, DoDEA/DoDDS, and Virgin Islands, as well as nonpublic-school students from 23 states and Virgin Islands. Grade 8 state samples for writing included public-school students from 36 states, the District of Columbia, DoDEA/DDESS, DoDEA/DoDDS, and Virgin Islands, as well as nonpublic-school students from 23 states and Virgin Islands. Results for a few of these states and jurisdictions were not reported because reporting guidelines were not met.

This chapter describes the design for the 1998 main and state assessments and gives an overview of the steps involved in its implementation, from the planning stage through the creation of edited data files. The major components of the implementation are presented here with references to other chapters in Part I that provide greater detail on each aspect of the assessment. The procedures used for the analysis of the data are summarized in the overview to Part II. The remaining chapters, in Parts III, IV, and V, detail the data analysis by each subject area. Excluded are details of the analyses of special studies of 50-minute writing, classroom-based writing, 1988-to-1998 trends in civics, and high school transcripts. The results from and analyses used in these special studies will be described in separate documents.

---

[1] Nancy L. Allen, James E. Carlson, and John R. Donoghue were responsible for the psychometric and statistical analysis of the 1998 national and state NAEP data. The authors are indebted to the authors of Chapters 2 through 8 for portions of this chapter.

[2] DoDEA is the Department of Defense Education Activity. Within the DoDEA, two jurisdictions are reported for NAEP: one for domestic schools (Department of Defense Domestic Dependent Elementary and Secondary Schools [DDESS]) and one for overseas schools (Department of Defense Dependents Schools [DoDDS]).

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

- Section 1.2 provides an overview of the NAEP design for 1998 and includes a description of the constituent samples. To provide background information, the section also includes the assessment schedule from the inception of NAEP in 1969 through the 1998 assessment.

- Section 1.3 provides a summary of the development of the objectives for each subject area in the assessment and a description of the development and review of the items written to fit those objectives. Details and results of the objective and item development processes appear in Chapters 2, 14, 18, and 22.

- Section 1.4 provides a summary of the sampling design used for the 1998 national and state assessments, with a fuller description provided in Chapters 3 (national) and 4 (state).

- Section 1.5 includes a discussion of the assignment of the cognitive and background questions to assessment booklets and a description of the complex block designs that were the basis for assigning cognitive items to assessment booklets and assessment booklets to individuals. Chapters 14, 18, and 22 provide detailed descriptions of the assessment booklets for the subject areas of reading, writing, and civics, respectively.

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

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

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

- Section 1.9 provides a summary of the creation of the database, the quality control of data entry, and lists the 1998 database products. This section also includes a description of the use of the Internet for dissemination of NAEP information. Further details appear in Chapter 8.

1.2 THE 1998 NAEP DESIGN

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

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

In order to meet the goals of measuring trends reliably and responding to changes in the current thinking about subject areas, NAEP has instituted a multicomponent assessment system where each component is itself a set of assessments designed to accomplish a specific goal. There are four components in the NAEP design: national main assessments, state assessments, national long-term-trend assessment in reading, writing, math and science, and special assessments. The national main and state assessments respond to changes in curriculum on a regular basis, as compared to the long-term trend assessments, which were administered in 1996 and will be administered again in 1999. The instruments that measure long-term trends are never changed and measure longer-term trends in a content domain that is constant over the years.

Several improvements were made in the design of NAEP in the 1984 and succeeding assessments. Until the 1984 assessment, NAEP was administered using matrix sampling and tape recorders; that is, by administering booklets of exercises using an aurally presented stimulus that paced groups of students through the individual assessment exercises in a common booklet. In the 1984 assessment, balanced incomplete block (BIB) spiraling, which does not include aural pacing, was introduced in place of taped matrix sampling. BIB spiraling is defined in Section 1.5 of this chapter. The NAEP design now includes sampling grade populations for national main and state assessments, as well as the age populations that NAEP originally assessed for long-term trend assessments. The definitions of student age and the time of year in which the assessment takes place have been made uniform so that students in the fourth, eighth, and twelfth grades are assessed. To shorten the timetable for reporting results, the period for national main data collection was decreased in assessments since 1990 from the five-month period used in 1990 to a three-month period in the winter (corresponding to the period used for the winter half-sample of the 1990 national assessment). To enhance the coverage of the subject areas assessed, the number of items measuring knowledge and skills was increased for NAEP assessments since 1990.

A special feature of the 1998 national main and state assessments of reading was the collection of data from students who were offered accommodations and from students who were not, while using the new rules (introduced in 1996) for inclusion of students with disabilities (SD) and limited English proficient (LEP) students in NAEP assessments. Figure 1-1 contains the layout of the pieces of the sample collected for each grade of the national main and state assessments of reading. In one sample (sample type 2 in Figure 1-1), accommodations were not offered to students. In the other sample (sample type 3 in Figure 1-1), students were offered accommodations. Both sample type 2 and sample type 3 schools selected for participation in the 1998 assessments used the new inclusion rules to determine whether students should be included in the assessment.

For all subject areas, the inclusion rules were applied and accommodations were offered only when a student had been categorized in his or her individualized education program (IEP) as a student with disabilities (SD) or as a limited English proficient (LEP) student; all other students were asked to participate in the assessment. The accommodations provided by NAEP in the national main and state assessments were meant to match those specified in the student's IEP or those ordinarily provided in the classroom for testing situations. The most common accommodation was extended time.

For the 1998 reading national main and state assessments, the sample of students selected for most analysis and reporting purposes consisted of students from two groups: those who were not categorized as SD or LEP students (A_2 and A_3 in Figure 1-1); and those who were categorized as SD or LEP students and who attended schools providing no accommodations (B_2 in Figure 1-1). Test results for students who were offered accommodations (B_3 in Figure 1-1) were not included in the analysis or

reporting of the national main and state assessment results for reading, although the results for students offered accommodations were studied in follow-up analyses. The advantage of the selected reporting sample is that it preserves trend with previous assessments and it makes use of most of the data from the assessment. For the writing and civics assessments, NAEP used the new inclusion rules and provided accommodations to identified students (sample type 3 in Figure 1-1). The information in Chapters 3, 4, and 5 applies to schools and students in all of the sample types, while the data analysis chapters reflect schools and students in reporting samples only.

Figure 1-1
Subsamples of the 1998 NAEP Reading Assessment

	GROUPS OF SCHOOLS	
GROUPS OF STUDENTS	**Sample Type 2** **- NO ACCOMMODATIONS -**	**Sample Type 3** **- ACCOMMODATIONS -**
NOT SD/LEP[1]	**A_2**	**A_3[2]**
INCLUDED SD/LEP[1]	**B_2[2]**	**B_3[2]**
EXCLUDED SD/LEP[1]	**C_2[3]**	**C_3[3]**

[1] Students with Disabilities/Limited English Proficient
[2] Results for students in subsample B_3 were not reported in *NAEP 1998 Reading: Report Card for the Nation and the States*.
[3] Students in subsamples C_2 and C_3 were not included in the assessment.

NAEP's design for 1998 required collecting 19 different samples in order to conduct the assessments. The various samples collected and reported for the 1998 assessment are summarized in Table 1-1.

Table 1-1

NAEP 1998 Student Samples[*]

Sample	Booklet IDs	Cohort Assessed	Reporting Sample Size[†]
4 [Reading–Main]	R1-R16	Grade 4	7,672
8 [Reading–Main]	R1-R18, R21	Grade 8	11,051
12 [Reading–Main]	R1-R18, R21-R22	Grade 12	12,675
4 [Reading–State]	R1-R16	Grade 4	112,138[‡]
8 [Reading–State]	R1-R18,R21	Grade 8	94,429[‡]
4 [Writing–Main]	W201-W240	Grade 4	19,816
8 [Writing–Main]	W201-W240	Grade 8	20,586
12 [Writing–Main]	W201-W237	Grade 12	19,505
8 [Writing–50-Minute]	W241-W243	Grade 8	6,009
12 [Writing–50-Minute]	W241-W243	Grade 12	5,804
4 [Writing–Classroom Study]	—[§]	Grade 4	2,395[**]
8 [Writing–Classroom Study]	—[§]	Grade 8	2,480[**]
8 [Writing–State]	W201-W240	Grade 8	97,589[‡]
4 [Civics–Main]	C301-C318	Grade 4	5,948
8 [Civics–Main]	C301-C332	Grade 8	8,212
12 [Civics–Main]	C301-C332	Grade 12	7,763
4 [Civics–Special Trend]	CT340[††]	Grade 4	2,088
8 [Civics–Special Trend]	CT340[††]	Grade 8	2,055
12 [Civics–Special Trend]	CT340[††]	Grade 12	2,193
Total without [Writing–Classroom Study][†]			**438,164**

[*] The 1998 assessment was administered January 5–March 27, 1998. Final makeup sessions were held March 30–April 3, 1998.

[†] The reporting samples for reading include students in groups A_2, A_3, and B_2 in Figure 1-1. Reporting and assessed samples for writing and civics include students designated by A_3 and B_3.

[‡] This sample size includes counts of students from distinct samples for each state or jurisdiction participating in the assessment.

[§] No booklets were administered in the [Writing–Classroom Study]; instead, examples of classroom-based writing were collected from students participating in this study.

[**] Because some of the students in this study were included in the [Writing–Main] and [Writing–50-Minute] samples and others were not included in these samples, the students in the [Writing–Classroom Study] who are counted here are not included in the reporting sample size total.

[††] These booklets were also administered as a part of the 1988 assessment of civics.

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

1. *Sample* is the sample identifier. The first part of the sample code is a number (the grade) representing the student cohort included in the sample; the second part, in brackets, denotes the specific sample type. For example, 4 [Reading–Main] is a national main assessment reading sample for grade 4. A full description of the purposes for the various sample types is given in Section 1.2.1.

2. *Booklet IDs* give the identifier numbers for the booklets used for the assessment of the particular sample.

3. The *cohort assessed* denotes the age, grade, or age/grade of the population being sampled. For example, a *grade 4* cohort represents students who are in the fourth grade; an *age 17* cohort consists of students (in any grade) who are 17 years old. Samples for the 1998 national main assessments were selected on the basis of grade only. The traditional NAEP samples used in long-term trend estimation were defined by age only. The definitions of age, and thus the corresponding grade, have changed in ways that are described in Section 1.2.2.

4. The *reporting sample size* is the number of students in the sample who were administered the assessment and whose results were used in the NAEP subject-area reports. SD/LEP students who were excluded from the assessment (C_2 and C_3 in Figure 1-1) are not included in the reporting samples. The reporting samples for the reading assessment include students who were not categorized as SD or LEP students (A_2 and A_3 in Figure 1-1), as well as students who were categorized as SD or LEP students and attended schools where no accommodations were offered (B_2 in Figure 1-1). The reporting sample for the writing and civics assessments include students who were not categorized as SD or LEP students (A_3 in Figure 1-1) and students who were categorized as SD or LEP students and attended schools where accommodations were offered (B_3 in Figure 1-1).

1.2.1 The 1998 NAEP Samples

The NAEP samples in 1998 consisted of three types: the main samples from the national assessment, samples from the state assessment, and the special studies samples from the national assessment. No data from long-term trend (LTT) for reading, writing, math, or science samples were collected in 1998.

***The National Main Assessment Samples*.** The national main NAEP samples are labeled in Table 1-1 as [Reading–Main], and [Writing–Main], and [Civics–Main]. The samples used complex spiraling procedures (defined in Section 1.5), and were intended to form the basis for future assessments. Each sample was assessed in the winter period. In these samples, only grade populations were sampled, although age/grade populations were assessed in previous assessment years for reading. The national main assessment samples, and their purposes, are as follows:

[Reading–Main] are grades 4, 8, and 12 national reading assessment samples used for measuring national reading achievement in 1998. The grade 4 and 8 samples also provided the comparison groups for the 1998 state assessment of reading in grades 4 and 8 [Reading–State]. These samples used print administration.

[Writing–Main] are grades 4, 8, and 12 national writing assessment samples used for measuring national writing achievement in 1998. The grade 8 samples also provided the comparison groups for the 1998 state assessment of writing in grade 8 [Writing–State]. These samples used print administration.

[Civics–Main] are grades 4, 8, and 12 civics national assessment samples used for measuring national civics achievement in 1998. Civics was not part of the state assessment in 1998. These samples used print administration.

The State Assessment Samples. In Table 1-1, [Reading–State] and [Writing–State] refer to samples of public- and nonpublic-school students from each of the states and jurisdictions participating in the NAEP 1998 state assessments of reading (at grades 4 and 8) and writing (at grade 8). The assessment booklets were the same print-administered booklets as those used for the matching national samples [Reading–Main] and [Writing–Main], but the administrative procedures varied from that of the main assessment in that state personnel collected the data.

The Special Studies Samples. Three sets of samples were collected as part of special NAEP studies. The samples used special innovative procedures to allow the study of specific aspects of writing and civics. Each sample was assessed in the winter period. In these samples, only grade populations were sampled. The special studies samples, and their purposes, are as follows:

[Writing–50-Minute] are samples of specially selected students in grades 8 and 12 who were administered 50-minute writing blocks in sessions separate from those in which 25-minute blocks were administered.

[Writing–Classroom Study] are samples of grade 4 and grade 8 students in intact classrooms within schools that participated in the national main writing assessment. Analyses of the data from the classroom-based writing study are described in the special report of results from this study. They are not described in this report.

[Civics–Special Trend] are samples of specially selected students in grades 4, 8, and 12 who were administered a booklet from the 1988 civics assessment.

In addition to these special study samples for which different analyses were conducted, the High School Transcript Study based on the full sample of twelfth grade students required special analyses. Westat conducted this study and is responsible for analysis of the data. Although the results of this study are not described in this technical report, documentation is available through Westat in Rockville, Maryland.

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 1998 subject areas of reading, writing, and civics, several other subject areas have been assessed over the years—mathematics, science, 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.

Table 1-2

National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969–1998

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*	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†						X			X	X	
1976–77	Science			X			X			X		
	Basic Life Skills†									X		
	Health†										X	
	Energy†										X	
	Reading†										X	
	Science†										X	
1977–78	Mathematics			X			X			X		
	Consumer Skills†									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	

* *Age 17 students who had dropped out of school or had graduated prior to assessment.*

† *Small, special-interest assessments conducted on limited samples at specific grades or ages.*

(continued)

Table 1-2 (continued)

National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969–1998

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[*]**	**Adult**
1981–82	Mathematics			X			X			X		
	Citizenship/Social Studies			X			X			X		
	Science[†]			X			X			X		
1983–84	Reading		X	X		X	X			X		
	Writing		X	X		X	X			X		
1985	Adult Literacy[†]											X
1986	Reading	X		X	X		X	X		X		
	Mathematics	X		X	X		X	X		X		
	Science	X		X	X		X	X		X		
	Computer Competence	X		X	X		X	X		X		
	U.S. History[†]							X		X		
	Literature†							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		

[‡] *It should be noted that 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.*

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

[†] *Small, special-interest assessments conducted on limited samples at specific grades or ages.*

(continued)

Table 1-2 (continued)

National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969–1998

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[*]	Adult
1988	Reading		X	X		X	X		X	X		
	Writing		X	X		X	X		X	X		
	Civics		X	X		X	X		X	X		
	U.S. History		X	X		X	X		X	X		
	Document Literacy†					X	X		X	X		
	Geography†								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	Reading		X	X		X	X		X	X		
	Mathematics		X	X		X	X		X	X		
	Science		X	X		X	X		X	X		
	Reading (long-term trend)		X	X		X	X	X		X		
	Writing (long-term trend)		X	X		X	X	X		X		
	Mathematics (long-term trend)			X			X			X		
	Science (long-term trend)			X			X			X		
	Trial State Mathematics					X						
1992	Reading		X	X		X	X		X	X		
	Writing		X	X		X	X		X	X		
	Mathematics		X	X		X	X		X	X		
	Reading (long-term trend)		X	X		X	X	X		X		
	Writing (long-term trend)		X	X		X	X	X		X		
	Mathematics (long-term trend)			X			X			X		
	Science (long-term trend)			X			X			X		
	Trial State Mathematics		X			X						
	Trial State Reading		X									

[‡] *It should be noted that 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.*

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

(continued)

Table 1-2 (continued)

National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969–1998

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[*]	Adult
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†					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						

[‡] *It should be noted that 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.*

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

[†] *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.*

Since its inception, NAEP has assessed 9-year-olds, 13-year-olds, and in-school 17-year-olds, although the age definitions changed in 1986 and again in 1988. Because of budget restrictions, NAEP no longer routinely assesses out-of-school 17-year-olds or young adults. (A separate assessment of young adults of ages 21 to 25 was conducted in 1985 under a separate grant.) Currently, NAEP assesses fourth- and eighth-grade students in the national and state assessments, and twelfth-grade students in the national assessment. Between 1980 and 1996, assessments were administered bi-annually, rather than annually, due to funding restrictions. National (main and/or long-term trend) assessments are now conducted annually, and state assessments continue to be conducted bi-annually.

The table also indicates that in 1984, NAEP began gathering data by grade as well as by age, a practice that had been continued in national main assessments up to 1994; the 1996 and 1998 national main assessments included data gathered by grade only. It should be noted that somewhat different age definitions were used in the 1984, 1986, and 1988 assessments. In the 1984 assessment, the two younger ages were defined on a calendar-year basis, while the 17-year-olds were defined on an October 1 to September 30 basis. This resulted in modal grades of 4, 8, and 11. To allow for age cohorts that were exactly four years apart, in the 1986 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, Beaton, Johnson, & Johnson, 1988) were conducted to measure the effect of the changes in age definition. Because of problems encountered in assessing third-graders, in 1988 the ages were redefined on a calendar-year basis, with the modal grades being 4, 8, and 12. These were the age definitions used in the 1990, 1992, and 1994 national main assessments.

1.3 DEVELOPMENT OF ASSESSMENT OBJECTIVES, ITEMS, AND BACKGROUND QUESTIONS

In 1998, NAEP conducted national assessments of students at all three grade levels in reading, writing, and civics. These assessments entailed the generation of a large number of cognitive items—items measuring knowledge and skills. In addition, a large number of background questions were asked of students. School, teacher, and instructional questions were asked of principals and teachers. Details on the item-development procedures for the 1998 national assessment are given in Chapter 2.

In addition to the cognitive items, several questionnaires were developed: a common student background questionnaire given to all assessed students of a given grade, a subject-specific background questionnaire, a school characteristics and policies questionnaire, and teacher questionnaires for teachers of fourth- and eighth-grade students in reading, writing, and civics. A questionnaire for which teachers or school officials provided information about students with disabilities (SD) or students with limited English proficiency (LEP) was also developed. Each of these questionnaires was developed through a broad-based consensus process.

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

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

1.4 THE 1998 SAMPLE DESIGN

The sample for the 1998 NAEP assessment was selected using a complex multistage sample design. The multistage sample design includes the sampling of students from selected schools within geographic areas (for national NAEP only), called primary sampling units (PSUs), across the United States. Additional stages in the design are the assignment of assessment sessions to schools and the assignment of students to sessions. Apart from the assignment of two types of samples in the reading assessment (one that provided accommodations to certain students and one that did not), the general sampling design for the 1998 assessment was similar in most respects to that of 1996. The design is described in detail by Westat, the firm contracted by NCES to select the sample, in the *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000). The following sections provide an overview of the steps used to draw NAEP samples using the multistage sample design. Further details are given in Chapters 3 and 4. Steps 3 and 4 describe the assignment of sample types and assessment sessions to the second sampling unit schools.

1.4.1 Step 1: Primary Sampling Units

National Assessment

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

Ninety-four of the PSUs were selected for the 1998 national assessment. Twenty-two PSUs were designated as certainty units (required to be in the sample) because of their size, and were included in the sample with certainty. The remaining smaller PSUs were not guaranteed to be selected and were accordingly designated as noncertainty PSUs. Within each major stratum, further stratification was achieved by ordering the noncertainty PSUs according to several additional socioeconomic characteristics, creating a second group of strata. Seventy-two PSUs were selected, one per stratum from each of the noncertainty strata, with probability proportional to size (total population from the 1990 census). To enlarge the samples of Black and Hispanic students, thereby enhancing the reliability of estimates for these groups, PSUs from the high-minority noncertainty strata were sampled at twice the rate of those from the other strata. This was achieved by creating smaller strata within the high-minority noncertainty strata.

State Assessment

For each jurisdiction in the state assessment, schools were the primary sampling units (PSUs).

1.4.2 Step 2: Selection of Schools

National Assessment

In the second stage of sampling for the national assessments, the public schools (including Bureau of Indian Affairs [BIA] schools and Department of Defense Education Activity [DoDEA] schools) and nonpublic schools (including Catholic schools) within each of the selected PSUs were listed according to the grade ranges associated with the three age classes. An independent sample of schools was selected separately for each of the grades so that some schools were selected for assessment of two grades, and a few were selected for all three. Schools within each PSU were selected (without replacement) with probabilities proportional to assigned measures of size with oversampling of nonpublic schools and of schools with high minority enrollment. Overall probabilities of selection for high-minority schools were twice those for other schools, while the probabilities of selection for nonpublic schools were triple those for low-minority public schools of the same size. The increased probabilities of selection enlarged the samples of Black and Hispanic students and the samples of students from nonpublic schools, thereby enhancing the reliability of estimates for these groups. Details of the probabilities used for school selection appear in Chapters 3 and 4. For the national samples, the overall school cooperation rate was 86 percent for grade 4, 83 percent for grade 8, and 79 percent for grade 12. In certain instances, refusing schools were replaced by substitutes according to the rules indicated in Chapters 3 and 4.

State Assessment

For the state samples, the stratification used for sample selection varied by school type (public or nonpublic). Stratification of public schools involved four primary dimensions, whereas the stratification of nonpublic schools involved three primary dimensions. Public schools were stratified hierarchically by small- or large-district status, school size class (measured by student enrollment), urbanization classification, and minority classification. Nonpublic schools were stratified by school size class, metro-area status, and school type (Catholic or other nonpublic). Public schools were further stratified implicitly by median household income (i.e., sorted in ascending or descending order) of the ZIP code area where the school was located, and nonpublic schools were further stratified implicitly by estimated grade enrollment in order to provide some control over these variables. Schools were randomly sampled within these stratification classifications.

1.4.3 Step 3: Assigning Assessment Session and Sample Type to Schools

National and State Assessments

Sessions were assigned to the selected schools found to be appropriate at the time of session assignment, as described in Chapters 3 and 4. Sessions were assigned to schools with three goals in mind. The first was to distribute students to the different session types across the entire sample for each grade so that the target numbers of assessed students would be achieved (in each sample type separately in the national main assessments). The second was to maximize the number of different session types that were administered within a given selected school, without creating unduly small sessions. The third was to give each student an equal chance of being selected for a given session type regardless of the number of sessions conducted in the school.

In order to determine the effect of using different criteria for excluding students from the assessment, three different sample types were assigned to the schools selected for the national main assessment in 1996. In sample type 1 schools, the inclusion criteria for the national main samples were identical to those used in 1990 and 1992. In sample type 2 schools, new 1996 inclusion criteria were

used. In sample type 3 schools, the new 1996 inclusion criteria were used and accommodations were offered to SD/LEP students. In the 1998 national main and state reading assessments, sample types 2 and 3 were assigned to schools. The writing and civics assessments were administered to sample type 3 schools only. More detailed information on assigning sample type to schools is provided in Chapters 3 and 4. Inclusion criteria and accommodations are described in Chapter 5.

1.4.4 Step 4: Sampling Students and Teachers

National and State Assessments

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

Step 4a: Excluded Students. Despite NAEP's goal to assess all selected students, certain selected students were judged by school authorities as being incapable of participating meaningfully in the assessment. For each student who was excluded, school staff who had knowledge of the student's capabilities completed an SD/LEP student questionnaire, listing the reason for exclusion and providing some background information. For each SD/LEP student who was included in the assessment, school staff also completed an SD/LEP student questionnaire.

As stated previously, for the national main NAEP samples, the procedures for assessing students with disabilities (SD) and students of limited English proficiency (LEP) varied by sample type. In sample type 2 schools (for reading), new 1996 inclusion criteria were used. In sample type 3 schools (for reading, writing, and civics), the new 1996 inclusion criteria were used and accommodations were offered to SD/LEP students. The new inclusion criteria were developed to more closely match the procedures used by many states and school districts in testing situations.

Step 4b: Sampling Teachers. Teachers of students assessed were identified and asked by the NAEP supervisor to complete a questionnaire (described in Chapter 2) about their background and instructional practices, by class, for any classes containing assessed students. If the questionnaire was not collected at the time of the assessment, teachers were asked to return the questionnaire in a postage-paid envelope.

Step 4c: The School Characteristics and Policies Questionnaires. Before the assessment, Westat mailed a School Characteristics and Policies Questionnaire to every sampled school for completion by the principal or school administrator. The Westat supervisor then collected the questionnaires and returned them to ETS. The school characteristics and policies questionnaire is described in Chapter 2.

1.5 ASSESSMENT INSTRUMENTS

Four types of instruments were used in the 1998 assessment:

- Student assessment booklets, containing cognitive items and background questions (demographic and subject-specific)
- Teacher questionnaires
- School characteristics and policies questionnaires
- SD/LEP questionnaires

For some assessments, NAEP uses a type of matrix sampling called focused balanced incomplete block (BIB) spiraling to assign blocks or groups of cognitive items to student booklets and to specific students. For other assessments, NAEP uses focused partially balanced incomplete block (PBIB) spiraling for the assignment of items to booklets and students. Because of BIB and PBIB spiraling, NAEP can sample enough students to obtain precise results for each question while generally consuming an average of about an hour and a half of each student's time.

The "focused" part of NAEP's matrix sampling method requires that each student answer questions from only one subject area. The "BIB" or "PBIB" part of the method ensures that students receive different interlocking sections of the assessment forms, enabling NAEP to check for any unusual interactions that may occur between different samples of students and different sets of assessment questions. "Spiraling" refers to the method by which test booklets are assigned to pupils, which ensures that any group of students will be assessed using approximately equal numbers of the different versions of the booklet.

In a BIB design, the cognitive blocks are balanced. Each cognitive block appears an equal number of times in every possible position. Each cognitive block is also paired with every other cognitive block in at least one test booklet. (The NAEP BIB design varies according to subject area.)

Table 1-3 presents a simplified example of a BIB design. The full sample of students is divided into seven equivalent groups, and each group of students is assigned one of the seven test booklets. In this design, each cognitive block appears only once in each of the three possible positions, and each block is paired once with every other block. (This example shows only the cognitive blocks, even though the test booklets also contain background blocks.) The booklets are spiraled in each packet of booklets, so students in each assessment session received each of the seven booklets.

Table 1-3
An Example of a BIB Design

Booklet Version	Position 1 Cognitive Block	Position 2 Cognitive Block	Position 3 Cognitive Block
1	A	B	D
2	B	C	E
3	C	D	F
4	D	E	G
5	E	F	A
6	F	G	B
7	G	A	C

In a PBIB design, one of the characteristics of a BIB design is not present. Table 1-4 presents a simplified example of a PBIB design, similar to the NAEP national and state reading assessment PBIB design. In this case, every block appears in the first and in the second position twice. All blocks containing items from a content area are paired with every other block with items from that content area, but is paired with only one block with items from the other content area. In this example, blocks A, B, C, and D contain items from Content Area 1, and blocks E, F, G, and H contain items from Content Area 2. The first six booklet versions pair Content Area 1 blocks, and the second six booklet versions pair Content Area 2 blocks. In the final four booklet versions, every block is paired with a block of items from the other content area.

For information on the design of specific assessment instruments, see Chapters 2, 14, 18, and 22.

Table 1-4
An Example of a PBIB Design

Booklet Version	Position 1 Cognitive Block	Position 2 Cognitive Block
1	A	C
2	B	A
3	C	D
4	D	B
5	A	D
6	B	C
7	H	E
8	E	F
9	F	G
10	G	H
11	G	E
12	H	F
13	C	G
14	D	H
15	E	B
16	F	A

1.6 FIELD OPERATIONS AND DATA COLLECTION

Field operations and data collection for the 1998 assessment were the responsibility of Westat, and are documented in Chapter 5 and in Westat's *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000). The field operation was conducted by a staff at Westat's home office and a larger staff in the field. The Westat home-office staff coordinated all activities related to field operations and managed materials distribution and home-office receipt of assessment reporting forms. The field staff consisted of area supervisors, assessment supervisors, and exercise administrators. The assessment supervisors, who were trained by Westat, were each responsible for the assessment activities in one or more PSUs. Although ETS made initial contact with participating school districts, each assessment supervisor was primarily responsible for making follow-up contacts with these districts, recruiting and training exercise administrators to work with them in administering the assessment sessions, arranging the assessment sessions, and selecting the sample of students to be assessed within each school. The assessment supervisors and the exercise administrators administered the assessments, filled out the

necessary forms, performed process control, and shipped the assessment booklets and forms to National Computer Systems (NCS), the subcontractor responsible for processing NAEP materials and data.

Gaining school cooperation was the joint responsibility of Westat and ETS. ETS made the preliminary contacts preparatory to obtaining school cooperation by first contacting the Chief State School Officers, informing them that schools within their states had been selected for the assessment, and in a later letter, listing the selected schools and districts. Later mailings were sent to superintendents of public schools and parochial schools and principals of other nonpublic schools for all schools selected in the assessment. These materials provided an explanation of NAEP, a list of the selected schools in the official's jurisdiction, and a cover letter explaining that a Westat district supervisor would contact them to set up an introductory meeting. Westat district supervisors then scheduled and conducted introductory meetings (both by telephone and in person), worked with the schools to schedule the assessments, and, with the exercise administrators, conducted the assessments. The unweighted school response rate for the national main assessments in 1998 was 86 percent overall. The final sample of cooperating schools included 733 schools at grade 4; 761 schools at grade 8; and 608 schools at grade 12. Further detail on school participation rates is given in Chapters 3 (national) and 4 (state). An automated management system tracked and recorded the progress of field work throughout the 1998 assessment period. In addition, progress was constantly monitored through telephone reports held between the area supervisors and the assessment supervisors and between the area supervisors and the home-office staff.

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

1.7 MATERIALS AND DATA PROCESSING

After completing an assessment session, Westat field supervisors and exercise administrators shipped the assessment booklets and forms from the field to NCS for entry into computer files, professional scoring, and creating the data files for transmittal to ETS. Careful checking assured that all data from the field were received. More than 500,000 booklets and questionnaires were received and processed for the national portion of the 1998 assessment. The extensive processing of these data is detailed in Chapter 6.

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

1.8 PROFESSIONAL SCORING

Items requiring a written response from the student (constructed-response items) were included in the national and state assessments in reading and writing and in the national assessment in civics. More than four million constructed responses were read and marked by the professional scoring staff for the national and state portions of the 1998 assessment. Image processing and scoring were again used in 1998. Images of students' responses to the constructed-response items were scanned into computerized form, then scored online by professional raters.

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

1.9 CREATION OF THE DATABASE

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

Chapter 2

DEVELOPING THE NAEP OBJECTIVES, ITEMS, AND BACKGROUND QUESTIONS FOR THE 1998 ASSESSMENTS OF READING, WRITING, AND CIVICS[1]

Terry L. Schoeps
Educational Testing Service

2.1 INTRODUCTION

In 1998, national main NAEP assessments were conducted in reading, writing, and civics. Additional data were gathered under the auspices of the state assessment programs in reading and writing. The state assessment in reading assessed representative samples of public- and nonpublic-school students from 43 jurisdictions at grades 4 and 8; the state assessment in writing assessed representative samples of public- and nonpublic-school students from 39 jurisdictions at grade 8 only.

From its inception, NAEP has developed assessments through a consensus process, and the 1998 instruments were no exception. Under the direction of the National Assessment Governing Board (NAGB), educators, scholars, and citizens representative of many diverse constituencies and points of view designed assessment frameworks for the writing and civics subject areas. The NAEP reading framework used in the 1992 and 1994 assessments served as the framework for the 1998 reading assessment. Copies of the frameworks for these assessments are available on the National Assessment Governing Board (NAGB) web site at *http://www.nagb.org*. Staff at Educational Testing Service (ETS) who are subject-area experts in their respective fields worked with subject-area consultants well versed in assessment methodology to develop assessment questions appropriate to the objectives. All questions underwent extensive reviews by subject-matter specialists and measurement specialists, both within and outside ETS. All questions were also reviewed for bias by staff specially trained in ETS's fairness review process. Questions were assembled and printed into booklets suitable for matrix sampling and then administered either by a trained field staff (for the national program) or by state or local school district staff (for the state assessment program) to stratified, multistage probability samples of students.

All 1998 assessment development efforts were governed by four major criteria:

1. Each assessment was required to match the content definitions included in the assessment frameworks, which had been developed through consensus processes conducted under the auspices of the NAGB.

2. As outlined in the ETS proposal for the administration of the NAEP cooperative agreement (ETS, 1992), the development of items was guided by an instrument development committee for each subject area.[2]

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

[1] Terry L. Schoeps coordinates the production of NAEP technical reports at Educational Testing Service.
[2] A list of the consultants who comprised the 1998 instrument development committees is included in Appendix K.

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

The following sections provide an overview of the process of setting objectives and developing items, as well as specific details about the development of subject-specific objectives and assessments.

2.2 OVERVIEW OF THE 1998 ASSESSMENT OBJECTIVES AND FRAMEWORKS

The subject-area objectives for each NAEP assessment are determined through a legislatively mandated consensus process. Once objectives are established, *frameworks* (matrices) are created, delineating the important content and process areas to be assessed. In addition to these broad frameworks, the Council of Chief State School Officers (CCSSO) and NAGB provide detailed descriptions of item types and the numbers of items to be selected for each category. The frameworks for the 1998 assessments are described below and in Chapters 14 (reading), 18 (writing), and 22 (civics).

The frameworks for the national main 1998 NAEP assessments were developed through consensus processes and were conducted by the CCSSO in reading and civics, and by the Center for Evaluation on Research Standards and Student Testing (CRESST) in writing, working under contract to NAGB. The process involved participation and review by many groups, including teachers, content-area scholars, educational policy makers, and members of the general public. In addition to people directly involved in the framework development processes, the documents were reviewed by state education and testing officials, by representatives of professional associations, and by researchers. In addition, the frameworks were the subject of testimony at public hearings arranged to allow the widest possible participation in the consensus process. The objectives resulting from these processes reflect neither a narrowly defined theoretical framework nor the view of every participant. They do, however, represent the thinking of a broad cross section of individuals who are deeply committed to improving American education.

The framework that governed the 1998 NAEP **reading** assessment was used for the 1992 and 1994 assessments. The NAEP reading assessment was developed in accordance with the *Reading Framework for the National Assessment of Educational Progress, 1992–1998* (NAGB, 1990), making this the third assessment cycle using this framework. The reading assessment was designed around questions requiring in-depth analysis of authentic reading materials. A mixture of multiple-choice, short constructed-response, and extended constructed-response questions made up the assessment. In aggregate, well over half of the student assessment time was spent answering constructed-response rather than multiple-choice questions.

The reading framework is organized according to four reading processes that characterize the ways in which readers gain meaning from text:

- Initial understanding
- Developing an interpretation
- Personal response
- Critical stance

In addition, the assessment was designed to measure the three global reading purposes:

- Reading for literary experience
- Reading to gain information
- Reading to perform a task

The assessment measured students' ability to read based on a variety of passages, including informational materials, documents, news articles, essays, and stories. Each student in the assessment was asked to complete either two 25-minute sets (at all three grades) or one 50-minute set (at grades 8 and 12) of reading passages and comprehension questions. A combination of multiple-choice and constructed-response questions is used to assess students' understanding of the assessment passages.

The 1998 **writing** assessment is structured in accordance with the *Writing Framework and Specifications for the 1998 National Assessment of Educational Progress* (NAGB, 1996b), the assessment measured three kinds of writing:

- Informative
- Narrative
- Persuasive

Because the 1998 writing assessment was based on a new framework, it represents the beginning of a new trend line. Participants responded either to two 25-minute passages or (for some students at grades 8 and 12) to one 50-minute passage. The writing assessment also contained a special study of classroom writing. In that study, 100 teachers at grade 4 and 100 teachers at grade 8 were interviewed about how they teach writing. In addition, for one of their classes, every student was asked to choose and submit the two best pieces of writing he or she had written for that class. Results of this study will published in a separate report. Unlike the reading assessment, the writing and civics assessments are reported along a single within-grade scale.

The framework for the 1998 **civics** assessment, titled *Civics Framework and Specifications for the 1998 National Assessment of Educational Progress* (NAGB, 1996a), is strongly related to the *National Standards for Civics and Government* developed by the Center for Civic Education (1994). Because the 1998 civics assessment was based on a new framework, it represents the beginning of a new trend line. A combination of multiple-choice, short constructed-response, and extended constructed-response questions made up the assessment. In addition to the national civics assessment, a special civics trend study was conducted, in which students were administered instruments from the 1988 NAEP civics assessment.

According to the framework, the civics assessment was designed to measure three interrelated components of civics proficiency: knowledge, intellectual and participatory skills, and civic dispositions. The knowledge component of the framework was divided into five content areas:

- Civic life, politics, and government
- The foundations of the American political system
- The Constitution and American government
- The United States and world affairs
- The roles of United States citizens

The framework also divided intellectual skills into three types, ranging roughly from simpler to higher order thinking skills:

- Identifying and describing
- Explaining and analyzing
- Evaluating, taking, and defending positions

The framework recommended that a special study in civics trend be conducted, in which a subsample of students participating in the national civics assessment would be administered an intact portion of the assessment instruments used in the 1988 civics assessment. Results for the portions administered could then be compared to results of corresponding portions from the 1998 assessment.

2.3 GENERAL OVERVIEW OF PROCEDURES FOR DEVELOPING COGNITIVE ITEMS

A carefully developed and tested series of steps, similar to those used for past NAEP assessments, was utilized to create assessment items that reflected reading, writing, and civics objectives and measured achievement related to them (see Chapters 14, 18, and 22 for information on assessment instruments for reading, writing, and civics, respectively). The item-development steps for each subject area were as follows:

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

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

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

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

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

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

7. Editorial and fairness reviews were conducted as required by the *ETS Standards for Quality and Fairness* (ETS, 1987).

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

9. A field test was conducted in many states, the District of Columbia, and Virgin Islands.

10. Representatives from state education agencies met and reviewed the field test results for all exercises selected for the state assessment.

11. Based on the field test analyses, new items for the 1998 assessment were revised or modified where necessary. The items once again underwent the full range of ETS reviews.

12. The instrument development committees approved the selection of items to include in the 1998 assessment.

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

Development of the reading, writing, and civics assessments are described in more detail in Chapters 14, 18, and 22, respectively.

2.4 DEVELOPING BACKGROUND ITEMS

As part of the assessment, a series of questionnaires was administered to students, teacher, and school administrators. Similar to the development of the cognitive items, the development of the policy issues and questionnaire items was a consensual process that involved staff work, field testing, and review by external advisory groups. A Background Questionnaire Panel drafted a set of policy issues and made recommendations regarding the design of the items. They were particularly interested in capitalizing on the unique properties of NAEP and not duplicating other surveys.

The Panel recommended a focused study that addressed the relationship between student achievement and instructional practices. The issues, items, and field test results were reviewed by the group of external consultants who identified specific items to be included in the final questionnaires. The items underwent internal ETS review procedures to ensure fairness and quality and were then assembled into questionnaires.

Detailed descriptions of the student and teacher questionnaires are given in Chapter 14 (reading), Chapter 18 (writing), and Chapter 22 (civics). In addition to these, two additional questionnaires were developed for use across subject areas.

- The *School Characteristics and Policies Questionnaire* was given to the principal or other administrator of each school that participated in NAEP. This questionnaire included questions about characteristics of the school, school enrollment, absenteeism, drop-out rates, tracking policies, curriculum, testing practices and use, special priorities and schoolwide programs, availability of resources, special services, community services, policies for parental involvement, and schoolwide problems.

- The *SD/LEP Questionnaire* was completed for each student who was selected to participate in the assessment sample and was classified as a student with a disability (SD), or was categorized as a limited English proficient (LEP) student. This questionnaire, which was completed by someone at the school knowledgeable about the student, asked about the student's background and the special programs in which the student participated. This questionnaire was completed for each SD, LEP, or SD/LEP student in the sample, whether or not that student included in the assessment.

Chapter 3

SAMPLE DESIGN FOR THE NATIONAL ASSESSMENT[1]

Keith F. Rust and Tom Krenzke
Westat

Jiahe Qian and Eugene G. Johnson
Educational Testing Service

3.1 INTRODUCTION

This chapter details sampling activities of the 1998 National Assessment of Educational Progress (NAEP). This introduction gives an overview of the sample design and selection activities and provides some highlights of the current design for the national assessments. Section 3.2 presents detailed documentation of the 1998 sampling of primary sampling units (PSUs) and of schools within PSUs. Section 3.3 discusses the allocation of sessions to schools and the assignment of sample types to schools, and Section 3.4 discusses student sampling within schools. Additional details on the sampling design and process can be found in Westat's *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

3.1.1 Brief Overview of the Sample Design and Sampling Activities

The sample for the 1998 national assessment was a multistage probability sample. Counties or groups of counties were the first-stage sampling units, and elementary and secondary schools were the second-stage units. The third stage of sampling involved the assignment of sessions by type and of sample types to sampled schools. The fourth stage involved selection of students within schools and their assignment to session types.

A total of 94 primary sampling units (PSUs) were included in the national sample; a sample of 733 schools actually participated in the assessment at the fourth grade, 761 schools at the eighth grade, and 608 schools at the twelfth grade. Various blocks or packages of exercises were administered in these schools to 36,104 fourth-graders, 48,797 eighth-graders, and 48,588 twelfth-graders, for a total of 133,489 assessed students. Sometimes schools selected for the sample could not participate in the NAEP assessments (e.g., the schools had closed or no longer taught the appropriate grade level). The participation rates of schools and students are discussed in Section 3.2.4. The use of partially balanced incomplete block (PBIB) designs in the assessment booklets, and spiraling in the assembling of booklets for the assessment is described in Chapter 1.

The weighting procedures for the 1998 NAEP included computing a student's base weight (i.e., the reciprocal of the overall probability that the student was invited to a particular type of session) and adjusting this base weight for nonresponse. The weights were further adjusted by a poststratification procedure. Counts of students in various regions and ethnic subclasses were estimated for the 1997–98 school year by age and grade on the basis of information from the Current Population Survey and Census Bureau tabulations of population distributions. The procedures of poststratifying weights are discussed in

[1] Keith F. Rust was responsible for overseeing all sampling activities; Tom Krenzke carried out most of the national sampling activities. Jiahe Qian, in consultation with Eugene G. Johnson, was responsible for the specification and coordination of the national sampling at ETS.

Section 10.2.5. The weights were then adjusted so that the aggregate NAEP estimates would agree with these estimated counts for each subclass. In all NAEP assessments, including 1998, weights were not poststratified to the Common Core of Data (CCD) for the following reasons:

- CCD contains only public schools.
- CCD data is not as current as census data.
- CCD collects data at the school level.
- CCD, at that time, did not collect data by grade and race.
- CCD, like other publicly available lists of schools, contains ineligible schools that were thought to be eligible at the time the CCD was produced.

The CPS estimates and census projections provide independent data sources (i.e., independent from the source of the NAEP sampling frame), which is commonly used for poststratification in national surveys.

Variances for NAEP are computed by the jackknife procedure. Westat computed estimates of summary measures for the samples and their sampling errors in the process of reviewing weights and weight adjustments. The principal estimates and their variances were computed at ETS.

3.1.2 Target Population and Sample Size

The target population for the 1998 assessment consisted of fourth-grade, eighth-grade, and twelfth- grade students enrolled in public and nonpublic elementary and secondary schools. Table 3-1 shows the target number of students to be assessed in each grade. The targets were intended to yield approximately 2,000 completed assessment booklets containing each block of items in the PBIB assessments for each grade. To allow for the derivation of reliable estimates for nonpublic-school students, the selection probabilities for nonpublic schools were larger than those of similarly sized public schools not designated high-minority (see Section 3.2.4.2).

Table 3-1
1998 NAEP National Samples and Target Sample Sizes

	Subject	Target Sample Size	
Total			**132,000**
Grade 4	Civics	6,000	
	Civics Special Trend	2,000	
	Reading	8,000	
	25-Minute Writing	20,000	
		Grade 4 Total	**36,000**
Grade 8	Civics	8,000	
	Civics Special Trend	2,000	
	Reading	11,000	
	25-Minute Writing	20,000	
	50-Minute Writing	6,000	
		Grade 8 Total	**47,000**
Grade 12	Civics	8,000	
	Civics Special Trend	2,000	
	Reading	13,000	
	25-Minute Writing	20,000	
	50-Minute Writing	6,000	
		Grade 12 Total	**49,000**

3.1.3 Highlights of Design Changes for the 1998 Assessment

The general sampling design plan for the 1998 assessment was similar in most respects to that of 1996. Four major changes were made:

- The long-term trend assessments of reading, writing, mathematics, and science were not administered in 1998.
- The samples consisted of three distinct session types (writing/civics, civics special trend, and reading) for each grade, four distinct subjects for grade 4, and five distinct subjects for each of grade 8 and 12 (as shown in Table 3-1). Writing and civics assessments were given in the same session.
- Two sample types (S2, S3) were assigned to subsamples by session in schools. For S2 students, accommodations were not provided for SD/LEP students, while for S3 students, accommodations were provided.
- While SD/LEP students were sampled at a higher rate than non-SD/LEP students, just as in 1996, Black and Hispanic students were also sampled at a higher rate within schools that were in low-minority geographic areas (see Section 3.4.5.1).

To aid the reader, a glossary of terms and abbreviations used in this chapter is provided at the end of the chapter.

3.2 THE SAMPLE OF PRIMARY SAMPLING UNITS AND SCHOOLS

The samples for the 1998 NAEP assessment were selected using a complex multistage sample design involving the sampling of students from selected schools within 94 selected geographic areas, called primary sampling units (PSUs), across the United States. The samples were designed to represent fourth-, eighth-, and twelfth-grade students enrolled in public and nonpublic elementary and secondary schools. The sample design had four steps in the selection process:

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

Steps 1 and 2 are documented in this section. Step 3 is discussed in Section 3.3. Step 4 is discussed in Section 3.4. For area sampling technique, see Kish (1965).

3.2.1 The Definition of Primary Sampling Units

The basic PSU sample design for 1994 NAEP to 2002 NAEP is a stratified probability sample with one PSU selected per stratum (for each round), with probability proportional to population. A PSU consists of a consolidated metropolitan statistical area (CMSA), a metropolitan statistical area (MSA), a New England County metropolitan area (NECMA), a county, or group of contiguous counties in the U.S. (including Alaska, Hawaii, and the District of Columbia). A total of 94 PSUs per round were selected.

The PSU sampling frame for 1994 NAEP to 2002 NAEP was constructed by grouping counties following specific rules as follows:

- Each 1990 CMSA, and each MSA that was not part of a CMSA, was considered a separate PSU. In New England, NECMAs were the metropolitan PSU unit.

- Non-MSA PSUs were made to consist only of non-MSA counties. Whenever possible, each non-MSA PSU contained geographically contiguous counties with a minimum 1990 total population of 60,000 persons in the Northeastern and Southeastern regions, and 45,000 persons in the Central and Western regions. The criteria of minimum population for a non-MSA PSU were determined by survey design to achieve similar numbers of PSUs across the regions.

- Region boundaries were not crossed in the definition of a PSU, not even in the case of MSAs. If a county in an MSA was in a separate region, it was taken out of the MSA and grouped with other contiguous counties in its region to define a PSU.

Checks were made to ensure that every county was included in one and only one PSU. The frame contained 1,027 PSUs: 290 MSAs and 737 non-MSAs.

3.2.2 Definition of PSU Strata

Eight major strata were formed by crossing region and MSA status. The PSUs were classified into four regions, each containing about one-fourth of the U.S. population. These regions were defined primarily by state (Table 3-2).

Table 3-2
Definition of NAEP Stratification and Reporting Regions

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

*Those counties and independent cities in Virginia that are part of the Washington, DC, MD-VA metropolitan statistical area are included in the Northeast region. The remainder of Virginia is included in the Southeast region.

The 22 largest PSUs were included with certainty because of their large sizes. The inclusion of these PSUs in the sample with certainty provided an approximately optimal and cost-efficient sample of schools and students when samples were drawn within them at the required national sampling rate. The 22 largest PSUs by region are presented in Table 3-3.

The remaining smaller PSUs were not guaranteed to be selected for the sample. These were grouped into a number of noncertainty strata (PSUs in these strata were not included in the sample with certainty), and one PSU was selected from each stratum. In each region, noncertainty PSUs were classified as MSA (metropolitan) or non-MSA (nonmetropolitan).

Table 3-3

The 22 Largest Primary Sampling Units, by Region, 1998 NAEP

Region	Primary Sampling Unit
Northeast	
	Baltimore, MD MSA
	Boston-Lawrence-Salem-Lowell-Brockton, MA NECMA
	New York-Northern New Jersey-Long Island, NY-NJ CMSA (excluding that part in CT)
	Philadelphia-Wilmington-Trenton, PA-DE-NJ-MD CMSA
	Pittsburgh-Beaver Valley, PA CMSA
	Washington, DC-MD-VA MSA
Southeast	
	Atlanta, GA MSA
	Miami-Fort Lauderdale, FL CMSA
	Tampa-St. Petersburg-Clearwater, FL MSA
Central	
	Chicago-Gary-Lake County, IL-IN-WI CMSA
	Cleveland-Akron, OH CMSA
	Detroit-Ann Arbor, MI CMSA
	Minneapolis-St. Paul, MN-WI MSA
	St. Louis, MO-IL MSA
West	
	Dallas-Fort Worth, TX CMSA
	Denver-Boulder, CO CMSA
	Houston-Galveston-Brazoria, TX CMSA
	Los Angeles-Anaheim-Riverside, CA CMSA
	Phoenix, AZ MSA
	San Diego, CA MSA
	San Francisco-Oakland-San Jose, CA CMSA
	Seattle-Tacoma, WA CMSA

Within each major stratum, further stratification was achieved by ordering the noncertainty PSUs according to several additional socioeconomic characteristics, yielding 72 strata. The number of such strata formed within each major stratum is shown in Table 3-4. The strata were defined so that the aggregate of the measures of size of the PSUs in a stratum was approximately equal for each stratum. The size measure used was the population from the 1990 Census. The characteristics available for all PSUs, that were used to define strata were the percent minority population, the percentage change in total population since 1980, the per capita income, the percent of persons age 25 or over with college degrees, the percent of persons age 25 or over who have completed high school, and the civilian unemployment rate. Up to four of these characteristics were used in any one major stratum. For each major stratum the characteristics used were chosen by modeling NAEP PSU-level mean reading scale scores for years 17, 19, and 21 (1988, 1990, and 1992). The characteristics chosen were the best predictors of PSU-level mean reading scale scores in these models.

Table 3-4
The Number of Noncertainty Strata in Each Major Stratum 1998 NAEP

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

3.2.3 Selection of Noncertainty PSUs

In the first stage of sampling, a sample of PSUs was drawn for the national NAEP samples for each of the survey years 1994, 1996, 1998, 2000, and 2002. For each survey year, 94 PSUs were selected. Of the 94 selected PSUs, 22 were included with certainty because they had the largest populations in the PSU universe. These 22 certainty PSUs were used in the sample for each of the survey years. The rest of the PSUs in each survey, 72 in total, were selected with a probability proportional to their respective population size. To select noncertainty PSUs, the remaining PSUs on the sampling frame were further stratified into 72 noncertainty strata.

Within each of the noncertainty strata, one PSU was selected with probability proportionate to its 1990 population size for each survey year. That is, within each stratum, a PSU's probability of being selected was proportional to its population size. The PSUs were selected with probability proportional to size (PPS) with the twin aims of obtaining approximately self-weighting samples of students and having approximately equal workloads in each PSU. PSUs were drawn to minimize overlap of the PSUs from one assessment to the next, except that certainty PSUs were retained in each assessment year, and some of the larger noncertainty PSUs are in the sample for more than one of these assessment years. Each sample of 94 PSUs was drawn from a population of about 1,000 PSUs. Primarily because of the use of MSAs as PSUs, PSUs varied considerably as to their probability of selection, since they varied greatly in size. In 1998, the 36 selected MSA PSUs had probabilities of selection ranging from 0.03 to 0.56, while the 36 selected non-MSA PSUs had probabilities ranging from 0.03 to 0.10. Parts of 44 states were included in the sample PSUs. Since one PSU was selected from each noncertainty stratum, the distribution of the noncertainty PSUs is the same as the noncertainty strata, as shown in Table 3-4.

Within each stratum the order of the PSUs was randomized. As detailed later in the section, the selection of PSUs within a stratum was not independent among the survey years. Ordering the PSUs within a stratum by size, geography, or other variables could have resulted in unintended and possibly detrimental correlations between survey estimates across years. Since only one PSU is selected for a given year, the PSU ordering has no effect on sampling variance.

For each PSU within a stratum a normalized measure of size was calculated by dividing the PSU's 1990 population by the aggregate 1990 population of all PSUs in the stratum. Next, a cumulative count of normalized measures of size was calculated for each PSU within a stratum. The cumulative count for the k^{th} PSU in the i^{th} stratum, denoted C_{ik}, was equal to $\sum_{j=1}^{k} \underline{NM}_{ij}$ where $\underline{NM}_{ij}$ represents the normalized measure for the j^{th} PSU in the i^{th} stratum.

For each stratum a random number between 0 and 1 was generated. Using this random number, denoted *r*, the following sequence of sample designation numbers was generated for the five survey years:

Survey Year	1994	1996	1998	2000	2002
Sample Designation Number	*r*	*r* + 0.4	*r* + 0.8	*r* + 0.2	*r* + 0.6

Only the noninteger part of any number in the sequence that exceeded 1.0000 was retained. For example, if *r* was equal to 0.326743, then *r* + 0.8 was equal to 1.126743 and 0.126743 became the sample designation number for 1998.

The first PSU in the stratum whose cumulative count was equal to or greater than *r* was designated the 1994 sample PSU. Similarly, the first PSUs in the stratum whose cumulative counts were equal to or greater than the noninteger portions of *r* + 0.4, *r* + 0.8, *r* + 0.2, and *r* + 0.6 were designated the 1996, 1998, 2000, and 2002 sample PSUs, respectively.

The purpose of having the sample designation numbers for 1996, 1998, 2000, and 2002 be functions of *r* was to attempt to minimize the overlap among the sets of sample PSUs chosen for the various survey years. In strata with smaller numbers of PSUs, some PSUs had large enough normalized measures of size so that they were drawn for two and sometimes even three survey years. As the spacing between the sample designation numbers for any two consecutive survey years was at least 0.4, selecting the same PSU in two consecutive survey years was rare.

3.2.4 School Sample

3.2.4.1 Frame Construction

The second-stage sampling is to select a sample of schools within each selected PSU. A list of schools was formed within each PSU, using a number of sources. The public schools (including Bureau of Indian Affairs [BIA] schools and Department of Defense Education Activity [DoDEA] schools) and nonpublic schools (including Catholic schools) were listed according to the three grades. The lists of schools were obtained from two sources. A list of public, BIA, and DoDEA schools, which is maintained by Quality Education Data, Incorporated (QED) and included information from the 1994–95 NCES Common Core of Data (CCD), was obtained in March of 1997. Regular public schools are schools with students who are classified as being in a specific grade (as opposed to schools having only "ungraded" classrooms). This includes statewide magnet schools and charter schools. Catholic and other nonpublic schools were obtained from the *1995-96 Private School Survey* (PSS) developed for the National Center for Education Statistics. The PSS list of schools is an on-going registry of private schools. The registry is updated prior to the survey through two sources. The first source, called the list frame, is a conglomeration of a number of lists from several associations, states, etc. Although the list frame attempts to have complete coverage of the private school universe, it needs to be supplemented with a second source. The second source uses an area frame to identify and represent schools not on the list frame. The area samples are conducted first by randomly selecting primary sampling units (PSUs), which are single counties or groups of counties from the area frame, which consists of all counties in the nation. Within each selected PSU, a complete list of schools is gathered from a variety of means, and schools not on the list frame are identified and are added to the list frame of nonpublic schools. The majority of the PSS list comes from complete enumeration of schools, a list of schools obtained from different resources. But a small portion of the PSS list was obtained from a sample of counties selected for the PSS. For details of PSS area sampling frame, see the *Private School Universe*

Survey, 1995-1996 (Broughman & Colaciello, 1998). The probabilities of selection for schools to be on the PSS list ranged from 0.06 to 1.00. A weight component was computed, as discussed in Chapter 10, so that these selected PSS nonpublic schools represent themselves, as well as the non-PSS nonpublic schools for non-PSS PSUs.

The ID variable NCESSCH is contained in the CCD file and is echoed by the QED file. This is the unique NCES-assigned school number. The variable NCESSCH is filled in for new schools that were added to the NAEP samples. It can be used to merge NAEP data back with CCD files. The schools that do not match will probably be the additional schools, and nonpublic schools.

Table 3-5 shows the numbers of schools included in the various sampling frame components. The population of eligible schools for each grade was restricted to the selected 94 PSUs. Any school having one or more of the eligible grades, and located within an appropriate PSU, was included in the sampling frame of schools (the list of schools from which the samples of schools were drawn) for a given sample. An independent sample of schools was selected for each of the grades.

Table 3-5
Number of Schools Eligible in QED and PSS Sampling Frame Components by Grade, 1998 Main NAEP

Sample	QED Public*	QED Nonpublic†	PSS Nonpublic
Grade 4	19,962	20	11,428
Grade 8	7,382	11	10,169
Grade 12	4,513	8	4,845

* Public schools, including state-run schools; does not include DoDEA, BIA schools.

† DoDEA, BIA, Catholic, and other nonpublic schools

For each school in each frame, estimates were made of the number of students who were eligible by grade. The QED and PSS files give total enrollment, enrollment by grade, and the grade range for each school, thus providing the average enrollment per grade.

A school would appear in the frame for a particular grade without regard to its eligibility status for either of the two other designated grades. As a result, there is considerable overlap among the three frames.

Before selecting schools, high-minority public schools were identified for oversampling. If the percentage of Hispanic and Black students was not reported or if it was less than 10%, the school was classified as not high-minority; otherwise, the school was classified as high-minority if the percentage of Hispanic and Black students was greater than 10% (15% for grade 12) and if the number of Hispanic and Black students was at least 10 (15 for grade 12).

3.2.4.2 Assigning Size Measures and Selecting School Samples

For each grade-level sample, schools were selected (without replacement) across all PSUs systematically from a sorted list, with probabilities proportional to assigned measures of size. The sorting variables included NAEP region, private/public classification, type of location, high/low minority classification, PSU stratum, and estimated grade enrollment. The order of the sort differed depending on

public and private school classification and certainty/noncertainty PSU classification. To increase cost-efficiency in sampling, samples were designed to include more nonpublic schools and high-minority public schools, and more relatively large schools. Therefore, a measure of size was assigned to each school according to the following scheme.

Let S_i denote the estimated number of grade-eligible students in school *i*. Let $L = 100$ for the assessment of grade 4, $L = 125$ for the assessment of grade 8, and $L = 150$ for the assessment of grade 12. The measure of size was:

$.25\ k_i$,	if S_i was less than 6;
$k_i\ S_i / 20$,	if S_i was greater than 5 but less than 20;
k_i,	if S_i was greater than 19 but less than 101 (grade 4) or 126 (grade 8) or 151 (grade 12); and
$k_i\ S_i / L$,	if S_i was greater than L;

where

k_i = 3, for nonpublic schools (other than BIA and DoDEA schools);

= 2, for high-minority public schools, and;

= 1, for low-minority public schools.

This procedure was used so as to obtain approximately self-weighting samples of students (i.e., students selected with approximately equal overall probabilities) within the oversampling domains at each grade. Three variations to the overall goal of self-weighting samples were implemented. First, schools with fewer than 20 estimated grade-eligible students were assigned somewhat lower measures of size, and thus lower probabilities of selection. This was designed to increase cost efficiency.

Second, each public school designated as high-minority was given double the measure of size of a public school of similar size not designated high-minority. Such high-minority schools were oversampled in order to enlarge the sample of Black and Hispanic students, thereby enhancing the reliability of estimates for these groups. For a given overall sample size, this procedure reduces somewhat the reliability of estimates for all students as a whole and for those not Black or Hispanic. Third, each nonpublic school was given triple the measure of size of a public school of similar size not designated high-minority. These greater probabilities of selection were used to ensure adequate samples of nonpublic-school students in order to allow the derivation of reliable estimates for such students.

The participation rates used to determine the school and student sample sizes are the participation and eligibility rates achieved in 1996. They are shown in Table 3-6. In addition, we inflated the resulting sample sizes by 1.05 to allow for the possibility of decreases in response rate, and for the inaccuracy of the estimated enrollments.

Table 3-6
Participation Rates in 1996 National NAEP

	Grade 4	Grade 8	Grade 12
School Participation Rate	0.86	0.83	0.79
School Eligibility Rate	0.93	0.95	0.96
Student Participation Rate	0.95	0.92	0.80
Overall Participation Rate	0.82	0.76	0.64

3.2.4.3 Updating the School Frame and Sample

The QED files do not contain schools that opened between 1996 and the assessment dates. Therefore, special procedures were implemented to be sure that the NAEP assessment represented students in new public schools. Small school districts, those that contained only one eligible school for a given grade, were handled differently from large school districts, which contained more than one eligible school for a given grade. In small school districts, the schools selected for a given grade were thought to contain all students in the district who were eligible for the assessment. Districts containing these schools were asked if other schools with the appropriate grades for the assessment existed, and if so, they were automatically included in the assessment.

The procedure for obtaining lists of new schools in large districts was coordinated with a similar procedure used for the 1998 state assessment. For large school districts a district-level frame was constructed from the schools on the QED file. Then districts were sampled systematically with probabilities proportional to a measure of size. In most cases, the measure of size was total district enrollment, but in very small districts a minimum measure of size was used. New schools in small districts were identified during school recruitment. Each sampled district was asked to update the list of eligible schools based on information in the QED files. Frames of eligible new schools were then constructed at each grade level, and samples of new schools were selected systematically with probability proportional to eligible enrollment using the same sampling rates as for the QED schools. As a result of this process, 10 new public schools were selected —four at grade 4, three at grade 8, and three at grade 12.

The number of sampled schools by major stratum is presented in Table 3-7. The counts are shown for each grade and include new schools selected in the new schools sampling process. It should be noted that the variables that comprise the major strata (i.e. region, MSA status) were used implicitly as sorting variables in the school sampling process. Additional counts by geographic and school characteristics are shown in Table A-4 (for respondent schools).

Table 3-7
Number of Schools in the Original Samples by Major Stratum

Grade	Region	MSA Certainty PSU	MSA Noncertainty PSU	Non-MSA Noncertainty PSU	Total
4	Northeast	125	54	17	196
	Southeast	27	105	61	193
	Central	78	80	59	217
	West	145	88	50	283
	Total	375	327	187	889
8	Northeast	142	60	18	220
	Southeast	29	110	70	209
	Central	90	84	62	236
	West	148	95	49	292
	Total	409	349	199	957
12	Northeast	122	45	19	186
	Southeast	29	101	79	209
	Central	68	59	55	182
	West	139	84	52	275
	Total	358	289	205	852

3.2.4.4 School Substitution

Potential substitute schools were selected for all sampled schools in the 1998 national NAEP where a close match could be identified by their attributes. An attempt was made to pre-select (before field processes began) a maximum of two substitute schools for each sampled public school (one in-district and one out-of-district) and each sampled Catholic school and one for each sampled non-Catholic private school. A nonparticipating school was replaced by a substitute when the participating school for a particular grade was considered a final refusal. To minimize bias, a substitute school resembled the original selection as much as possible.

Substitutes were assigned by matching approximately on the following attributes:

- Affiliation
- Estimated number of grade-eligible students
- Minority composition

A substitute was always selected from the same PSU as the refusing school. When school non-participation was due to district refusal, none of the schools in the refusing district were considered substitute candidates. However, when substituting for refusals due to a principal's refusal, preference was given to substitute candidates in the same district.

The net numbers of substitutes added to the sample by the above procedure are shown in Table 3-8. The number of substitutes was substantially higher than in recent previous rounds of NAEP because of the efficient preselection method of assigning substitutes. The identity of the substitute schools was unknown to the field staff until after the corresponding original selection was designated as a final refusal. This was to protect against any temptation to move on to an "easier" substitute school.

A retrofitting procedure, which used the same criteria as used for the initial substitution procedure, was implemented midway through the data collection process. This method identified nonresponding schools that needed substitutes and assigned to them unused substitute schools. Unused substitute schools are those schools that were initially linked to cooperating original sampled schools. The same matching rules that were used for assigning substitutes were used in the retrofitting procedure.

3.2.4.5 School Participation Experience

Overall, the 1998 before-substitution school participation rates were lower than school participation rates encountered in previous years. However, the after-substitution participation rates were higher than in previous years. Table 3-8 presents a detailed breakdown by participation status of all schools contacted; 1992, 1994, and 1996 participation rates are also shown based on the same computations.

Table 3-8

Summary of School Participation Experience for 1998 National NAEP, Unweighted

	Grade 4	Grade 8	Grade 12	Total	Public*	Nonpublic†
Total Original Sample	889	957	852	2,698	1,581	1,117
Out-of-Range or Closed	54	79	103	236	29	207
No Eligibles Enrolled	7	7	4	18	0	18
State Tested All Students	1	0	0	1	1	0
District Refused	52	50	50	152	151	1
School Refused	104	118	135	357	162	195
Cooperating	671	703	560	1,934	1,238	696
Cooperation Rate Before Substitution‡	81%	81%	75%	79%	80%	78%
(1996)	86%	83%	79%	83%	85%	80%
(1994)	86%	86%	79%	83%	82%	85%
(1992)	86%	85%	81%	84%	86%	82%
Cooperating Replacement for Refusals	62	58	48	168	109	59
Total Cooperating Schools	**733**	**761**	**608**	**2,102**	**1,347**	**755**
Cooperation Rate After Substitution	89%	87%	82%	86%	87%	85%
Total Students Assessed	**36,104**	**48,797**	**48,588**	**133,489**	**110,825**	**22,664**

* Public schools including state-run schools; does not include DoDEA, BIA schools.

† DoDEA, BIA, Catholic, and other nonpublic schools.

‡ The percentages shown on this row take into account situations in which a school was cooperative but was unable to participate at a given grade, because no eligible students were enrolled in that grade at the time of assessment.

3.3 ASSIGNMENT OF SESSIONS AND SAMPLE TYPES TO SCHOOLS

The process of assigning sessions and sample types to schools differed by grade. For grade 4, sessions and sample types were assigned in the same process, while for grades 8 and 12, sessions were assigned first, then sample types. For simplicity, allocation of sessions will be explained first, followed by an explanation of the assignment of sample types.

3.3.1 Description of Session Types

Three different session types were conducted at all grades: writing/civics, reading, and civics special trend. The writing/civics session type contained two subjects for grade 4 (25-minute writing and civics), and three subjects for grades 8 and 12 (25-minute writing, 50-minute writing, and civics). The special civics trend and reading session types contained only one subject in each session type, respectively.

In the 1998 reading assessment, sample types 2 and 3 were assigned to schools. The writing and civics assessments were administered to sample type 3 schools only. More detailed information on assigning sample type to schools is provided in Section 3.3.3.

3.3.2 Allocation of Sessions

The method of determining the number and type of sessions to be administered in a given selected school varied slightly by grade. Sessions were randomly assigned to the selected schools found

to be appropriate at the time of session assignment. First, the number of sessions per school was established. Four sessions per school were specified for grade 4, and five sessions per school were specified for grades 8 and 12. This was the maximum number of sessions that could be administered without creating unduly small session sizes with few eligible students. Schools with fewer than 25 (30 for grade 12) eligible students were asked to conduct only a single session.

Sessions were assigned to schools with two aims in mind. The first was to distribute students to the different session types across the whole sample for each grade so that the target numbers of assessed students would be achieved in each sample type separately. The second was to maximize the number of different session types that were administered within a given selected school, without violating the minimum session sizes discussed above.

3.3.2.1 Grade 4 Allocation of Sessions

For grade 4, sessions were allocated to schools in the following way. First, each school was allocated a number of sessions, based on the estimated number of grade-eligible students, as shown here:

Estimated Number of Grade-Eligible Students	Number of Sessions Allocated
1 – 25	1
26 – 50	2
51 – 75	3
76 or More	4

The sessions were allocated to schools by placing schools in the order used for sampling, and allocating the appropriate number of sessions from the following repeated sequence (W denotes writing/civics, R denotes reading, and C denotes civics special trend): R, W, W, W, R, W, W, W, R, W, W, W, R, W, W, C, W, W. This sequence contains 13 W, 4 R, and 1 C. This sequence was designed to ensure the maximum feasible spread of assessment types among schools, while ensuring that close to 72 percent of the selected students were assigned to writing/civics, 22 percent of the selected students were assigned to reading, and 6 percent were assigned to civics special trend.

Schools with 26 or more eligible students were always assigned writing/civics. Schools with 76 or more eligible students were almost always assigned reading. Many schools were awarded "multiple" sessions of writing/civics. This did not necessarily mean that the school had to conduct physically multiple sessions of writing/civics, but the assignment of session types determined the proportions of selected students within the school that were assigned to each session type.

3.3.2.2 Grade 8 Allocation of Sessions

For grade 8, sessions were allocated to schools in the following way. First, each school was allocated a number of sessions, based on the estimated number of grade-eligible students, as shown here:

Estimated Number of Grade-Eligible Students	Number of Sessions Allocated
1 – 25	1
26 – 50	2
51 – 75	3
76 – 100	4
101 or more	5

The sessions were allocated to schools by placing schools in the order used for sampling, and allocating the appropriate number of sessions from the following repeated sequence (W denotes writing/civics, R denotes reading, and C denotes civics special trend): R, W, W, W, R, W, W, W, R, W, W, W, R, W, W, C, W, W, R, W, W, W, R, W, W, W, R, W, W, W, R, W, W, W, R, W, W, W, R, W, W, C, W, W, R, W, W. This sequence contains 34 W, 11 R, and 2 C. This sequence was designed to ensure the maximum feasible spread of assessment types among schools, while ensuring that close to 72 percent of the selected students were assigned to writing/civics, 23 percent of the selected students were assigned to reading, and 4 percent were assigned to civics special trend.

Schools with 26 or more eligible students were always assigned writing/civics. Schools with 76 or more eligible students were almost always assigned reading. Many schools were awarded "multiple" sessions of the same type. This did not necessarily mean that the school had to conduct physically multiple sessions of a given assessment type, but the assignment of session types determined the proportions of selected students within the school that were assigned to each session type.

3.3.2.3 Grade 12 Allocation of Sessions

In the same manner, sessions were allocated to grade 12 schools. First, each school was allocated a number of sessions, based on the estimated number of grade-eligible students, as shown here:

Estimated Number of Grade-Eligible Students	Number of Sessions Allocated
1 – 30	1
31 – 60	2
61 – 90	3
91 – 120	4
121 or more	5

The sessions were allocated to schools by placing schools in the order used for sampling, and allocating the appropriate number of sessions from the following repeated sequence (W denotes writing/civics, R denotes reading, and C denotes civics special trend): R, W, W, R, W, W, R, W, W, R, W, W, C, W, W, R, W, W, R, W, W, W, R, W, W, W, R, W, W, R, W, W, R, W, W, C, W, W, R, W, W, R, W, W, W, R, W, W, W. This sequence contains 34 W, 13 R, and 2 C. This sequence was designed to ensure the maximum feasible spread of assessment types among schools, while ensuring that close to 69 percent of the selected students were assigned to writing/civics, 27 percent of the selected students were assigned to reading, and 4 percent were assigned to civics special trend.

Schools with 31 or more eligible students were always assigned writing/civics. Schools with 91 or more eligible students were almost always assigned reading. Many schools were awarded "multiple" sessions of the same type. This did not necessarily mean that the school had to conduct physically multiple sessions of a given assessment type, but the assignment of session types determined the proportions of selected students within the school that were assigned to each session type.

3.3.3 Assignment of Sample Types

In order to determine the effect of using different criteria for excluding students from the assessment, two different sample types (S2 and S3) were assigned to the subsamples by session in sampled schools. In sample type 2 schools, the 1996 exclusion criteria were used, but no accommodations were offered. In sample type 3 schools, the 1996 exclusion criteria were used and

accommodations were offered to students with disabilities (SD) and students of limited English proficiency (LEP). For writing and civics sessions, there was only sample type, S3. For more details of the exclusion criteria and their implementation, and the accommodations offered students, see Exhibits 4-1 and 4-2 in *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000). The information in this chapter applies to both sample types or subsamples.

Sample type was assigned to schools separately for each grade so that 50 percent of the schools assigned reading were assigned sample type 2 and 50 percent were assigned sample type 3. Then, for schools that were also selected for the state assessment program, sample type was revised as explained in Section 3.3.3.4.

3.3.3.1 Grade 4 Assignment of Sample Types

At grade 4, sample type was assigned when allocating sessions to schools. Section 3.3.2 presented the session allocation sequence. The assignment of sample type to the subsamples by session was incorporated into the sequence as follows: R2, W, W, W, R3, W, W, W, R2, W, W, W, R3, W, W, C, W, W, where R2 means the school was allocated a reading session and assigned sample type 2, and R3 means the schools was allocated a reading session and assigned sample type 3. Thus, the sequence contained two reading sessions for sample type 2 (R2) and two reading sessions for sample type 3 (R3). In this manner, sample type was assigned so that a variety of schools with respect to region, school type, urbanization, and size were in each sample type.

3.3.3.2 Grade 8 Assignment of Sample Types

For grade 8, the schools were placed in the order of sampling, then sample types were assigned to subsamples for reading session by alternating sample types 2 and 3. Sample type was assigned so that a variety of schools with respect to region, school type, urbanization, and size were in each sample type.

3.3.3.3 Grade 12 Assignment of Sample Types

The assignment of sample type to grade 12 schools was done in the same manner as for grade 8.

3.3.3.4 Schools Selected in Both National and State Assessments

For schools selected in both the national samples and state assessment program within the same grade (only grades 4 and 8 applied), sample type was initially assigned as described above, and then reassigned for the national samples to be consistent with the state assessment. That is, schools were ultimately assigned the same sample type as for the state assessment.

3.4 STUDENT SAMPLE

The sample of students within sampled schools was drawn by systematic sampling from school-prepared lists of eligible students. Student listing forms (SLF) were prepared for each participating school in a given grade; all enrolled students of the specified grade were to be entered on the SLFs. For details, see Exhibit 1 of Appendix B in the *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000). Student samples that also included oversampling of Black and Hispanic students in low-minority areas, and oversampling SD/LEP students in public schools assigned to reading, were specified through the use of session assignment forms (SAF).

3.4.1 Updating Estimates of Grade-Eligible Students

All assessment components were administered to grade-eligible students. Target numbers of completed assessment booklets by booklet number played an important role in the sample design. Preliminary projections of completed test booklets by school were made as a part of the school sample selection procedure based on estimates of eligible students from frame data (see Section 3.2.4.1).

Up-to-date information on grade enrollment was obtained for sampled schools through two field processes. Scheduling assessment dates with schools and being on site at the school at the time of assessment allowed field staff to obtain updated information on the number of grade-eligible students.

3.4.2 Within-School Sampling Rates

Let

M_A = Maximum allowable sample size from an individual school (100, grade 4; 125, grade 8; 150, grade 12); and

G_i = Revised estimate of grade-eligible students for school i.

Then the sampling rate applied to the list of eligible students to select the sample was given by:

$$R = \frac{M_A}{G_i}$$

if $G_i > (M_A + 10)$, for grades 4 and 8; or
$> (M_A + 20)$, for grade 12;

or $R = 1$, otherwise.

Students were assigned to the sessions systematically, in proportion to the number of sessions of each type allocated to the school, as described in Section 3.3.2. Thus, for example, a grade 8 sample school with an estimated 125 grade-eligible students, assigned sessions W, W, R, W, W, would have four-fifths of the selected students allocated to writing/civics and one-fifth of the selected students allocated to reading.

3.4.3 The Session Assignment Form (SAF)

To control the student sampling operations as closely as possible, Westat generated a session assignment form for each school where sampling was to be carried out. This computer-generated form specified:

- The types of sessions that were to be administered at the school
- The line numbers (from the SLF) specifying the students to be drawn into the sample

- The minimum and maximum number of students listed on the SLF that could be accepted without requiring revision to the within-school sampling rates
- Notification of whether there were to be accommodations offered to SD/LEP students
- Directions and line numbers for oversampling Black and Hispanic students in public schools with low minority enrollment and SD/LEP students in schools assigned reading, and
- Special instructions as appropriate for the teacher survey (see Section 3.4.9), the SD/LEP questionnaire, the NAEP Classroom-Based Writing Study, and the High School Transcript Study (separate, but related to NAEP).

3.4.4 Updating Session Allocation When Generating SAFs

Due to the presence of updated grade enrollment numbers, it became necessary to revise the session allocation structure for some smaller-than-expected schools with more than one session type initially assigned. Smaller-than-expected schools were defined as having a potential of less than 12 students assigned to any particular session type. For example, if two writing/civics and one reading session were assigned, and the number of grade-eligible students was updated to 30, then there would be only 10 assessed in reading. In this case, and in general, for smaller-than-expected schools where the number of grade-eligible students per session type assigned (without regard to the number of sessions assigned for each type) was 12 or more (15 in the example), all session types were kept and students were split evenly across the session types. Thus, in the example given here, 15 students would be assigned to reading and 15 to writing, rather than the initial sample allocation number of 10 and 20, respectively. If the number of grade-eligible students per session type assigned was less than 12, just one session type was kept at random, and a weight adjustment factor was computed as the ratio of the number of sessions assigned to the number of sessions assigned for the session type that was kept. This weight adjustment accounts for dropping one or more session types.

3.4.5 Sample Selection

In the field operations of sample selection, the district supervisor generally carried out the sampling of students a week prior to the assessment. Student listing forms (SLF) were prepared for the applicable grade in each participating school. All enrolled students of the specified grade were to be entered on the SLF in any order convenient to the school, or the school could produce a computer-generated list. Before carrying out the sampling, the district supervisor reviewed the form and made comparisons with other information in an effort to make sure that the list included all eligible students. The sample SLF can be found in *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

The sampling was carried out according to very specific instructions described in the supervisor's manual. The sampling statisticians were available by telephone to assist in the resolution of sampling problems and to generate revised SAFs when necessary.

Briefly, the sample selection procedures involved the following:

- Numbering sequentially the lines listed on the SLF or computer-generated list
- Using the line numbers associated with each session type on the SAF, indicating the sample selection for each session type on the SLF for every student whose line number corresponded to the line numbers given on the SAF

3.4.5.1 Oversampling Black and Hispanic Students

As discussed in Section 3.2, in public schools with high-minority (Black and Hispanic) enrollments, schools were assigned a measure of size twice the size of other low-minority schools, therefore increasing their probability of selection, and indirectly increasing the number of Black and Hispanic students in the sample.

In public schools with low minority enrollment, an oversample of Black and Hispanic students was selected. The procedure was as follows. After the initial sample was selected, as discussed in Section 3.4.5, the nonselected Black and Hispanic students were identified and listed. All such extra Black and Hispanic students were sampled to a total that, as expected, was the same number of Black and Hispanic students as were already selected. In practice, if the number of nonselected students was less than the number of selected students, then all nonselected Black and Hispanic students were to be assessed also. Otherwise, Black and Hispanic students were sampled so that their overall within-school probability of selection was twice the rate of other students.

Line numbers were generated to split the additional sample of Black and Hispanic students into sessions as the session allocation rates applied to the initial sampling procedure. Thus, if the school was assigned two sessions of writing/civics and one of civics special trend, two-thirds of these extra Black and Hispanic students were assigned to writing/civics, and one-third to civics special trend.

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

Since the aim was to oversample by a factor of two where possible, but never more than two, the overall rate of oversampling of Black and Hispanic students was instead less than two. That is because in smaller low-minority schools there were no students remaining who had not already been assigned to a session. The weighting procedures ensured that the results were not biased as a result of the relative underrepresentation of Black and Hispanic students from smaller low-minority schools.

3.4.5.2 Oversampling SD/LEP Students in Reading

As noted in Section 3.1.3, in the reading assessments, the procedures for assessing SD and LEP students varied by sample type. SD/LEP students in sample type 3 were offered accommodations not available to other students or to SD/LEP students in sample type 2.

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

The general intent of oversampling within each school that was assigned at least one reading session was to select SD/LEP students at twice the rate at which non-SD/LEP students were sampled (or to include all SD/LEP students if there were not sufficient numbers to permit sampling at twice the rate). There was no oversampling of schools as part of the procedure.

The procedure was as follows. In each school where oversampling of SD/LEP students was to occur, the initial desired sample of students was drawn for each session assigned, from the full list of eligible students. In addition, in public schools in low-minority areas, oversampling of Black and Hispanic students occurred. Among those students not selected for either of the two prior sampling operations for this school, the SD/LEP students were identified. A sample from among these was drawn, using a sampling rate that would achieve the double sampling rate required overall. In most cases in grade 4, this involved selecting all such SD/LEP students in the school. Again, the weighting procedures ensured that the results were not biased as a result of the relative underrepresentation of SD/LEP students from smaller schools.

As with the oversampling of Black and Hispanic students, the sampling of additional SD/LEP students was carried out using designated line numbers.

Table 3-9 shows the results of the oversampling efforts relating to SD/LEP students for each grade and sample type for reading. The weighted results show the proportion of the sample that would have been SD/LEP students had no oversampling been attempted. The focus is on sample types 2 and 3 for reading, since this is where the oversampling of SD/LEP students occurred. The extent to which the unweighted percentage of SD/LEP students exceeds the weighted percentage is a measure of the effectiveness of the oversampling.

Table 3-9
Percentage of Assessed and Absent Students Who Were Specified as SD/LEP
National 1998 Reading Samples

Sample	Grade 4		Grade 8		Grade 12	
Type	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
2	11.0	8.3	12.2	7.2	9.4	4.8
3	13.9	10.8	16.0	9.9	10.5	5.9
Total	12.5	9.5	14.0	8.5	9.9	5.3

As can be seen, the procedure was effective in increasing the sample of SD/LEP students considerably at grades 8 and 12, and was effective to a lesser extent at grade 4. To increase the sample of SD/LEP students further at grade 4 would require the assessment of additional schools. The differences in rates between sample types 2 and 3 show the effects of accommodations being offered. It was expected that if no accommodations were offered, the rates would be equal; however, since accommodations were offered in sample type 3, more SD/LEP students were assessed.

3.4.6 Supporting the Field Staff on Sampling Issues

The completed SLF generally contained a number of students, which was different from the number used in operating the SAF. In order to control the total number of students tested per school, an acceptable range for that number was specified. Whenever the total number of students listed on the SLF was outside the specified range, the supervisor used a laptop computer to generate a new set of line

numbers. Based on revised sampling rates, a revised SAF was produced. The revised sampling rates were sent in from the field supervisors and were entered on the weight file.

In order to gain cooperation in some schools, we occasionally granted principals' special requests. For example, some large schools divided their students into clusters, and to minimize disruption among all students in the sampled grade, samples were administered to students within one randomly selected cluster. Students in the sampled cluster were listed on the SLF and new line numbers were generated using the cluster's enrollment. The revised sampling weights were entered on the weight file to account for sampling one cluster.

Table 3-10 shows the distribution of the number of students per school who were assessed for each assessment.

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

3.4.7 Excluded Students

The 1998 assessment, as did previous assessments, excluded students who were functionally handicapped to the extent that they could not participate in the assessment as it was normally conducted. Specific groups excluded were:

- Some students identified as having student disability (SD) or equivalent,
- Some students with limited English proficiency (LEP).

Any sample students who were classified SD or LEP (or both) were identified. The school completed an SD/LEP student questionnaire for each student with this designation. This was a change from assessments prior to 1996, in which these questionnaires, then called excluded student questionnaires, were completed only for students who were actually excluded. Then school personnel determined whether any of these students should be excluded from the assessment based on the criteria for excluding students.

According to Table 3-10, for the reading reporting population, about half of the SD/LEP students in grade 4 were excluded. However, for grades 8 and 12, less than half of the SD/LEP students were excluded. Rates of excluded SD/LEP students are also shown by sample type. Recall from Section 3.3 that students in sample type 2 (S2) were not offered accommodations, while students in sample type 3 were offered accommodations. The exclusion rates for SD/LEP students in sample type 2 are similar to that of the reporting population. This is because sample type 2 and the reporting populations contain the same group of SD/LEP students (numerator), but their denominator for the rate calculation differs slightly due to differing groups of non-SD/LEP students. For students in sample type 3, the rates of excluded SD/LEP students are lower.

This data collection effort permits national estimates of statistics for SD, LEP, and excluded students. Table 3-11 shows the distribution of excluded students by reason for exclusion for the three grades. The dominant reason for exclusion from NAEP across all grades and subjects was a student disability. The proportion attributable to student disability increased with grade, while the proportion attributable to limited English proficiency, the second reason, decreased with grade. Table 3-12 presents

the weighted student exclusion rates for each grade and subject by school type and sample type. The exclusion rate decrease as grade increases. The rate for writing and civics are lower than that of civics special trend, since accommodations were offered if necessary. Likewise, the reading sample type 3 rate was lower than that of sample type 2 because accommodations were offered. The rates for public schools are much higher than for private schools.

Table 3-10

*Number of Students Per School for Each Subject Type for 1998 National Assessments**

Sample	Subject Type	Number of Assessed Students	Number Of Schools	Distribution of Students Per Assessment Per School				Mean Number of Students Per Item Per School
				Mean	Median	Minimum	Maximum	
Grade 4	25-Minute Writing	19,816	678	29.2	28.5	1	73	2.9
	Civics	5,948	670	8.9	9.0	1	22	3.0
	Reading/S2	4,048	217	18.7	19.0	2	30	4.7
	Reading/S3	4,204	217	19.4	20.0	1	44	4.8
	Civics Special Trend	2,088	111	18.8	19.0	5	31	18.8
Grade 8	25-Minute Writing	20,586	702	29.3	30.0	1	165	2.9
	50-Minute Writing	6,009	694	8.7	9.0	1	48	2.9
	Civics	8,212	697	11.8	12.0	1	66	2.9
	Reading/S2	6,225	248	25.1	22.0	5	62	4.6
	Reading/S3	5,710	235	24.3	23.0	1	73	4.4
	Civics Special Trend	2,055	104	19.8	20.0	6	30	19.8
Grade 12	25-Minute Writing	19,505	569	34.3	35.0	1	111	3.4
	50-Minute Writing	5,804	564	10.3	10.5	1	34	3.4
	Civics	7,763	566	13.7	14.0	1	43	3.4
	Reading/S2	6,600	245	26.9	24.0	1	85	3.9-4.1[†]
	Reading/S3	6,723	241	27.9	25.0	1	64	3.7-4.3[†]
	Civics Special Trend	2,193	102	21.5	21.0	7	79	21.5

[*] The numbers in this table reflect the full samples, including S2 and S3 for reading.

[†] The number varied because reading for grades 8 and 12 was split into 25-minute reading and 50-minute reading. There was a higher proportion of students assigned to 25-minute reading, and also a larger number of booklets. At grade 8, the number of students per item for the 25-minute reading was equal to that of 50-minute reading.

Table 3-11

Weighted Percentages of Students Excluded (SD and LEP) from 1998 National Reading Assessment[*]

Population	Grade	Type	Total % of Students Identified SD or LEP	Total % of Students That Were Excluded	% of Students Identified w/SD	% of Students That Were Excluded and SD	% of Students Identified w/LEP	% of Students That Were Excluded and LEP
Reporting	4	Overall	17.12	9.61	10.05	5.29	7.55	4.71
		Public	*18.41*	*10.55*	*10.63*	*5.78*	*8.31*	*5.19*
		Nonpublic	*4.84*	*0.68*	*4.59*	*0.55*	*0.25*	*0.13*
Reporting	8	Overall	12.39	5.38	9.41	4.63	3.39	1.00
		Public	*13.51*	*5.96*	*10.22*	*5.13*	*3.75*	*1.11*
		Nonpublic	*2.23*	*0.11*	*2.11*	*0.11*	*0.12*	*0.00*
Reporting	12	Overall	7.86	3.08	5.99	2.77	2.14	0.48
		Public	*8.52*	*3.33*	*6.46*	*3.00*	*2.32*	*0.50*
		Nonpublic	*1.61*	*0.69*	*1.47*	*0.62*	*0.36*	*0.29*
S2	4	Overall	17.03	9.56	10.00	5.26	7.50	4.68
		Public	*18.29*	*10.48*	*10.56*	*5.75*	*8.25*	*5.15*
		Nonpublic	*4.85*	*0.68*	*4.61*	*0.55*	*0.25*	*0.13*
S2	8	Overall	12.01	5.21	9.12	4.49	3.29	0.96
		Public	*13.14*	*5.80*	*9.94*	*4.99*	*3.65*	*1.07*
		Nonpublic	*2.11*	*0.10*	*2.00*	*0.10*	*0.11*	*0.00*
S2	12	Overall	7.71	3.02	5.88	2.72	2.10	0.47
		Public	*8.39*	*3.28*	*6.37*	*2.95*	*2.29*	*0.50*
		Nonpublic	*1.53*	*0.66*	*1.40*	*0.59*	*0.34*	*0.27*
S3	4	Overall	16.57	6.48	10.60	4.40	6.46	2.42
		Public	*18.09*	*7.10*	*11.54*	*4.80*	*7.09*	*2.67*
		Nonpublic	*1.82*	*0.49*	*1.45*	*0.49*	*0.38*	*0.00*
S3	8	Overall	13.24	3.70	10.02	2.95	3.67	0.97
		Public	*14.40*	*4.07*	*10.89*	*3.23*	*4.00*	*1.07*
		Nonpublic	*2.34*	*0.29*	*1.83*	*0.29*	*0.51*	*0.00*
S3	12	Overall	7.84	2.10	5.78	1.86	2.19	0.31
		Public	*8.50*	*2.29*	*6.25*	*2.04*	*2.40*	*0.33*
		Nonpublic	*1.32*	*0.13*	*1.18*	*0.00*	*0.13*	*0.13*

[*] The numbers in this table reflect the full samples, including sample type 2 (S2), and sample type 3 (S3) for reading.

Table 3-12

*Weighted and Unweighted Distribution of Students Excluded for 1998 National Assessments, by Reason for Exclusion, Subject, and Grade**

	Grade 4			Grade 8			Grade 12		
Reason by Subject	**Unweighted Count**	**Weighted Count**	**Weighted Percent**	**Unweighted Count**	**Weighted Count**	**Weighted Percent**	**Unweighted Count**	**Weighted Count**	**Weighted Percent**
25-Minute Writing									
SD	717	138,905	64.8	625	116,229	79.2	532	67,450	85.8
LEP	656	66,657	31.1	213	25,797	17.6	95	8,111	10.3
SD and LEP	74	8,044	3.8	33	3,611	2.5	16	1,308	1.7
Other	3	603	0.3	6	1,125	0.8	15	1,779	2.3
Total	1,450	214,210	100.0	877	146,762	100.0	658	78,648	100.0
50-Minute Writing									
SD	—	—	—	186	110,258	78.2	159	72,355	83.3
LEP	—	—	—	71	27,481	19.5	34	11,015	12.7
SD and LEP	—	—	—	8	2,753	2.0	3	1,154	1.3
Other	—	—	—	1	459	0.3	6	2,365	2.7
Total	—	—	—	266	140,951	100.0	202	86,888	100.0
Civics									
SD	195	125,958	63.0	233	108,922	77.7	201	65,236	85.5
LEP	197	67,727	33.9	94	27,955	20.0	36	8,841	11.6
SD and LEP	14	5,900	3.0	14	3,221	2.3	6	1,420	1.9
Other	1	236	0.1	0	0	0.0	4	836	1.1
Total	407	199,822	100.0	341	140,098	100.0	247	76,333	100.0
Reading†									
SD	228	223,674	62.7	490	178,076	85.1	340	85,027	86.2
LEP	299	122,640	34.4	103	23,461	11.2	87	9,742	9.9
SD and LEP	11	6,435	1.8	14	2,916	1.4	12	1,753	1.8
Other	7	3,798	1.1	16	4,694	2.2	3	2,152	2.2
Total	545	356,547	100.0	623	209,148	100.0	448	98,674	100.0
Civics Special Trend									
SD	116	200,458	75.9	71	131,949	81.7	89	109,674	91.1
LEP	54	58,115	22.0	21	28,631	17.7	12	9,479	7.9
SD and LEP	6	5,596	2.1	0	0	0.0	2	1,190	1.0
Other	0	0	0.0	1	998	0.6	0	0	0.0
Total	176	264,169	100.0	93	161,578	100.0	103	120,343	100.0

* Weighted counts and percents may not add up exactly to the totals due to rounding.

† Represents the reporting population

Table 3-13

Student Exclusion Rates for 1998 National Assessments By Grade, School Type, and Sample Type, Weighted

	Grade 4			Grade 8			Grade 12		
Subject/Sample Type	**Public**	**Nonpublic**	**Total**	**Public**	**Nonpublic**	**Total**	**Public**	**Nonpublic**	**Total**
25-Minute Writing	6.5%	0.3%	5.8%	4.2%	0.4%	3.8%	2.7%	0.0%	2.5%
50-Minute Writing*	—	—	—	4.2%	0.1%	3.8%	3.0%	0.0%	2.7%
Civics	6.1%	0.2%	5.5%	4.0%	0.3%	3.7%	2.6%	0.0%	2.4%
Reading/S2	10.5%	0.7%	9.6%	5.8%	0.1%	5.2%	3.3%	0.7%	3.0%
Reading/S3	7.1%	0.5%	6.5%	4.1%	0.3%	3.7%	2.3%	0.1%	2.1%
Civics Special Trend	7.6%	0.0%	6.9%	4.4%	0.0%	4.1%	4.2%	0.4%	3.8%

* 50-minute writing blocks were administered at grades 8 and 12 only.

3.4.8 Student Participation Results

The NAEP sample was designed to yield a target number of each of the various assessment components. Table 3-14 compares the target assessments to the actual assessments for the three grades. The targets were quite closely met in all cases. Achieving sampling goals precisely is dependent on many factors, including the reliability of frame enrollment data, and the actual response and exclusion rates encountered.

Table 3-14

Comparison of Target Assessments to Actual Assessments for 1998 National Samples, by Grade

	Grade 4		Grade 8		Grade 12	
Assessments	**Target**	**Actual**	**Target**	**Actual**	**Target**	**Actual**
Total	36,000	36,104	47,000	48,797	49,000	48,589
25-Minute Writing	20,000	19,816	20,000	20,586	20,000	19,505
50-Minute Writing*	—	—	6,000	6,009	6,000	5,805
Civics	6,000	5,948	8,000	8,212	8,000	7,763
Reading	8,000	8,252	11,000	11,935	13,000	13,323
Civics Trend	2,000	2,088	2,000	2,055	2,000	2,193

* 50-minute writing blocks were administered at grades 8 and 12 only.

Table 3-15 shows the unweighted student participation rates of invited students. The set of invited students consists of the selected students, after removing the excluded students. For a given session, a makeup session was called for when, for various reasons, more than a predetermined tolerable number of invited students were absent from the originally scheduled session to which they were invited. The participation rates given in the table express the number finally assessed as a percentage of those initially invited in the participating schools. Participation rates are shown for public and nonpublic schools separately.

Table 3-15

Unweighted Student Participation Rates for National Assessments, by Grade and School Type

	1998 Public		1998 Nonpublic		1998 Combined		1996
Grade	**Number Invited**	**Participation Rate**	**Number Invited**	**Participation Rate**	**Number Invited**	**Participation Rate**	**Participation Rate**
4	31,400	95.0	6,545	95.8	37,945	95.1	95.4
8	44,171	91.7	8,639	95.9	52,810	92.4	91.5
12	52,148	77.6	8,871	91.4	61,019	79.6	79.9

Overall participation rates are also shown for comparable samples from the 1996 NAEP assessment. The table shows that student participation rates in 1998 are similar to those experienced in 1996. The rates increased slightly at grade 8, and remained fairly steady for the other grades. At all grades, the participation rate of nonpublic-school students exceeds that of public-school students, with the difference, both relative and absolute, increasing with grade.

The combined impact of school nonparticipation and student absenteeism from sessions within participating schools is summarized in Table 3-16. The table shows the percentages of students assessed, from among those who would have been assessed if all initially selected schools had participated and if all invited students had attended either an initial or make-up session. The results show that, consistent with

earlier rounds of NAEP, the overall level of participation decreases substantially with the increase in the grade of the students.

Table 3-16
Overall Unweighted Participation Rates (School and Student Combined) for 1998 National Assessments, by Grade

1998 Sample	Grade 4	Grade 8	Grade 12	Overall
School Participation				
Before Substitution	81.1%	80.7%	75.2%	79.2%
After Substitution	88.6%	87.3%	81.6%	86.0%
Student Participation	95.1%	92.4%	79.6%	88.0%
Overall Student Participation	84.3%	80.7%	65.0%	75.7%
Number of Participating Students	36,104	48,797	48,589	133,490

So far in this section, only unweighted participation rates by grade and school type have been presented. However, analysis is typically performed separately by grade and subject type, and NCES standards regarding acceptable potentials for bias are expressed in terms of weighted participation rates. Therefore, Table 3-17 shows weighted participation rates by grade and subject type. The sample rates are for students in the reporting populations. Note that the school and student participation rates decrease as grade increases for different session types. At the school level, session types were assigned, and the writing/civics session contained two subject types in grade 4 and three subject types in grades 8 and 12, to which students were assigned. Therefore, the school participation rates for 25-minute writing, 50-minute writing (grades 8 and 12) and civics are identical. The school participation rates (before and after substitution) are fairly similar across subject types. The overall participation rates are relatively low for twelfth grade samples.

The procedures for taking into account nonparticipating schools and for taking into account absent students through weighting were designed (so far as feasible) to reduce the biases resulting from school and student nonparticipation. These procedures are discussed in Chapters 10 and 11.

Table 3-17
Weighted Participation Rates by Grade and Subject Type for the 1998 National Reporting Samples

Participation (Sample Type)	25-Minute Writing	50-Minute Writing	Civics	Reading	Civics Special Trend
Grade 4					
School Participation					
Before Substitution	79.7%	—	79.7%	81.0%	81.1%
After Substitution	88.6%	—	88.6%	89.4%	90.0%
Student Participation	94.9%	—	94.8%	96.0%	95.4%
Overall Participation	84.1%	—	84.0%	86.0%	86.1%

(continued)

Table 3-17 (continued)
Weighted Participation Rates by Grade and Subject Type for the 1998 National Reporting Samples

Participation (Sample Type)	25-Minute Writing	50-Minute Writing	Civics	Reading	Civics Special Trend
Grade 8					
School Participation					
Before Substitution	77.1%	77.1%	77.1%	76.7%	77.1%
After Substitution	84.6%	84.6%	84.6%	84.1%	90.7%
Student Participation	92.2%	93.0%	92.3%	92.7%	92.3%
Overall Participation	78.0%	78.7%	78.1%	77.9%	83.7%
Grade 12					
School Participation					
Before Substitution	69.7%	69.7%	69.7%	69.7%	68.3%
After Substitution	78.0%	78.0%	78.0%	78.2%	83.4%
Student Participation	79.7%	80.4%	79.4%	80.1%	82.0%
Overall Participation	62.1%	62.7%	61.9%	62.6%	68.4%

3.4.9 Teacher Survey

For the grade 4 and grade 8 samples, a survey of teachers was conducted to obtain information about the teachers, their classes, and those of their students who participated in the assessment using the relevant booklet. The questionnaire gathered information about the teaching practices of teachers of sampled students in each of the subject areas that were assessed (i.e., reading, writing, and civics) at grades 4 and 8. The teacher survey was not administered to civics special trend assessments or for assessments in grade 12. Teachers were asked to complete the questionnaires in order that teachers' background instructional practices can be linked to student achievement data.

GLOSSARY

AS:	The administration schedule was prepared for each session to be held in the school and served as a student roster to be used by the school coordinator and exercise administrator (EA) to carry out the session.
BIB design:	A design in which all the exercises in the assessment for an age class are divided up into small blocks. Each exercise block is then assigned to a number of assessment packages (booklets) such that each block is paired with every other block in some booklet the same number of times in a balanced incomplete block (BIB) design. Variants of this design are called partially balanced incomplete block (PBIB) designs.
PSS:	Enrollment grade span and other data for individual private schools were aggregated into data for use in sampling PSUs and schools, and in preliminary session allocation. These data were obtained from a computer file of schools from the Private School Survey conducted by NCES.
PSU:	Primary sampling units are metropolitan statistical areas, counties, or groups of contiguous counties in the U.S. that served as the first-stage sampling units (see Section 3.2.1).
QED:	Enrollment grade span and other data for individual public schools was aggregated into data for use in sampling PSUs and schools, and in preliminary session allocation. These data were obtained from a computer file of schools and school districts from Quality Education Data, Inc.
SAF:	The session assignment form was generated for each cooperating school. It identified the subjects to be administered and the line numbers on the SLF that identified the sampled students to be included in each subject.
Session:	A group of students reporting for the administration of an assessment. A distinction was made between the number of invited students and the number completing the assessment.
SLF:	The student listing forms were the forms used by the school (or supervisor) to list eligible students. Students were sampled from these lists.
Spiraling:	A procedure for assigning tests to students whereby the test packages that are included in the spiral administration procedure are systematically interspersed, and are assigned for testing in this arrangement.
Type of Locale:	The type of locale (TOL) code is a Westat code for the location of a school relative to populous areas.

Chapter 4

SAMPLE DESIGN FOR THE STATE ASSESSMENT[1]

Keith F. Rust and Leslie Wallace
Westat

Jiahe Qian
Educational Testing Service

4.1 INTRODUCTION

This chapter describes sampling activities for the 1998 NAEP state reading and writing assessments, in which 333,624 students were assessed (see Table 5-4). The 1998 state assessment program in *reading* included assessments of fourth- and eighth-grade students. The 1998 state assessment program in *writing* was conducted in grade 8 only. *Civics* was not assessed at the state level. The details of the sample design and selection procedure can be found in the *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000). For the eighth grade, the samples selected for both the reading and writing assessments were selected as part of the same process; and in some schools in the eighth-grade sample, both sessions of reading and writing were assigned. A representative sample of public- and nonpublic-school students was drawn in each participating jurisdiction. The samples in each jurisdiction were selected in two stages, with schools selected at the first stage and students selected at the second stage. This design was intended to produce aggregate estimates as well as estimates for various subpopulations of interest for all the participating jurisdictions. The sample for the fourth- and eighth-grade public-school assessments in each jurisdiction consisted of about 3,150 students (before attrition) in each subject from about 100 public schools in each case. The target for nonpublic-school students varied by jurisdiction and was proportional to their representation in the jurisdiction.

The target population for the 1998 state assessment program included students in public and nonpublic schools who were enrolled in the fourth and eighth grade at the time of assessment. The sampling frame included public and nonpublic schools having the relevant grade levels in each jurisdiction. The samples were selected based on a two-stage sample design; selection of schools within participating jurisdictions, and selection of students within schools. The first-stage samples of schools were selected with probability proportional to a measure of size based on the estimated grade-specific enrollment in the schools. Special procedures were used for jurisdictions with many small schools (see Section 4.4.2), and for jurisdictions having small numbers of grade-eligible schools (See Section 4.4.4). Note that the 1998 *national* sample was a four-stage probability sample and the first-stage sampling units were counties or groups of counties.

Stratification variables were added to the sampling frame prior to sample selection. Public schools were stratified by urbanization and minority class and nonpublic schools were stratified by metro area status and type of nonpublic school . The urbanization strata were defined in terms of large or midsize central city, urban fringe of large or midsize city, large town, small town, and rural areas. Within urbanization strata, public schools were further stratified explicitly on the basis of minority enrollment in those jurisdictions with substantial Black or Hispanic student population. Minority enrollment was defined as the total percent of Black and Hispanic students enrolled in a school. Within minority strata, public schools were sorted by median household income of the ZIP code area where the school was

[1] Keith F. Rust was responsible for overseeing all sampling activities; Leslie Wallace carried out most of these activities. Jiahe Qian was responsible for the specification and coordination of the state sampling at ETS.

located. Metro area status was determined by U.S. Bureau of Census definitions as of June 30, 1993. Other stratification variables were obtained from Quality Education Data, Inc. (QED) and the National Center for Education Statistics' Common Core of Data (CCD). For details, see Sections 4.2.2 and 4.3.2. School type was a dichotomous variable (public, and Catholic or other nonpublic). Within school type, nonpublic schools were sorted by estimated grade enrollment.

From the stratified frame of public and nonpublic schools within each jurisdiction, a systematic random sample of grade-eligible schools was drawn with probability proportional to a measure of size based on the estimated grade-specific enrollment of the school. One or more sessions were sampled within each school. The number of sessions selected depended on the school's estimated grade-specific enrollment, though the overwhelming majority of schools at grade 4 were allocated a single session. In selection of schools, two sets of inclusion rules for SD/LEP students (S2 and S3 subsamples) were applied in the state assessment.

For jurisdictions that participated in an earlier trial state assessment, 25 percent of the selected public and nonpublic schools were designated at random to be monitored during the assessment field period so that reliable comparisons could be made between sessions administered with and without monitoring. For jurisdictions that did not participate in an earlier assessment, 50 percent of the selected public and nonpublic schools were designated to be monitored.

Approximately 3,150 public-school students were targeted for selection for a given grade and subject in a given jurisdiction. For nonpublic schools, the target for each grade and subject varied by jurisdiction. On average, 105 public schools and 19 nonpublic schools were selected for fourth grade in each jurisdiction and 99 public schools and 31 nonpublic schools were selected for eighth grade in each jurisdiction. The maximum numbers of public and nonpublic schools sampled in a participating jurisdiction were 121 and 36, respectively, for fourth grade. The minimum numbers of public and nonpublic schools sampled in a participating jurisdiction were 24 and 10, respectively, for fourth grade. The maximum numbers of public and nonpublic schools sampled for eighth grade were 125 and 46, respectively, for eighth grade. The minimum numbers of public and nonpublic schools sampled in a participating jurisdiction were 6 and 14, respectively, for eighth grade. Each selected school provided a list of eligible enrolled students, from which a systematic sample of students was drawn. Where possible, 30 students were selected for each session.

For the information of state school samples, Tables B-1 through B-6 in Appendix B provide the weighted participation rates and the mean values of certain school characteristics for both public and nonpublic schools, both before and after nonresponse for grade 4 reading, grade 8 reading, and grade 8 writing, respectively. Tables B-15 through B-18 provide the distributions of selected schools by sampling strata by grades for both public and nonpublic schools.

For the characteristics of interest for state student samples, Tables B-7 through B-12 in Appendix B provide the weighted student participation rates and a different set of statistics for public schools and all schools, for both full samples and assessed samples of the state assessments. The information of the unweighted and final weighted counts of assessed and excluded students can be found in Tables 11-1 through 11-6 in Chapter 11, both for public and nonpublic schools for each jurisdiction, grade and subject. For weighting procedures for state samples, including those for excluded students, see Chapter 11.

The rest of this chapter documents the procedures used to select schools for the 1998 state assessment. Section 4.2 describes the construction of the sampling frames, including the sources of school data, missing data problems, and definition of appropriate schools. Section 4.3 includes a description of the various steps in stratification of schools within participating jurisdictions. Section 4.4

describes school sample selection procedures (including new and substitute schools). Section 4.4.6 provides information about the subject sessions, sample type, and monitor status. Section 4.5 includes the steps involved in selection of students within participating schools.

4.2 TARGET POPULATIONS AND SAMPLING FRAME FOR THE 1998 STATE ASSESSMENT

4.2.1 Target Population

The target population for the 1998 state assessment included students in public and nonpublic schools who were enrolled in the fourth or eighth grade. Nonpublic schools included Catholic and other religious schools, private schools, DoDEA/DDESS (Department of Defense Education Activity/Department of Defense Domestic Dependent Elementary and Secondary Schools), and Bureau of Indian Affairs (BIA) schools. Special education schools were not included. Both subsamples of sample type S2, where accommodations were not offered to SD/LEP students, and sample type S3, where accommodations were offered, shared this target population.

4.2.2 Sampling Frame

In order to draw the school samples for the 1998 state assessment, it was necessary to obtain a sampling frame, a comprehensive list of public and nonpublic schools, in each jurisdiction. For each school, useful information for stratification purposes, reliable information about grade span and enrollment, and accurate information for identifying the school to the state coordinator (district membership, name, address) were required.

Based on prior experience with the 1992, 1994, and 1996 trial state assessments, and national assessments from 1984 to 1996, the file made available by QED was elected as the primary sampling frame. The QED list covers all U.S. states but not the territories. The CCD school file was used to obtain schools in Guam and Virgin Islands, and was used to check the completeness of the QED file.

The version of the QED file used was released in early 1997, in time for selection of the school sample. However, for some schools, the file was missing racial/ethnic minority enrollment and urbanization data (due to the inability of QED to match these schools with the corresponding CCD file). Since these variables were to be used for stratification, considerable efforts were undertaken to obtain these variables for all schools in jurisdictions. These efforts are described in the next section.

For 1998 state assessment, the files of the Private School Universe Survey (PSS), which was administered by the National Center for Education Statistics, were used as the sampling frame for nonpublic schools. The QED list was not used to form the sampling frame for nonpublic schools as had been done in the past. Following the very intensive work of unduplicating these two lists in 1996 and an evaluation of the 1996 NAEP nonpublic-school sample, it was decided to use PSS as the sole source for the sampling frame of nonpublic schools.

Tables 4-1 and 4-2 show the distribution of fourth- and eighth-grade schools as well as enrollment within schools as reported in the combined frame. Grade-specific enrollment was estimated for each school as the quotient of total school enrollment and the number of grades in the school.

Table 4-1

Distribution of Fourth-Grade Schools and Enrollment in Combined Sampling Frame for 1998 NAEP State Assessments

	Public Schools		Nonpublic Schools	
Jurisdiction	**Total Schools**	**Total Enrollment**	**Total Schools**	**Total Enrollment**
Total	40,139	2,877,001	11,487	246,708
Alabama	764	58,729	261	6,154
Arizona	719	62,633	260	4,689
Arkansas	533	35,859	166	2,733
California	4,989	445,937	2,872	61,625
Colorado	808	51,882	277	4,779
Connecticut	571	42,507	253	5,484
Delaware	52	7,983	86	2,126
District of Columbia	113	6,330	68	1,476
DoDEA/DDESS	39	3,215	N/A	N/A
DoDEA/DoDDS	103	6,777	N/A	N/A
Florida	1,487	173,855	1,073	24,346
Georgia	1,056	108,774	448	9,469
Hawaii	177	15,343	99	2,589
Illinois	2,268	152,948	1,195	27,633
Iowa	752	37,515	224	4,677
Kansas	798	36,548	191	3,747
Kentucky	782	47,576	289	6,717
Louisiana	793	60,398	377	11,794
Maine	385	17,128	106	1,213
Maryland	804	62,012	459	10,818
Massachusetts	1,039	74,564	473	9,836
Michigan	1,919	130,496	909	18,291
Minnesota	844	64,029	469	8,647
Mississippi	458	40,674	166	4,163
Missouri	1,123	68,180	529	11,236
Montana	455	13,485	75	932
Nebraska	883	22,147	194	3,753
Nevada	254	23,038	59	1,167
New Hampshire	266	16,562	93	1,374
New Mexico	387	25,607	176	2,855
New York	2,250	207,021	1,656	42,214
North Carolina	1,140	97,817	429	7,963
Oklahoma	941	50,649	128	2,389
Oregon	751	42,503	247	3,738
Rhode Island	181	12,086	89	1,933
South Carolina	554	50,729	256	4,971
Tennessee	926	71,198	370	6,557
Texas	3,304	291,812	970	21,139
Utah	441	35,513	54	934
Virgin Islands	24	1,831	27	543
Virginia	1,051	86,583	384	7,729
Washington	1,065	74,783	390	7,122
West Virginia	532	23,168	118	1,305
Wisconsin	1,137	66,170	846	14,256
Wyoming	221	7,654	33	319

Table 4-2

Distribution of Eighth-Grade Schools and Enrollment in Combined Sampling Frame for 1998 NAEP State Assessments

	Public Schools		Nonpublic Schools	
Jurisdiction	**Total Schools**	**Total Enrollment**	**Total Schools**	**Total Enrollment**
Total	17,660	2,796,611	5,378	121,361
Alabama	484	56,743	232	5,443
Arizona	364	59,746	235	4,355
Arkansas	352	36,434	126	1,968
California	1,719	393,472	2,417	53,298
Colorado	342	51,100	229	3,929
Connecticut	208	36,775	250	5,754
Delaware	30	8,506	78	1,951
District of Columbia	33	4,421	64	1,438
DoDEA/DDESS	12	1,625	N/A	N/A
DoDEA/DoDDS	65	5,093	N/A	N/A
Florida	499	168,930	911	21,194
Georgia	420	104,295	399	8,357
Hawaii	52	13,183	85	3,127
Illinois	1,370	144,236	1,121	26,481
Kansas	421	36,269	147	2,958
Kentucky	347	50,454	254	5,986
Louisiana	441	59,009	367	13,757
Maine	232	16,617	101	1,168
Maryland	239	60,756	426	10,218
Massachusetts	401	65,981	468	10,452
Minnesota	448	64,025	358	7,073
Mississippi	780	121,964	140	3,848
Missouri	652	67,282	477	10,696
Montana	319	13,277	69	841
Nebraska	580	23,402	160	3,400
Nevada	93	21,028	50	1,061
New Mexico	154	25,227	131	2,393
New York	1,020	192,295	1,496	40,224
North Carolina	521	92,213	368	6,347
Oklahoma	613	49,440	107	2,103
Oregon	338	41,762	228	3,376
Rhode Island	52	11,409	91	2,327
South Carolina	255	51,632	220	4,186
Tennessee	532	67,373	347	6,618
Texas	1,519	284,146	756	16,975
Utah	154	38,971	57	1,022
Virgin Islands	6	2,368	20	411
Virginia	343	84,608	343	7,397
Washington	430	73,529	326	6,115
West Virginia	206	23,826	99	1,143
Wisconsin	520	64,855	751	12,815
Wyoming	94	8,334	28	234

4.3 STRATIFICATION OF SCHOOLS IN THE SAMPLING FRAME

4.3.1 Stratification Variables

The stratification used for sample selection varied by school type (public or nonpublic), because the availability of information and the feasibility of performing sampling are different for public and nonpublic schools. Stratification of public schools involved four primary dimensions, whereas the stratification of nonpublic schools involved three primary dimensions. Public schools were stratified hierarchically by small or large district status, school size classification (measured by student enrollment), urbanization classification, and minority classification. For details of the resources for stratification variables, see Section 4.3.3. Nonpublic schools were stratified by school size classification, metro area status, and school type (Catholic or other nonpublic).

Public schools were further stratified implicitly by median household income (i.e., sorted in ascending or descending order) of the ZIP code area where the school was located, and nonpublic schools were further stratified implicitly by estimated grade enrollment, in order to provide some control over these variables.

Prior to the selection of the school samples, the public schools were sorted by their four stratification variables (small or large district status, school size classification, urbanization classification, and minority classification) in an order such that changes occur on only one variable at a time (also known as a serpentine order). This is accomplished by alternating between ascending and descending sort order on each variable successively through the sort hierarchy. Within this sorted list, the schools were sorted, in serpentine order, by the median household income. This final stage of sorting resulted in implicit stratification of median household income.

The counts of sampled schools by the primary stratification variables can be found in Tables B-15 through B-18 in Appendix B.

4.3.2 Missing Stratification Variables

As stated earlier, the sampling frame for the 1998 state assessment was the combination of the most recent version of the QED file available and the 1995 PSS list of nonpublic schools. The CCD file was used to extract information on urbanization (“type of location”) for public schools where this information was missing on the QED file. Any public schools with remaining missing values in urbanization or minority enrollment had their data imputed.

Schools with missing values in urbanization data were assigned the urbanization of other school records within the same state, county, and city when urbanization did not vary within the given city. Any schools still missing urbanization were assigned the modal value of urbanization within their city. Any remaining missing values were assigned individually based on city, using U.S. Bureau of Census publications.

Schools with missing values in minority enrollment data were assigned the average minority enrollment within their school district. Any schools still missing minority enrollment data were assigned values individually, using ZIP code and U.S. Bureau of Census data. The minority data were extracted only for those schools in jurisdictions in which minority stratification was performed.

Metro area status was assigned to each nonpublic school based on U.S. Bureau of Census definitions as of June 30, 1993, based on Federal Information Processing Standard (FIPS) county code,

and was found for all schools in the sampling frame. The Catholic school flag was assigned to each nonpublic school based on the PSS school type and was found for all schools in the sampling frame.

Median household income was assigned to every school in the sampling frame by merging on ZIP code with a file from Donnelly Marketing Information Services. Any schools still missing median household income were assigned the mean value of median household income for the three-digit ZIP code prefix or county within which they were located.

4.3.3 Resources for Stratification Variables

The procedures used to compile or create the stratification variables for sampling schools are described below. The resulting classifications for urbanization, minority stratification, metro area status, and school type for schools used within each participating jurisdiction can be found in Tables B-15 through B-18 in Appendix B.

4.3.3.1 Urbanization Classification

Urbanization classification was created based on the NCES type of location variable. The type of location variable contains at most seven levels:

1. *Large Central City*: A central city of a metropolitan statistical area (MSA) with a population greater than or equal to 400,000, or a population density greater than or equal to 6,000 persons per square mile;

2. *Midsize Central City*: A central city of an MSA but not designated as a large central city;

3. *Urban Fringe of Large City*: A place within an MSA of a large central city and defined as urban by the U.S. Bureau of Census;

4. *Urban Fringe of Midsize City*: A place within an MSA of a midsize central city and defined as urban by the U.S. Bureau of Census;

5. *Large Town*: A place not within an MSA, but with a population greater than or equal to 25,000 and defined as urban by the U.S. Bureau of Census;

6. *Small Town*: A place not within an MSA, with a population less than 25,000, but greater than 2,499 and defined as urban by U.S. Bureau of Census; and

7. *Rural*: A place with a population of less than 2,500 and defined as rural by the U.S. Bureau of Census.

Urbanization classification was created by collapsing type of location categories as necessary and according to specific rules until each urbanization stratum included a minimum of 10 percent of eligible students in the participating jurisdiction. The specific rules used were to first try collapsing categories 1 and 2, 3 and 4, or 5 and 6. If that did not work, categories 1-4 or 5-7 were collapsed. For an explanation of the rules used, see Westat's *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

4.3.3.2 Minority Classification

Minority classification was created within urbanization strata and was based on a school's percentages of Black and Hispanic students. Three different minority classification schemes were used and are described as follows:

- *Case 1:* Urbanization strata with less than 10 percent Black students and 7 percent Hispanic students were not stratified by minority enrollment (Level 0).

- *Case 2:* Urbanization strata with greater than or equal to 10 percent Black students or 7 percent Hispanic students, but not more than 20 percent of each, were stratified by ordering percent minority enrollment (Black plus Hispanic) within the urbanization classes and dividing the schools into three groups with about equal numbers of students per minority classification (Levels 1, 2, and 3).

- *Case 3:* In urbanization strata with greater than 20 percent of both Black and Hispanic students, minority strata were formed with the objective of providing equal strata with emphasis on the minority group (Black or Hispanic) of higher concentration. The stratification was performed as follows. The higher percentage minority group provided the primary stratification variable; the other group gave the secondary stratification variable. Within urbanization class, the schools were first sorted based on the primary stratification variable; then they were divided into two groups of schools containing approximately equal numbers of students based on estimated grade enrollment. Within each of these two groups, the schools were sorted by the secondary stratification variable and subdivided into two subgroups of schools containing approximately equal numbers of students. As a result, within urbanization strata there were four minority classifications (e.g., low Black/low Hispanic, low Black/high Hispanic, high Black/low Hispanic, and high Black/high Hispanic (Levels 4, 5, 6, and 7).

The minority groups and classifications were formed solely for the purpose of creating efficient stratification design at this stage of sampling. These classifications are not directly used in analysis and reporting of the data, but will act to reduce sampling errors for scale score estimates.

4.3.3.3 Median Household Income

The data on median household income was related to the ZIP code area in which the school is located. The data were derived from the 1990 Census and were obtained from Donnelly Marketing Information Services.

4.3.3.4 Metro Area Status

All schools in the sampling frame were assigned a metro area status based on their Federal Information Processing Standard (FIPS) county code and Office of Management and Budget (OMB) metropolitan area Definitions as of June 30, 1993. This field indicated if a school was located within a metropolitan area or not.

4.3.3.5 School Type for Nonpublic Schools

All nonpublic schools were assigned a school type (Catholic or other nonpublic) based on their PSS school-type variable.

4.4 SCHOOL SAMPLE SELECTION

When the public and nonpublic schools in the sampling frame were stratified within each jurisdiction, a sample of about 100 grade-eligible schools was drawn with probability proportional to a measure of size (PPS) based on the estimated grade-specific enrollment of the school. In practice, the PPS sampling was implemented by the PPS systematic sampling. The number of schools selected generally did not vary by the sizes of jurisdictions. In each selected school, students were selected by systematic sampling. The PPS sampling schools and systematic sampling for students would give each student an equal probability of selection (Kish, 1965).

One or more sessions were sampled within each school. The number of sessions selected depended on the school's estimated grade-specific enrollment, though the overwhelming majority of schools at grade 4 were allocated a single session.

4.4.1 Measure of Size and Sample Selection

For each grade-eligible school, an estimated grade enrollment (EGE) was obtained by dividing the school's total student enrollment by the school's number of grades. Based on previous assessments, the EGE provided appropriate estimates for the sampling process. The estimated grade enrollment was not used directly in sample selection as the measure of size of grade students in schools. Instead, the measure of size was based on the following function of estimated grade enrollment. Tables 4-3 and 4-4 define the relationship between the estimated grade enrollment and measure of size in sample selection for grades 4 and 8.

Table 4-3
Estimated Grade Enrollment and Measure of Size, Grade 4

Estimated Grade Enrollment (EGE)	Measure of Size
$EGE < 10$	15
$10 \leq EGE < 20$	$1.5 \times EGE$
$20 \leq EGE < 33$	30
$33 \leq EGE$	EGE

Table 4-4
Estimated Grade Enrollment and Measure of Size, Grade 8

Estimated Grade Enrollment	Measure of Size
$EGE < 10$	30
$10 \leq EGE < 20$	$3 \times EGE$
$20 \leq EGE < 65$	60
$65 \leq EGE$	EGE

Schools were designated as being in "small" or "large" districts and were assigned to one of two school size classifications. A large district was defined as a district containing 20 percent or more of a jurisdiction's student population. All other districts were considered small. Schools were assigned to the large school size classification if their estimated grade enrollment was greater than 19 students. Otherwise, schools were assigned to the small school size classification.

A sample of schools was then selected for each jurisdiction with probability proportional to each school's measure of size. The sampling frame of schools was sorted in systematic order prior to sample selection, as follows:

- Public schools
  - Small or large district status
  - School size classification
  - Urbanization stratum
  - Minority stratum
  - Median household income
- Nonpublic schools
  - School size classification
  - Metro area status
  - Catholic/nonCatholic
  - Estimated grade enrollment

Sorting the sampling frame in a specific order prior to systematic sample selection ensures that the sampled schools represent a variety of population subgroups. Tables B-15 through B-18 in Appendix B provide the distributions for the counts of selected schools by sampling strata by grades for both public and nonpublic schools. Tables B-19 through B-22 show weighted school participation rates and counts of sampled schools by jurisdiction, grade, and subject for both public and nonpublic schools.

4.4.2 Sparse State Sample Option

The standard NAEP sample design requirements are burdensome for jurisdictions whose student populations are largely concentrated in small schools. In these jurisdictions, large numbers of schools must be selected in order to reach the required student sample sizes. Thus these jurisdictions bear an exceptionally large burden in school recruitment and assessment administrations, but are not eligible for any reduction in sample size under the reduced sample option, which is described in Section 4.5.2. In an effort to address this problem, while at the same time ensuring that adequate sampling standards for representation and precision were assured, the sparse state sample option was offered to qualifying jurisdictions for the first time in 1998. The jurisdictions eligible for this option were those that would have had at least 120 public schools selected under the full sample. Under the option, a proportional sample of schools was selected and the school and student sample sizes were reduced such that the following conditions held:

1. The number of schools selected was at least 115 (noting that many states have been assigned sample sizes close to this in the past).

2. The number of schools selected for each individual subject was at least 80 (so as to assure reliable sample inferences can be made for each subject).

3. The sampling probability of each individual school was at least half as great as for a full sample (this is to ensure that all parts of the jurisdiction's student population are adequately represented).

4. The largest schools were all retained in the sample, and the student sample sizes in these schools were also retained.

Note that the third and fourth conditions taken together imply that all of the large schools were retained and at least half of the small schools were retained. In practice, this usually meant that jurisdictions had their samples reduced from over 120 schools to 115, since the first condition is usually the most restrictive. Also, the student sample would be at least a half sample, and usually was substantially more than that. The eligible jurisdictions were Alaska, Kansas, Montana, Nebraska, North Dakota, Oklahoma, and South Dakota at grade 4; and Alaska, Montana, Nebraska, North Dakota, South Dakota, Vermont, and Wyoming at grade 8. The effect of the Sparse State Sample Option on sample sizes is shown in Table 4-5 for participating jurisdictions exercising the option. Note that Alaska, Nebraska, and North Dakota at grade 4, and Nebraska and North Dakota at grade 8 also requested the option, but later decided not to participate (at least in the public-school portion of the assessment).

Table 4-5
The Effect of the Sparse State Option on Sample Sizes, by Grade for Jurisdictions Exercising the Option

Grade	Jurisdiction	Original School Sample	Reduced School Sample	Reduced Student Sample as a Percentage of the Original Student Sample
4	Montana	132	115	88%
8	Montana	139	116	89%
8	Oklahoma	130	115	89%

4.4.3 Control of Overlap of School Samples for National Educational Studies

The issue of school sample overlap has been relevant in all rounds of NAEP in recent years. To avoid excessive burden on individual schools, NAEP has developed a policy for 1998 of avoiding overlap between national and state samples. This was to be achieved without unduly distorting the resulting samples by introducing bias or substantial variance. The procedure used was an extension of the method proposed by Keyfitz (1951). The general approach is given in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading* (Mazzeo, Allen, & Kline, 1995). It is summarized briefly as follows.

To control overlap between NAEP state and national samples, a procedure was used that conditions on the national NAEP PSU sample. This simply means that national school selection probabilities that were conditional on the selection of national sample PSUs (i.e., within PSU school selection probabilities) were used in determining state NAEP school selection probabilities. No adjustments were made to state NAEP school selection probabilities in jurisdictions where there were no national NAEP PSUs selected. This procedure reduces the variance of the state samples, although it leads to a greater degree of sample overlap than if unconditional national selection probabilities had been used in the procedure for controlling overlap between state and national samples. The procedure also recognizes the impact of the heavy within-PSU sampling in noncertainty PSUs in some jurisdictions. Schools to be included with certainty in the state sample are not subject to overlap control, as such schools are self-representing in the state sample. Excluding such schools on a random basis would add

extra variance to the state estimates. For actually drawing the state samples, a conditional probability of selection was used that was conditional on the selection of PSUs for the national NAEP samples. This procedure in general gave state NAEP conditional selection probabilities that are smaller than the unconditional state selection probabilities for schools that had been selected for the national sample. The state NAEP conditional selection probabilities were such that the unconditional probabilities obtained by integrating over the national sampling process were the required state NAEP probabilities, had overlap control not been implemented. Thus, a school's unconditional probability of selection for state NAEP was the same regardless of whether overlap control had been implemented. Counts of school selection for both state and national NAEP are found in Table 4-6.

Table 4-6
Number of Schools Selected for Both State and National NAEP, by Grade and School Type

State NAEP		National NAEP Grade		
Grade	**School Type**	**4**	**8**	**12**
4	Public	11	4	2
4	Nonpublic	0	18	4
8	Public	6	38	9
8	Nonpublic	15	3	28

4.4.4 Selection of Schools in Small Jurisdictions

All schools in jurisdictions with small numbers of public schools were selected. This was also true for the nonpublic schools in two jurisdictions. The jurisdictions and grades are shown in Table 4-7.

Table 4-7
Jurisdictions Where All Schools Were Selected, by Grade and School Type

	Public		Nonpublic	
Jurisdiction	**Grade 4**	**Grade 8**	**Grade 4**	**Grade 8**
Delaware	*	*	—	—
District of Columbia	*	*	—	*
DoDEA/DDESS	*	*	—	—
DoDEA/DoDDS	*	*	—	—
Hawaii	—	*	—	—
Rhode Island	—	*	—	—
Virgin Islands	*	*	*	*

4.4.5 Selection of New Public Schools

A sample of new public schools was drawn to properly reflect additions to the target population occurring after the sampling frame building information was created. A district-level file was constructed from the QED school-level file. The district-level file was divided into a "small" districts file that was not used in the selection of new schools, and a "medium and large" districts file that was used for this purpose. Small districts consisted of those districts in which there were at most three schools on the aggregate frame and no more than one fourth-, one eighth-, and one twelfth-grade school. New schools in

small districts were identified during school recruitment. The remainder of districts were denoted as "medium and large" districts.

A sample of medium and large public-school districts was drawn in each jurisdiction. All districts were selected in Delaware, the District of Columbia, Hawaii, and Rhode Island. The remaining jurisdictions in the file of medium and large districts (eligible for sampling) were divided into two files within each district. Two districts were selected per jurisdiction with equal probability among the smaller districts with combined enrollment of less than or equal to 20 percent of the state enrollment in the medium and large districts file. From the rest of the file, eight districts were selected per jurisdiction with probability proportional to enrollment. The breakdown given above applied to all jurisdictions that had at least eight large districts. In jurisdictions with fewer than 8 large districts, all of the large districts were selected, and then enough small districts were selected to make 10 districts selected altogether. The 10 selected districts in each jurisdiction were then sent a listing of all their schools that appeared on the file, and were asked to provide information about the new schools not included in the file. These listings, provided by selected districts, were used as sampling frames for selection of new public schools.

The eligibility of a school was determined based on the grade span. A school was also classified as "new" if a change of grade span was such that the school status changed from ineligible to eligible. The average grade enrollment for these schools was set to the average grade enrollment before the grade-span change. The schools found eligible for sampling due to the grade-span change were added to the new school selection frame.

The probability of selecting a school was

$$\text{minimum}\left\{\frac{\text{sampling rate}\cdot\text{measure of size}}{P(\text{district})},\ 1\right\},$$

where *P*(district) was the probability of selection of a district and the sampling rate was the rate used for the particular jurisdiction in the selection of the original sample of schools. For example, in a state where the sampling rate is .005, a school with 100 eligible students in a district selected with probability .75 would have a probability of selection of .67 [(.005 x 100)/.75].

In each jurisdiction, the sampling rate used for the main sample of grade-eligible schools was used to select the new schools. Additionally, all new eligible schools coming from small districts (those with at most one grade 4 and one grade 8 school and at most three schools on the aggregate frame) that had a school selected in the regular sample for the fourth grade were included in the sample with certainty. In the 1998 state assessment, there were no such schools.

Table 4-8 shows the number of new schools coming from the medium and large and small districts for the fourth- and eighth-grade samples.

Table 4-8

*NAEP 1998 Distribution of New Schools Coming from Districts Designated as "Medium" or "Large"**

Jurisdiction	Grade 4 Samples	Grade 8 Samples
Total	70	49
Alabama	2	3
Alaska	—	0
Arizona	5	5
Arkansas	0	1
California	1	1
Colorado	3	2
Connecticut	3	0
Delaware	13	2
District of Columbia	1	5
Florida	0	0
Georgia	0	0
Guam	—	0
Hawaii	2	4
Illinois	0	—
Indiana	—	0
Iowa	1	0
Kansas	0	—
Kentucky	1	0
Louisiana	4	4
Maine	3	2
Maryland	0	2
Massachusetts	6	1
Michigan	1	0
Minnesota	1	0
Mississippi	0	0
Missouri	2	2
Montana	0	0
Nebraska	4	2
Nevada	6	1
New Hampshire	0	0
New Jersey	—	0
New Mexico	1	1
New York	1	3
North Carolina	1	2
North Dakota	—	0
Oklahoma	0	—
Oregon	0	0
Rhode Island	0	0
South Carolina	2	0
Tennessee	0	0

* In the 1998 assessment, there were no sampled schools designated "small".

(continued)

Table 4-8 (continued)
*NAEP 1998 Distribution of New Schools Coming from Districts Designated as "Medium" or "Large"**

Jurisdiction	Grade 4 Samples	Grade 8 Samples
Texas	1	3
Utah	1	0
Vermont	—	0
Virgin Islands	1	—
Virginia	0	0
Washington	0	0
West Virginia	0	0
Wisconsin	0	1
Wyoming	0	2
DoDEA/DDESS	2	0
DoDEA/DoDDS	1	0

* In the 1998 assessment, there were no sampled schools designated "small".

4.4.6 Assigning Subject, Sample Type, and Monitor Status

For the sampled schools, one or more subject sessions were assigned within each school. The number of sessions selected depended on the school's estimated grade-specific enrollment, though the overwhelming majority of schools at grade 4 were allocated a single session.

Rules for assigning subjects (reading at grades 4 and 8; writing at grade 8 only) varied by grade. All fourth-grade schools were assigned to participate in reading assessments. All eighth-grade schools with 25 or more students were assigned to participate in both reading and writing assessments. Schools with fewer than 25 students were assigned one randomly selected subject.

The 1998 state assessment used the inclusion rules from 1996 for SD/LEP students (see Chapter 3) for two different sets of schools (S2 and S3 subsamples). The S2 subsample was not given the option of taking the assessment with accommodations. The S3 subsample was given the option of offering SD/LEP students accommodations. A sample type variable was created to reflect which set of rules to use within a given school. The sample type variable applied to reading only because writing was always administered using S3 rules including accommodations.

The schools assigned reading were sorted by stratum (public and nonpublic) and school ID and then assigned sample type in an alternating pattern within the sorted list. The inclusion rules for SD/LEP students are described in Chapter 3.

Since the state assessments were given by local administration, Westat monitored field assessments in some of the schools in the state assessments as they did in the national assessments to make reliable comparisons between both assessments. Jurisdictions received 25 or 50 percent monitoring of sessions depending on previous participation in the state assessments. All jurisdictions received 25 percent monitoring except Kansas, where 50 percent monitoring was used. The sampled schools were sorted by stratum, subject, sample type, and school ID and then assigned the two levels of monitoring in an alternating pattern.

4.4.7 School Substitution and Retrofitting

A substitute school was assigned to each sampled school (to the extent possible) prior to the field period through an automated substitute selection mechanism that used distance measures as the matching criterion. Schools were also required to be of the same type (i.e., public, nonpublic, BIA, and DoDEA schools were only allowed to substitute for each other), and substitutes for nonpublic, BIA, and DoDEA schools were required to come from within the same district. Public-school substitutes were required to come from different districts. Two passes were made at the substitution, with the second pass raising the maximum distance measure allowed and removing the different district assignment requirement for public schools. This strategy was motivated from the fact that most public-school nonresponse occurs at the school district level.

A distance measure was used in each pass and was calculated between each sampled school and each potential substitute. The distance measure was equal to the sum of four squared standardized differences. The differences were calculated between the sampled and potential substitute school's estimated grade enrollment, median household income, percent Black enrollment and percent Hispanic enrollment. Each difference was squared and standardized to the population standard deviation of the component variable (e.g., estimated grade enrollment) across all grade-eligible schools and jurisdictions. The potential substitutes were then assigned to sampled schools by order of increasing distance measure. An acceptance limit was put on the distance measure of .60 for the first pass. A given potential substitute was assigned to one and only one sampled school. Some sampled schools did not receive assigned substitutes (at least in the first pass) because the number of potential substitutes was less than the number of sampled schools or the distance measure for all remaining potential substitutes from different districts was greater than .60.

In the second pass, the different district constraint for public schools was lifted and the maximum distance allowed was raised to .75. This generally brought in a small number of additional assigned substitutes. Although the selected cutoff points of .60 and .75 on the distance measure were somewhat arbitrary, they have been used since 1994 after being decided upon for the 1994 trial state assessment by a group of statisticians reviewing a large number of listings beforehand and finding a consensus on the distance measures at which substitutes began to appear unacceptable.

Jurisdictions that did not receive substitutes for all selected schools were allowed to retrofit unused substitutes after part of the field period elapsed. Substitutes that were assigned to cooperating or ineligible original selections were free to be assigned to other original selections that did not receive substitutes. These free substitutes were put back into the substitute selection mechanism described above and allowed to pair up with other original selections.

The information about the number of substitutes provided and the number participating in each jurisdiction can be found in the report *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000). Of the 45 participating jurisdictions, 42 were provided with at least one substitute at grade 4, and 41 were provided with at least one substitute at grade 8. Among jurisdictions receiving no substitutes, the majority had 100 percent participation from the original sample. The total number of substitutes associated with nonparticipating original schools were 524, 600, and 400 for grade 4 reading, grade 8 reading, and grade 8 writing, respectively. The numbers of substitutes that participated were 153, 93, and 97, respectively.

4.5 STUDENT SAMPLE SELECTION

4.5.1 Student Sampling and Participation

To select a student sample, schools initially sent a complete list of students to a central location in November 1997. They were not asked to list students in any particular order, but were asked to implement checks to ensure that all grade-eligible students were listed. Based on the total number of students on this list, the student listing form, sample line numbers were generated for student sample selection. To generate these line numbers, the sampler entered the number of students on the form and the number of sessions into a personal computer that had been programmed with the sampling algorithm. The program generated a random start that was used to systematically select the student line numbers (30 per session). To compensate for new enrollees not on the student listing form, extra line numbers were generated for a supplemental sample of new students.

After the student sample was selected, the administrator at each school identified students who were incapable of taking the assessment either because they were identified as students with disabilities (SD) or because they were classified as being of limited English proficiency (LEP). New inclusion rules, which were first used in 1996, were used. These rules were meant to clarify the procedure for identifying whom to exclude from NAEP and to provide wider inclusion of SD and LEP students. More details on the procedures for student exclusion are presented in Chapter 5 of this report and in Westat's *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

When the assessment was conducted in a given school, a count was made of the number of nonexcluded students who did not attend the session. If this number exceeded three students, to reduce nonresponse error, the school was instructed to conduct a makeup session, to which all students who were absent from the initial session were invited. A summary of the distribution of the student samples, student exclusion rates, and response rates by grade, school type, and jurisdiction can be found in Tables B-23 to B-28 in Appendix B.

4.5.2 The Reduced Sample Option

Jurisdictions with fewer than 100 schools, and schools assigned more than two sessions at grade 4 or more than three sessions at grade 8 were given the option to reduce the expected student sample size in order to reduce testing burden and the number of multiple-testing sessions for participating schools. If jurisdictions chose to exercise this option, the estimates obtained from the assessment were more variable than they otherwise would have been. In general, jurisdictions could reduce student sample sizes by adjusting the number of sessions with participating schools subject to the following constraints:

- The minimum number of sessions per school had to be equal to 1.
- The maximum number of sessions per school had to be equal to 2 at the fourth grade and 3 at the eighth grade.
- The expected student size from the reduced sample was greater than or equal to half of the original student sample size.

To reduce testing burden and the number of testing sessions for participating schools, Delaware exercised the reduced sample option at both grade levels.

Chapter 5

FIELD OPERATIONS AND DATA COLLECTION[1]

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

5.1 INTRODUCTION

This chapter describes the field operations and data collection activities for the 1998 National Assessment of Educational Progress (NAEP). Traditionally, NAEP is comprised of main national samples, long-term trend (LTT) national samples, and state samples. For 1998, LTT was not scheduled, however, so the 1998 assessment program consisted of main, national, and state samples, as described in this chapter. The national NAEP component typically involves new assessment items, and may include new subject areas and innovative features. The national assessments are based on national probability samples of schools and students that allow for regional and national reporting only. The state assessment, the other major component of NAEP for 1998, comprises the state program that uses national NAEP assessment materials and involves much larger sample sizes per state (or jurisdiction), so that results can be reported for each participating state or jurisdiction.

The organization and operation of 1998 NAEP field activities are described in the remaining sections of this chapter. For all components, NAEP guarantees the anonymity of participants, and student or teacher names are never recorded on assessment booklets nor removed from the schools. NAEP results are reported on the national level, by region of the country, by state, or by demographic subgroup.

5.1.1 Organization of the National Assessment for 1998

The 1998 national assessment was conducted in a sample of approximately 2,700 public and nonpublic schools located in 94 geographic areas called primary sampling units (PSUs) throughout the states and the District of Columbia. The PSUs were selected by Westat to represent the nation as a whole.

Assessments for national NAEP were conducted from January through March at grades 4, 8, and 12. Students were assessed in reading, writing, and civics, and this included a special assessment in civics only, which established a trend line (but not long-term trend) from the earlier civics assessment in 1988. The civics special trend assessment was conducted at the same time and in some of the same schools as national NAEP. Three session types were administered in 1998:

- *Reading:* The reading assessment was based on the existing frameworks, which established a new trend line in 1992 (NAGB, 1990). The reading booklets included the background questions in the front of the booklet.

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

- *Writing/Civics:* The writing and civics assessments were combined into one session, with the different booklets spiraled together. These assessments were based on new frameworks developed for the 1997 field test (Center for the Evaluation, Standards, and Student Testing [CRESST], 1996; Council of Chief State School Officers [CCSSO], 1996)
- *Civics Special Trend:* The civics special trend study was based on the frameworks developed for the 1988 assessment (CCSSO, 1996), and was distinct from the civics assessment included with the writing tests. These sessions used the same materials used in 1988, including an answer sheet separate from the test booklets.
- Most schools had two of the possible three types of sessions administered in 1998 (reading, writing/civics, and/or civics special trend). In some of the smallest schools, only one of the types of sessions was administered. Following the precedent established in 1996, accommodations (described in Section 5.1.1.2) were offered for the writing/civics sessions and for half of the reading sessions, but none for the civics special trend.

In order to reduce the burden on the participating schools, NAEP field staff performed most of the work associated with the assessments. Introductory contacts and meetings (if needed) occurred in the fall of 1997 to enlist cooperation and explain the assessment procedures to district and school representatives and to set a mutually agreed-upon assessment date for each school. The assessment supervisor visited the school a week or two before the assessment to select the sample of students. The assessment sessions were conducted by exercise administrators, also members of the NAEP field staff, under the direction of the assessment supervisor. At the conclusion of the assessment in a school, field staff coded demographic information on the booklet covers and shipped the completed materials to National Computer Systems (NCS), the processing subcontractor for NAEP (see Chapter 6 for more detailed information on processing assessment materials). For reference, the national NAEP field staff administrative structure is summarized in the chart below.

WESTAT NATIONAL NAEP FIELD STAFF ADMINISTRATIVE STRUCTURE
Field Director *Oversees all aspects of field operations*
Field Managers *Report to Westat Field Director and oversee supervisors who have direct contact with schools*
Field Supervisors *Report to a specific field manager, gain cooperation of schools, select student samples, arrange and supervise assessments, assigning assessments to exercise administration*
Exercise Administrators *Conduct assessment sessions and assist with field paperwork/record keeping under direct supervision of a field supervisor*

5.1.1.1 Additional Special Studies

Apart from the civics special trend study, two other special studies, each requiring additional interaction with school personnel, were carried out in conjunction with the national 1998 assessment. A classroom-based writing study was designed to explore methods of assessing students' writing abilities at grades 4 and 8 by using written assignments that students had completed as part of their school curriculum. A High-School Transcript Study, similar to the transcript study that took place in 1994, was conducted in a number of grade 12 schools included in the main assessment.

These results from these two studies will be available in forthcoming reports. More information about the studies is provided in section 5.3.2.

5.1.1.2 Exclusions and Accommodations for Students

Historically, a small proportion (less than 10%) of the sampled students have been "excluded" from NAEP assessment sessions because, according to school records, they are students with either disabilities (SD) or limited English language proficiency (LEP) who have been determined to be incapable of participating meaningfully in the assessment. More recently, especially with the passage of the Individuals with Disabilities Education Act, increased attention has been given to these students and to including as many of them as possible in NAEP sessions. NAEP addressed these concerns through a 1996 special study (Mazzeo, Carlson, Voelkl, & Lutkus, 1999) that used both old and new "inclusion" criteria and (in some schools) offered accommodations for testing students with disabilities, limited English proficiency, or both (SD/LEP).

Results of the 1996 assessment indicated that the revision of the criteria for including students had little impact on the numbers of students included; therefore, for 1998 and beyond, the revised criteria were used because they are most current. The 1996 data also indicated that providing accommodations resulted in greater inclusion of students who might previously have been excluded from NAEP.

The inclusion criteria used in the 1998 NAEP assessments fell into two categories—students with disabilities (SD) and students with limited English proficiency (LEP). A student identified as having a disability (SD), that is, a student with an Individualized Education Plan (IEP) or equivalent classification, was to be excluded from the NAEP assessment if any of the three following conditions applied:

- The IEP team or equivalent group determined that the student was unable to participate in assessments such as NAEP.
- The student's cognitive functioning was so severely impaired that he or she could not participate.
- The student's IEP required that the student be tested with an accommodation that is not permitted by NAEP, and the student could not demonstrate his or her proficiency in reading, writing, or civics without that accommodation.

A student who was identified as limited English proficient (LEP) and was a native speaker of a language other than English was to be excluded from the NAEP assessment only if both of the following conditions applied:

- The student received language arts instruction primarily in English for less than three school years including the current year.

- The student was unable to demonstrate his or her proficiency in reading, writing, or civics, even with an accommodation permitted by NAEP.

Decisions on exclusion were made by the assessment supervisor in consultation with school staff and were guided by the SD/LEP questionnaires completed by the school staff. This questionnaire, which was completed for each SD/LEP student in the sample by someone at the school knowledgeable about the student, asked about the student's background and the special programs in which the student participated.

Because the 1998 reading assessment results were to be compared to those from the 1992 assessment, one group of students was assessed under conditions similar to those in 1992. Thus, in half of the 1998 reading sessions, accommodations were not permitted. To be able to evaluate the differences in results that occur when students are assessed with accommodations, accommodations *were* permitted in the other half of the reading sessions.

For the writing/civics sessions, because new trend lines are being established, accommodations were made available to all students, if needed or appropriate. Finally, for civics special trend sessions, accommodations were not permitted for any students.

Accommodations included but were not limited to extended time to answer the test questions, large-print booklets, bilingual dictionaries, scribe or use of computer to record answers, session in which the test administrator would read the test questions aloud, sessions with a smaller number of students than in the regular sessions, and one-on-one test administrations.

5.1.2 Organization of the State Assessment for 1998

Forty-four states, the District of Columbia, Virgin Islands, and Guam volunteered for the 1998 state assessment, as did the Department of Defense Domestic Dependent Elementary and Secondary Schools (DoDEA/DDESS) and the Department of Defense Dependents Schools (DoDEA/DoDDS).

Table 5-1 identifies the jurisdictions participating in the state assessment. For the state program, assessments were conducted in one subject, reading, at the fourth grade and in reading and writing at the eighth grade.

Data collection for the 1998 state assessment involved a collaborative effort between the participating jurisdictions and the NAEP contractors, especially Westat, the field administration contractor. Westat's responsibilities included:

- Selecting the sample of schools and students for each participating jurisdiction
- Developing the administration procedures and manuals
- Training state and school personnel to conduct the assessments, and
- Conducting an extensive quality assurance program which involves observing and monitoring 25 percent of the state NAEP sessions conducted by school staff.

Table 5-1

Jurisdictions Participating in the 1998 State Assessment Program

Alabama	Guam	Missouri	South Carolina
Alaska	Hawaii	Montana	Tennessee
Arizona	Illinois[2]	Nebraska	Texas
Arkansas	Indiana	Nevada	Utah
California	Iowa	New Hampshire	Vermont
Colorado	Kentucky	New Jersey	Virginia
Connecticut	Louisiana	New Mexico	Washington
Delaware	Maine	New York	West Virginia
DoDEA/DDESS[1]	Maryland	North Carolina	Wisconsin
DoDEA/DoDDS[1]	Massachusetts	North Dakota	Wyoming
District of Columbia	Michigan	Oregon	
Florida	Minnesota	Pennsylvania	
Georgia	Mississippi	Rhode Island	

[1] DoDEA refers to the Department of Defense Education Activity. Its domestic schools (Department of Defense Domestic Dependent Elementary and Secondary Schools [DDESS]) and its overseas schools (Department of Defense Dependents Schools [DoDDS]) participated in the state assessment program.
[2] Illinois participated in the assessment; however, results were not reported due to low school participation rates prior to the addition of substitute schools.

Each jurisdiction volunteering to participate in the 1998 program was asked to appoint a state coordinator. In general, the coordinator was the liaison between NAEP/Westat staff and the participating schools. In particular, the state coordinator was asked to:

- Gain the cooperation of the selected schools
- Assist in the development of the assessment schedule in the selected schools
- Receive the lists of all grade-eligible students from the schools
- Coordinate the flow of information between the schools and NAEP
- Provide space for the Westat state supervisor to use when selecting the samples of students
- Notify assessment administrators about training and send them their assessment manuals, and
- Send the lists of sampled students to the schools.

Westat hired and trained six field managers for the state assessment. Each field manager was responsible for working with the state coordinators of seven to eight jurisdictions and for overseeing assessment activities. The primary tasks of the field managers were to:

- Obtain information from state coordinators about cooperation and scheduling
- Make sure the arrangements for the assessments were set and assessment administrators identified, and
- Schedule the assessment administrator training sessions.

Westat also hired and trained a state supervisor for each jurisdiction. The 1998 state assessment involved about the same number of state supervisors (Westat staff) as the 1992, 1994, and 1996 assessments, since approximately the same number of jurisdictions were involved each year. In addition, three troubleshooters were trained in case any state supervisor was unable to complete their assignment. The primary tasks of the state supervisor were to:

- Select the samples of students to be assessed
- Recruit and hire the quality control monitors throughout their jurisdiction
- Conduct in-person assessment administration training sessions, and
- Coordinate the monitoring of the assessment sessions and makeup sessions.

At the school level, an assessment administrator(s) was appointed (by the school), and this person, often a teacher, was responsible for preparing for and conducting the assessment session(s) in one or more schools. These individuals were usually school or district staff and were trained by Westat staff. The assessment administrator's responsibilities included:

- Receiving the list of sampled students from the state coordinator
- Identifying sampled students who should be excluded
- Distributing assessment questionnaires to appropriate school staff and collecting them upon their completion
- Notifying sampled students and their teachers
- Administering the assessment session(s)
- Completing assessment forms, and
- Preparing and shipping the completed assessment materials.
- Decisions on exclusion of students (if any) were made in consultation with school staff and were guided by the SD/LEP questionnaires completed by the school staff.

In addition, Westat hired several quality control (QC) monitors in each jurisdiction to monitor assessment sessions. The number of QC monitors varies, from about 4 to 6, by state according to the number of schools samples in a state. The QC monitors report to Westat supervisors and are responsible for observing a subset of the state NAEP sessions conducted by the school staff. For reference, the state NAEP field staff administrative structure is summarized in the following chart.

WESTAT STATE NAEP FIELD STAFF ADMINISTRATIVE STRUCTURE
Field Director *Oversees all aspects of field operations*
Field Managers *Work directly with state coordinators on gaining cooperation of schools and oversee state supervisors (Westat staff) who select student samples and supervise QC monitors*
Field Supervisors *Select student samples at state coordinators office, train assessment administrators (chosen by schools) to conduct assessments, schedule and oversee assessment observation visits made by quality control monitors*
Assessment Administrators *Are school (or district) staff appointed by the school to conduct one or more state NAEP assessment sessions in that school*
Quality Control Monitors *Are hired and trained by Westat field managers and field supervisors, interview each school for feedback on the assessment and to visit a specific subsample of schools to observe the administration of the NAEP session by school staff; report directly to field supervisor*

5.2 PREPARING FOR THE ASSESSMENTS

5.2.1 Gaining the Cooperation of Sampled Schools

The process of gaining cooperation of the schools selected for the NAEP assessments, both national and state, began in August 1997 with a series of letters and contacts with state and district-level officials. The National Center for Education Statistics (NCES) first sent each jurisdiction a letter announcing NAEP plans for 1998. Westat then contacted the state test directors or NAEP state coordinators in each sampled state to notify them of the districts and schools selected in their states. In the 41 jurisdictions participating in the state assessment that also had schools sampled for the national assessment, the state received the list of districts and schools sampled for both the national and state assessments.

From September through early December 1997, Westat sent lists of schools sampled for the assessments and other NAEP materials to district superintendents, diocesan superintendents of Catholic schools, and principals or heads of schools in other nonpublic schools, inviting their participation. These initial mailings paved the way for telephone contacts by NAEP field supervisors who were assigned the task of gaining cooperation and scheduling assessment dates.

The schedule for project activities for the 1998 national and state assessments was as follows:

August 1997	*Department of Education sent first letter to chief state school officers and state test directors.*
	Westat sends state coordinators the lists of schools selected for 1998 state assessments along with informational materials. Similar mailings continue, to state test directors, through mid-September 1997 for national NAEP schools.
August/September 1997	*Westat field managers visit states to train state coordinators to use computerized state NAEP field management system for recording participation status of the state NAEP schools.*
September 24–27, 1997	*Training session held for national assessment schedulers.*
Mid-to-Late September 1997	*Westat sent samples and informational materials to school districts, if not already sent by state coordinators.*
Mid-September – December 1, 1997	*Supervisors contacted districts and schools to secure cooperation and to schedule assessments in national NAEP schools.*
	Supervisors conducted introductory meetings for the national NAEP assessment, by telephone (or in person if requested by districts or schools). Westat selected substitutes for refusals.
	Supervisors recruited, hired, and trained exercise administrators for national NAEP.
September – November 1997	*State coordinators obtained cooperation from districts and public schools for state NAEP samples. State coordinators reported participation status to Westat field managers via hardcopy lists or computer files.*
	Westat field staff secured cooperation from sampled nonpublic schools (for state NAEP samples).
	State coordinators sent summary of school tasks, student listing forms, and new enrollee student listing forms to participating public schools in state NAEP samples.
October 6 – November 12, 1997	*Westat sent student listing forms and new enrollee listing forms to participating nonpublic schools in state NAEP samples.*

November 5 – 8, 1997	*Training session for state NAEP supervisors.*
Early December 1997	*Supervisors sent informational materials to principals and school coordinators and Westat send letters confirming assessment schedules to each national NAEP school.*
December 1 – 12, 1997	*State NAEP supervisors visited state coordinator offices to select student samples and prepare administration schedules listing the students selected for each session in public schools selected for state NAEP. The state supervisor prepared a package to be sent to each public school containing the administration schedules and the instructions for assessing students with disabilities and/or limited English proficiency.*
December 1 – 5, 1997	*Westat provided schedule of state NAEP assessment administrator (AA) training sessions and copies of the Manual for Assessment Administrators to state coordinators for distribution.*
	Westat distributed state NAEP AA training schedules and manuals directly to nonpublic schools.
December 8, 1997 – January 2, 1998	*State coordinator notified state NAEP AAs of the date and time of training and sent each a copy of the Manual for Assessment Administrators.*
December 9 – 15, 1997	*National NAEP assessment supervisor training session was held.*
January 5 – March 27, 1998	*Student samples were selected for national NAEP and assessments were administered. Makeup sessions, if needed, were held from March 30 to April 3, 1998.*
January 7 – 10, 1998	*Training session was conducted for quality control monitors (see Section 5.4.2) who observe state NAEP AAs in 25% of state NAEP sessions.*
January 12 – 30, 1998	*Westat state NAEP supervisors conducted assessment administrator training sessions.*
	Student samples were selected for nonpublic schools in state NAEP training sessions for state NAEP AAs.

January 19 – February 13, 1998	*State coordinators sent packages containing administration schedules and instructions for assessing students with disabilities and/or limited English proficiency to each public school two weeks before the scheduled assessment date for state NAEP.* *NCS sent assessment materials to each school two weeks before the scheduled assessment date for state NAEP.*
February 2 – 27, 1998	*State NAEP assessments were conducted and monitored, with makeup sessions held the week of March 2–6, 1998.*

5.2.2 Supervisor Training

Training for assessment supervisors was multiphased and involved separate sessions conducted in August, September, and December 1997. In addition, a large state NAEP training session for quality control monitors was held in early January 1998. All training was conducted by the Westat project director, field director, and home office staff. Also in attendance were representatives from Educational Testing Service (ETS), NCS, and NCES.

The first training session was held September 24 – 27, 1997 for 40 field staff assigned to gaining cooperation phase of the project. After an introduction to the study, which included the background and history of NAEP, an overview of the 1998 assessments, and the 1997–1998 assessment schedule, the training continued with a thorough presentation of NAEP's activities for contacting schools and gaining their cooperation. This is a lengthy process of contacting states, districts, and schools regarding their participation in and scheduling for NAEP; several demonstration phone calls, role plays, and exercises were used to provide some practical experience during this part of the training. Other training topics included: supervisory responsibilities, setting the assessment schedule, recruiting and training exercise administrators, and administrative forms and procedures. The scheduling supervisors also received a full day of training on using the reporting system installed on the laptop computers assigned to each of them for the gaining cooperation and scheduling phase. The reporting system is Westat's computerized field system used throughout national NAEP to record and update the participation status of each school and the attendance at each assessment session.

The 75 NAEP supervisors who were responsible for national NAEP assessment activities were trained again, in a second session, held December 9–15, 1997. The training began with a review of the preliminary activities during the fall, including results of gaining cooperation with districts and schools, scheduling of assessments, and the status of exercise administrator (EA) recruitment. (The role of EAs who conduct the assessments is discussed in Section 5.2.4.) The main focus of the training was a thorough discussion of assessment activities: sampling procedures, inclusion of SD/LEP students, teacher surveys, providing testing accommodations, conducting the sessions, and administrative forms and procedures. Westat's classroom management videotape, which is a 40-minute presentation on student behavior/attitudes and suggested approaches to "handling" students at various grade levels, was also shown at this training session. Key portions of the December training were devoted to carefully presenting the procedures involved in each of the two special studies, and each of these studies required a full day of training. These special studies, High School Transcript and Classroom-Based Writing, were initiated during the sampling visit to each school and continued on the assessment day, with certain

follow-up activities performed after the assessments. A full day of training on Westat's computerized NAEP field reporting system was also offered at the December training session.

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

The state NAEP supervisors attended a training session held November 5–8, 1997. This training session focused on the state supervisors' immediate tasks—selecting the student samples and hiring quality control monitors. Supervisors were given the training script and materials for the assessment administrators' training sessions they would conduct in January so they could become familiar with these materials.

Approximately 400 quality control monitors were trained for state NAEP in a session held in early January 1998. The first day of the training session was devoted to a presentation of the assessment administrators' training program by the state supervisors, which not only gave the monitors an understanding of what assessment administrators were expected to do, but gave state supervisors an opportunity to practice presenting the training program. The remaining days of the training session were spent reviewing the quality control monitor observation form and the role and responsibilities of the quality control monitors.

5.2.3 Contacting Districts and Nonpublic Schools

Once the supervisors were trained in September 1997, they began working on obtaining cooperation for national NAEP. In the states both sampled for national NAEP and participating in the state assessment, the national NAEP supervisor first spoke with the state NAEP field manager to determine what contacts, if any, had already been made with districts about NAEP. The approach the supervisors took when calling superintendents depended on whether the district had been notified about national NAEP by the state coordinator and whether the district also had schools selected for the state assessment. For districts that had been contacted by the state coordinator, the supervisor began by referring to that contact. Gaining specific cooperation in "state NAEP" schools was the responsibility of the state coordinators, while the Westat supervisors gained cooperation from all other schools, that is, the national NAEP schools and the nonpublic schools in state NAEP.

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

Over the years, however, these meetings have become somewhat redundant, since many districts have fallen into the national sample more than one time. It has also become more and more difficult to schedule these meetings, as district and school officials find it harder to allot time away from their offices. Thus, during the fall preparations for both the 1996 and 1998 NAEP studies, the material was almost always presented to the superintendents and principals during telephone calls rather than in formal meetings. Generally, an in-person meeting was held only if specifically requested by the district or school officials, or if the supervisor felt that such a meeting would provide a better chance for convincing a district to participate.

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

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

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

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

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

5.2.4 Recruiting, Hiring, and Training Exercise Administrators

During the fall, while the supervisors were contacting schools and scheduling assessments, their other major responsibility was to recruit and hire exercise administrators, who would administer the assessment sessions for national NAEP (for state NAEP, the school or district provides the assessment staff, known as assessment administrators). Exercise administrators for national NAEP were recruited from many sources. Each supervisor was given a PSU-by-PSU computerized list of exercise

administrators and other field staff who had worked previously on education studies for Westat. People who had served as exercise administrators before, with good evaluations from their previous supervisors, were usually the first considered for hiring. Subsequently, during contacts with the schools, the supervisors asked the school principals and other staff to recommend potential exercise administrators. These referrals were frequently retired teachers or substitutes. Finally, where necessary, ads were placed in local newspapers and the employment service was notified.

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

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

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

Each exercise administrator received an exercise administrator manual, which covered the full range of their job responsibilities. After studying the manual, they attended a half-day training session. During the training, the supervisor reviewed all aspects of the exercise administrators' job, including preparing materials, booklets, and administration schedules for assessments; the actual conduct of the session; post-assessment collection of materials; coding booklet covers; recordkeeping; and administrative matters. In January 1998, each exercise administrator attended a shorter, refresher training session, conducted by the assessment supervisor, to gain further experience with the specific procedures and materials to be used in the assessment sessions.

For state NAEP, assessment administrators (AAs), rather than exercise administrators, conducted the NAEP sessions in each school. These persons were appointed by the school (or the district), usually from school staff, at the request of the state coordinator who gained cooperation and established the assessment arrangements for state NAEP schools. All of these arrangements were made during October–December 1997. Manuals on conducting the assessment were shipped to AAs by the state coordinators. Then, in January 1998, each AA attended a half-day assessment administrator training conducted by Westat supervisors for state NAEP. Many of the assessment procedures addressed in these AA training sessions are thoroughly demonstrated in person via film and through exercises.

5.3 SELECTING THE STUDENT SAMPLES

5.3.1 Selecting the National NAEP Student Samples

After securing cooperation from the school, the first scheduled visit to each national NAEP school was made to select the sample of students to take part in the national assessments, and to conclude the arrangements for the actual testing. This visit was made in January by the supervisor responsible for the assessments in the school. Upon arriving at the school (rarely, sampling was done at the district office instead of in the school), the supervisor first reviewed the list of grade-eligible students and confirmed

verbally with the school coordinator that all eligible students were listed. If any eligible students were omitted, sampling could not proceed until the list was completed. Instructions for preparing the student list, which essentially should contain all students (even those not normally tested) enrolled in the grade to be assessed, are mailed to schools late in the fall term prior to the national assessments.

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

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

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

The school coordinator was also asked to determine whether any of these students should be excluded from national sessions based on the criteria for assessing SD/LEP students (the use of the criteria for each NAEP session type are discussed more specifically in Section 5.1.1.2). If the school coordinator could not identify the excluded students while the supervisor was at the school, the instructions were left with the coordinator along with blank copies of the SD/LEP student questionnaire. In those cases, the coordinator consulted with other school officials and informed the supervisor as to who was to be excluded when the coordinator returned for the national assessment.

For the 1998 assessment, the sampling process generated, in total, 149,880 students to be assessed in those schools cooperating in national NAEP. These counts include the SD/LEP students whom the schools determined should participate in the assessments. Accommodations were provided for an estimated 3,270 students. The most frequently provided accommodations were small-group, extended-time (untimed testing), and one-one-one testing. Detailed information on SD/LEP results and on the specific numbers of students actually assessed are provided earlier in Chapter 3 of this report, beginning with Table 3-8 and continuing in subsequent tables.

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

5.3.2 Selecting the Special Studies Samples

Two special studies, requiring added steps in the sampling process, were included in the national assessment for 1998. One of these special studies involved some of the students in writing assessments. The other involved collecting high school transcripts for grade 12 students. In the case of both studies, no student names or other identifiers were taken out of the schools.

The classroom-based writing study involved the random selection during the national NAEP sampling visit of one English/language arts classroom from each fourth- and eighth-grade school in which a writing assessment was to be conducted. At the same time, the students in that classroom were listed on a writing study linkage form so that the classroom students who also took the national writing assessment could be identified. The classroom's English/language arts teacher was asked to work with the students and have them select two examples of their best classroom writing. The students were asked to answer a few questions about each selection. The teachers completed an interview with the supervisor who collected the writing materials after the assessment. A full report on this study is due to be published in the year 2001.

The High School Transcript Study (HSTS) involved a subsample of most of the NAEP public high schools and one-third of the private high schools selected for the original 1998 national NAEP sample. This subsample comprised approximately 350 schools. Sampled schools were included regardless of whether they participated in national NAEP in order to minimize nonresponse bias. The HSTS student sample included all eligible twelfth-grade students who were sampled for the 1998 national assessment. This included students who were either excluded or absent, though not those who had withdrawn or were ineligible. Approximately 23,000 student transcripts were collected in this sample. Seven steps of the HSTS process were completed by Westat field supervisors at the time of the NAEP sampling visit, and these seven steps are as follows:

- Discuss the HSTS with the school coordinator prior to sampling visit.
- Complete the school information form concerning the organization of course offerings and course credits at this school, in an interview with school coordinator.
- Obtain and review course catalogs.
- Complete the course catalog check sheet.
- Obtain and review three examples of student transcripts.
- Mask all identifiers on the sample transcripts.
- Identify and mark the sampled students' files.

The actual collecting of the transcripts for the sampled twelfth-grade students was performed after the end of the 1997–1998 school year. The HSTS is conducted periodically to provide educational policy makers with information regarding course offerings and course-taking patterns, including links to the NAEP assessment results, in the nation's secondary schools. The 1998 results will be provided in detail at a later date in a separate HSTS report prepared by Westat.

5.3.3 Selecting the State NAEP Student Samples

Following their November training, the state NAEP supervisors' first task was to complete the selection of the sample of students who were to be assessed in each school. All participating schools were asked to send a list of their grade-eligible students to the state coordinator by November 14. Sample- selection activities were conducted in the state coordinator's office unless the state coordinator preferred that the lists be taken to another location.

Using a sampling package on their laptop computers, the supervisors generally selected a sample of 30 students per session type per school, with three exceptions: in schools with fewer than 30 students in the grade to be assessed, all of the students were selected; in schools in which more than one session was scheduled, 60 students (or some multiple of 30 students) were selected; and in schools with no more than 33 students in the grade, all students were selected for the assessment.

After the sample was selected, the supervisor completed an administration schedule for each session, listing the students to be assessed. The administration schedules for each school were put into an envelope and given to the state coordinator to send to the school two weeks before the scheduled assessment date. Included in the envelope were instructions for sampling students who had enrolled at the schools since the creation of the original list.

5.4 CONDUCTING THE ASSESSMENT SESSIONS

5.4.1 Conducting the National Assessments

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

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

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

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

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

5.4.2 Conducting the State Assessments

During the months of November and December 1997, the state supervisors also recruited and hired quality control monitors to work in their jurisdictions. It was the quality control monitor's job to observe the sessions designated to be monitored, to complete an observation form on each session, and to intervene when the correct procedures were not followed. Because earlier results indicated little difference in performance between monitored and unmonitored schools, and in an effort to reduce costs, the percentage of public schools to be monitored was maintained at 25 percent (i.e., the reduced monitoring rate initiated in 1994). The monitoring rate for nonpublic schools was also maintained at 25 percent (and reduced from the 50% rate used in 1994, which was the first year that nonpublic schools were assessed by NAEP). As has been customary in the past, monitoring was conducted at 50 percent for jurisdictions that were new to the state assessment in 1998. The schools to be monitored were known only to contractor staff; it was not indicated on any of the listings provided to state staff.

Almost immediately following the quality control monitor training, supervisors began conducting training for assessment administrators. Each quality control monitor attended at least two training sessions, to assist the state supervisor and to become thoroughly familiar with the assessment administrator's responsibilities. To ensure uniformity in the training sessions, Westat developed a highly structured three-day training program involving a script for trainers, a videotape, and a training example to be completed by the trainees. The training package, developed for previous state assessments, was revised to reflect the subjects and grades assessed in 1998. The supervisors were instructed to read the script verbatim as they proceeded through the training, ensuring that each trainee received the same information. The script was supplemented by the use of overhead transparencies, displaying the various forms that were to be used and enabling the trainer to demonstrate how they were to be filled out.

Two weeks prior to the scheduled assessment date, the state NAEP assessment administrator received the administration schedule and assessment questionnaires and materials. Five days before the assessment, the quality control monitor made a call to the administrator and recorded the results of the call on the quality control form for monitored schools, because the assessment administrators were not supposed to know in advance which sessions were designated to be monitored. The preassessment call was conducted in exactly the same way regardless of whether the school was to be monitored or not. For example, directions to the school were obtained even if the school was in the unmonitored sample. Most of the questions asked in the preassessment call were designed to gauge whether the assessment administrator had received all materials needed and had completed the preparations for the assessment.

If the sessions in a school were designated to be monitored, the quality control monitor was to arrive at the school one hour before the scheduled beginning of the assessment to observe preparations for the assessment. To ensure the confidentiality of the assessment items, the booklets were packaged in shrink-wrapped bundles and were not to be opened until the quality control monitor arrived or 45 minutes before the session began, whichever occurred first.

In addition to observing the opening of the bundles, the quality control monitor used the quality control form to check that the following had been done correctly: sampling newly enrolled students,

reading the script, distributing and collecting assessment materials, timing the booklet sections, answering questions from students, and preparing assessment materials for shipment. After the assessment was over, the quality control monitor obtained the assessment administrator's opinions of how the session went and how well the materials and forms worked.

If four or more students were absent from the session, a makeup session was to be held. If the original session had been monitored, the makeup session was also monitored. This required coordination of scheduling between the quality control monitor and assessment administrator.

5.4.3 Participation of Department of Defense Education Activity Schools in State NAEP

The schools run by the Department of Defense at military bases and other installations around the world participated in the NAEP state assessment for the third time in 1998. The participation of the selected schools was mandated by the Department of Defense Education Activity (DoDEA) schools. To accommodate the geographic diversity of DoDEA schools, some minor adaptations were made in the preparatory activities used for the other jurisdictions.

For 1998, as in 1996, the data collection in DoDEA schools was expanded from the 1994 model so that both the DoDEA's Department of Defense Elementary and Secondary Schools (DDESS), which includes domestic schools, and the DoDEA's Department of Defense Dependents Schools (DoDDS), which includes overseas schools, were surveyed. In 1994, only the schools at overseas installations were sampled as part of the state assessment.

Many of the quality control monitors hired for the DoDEA schools were based overseas, and many had previous experience working within the DoDEA system. They were referred to Westat by DoDEA. All quality control monitors for the DoDEA schools attended the quality control training in Los Angeles and several assessment administrator training sessions in the geographic areas in which they worked.

The samples of students to be assessed in the DoDEA schools were selected in the Westat home office, using standard NAEP procedures, from lists of students produced in the DoDEA offices in northern Virginia. Due to privacy concerns, only student ID numbers and not student names appeared on the DoDEA lists. Thus, after sampling, the administration schedules contained only the ID numbers, and the assessment administrators consulted school records and added the names of the students to the administration schedules prior to the assessments.

Two field supervisors were hired specifically to conduct assessment administrator trainings and monitor quality control monitors in the DoDEA/DoDDS schools. The DoDEA liaison in northern Virginia, who essentially functioned as the state coordinator, arranged the assessment administrator training sessions, all of which were held in schools or other facilities on the bases. In many cases, the quality control monitors were required to obtain special clearances through DoDEA to visit the bases for training and the assessments. The assessments in DoDEA schools were conducted using the same procedures as in all state assessment schools.

5.5 RESULTS OF THE NATIONAL NAEP ASSESSMENT

5.5.1 School and Student Participation

The unweighted school response rate for the national assessments in 1998 was 86 percent overall. This rate reflects the final sample of cooperating schools including 731 schools at grade 4; 753 schools at grade 8; and 599 schools at grade 12. Table 3-8 in Chapter 3 provides detailed counts and response rates.

The school response rates increased for 1998, which reverses the small declines in national assessment school response rates that occurred between 1990 and 1996. The gains were most likely due to persistent efforts to convert schools and districts that indicated that they were not interested in participating in the assessments. Both Westat field managers and ETS staff were employed in these conversion efforts.

Although school response rates for 1998 reached their highest levels since 1990, the most frequently stated reason for school and district refusals, historically, has been the increase in testing throughout the jurisdictions and the resulting difficulty in finding time in the school schedule to conduct the NAEP assessments. With so many states now mandating their own testing, school schedules are becoming tighter, and administrators are finding it increasingly difficult to accommodate outside testing. Despite the increased visibility and publicity surrounding NAEP, schools are reluctantly finding it necessary to decline participation as a result of the increasing demands on their students' time.

Of the 160,480 students sampled for the 1998 assessment, roughly 5 percent overall were excluded by schools. Altogether, 133,489 students were assessed across all three grades: 36,104 students were assessed at fourth grade, 48,797 were assessed at eighth grade, and 48,588 students were assessed at twelfth grade. The final student participation rate was 89 percent and this reflects students who participated in the NAEP session, based on "students to be assessed", that is, after eliminating any students withdrawn from the school, not eligible, or excluded by the school.

The student response rate at which supervisors were required to conduct a makeup session was 90 percent (lower rates were used prior to 1996); that is, any session (or group of sessions within the same subject area) at which fewer than 90 percent of the eligible students were assessed would require a makeup session. For 1998 NAEP sessions, about 23,200 of the roughly 150,000 students to be assessed were absent from the original sessions. Almost 7,000 of the absent students were assessed in makeup sessions, which represents about 30 percent of those absent from the original sessions. The makeup assessments added an estimated 4.5 percentage points to the overall student response rate for all grades combined, and it is further estimated that the makeups were conducted in 25 to 30 percent of the schools, with some variation according to the grade level assessed.

5.5.2 Assessment Questionnaires

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

For fourth grade and eighth grade, selected teachers in the subject areas of language arts and civic education were asked to fill out teacher questionnaires. The teachers asked to participate were the reading, writing, or civics teachers of those students selected for the assessment so that the teacher data could be linked to student performance data. The teacher questionnaire for grade 4 was combined into

one form, since it is recognized that at this grade level the same teacher would probably teach all of the subjects. For grade 8, there were two distinct questionnaires, one for language arts teachers and the other for civics teachers. At grade 12, teacher questionnaires were not used in 1998 NAEP.

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

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

Table 5-2
*Background Questionnaires Received for Schools, Teachers, and SD/LEP Students in the 1998 National Assessment**

		Teacher Questionnaires			
	School Questionnaire	Language Arts/Civics (Grade 4 only)	Language Arts	Civics	SD/LEP Student Questionnaire
Grade 4					
Number Expected	731	2,145	—	—	7,066
Number Received	700	2,081	—	—	6,830
Percent Received	96%	97%	—	—	97%
Grade 8					
Number Expected	753	—	2,303	1,594	7,942
Number Received	722	—	2,170	1,489	7,575
Percent Received	96%	—	94%	93%	95%
Grade 12					
Number Expected	599	—	—	—	6,588
Number Received	570	—	—	—	6,214
Percent Received	95%	—	—	—	94%

* Every cooperating school was given a school questionnaire, but some schools failed to complete their questionnaires, so that the number of completed questionnaires is smaller than the number of participating schools.

5.6 RESULTS OF THE STATE NAEP ASSESSMENT

5.6.1 School and Student Participation

Table 5-3 shows the results of the state coordinators' efforts to gain the cooperation of the schools selected for state NAEP.

Overall, for the 1998 state assessment in reading, 4,594 public schools and 570 nonpublic schools for grade 4 participated. For eighth grade, 3,805 public schools and 453 nonpublic schools participated in reading, and 3,688 public and 450 nonpublic participated in writing assessments.

Participation results for students in the 1998 state assessments are given in Table 5-4. Nearly 139,000 fourth-grade students and over 237,000 eighth-grade students were sampled. As can be seen from the table, the original sample, which was selected by the NAEP state supervisors, comprised approximately 135,000 (or 97%) of the total number of students sampled for grade 4, and approximately 231,500 (or 98%) of the total number of students sampled for grade 8. The original sample size was increased somewhat after the supplemental samples had been drawn (from students newly enrolled since the creation of the original list of students).

When queried, the quality control monitors felt most positive about the attitudes of the assessment administrators and somewhat less positive about the attitudes of other school staff and the students toward the assessment. The QC monitors' evaluations, impressions, and observations are recorded in the QC monitoring form provided to them for each school.

Quality control monitors concluded the summary section of their QC monitoring form by assigning a final rating of the assessment administrator's performance. With this rating, the quality control monitor reconsidered the session from the vantage point of how well it would have gone without the quality control monitor's presence. Eighty-four percent of the assessment administrators in monitored sessions were self-reliant or needed to consult the quality control monitors for only one or two minor items. Between four and five percent cited serious difficulty conducting the session (that is, relied on the quality control monitor to initiate procedures or conduct the session).

Table 5-3
*School Participation, 1998 State Assessment**

	Grade 4 Reading		Grade 8 Reading		Grade 8 Writing	
	Public	**Nonpublic**	**Public**	**Nonpublic**	**Public**	**Nonpublic**
Schools in original sample	4,594	570	3,805	453	3,688	450
Schools not eligible (closed or no sampled grade)	73	68	85	71	93	65
Eligible schools in original sample	4,521	502	3,720	382	3,595	385
Noncooperating†	440	131	397	90	362	107
Cooperating	4,081	371	3,323	292	3,233	278
Participating substitutes for noncooperating schools	125	27	84	8	86	11
Total of schools participating (after substitution)	4,206	398	3,407	300	3,319	289

* Corresponding data for national NAEP schools are provided in Chapter 3 of this report.
† e.g., school, district, or state refusal

Table 5-4

*Student Participation, 1998 State Assessment**

	GRADE 4 READING		GRADE 8 READING		GRADE 8 WRITING	
	Public	**Nonpublic**	**Public**	**Nonpublic**	**Public**	**Nonpublic**
Number Sampled	130,230	8,621	113,789	5,922	111,535	5,939
Original Sample	126,414	8,551	110,995	5,880	108,728	5,897
Supplemental Sample	3,816	70	2,794	42	2,807	42
Percent Increase in Original Sample	3.0%	0.8%	2.5%	0.7%	2.6%	0.7%
Number of Originally Sampled Students Withdrawn	5,628	88	5,357	57	5,347	63
Percent of Originally Sampled Students Withdrawn	4.4%	1.0%	4.8%	1.0%	4.9%	1.1%
Number of Students Excluded†	9,186	64	6,068	43	4,872	27
Number of Sampled Students Identified as SD	15,040	210	12,750	157	12,342	159
Percent of Sampled Students Identified as SD	11.5%	2.4%	11.2%	2.7%	11.1%	2.7%
Number of Sampled Students Excluded as SD	7,181	54	5,039	27	3,898	13
Percent of Sampled Students Excluded as SD	5.5%	0.6%	4.4%	0.5%	3.5%	0.2%
Number of Sampled Students Identified as LEP	5,514	53	3,338	64	3,329	63
Percent of Sampled Students Identified as LEP	4.2%	0.6%	2.9%	1.1%	3.0%	1.1%
Number of Sampled Students Excluded as LEP	2,406	13	1,260	19	1,187	15
Percent of Sampled Students Excluded as LEP	1.8%	0.2%	1.1%	0.3%	1.1%	0.3%
Number of Students To Be assessed	115,416	8,469	102,364	5,822	101,316	5,849
Number of Students Assessed	109,149	8,101	93,229	5,554	91,998	5,593
Original Sessions	108,145	8,020	91,614	5,511	90,410	5,557
Makeup Sessions	1,004	81	1,615	43	1,588	36
Student Participation Rates – Before Makeups	93.7%	94.7%	89.5%	94.7%	89.2%	95.0%
Student Participation Rates – After Makeups	94.6%	95.7%	91.1%	95.4%	90.8%	95.6%

* Corresponding data for national NAEP schools are provided in Chapter 3 of this report.

† To be excluded, a student had to be designated as SD or LEP and judged incapable of participating in the assessment. A student could be identified as both SD and LEP, resulting in this number being less than the sum of the students excluded as SD or LEP.

5.6.2 Results of the Observations

During the state NAEP assessment sessions, the quality control (QC) monitors observed whether the assessment environment was adequate or inadequate based on factors such as room size, seating arrangements, noise from hallways or adjacent rooms, and lighting. (If the room was unsuitable, however, the quality control monitors did not routinely ask the assessment administrator to make other arrangements.) Of the approximately 3,300 monitored assessment sessions, the quality control monitors felt that at least 96 percent of the sessions were held in suitable surroundings. This evaluation of the assessment environment is recorded in the QC monitoring form provided to them for each school observed, that is, the QC monitors' observations are recorded systematically in the pre-printed form during their observations of the sessions.

The Manual for Assessment Administrators encouraged assessment administrators to use an assistant during the assessment session, a suggestion that came from the earliest state assessment in 1990. To measure how frequently that advice was heeded, quality control monitors noted whether an assistant was used in the monitored sessions. The results indicate that assistants were used for about 52 percent of the public-school sessions. In nonpublic schools, however, an assistant was employed less often (19–29% of the time), which is possibly a reflection of fewer staff resources and generally smaller session sizes in nonpublic schools; the largest occurrence of assistants in public schools (29%) was at grade 4. Assessment administrators used assistants in varying capacities. The Manual for Assessment Administrators was very emphatic that only a NAEP-trained person could actually administer the assessment session. In most cases, assistants helped to supervise the session and to prepare, distribute, and collect assessment materials and booklets.

The assessment administrators were asked to estimate the total time that they spent on the preparations for and the conduct of the assessment, including their attendance at the training session. Estimates for 1998 were similar to those for previous years. In 1998, a majority of the assessment administrators with grade 4 sessions (73% in public schools and 90% in nonpublic schools) stated that they spent less than 20 hours on the assessment. For grade 8, however, only 40 percent of the assessment administrators in public schools, compared to 88 percent of those in nonpublic schools, spent fewer than 20 hours. The variation in time distribution for grade 8 public schools, particularly compared to public schools at grade 4, is most likely due to the fact that two session types (reading and writing) were usually conducted in each grade 8 school for state NAEP, but only one session type (reading) was held at grade 4. This does not appear to hold true for nonpublic schools, however, where the distribution of time spent is more similar for grades 4 and 8. It is evident that assessment administrators in nonpublic schools spent fewer hours overall on the assessment than did assessment administrators in public schools. Potential explanations might be the generally smaller sessions sizes in nonpublic schools (i.e., fewer materials to prepare and ship) and the possibility that some grade 8 schools may have used more than one assessment administrator, with each assessment administrator conducting one session (but compiling a larger total time for all sessions combined).

Quality control monitors observed that assessment booklet bundles were opened at the proper time in about 98 percent of sessions. In a few sessions, however, the bundle opening was not observed due to quality control monitor error (e.g., the quality control monitor was late, in the wrong place, or miscommunicated with the assessment administrator); presumably, some (or probably most) of these bundles were opened at the correct time. For a few other sessions, the quality control monitors were unable to observe the bundle opening that occurred early due to assessment administrator error (e.g., the assessment administrator misunderstood the procedures, felt more time was needed, had scheduling conflicts, or needed to prepare for multiple sessions starting at the same time).

After the conclusion of the state NAEP assessment sessions, Westat mailed state coordinators a short survey to obtain their reactions to the operations associated with the 1998 state assessment and any

suggestions they had for improving the program. Thirty-one of the forty-four state coordinators who were mailed the survey (or about 70 percent) responded by returning the survey or by providing their responses over the telephone. A detailed summary of the state coordinators' responses is contained in the *Report on Data Collection Activities for All States* (Westat, 1998), which was distributed to state coordinators in October 1998. Some of the responses from the state coordinators included:

- Eleven of the 31 reporting jurisdictions mandated participation in the 1998 state assessment.
- Only two jurisdictions reported that they helped gain the cooperation of nonpublic schools. One had success contacting parochial schools, but requested assistance from NAEP staff for recruiting other nonpublic schools. Most coordinators preferred that NAEP staff contact the nonpublic schools.
- All 31 jurisdictions responding (of the 44 jurisdictions sampled) used the computer system during the field period. Five jurisdictions used the system initially but not necessarily during the entire assessment period. The jurisdictions seemed to be comfortable with the computer system and were able to use it effectively. Typically, the reason for discontinuing use of the computer was that coordinators had completed their data-entry tasks and had turned responsibility back to the state supervisor who was coordinating requests for assessment date changes.
- Of the jurisdictions reporting on staff time devoted to NAEP, state coordinators spent an average of 28 days on NAEP activities, and in addition, other staff spent an average of 25 days.
- Reactions to the 1998 state assessment were quite positive. Most of the state coordinators who expressed an opinion said that the assessments went "very well" or "well"—with very few problems.

5.7 FIELD MANAGEMENT

Two field managers monitored the work of about 25 scheduling supervisors who worked during fall 1997 to gain cooperation of districts and schools for the national assessment. During the national assessment period, these staff were expanded to about 80 supervisors and 5 field managers. All supervisors reported directly to their field managers who, in turn, reported to Westat's field director. These contacts were made at least weekly.

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

These reports were an invaluable tool for the sampling statisticians as well as for the field director and field management staff. They provided the statisticians with the information needed to determine whether or not the response rates were high enough for the sample of schools to produce

representative results. Based on the information contained in these reports, the sampling statisticians selected substitute schools to replace some of the noncooperating schools.

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

Progress of the national NAEP assessments was constantly monitored through telephone reports held between NAEP supervisors, field managers, and home office staff. During these phone conversations, the supervisors' schedules were reviewed and updated, and any problems that the supervisors were experiencing were discussed. Progress of the fieldwork was also monitored during quality control visits made to the field by Westat and ETS office staff.

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

Chapter 6

PROCESSING ASSESSMENT MATERIALS[1]

Connie Smith, Charles Brungardt, and Timothy Robinson
National Computer Systems

6.1 INTRODUCTION

In the spring of 1998, the National Assessment of Educational Progress (NAEP) assessed students in reading, writing, and civics at grades 4, 8, and 12 at the national level. At the state level, reading was assessed at grades 4 and 8, and writing was assessed at grade 8 only. Civics was not assessed at the state level. National Computer Systems (NCS), under subcontract to Educational Testing Service (ETS), completed the following activities related to test-materials processing for both the national and state components of the 1998 assessment:

- Printing of test booklets and questionnaires
- Materials packaging and distribution
- Receipt control
- Data capture through image and optical mark recognition scanning
- Data editing and validation
- Performance scoring of constructed-response (open-ended) items
- Data file creation
- Inventory control and materials storage

NCS received and processed a total of 447,377 assessed student booklets and 113,676 questionnaires for the three grades and subjects assessed. A total of 4,272,139 readings of student constructed responses were conducted via image-based on-line scoring. This allowed for item-by-item scoring and on-line, real-time monitoring of both interrater reliabilities and the performance of each individual reader. Session and booklet information for the 1998 national and state assessments is given in Table 6-1. Table 6-2 provides information on questionnaires expected, received, and processed. Further detail is provided in NCS's *1998 NAEP Assessment Report of Processing and Professional Scoring Activities (*National Computer Systems, 1998).

6.2 PRINTING

For the 1998 assessments, 284 unique documents were designed. NCS printed more than 1,500,000 booklets and forms, totaling more than 60 million pages. This was a collaborative effort involving staff from ETS, Westat, and NCS. ETS created camera-ready blocks using NCS's DesignExpert™ software for the test booklets and questionnaires. Using ETS's booklet maps, which specified the order of blocks in each booklet, NCS assembled electronic components into complete

[1] Connie Smith was the NCS project manager for 1998 NAEP, Charles Brungardt was the NCS project director for 1998 NAEP scoring, and Timothy Robinson was the NCS senior processing coordinator for 1998 NAEP.

booklets. NCS then forwarded proofs to ETS, while conducting simultaneous quality control itself. Upon approval, final-form test booklets and questionnaires were produced and accounted for in the NCS inventory control system.

Table 6-1
Number of Sessions and Student Booklets Processed for the 1998 National and State Assessments

	Grade	Session Type	Number of Sessions	Assessed Booklets	Absent Booklets	Excluded Booklets
National						
	4					
		Reading	470	8,280	330	924
		Writing	1,519	25,816	1,317	1,880
		Civics	116	2,088	98	180
		Total	**2,105**	**36,184**	**1,745**	**2,984**
	8					
		Reading	623	11,970	937	977
		Writing	1,925	34,858	2,827	1,508
		Civics	114	2,055	161	96
		Total	**2,662**	**48,833**	**3,925**	**2,581**
	12					
		Reading	694	13,417	3,393	729
		Writing	1,769	33,106	8,373	1,207
		Civics	114	2,193	500	100
		Total	**2,577**	**48,716**	**12,266**	**2,100**
State						
	4					
		Reading	4,915	117,237	6,363	9,317
		Total	**4,915**	**117,237**	**6,363**	**9,317**
	8					
		Reading	4,389	98,776	9,236	6,176
		Writing	4,375	97,603	9,338	97,603
		Total	**8,764**	**196,479**	**18,574**	**103,799**

Table 6-2
Questionnaire Totals for the 1998 NAEP Assessment

		Expected	Received	Percent
National				
	Grade 4			
	Language Arts/Civics Teacher Questionnaire	2,145	2,081	97.0%
	School Questionnaire	731	700	95.8%
	SD/LEP Questionnaire	7,066	7	96.7%
	Grade 8			
	Language Arts Teacher Questionnaire	2,303	2,170	94.2%
	Civics Teacher Questionnaire	1,594	1,489	93.4%
	School Questionnaire	753	722	95.9%
	SD/LEP Questionnaire	7,942	7,575	95.4%
	Grade 12			
	School Questionnaire	599	570	95.2%
	SD/LEP Questionnaire	6,588	6,214	94.3%
State				
	Grade 4			
	Language Arts Teacher Questionnaire	16,597	16,339	98.4%
	School Questionnaire	4,593	4,550	99.1%
	SD/LEP Questionnaire	18,711	18,310	97.8%
	Grade 8			
	Language Arts Teacher Questionnaire	14,854	14,370	96.7%
	School Questionnaire	3,935	3,858	98.0%
	SD/LEP Questionnaire	28,515	27,798	97.5%

6.3 PACKAGING AND DISTRIBUTION

The distribution effort for the 1998 NAEP assessment involved packaging and mailing documents and associated forms and materials to the Westat supervisors for the national assessment and to individual schools for the state assessment. The NCS materials distribution system (MDS) was utilized again in 1998. Files in the MDS system contained shipping addresses, scheduled assessment dates, and a listing of all materials available for use by a participant in a particular subject area. Changes to any of this information were made directly in the MDS file either manually or via file updates provided by Westat.

Bar code technology continued to be utilized in document control, as has been done since the 1990 NAEP assessment. NCS identified each document with a unique 10-digit identification number. This number consisted of the 3-digit booklet number or form type, a 6-digit sequential number, and a check digit. Each form was assigned a range of identification numbers. Bar codes reflecting this identification number were applied to the front covers of documents by NCS bar code processes and high-speed ink-jet printers.

Spiraling of the NAEP booklets was done according to the pattern specified by ETS (see Section 1.5) to capture the sample size needed for each subject per grade. One booklet type from each grade and subject was designated as an accommodation booklet. These booklets were grouped in bundles of three.

Using sampling files provided by Westat, NCS assigned bundles to sessions and customized the packing lists. File data was coupled with the file of bundle numbers and the corresponding booklet numbers. This file was then used to preprint all booklet identification numbers, school name, school number, and session type directly onto the scannable administration schedule. This increased the quality level of the booklet accountability system by enabling NCS to identify where any booklet should be at any time during the assessments. To assist Westat supervisors with sampling in the schools, NCS distributed the preprinted administration schedules and questionnaires for the national assessment in December 1997. Preprinted administration schedules for the state assessment were sent to the appropriate state supervisor for distribution during training of the assessment administrators in January and February 1998.

NCS was also responsible for packaging and distributing bulk and session materials to Westat supervisors for the national assessment. Bulk shipments included materials that could be used by supervisors from one session to another, such as ancillary items and additional booklets.

Distribution of materials for the national assessment was accomplished in two phases. In the first phase, bulk supplies of materials were distributed to each supervisor. The second phase was the distribution of session specific materials by supervisor region and primary sampling unit (PSU). Each session box of materials contained the assigned bundles of booklets and the appropriate ancillary items. For additional materials, Westat supervisors were instructed to contact NCS using the NAEP toll-free line or the NAEP e-mail address.

Session materials were sent to individual schools in the NAEP state assessment. Distribution of materials was accomplished in five waves of shipment dates. Except for wave “zero,” session materials were sent to schools two weeks before their scheduled assessment date. All school materials were sent directly to an assessment administrator at a school or school district. Materials for Hawaii, Virgin Islands, and DoDEA/DoDDS (Department of Defense Education Activity’s Department of Defense Dependents Schools) were distributed in wave “zero”. These shipments required an alternate carrier to ensure timely delivery.

Initially, 6,933 individual sessions were shipped to 3,814 schools for the national assessment. For the state assessment, 13,586 sessions were mailed to 12,253 schools. Approximately 450 additional shipments of booklets and miscellaneous materials were also sent out for the national assessment and 3,000 for the state assessment.

To request additional materials for the 1998 NAEP assessment, Westat supervisors used either the NCS/NAEP toll-free telephone number or the NCS/NAEP e-mail address. After all the appropriate information had been entered, the system produced a packing list and mailing labels for NCS’s packaging staff, who filled and sent the order.

State assessment administrators (AAs) were given two options also, a toll-free telephone number or a toll-free fax number. This year NCS created a materials request form and included it in the school shipment to be used either as a guide for ordering materials over the phone or as a fax order form. A form was created for each grade and great care was taken to group items by session type to simplify the process for the AAs.

NCS clerical staff also responded to calls or e-mail concerning shipment delivery dates, lost shipments, and general questions concerning the NAEP assessment.

6.4 PROCESSING

NCS staff created a set of predetermined rules and specifications that was to be followed by the processing departments within NCS. Project staff performed a variety of procedures on materials received from the assessment administrators before releasing these materials into the NCS/NAEP processing system. Control systems were used to monitor all NAEP materials returned from the field. The NAEP Process Control System (PCS) contained the status of sampled schools for all sessions and their scheduled assessment dates. As materials were returned, the PCS was updated to indicate receipt dates, to record counts of materials returned, and to document any problems discovered in the shipments. As documents were processed, the system was updated to reflect processed counts. NCS report programs were utilized to allow ETS, Westat, and NCS staff to monitor progress in the receipt control operations. An alerts process was utilized to record, monitor, and categorize all discrepant or problematic situations. Throughout the processing cycle, alert situations were identified based on the processing specifications.

NCS's Work Flow Management system (WFM) was used to track batches of student booklets through each processing step, allowing project staff to monitor the status of all work in progress. It was also used by NCS to analyze the current work load, by project, across all workstations. Through routine monitoring of this data, NCS's management staff was able to assign priorities to various components of the work and to monitor all phases of the data receipt and processing.

6.4.1 Document Receipt and Opening

Shipments were to be returned to NCS packaged in their original boxes. The bar-coded label applied during the distribution phase containing the NAEP school identification number was scanned into a personal computer (PC) file upon receipt. The PC file was then transferred to the mainframe, and the shipment receipt date was applied to the appropriate school within the PCS system. This provided the status of receipts regardless of any processing delays. Each receipt was reflected on the PCS status report provided to the NCS receiving department and supplied to Westat weekly via electronic file transfer and in hard-copy format. ETS also received a hard copy. The PCS file could be manually updated to reflect changes. The shipment was then forwarded to the opening area.

Opening personnel checked the shipment to verify that the contents of the box matched the school and session indicated on the label. Each shipment was checked for completeness and accuracy. Any shipment not received within three days of the scheduled assessment date was flagged in the PCS system and annotated on the PCS report. The administration status of these delayed shipments was checked, and in some cases a trace was initiated on the shipment.

NCS was required to open all shipments within 48 hours of their receipt and to key-enter preliminary processing information into the PCS system from the administration schedule. The preliminary information was written on the administration schedule by Westat assessment administrators and consisted of the following:

- School number
- Session number
- Original test date
- Total number of students to be assessed
- Total number of students assessed
- Completeness flag

This preliminary information, used to provide Westat with timely student response rates, was updated with actual data when materials passed error-free through processing. The shipment was checked by NCS opening staff to see if any part of the shipment was missing, held for makeup, not administered, or refused. The shipment was also checked to verify that all booklets whose numbers were preprinted or handwritten on the administration schedule were returned with the shipment and that all administration codes matched from booklet cover to the administration schedule.

For all makeup sessions and for any missing materials not returned, the documents were placed on holding carts until the other documents arrived. These sessions were flagged on the PCS system and Westat was informed of this information. If the materials were not being returned, processing continued and the appropriate administration code was applied to the administration schedule. All questionnaires received were matched against the roster of questionnaires, which was a checklist of all types of questionnaires used in the assessment.

6.4.2 Batching of Booklets

Once all student booklets listed on the administration schedule for a session were verified as being present, the entire session (both the administration schedule and booklets) was forwarded to the batching administration area. Booklet batches were created by grade level, subject area, and session type. Each batch was assigned a unique batch number. This number, created on the Image Capture Environment (ICE) system for all image-scannable documents, facilitated the internal tracking of the batches and allowed departmental resource planning. All other scannable documents—school questionnaires, teacher questionnaires, SD/LEP (students with disabilities/limited English proficient) questionnaires, and the roster—were batched by document type in the same manner.

6.4.3 Scanning of Documents

The 1998 NAEP assessment used four rosters—one for each grade and one supplemental SD/LEP roster—to account for all questionnaires. Rosters of questionnaires were used to record the distribution and return of SD/LEP questionnaires, teacher questionnaires, and school questionnaires. Batches of school questionnaires and rosters, which are image scannable documents, were created on the ICE system. Batches of teacher and SD/LEP questionnaires, image scannable for the first time in the 1998 NAEP cycle, were also created on the ICE system. Batches were then forwarded to scanning, where all information on the rosters or questionnaires was scanned into the system.

6.4.4 Data Transcription

The transcription of the student response data into machine-readable form was achieved through the use of the following two systems: data entry (image scanning, intelligent character recognition [ICR], and key entry), and data validation (edit). NCS used the same format as in prior NAEP assessments and field tests to set up the document definition files for the number of unique documents used in the 1998 assessment. To do the proper edits, a detailed document definition procedure was designed to allow NCS to define an item once and use it in many blocks and to define a block once and use it in many documents.

6.4.4.1 Data Entry

The data-entry process was the first point at which booklet-level data were directly available to the computer system. Depending on the NAEP document, one of three methods was used to transcribe NAEP data to a computerized form. The gridded data on scannable documents were collected using NCS optical-scanning equipment, which also captured images of the constructed-response (open-ended) items and ICR fields in a single pass.

Optical Mark Recognition (OMR) Scanning. The data values were captured from the booklet covers and administration schedules and were coded as numeric data. Unmarked fields were coded as blanks and editing staff were alerted to missing or uncoded critical data. Fields that had multiple marks were coded as asterisks (*). The data values for the item responses and scores were returned as numeric codes. The multiple-choice single-response format items were assigned codes depending on the position of the response alternative; that is, the first choice was assigned the code "1," the second "2," and so forth. The mark-all-that-apply items were given as many data fields as response alternatives; the marked choices were coded as "1," while the unmarked choices were recorded as blanks.

Image Scanning. The images of constructed-response (open-ended) items were saved as a digitized computer file. The area of the page that needed to be saved was defined prior to scanning through the document definition process. The fields from unreadable pages were coded "X" as a flag for resolution staff to correct. Any image document or sheet unreadable by the image scanning system was taken to a flat-bed scanner to be scanned into the system. In addition to capturing the student responses, the bar code identification numbers used to maintain process control were decoded and transcribed to the NAEP computerized data file.

Intelligent Character Recognition. The intelligent character recognition (ICR) engine was again utilized to read various hand and machine printing on the front cover of the booklet and supervisor documents for the 1998 assessment. Some information from student documents, administration schedule, roster of questionnaires, and some questions in the school questionnaires, were read by the ICR engine and verified by an on-line key-entry operator. In all, the ICR engine read 1,994,416 characters for the 1998 assessment. Use of the ICR engine saved NAEP field staff a significant amount of time, since they did not have to grid rows and columns of data.

In all three cases, the data were edited, and suspect cases were resolved before further processing.

6.4.4.2 Data Validation

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

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

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

When the entire document was processed, the completed string of data was written to the data file. When all the documents in the batch were processed, the program generated an edit listing for nonimage and key-entered documents. Image-scanned items that required correction were displayed at an on-line editing terminal.

For rapid resolution, the edit criteria for each item in question appeared on the screen along with the suspect item. Corrections were made immediately. The system employed an edit/verify system that ultimately meant two different people viewed the same suspect data and operated on it separately. The verifier made sure the two responses (one from either the entry operator or the ICR engine) were the same before the system accepted that item as being correct. If the editor could not determine the appropriate response, he or she escalated the suspect situation to a supervisor. For errors or suspect information that could not be resolved by supervisory staff, a product-line queue was created, allowing supervisors in the processing area to escalate edits to project staff for resolution.

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

With the satisfactory conclusion of the count-verification phase, the edited batch file was uploaded to the mainframe, where it went through yet another edit process. A paper edit log was produced and, if errors remained, was forwarded to another editor. When this edit was satisfied, the PCS and WFM tracking systems were updated.

The teacher and SD/LEP questionnaires were edited on paper. Machine edits performed during data capture verified that each sheet of each document was present and that each field had an appropriate value.

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

Suspect data that were investigated during the edit phase consisted of, but were not limited to, the following by document types:

Administration Schedule

a) Verification that all assessed student booklets are present in a processed batch: If an administration code of 10-14, 20-24, or 71-79 was present on the administration schedule, the editor verified that a booklet was present. If the booklet was missing, the booklet was located and processed before the batch can continue to be processed.

b) Verification that the booklet bar code number was valid: NAEP booklet bar code numbers for the 1998 assessment were 10-digits long and fell within a certain range of numbers by grade. If, on a hand-written administration schedule, the booklet bar code written was less than 10-digits or out of range for the grade being processed, NAEP project staff corrected the bar code number as appropriate to match the booklet being processed.

c) Verification that the School number was valid: If the school number was blank or not on the PCS file, the school number was corrected by NAEP project staff.

Student Booklets

a) Investigating suspect bar codes, duplicate bar codes, or invalid check-digits: If the bar code number was read incorrectly by the scanner, the bar code was corrected to match the bar code on the booklet in question.

b) Investigating suspected absent students: If a booklet had an administration code indicating an assessed student, yet no multiple-choice responses were read by the scanning equipment, the editor manually checked the booklet for any multiple-choice responses. If a student had penciled in his or her multiple-choice responses too lightly for the scanners to read, the editor key entered the responses into the student data record. If no multiple-choice responses were present, but open-ended responses were, the booklet was sent through processing unchanged. If no multiple-choice or open-ended responses were present, the administration code was changed to indicate that there were no responses in the booklet, and the booklet was sent through processing with the updated administration code.

c) Investigating responses within the valid range: An example of a range check would be verifying that the birth month of the respondent falls with the range of 01-12. If the birth month is not within the valid range and a correct birth month can be determined from either the administration schedule or booklet cover, the birth month is corrected. If a valid response cannot be determined, the birth month is blanked out. The same type of range check is done for the birth year when specific years are valid by grade.

A computerized edit list, produced after NAEP documents were scanned, and all the supporting documentation sent from the field were used to perform the first phase of the edit function. The hard-copy edit list contained all the vital statistics about the batch: number of students, school code, type of document, assessment code, suspect cases, and record serial numbers. Using the information, the data editor verified that the batch had been assembled correctly and that each school number was correct. During data entry, counts of processed documents were generated by type. These counts were compared against the information captured during scanning.

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

(a) *Correctable error*: If the error was correctable by the editing staff according to the editing specifications, the correction was noted on the edit log for later correction via key entry.

(b) *Alert*: If an error was not correctable according to the specifications, an alert was issued to NAEP project staff for resolution. Once the correction information was obtained, the correction was noted on the edit log for key-entry correction.

(c) *Noncorrectable error*: If a suspected error was found to be correct as stated and no alteration was possible according to the source document and specifications, no corrective action was taken. The programs were tailored to allow this information to be accepted into the data record.

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

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

To provide another quality check on the image scanning and scoring system, NCS staff implemented a quality check process by creating a stamp with a valid score designated on it. Each unique document type scored via the image system had two quality assurance documents stamped with valid scores for the items present. The QA booklets were batched and processed together with student documents of the same type. During the process of scoring, valid score points could be changed or dropped due to revision in the scoring rubrics. NCS provided ETS with documentation as to what score points on these items were no longer valid. When an image quality assurance stamp was displayed to a reader that contained a score point that was no longer valid, the reader assigned the response a score point of OT (off-task).

NCS also produced various status reports. The Receipt Control Status Report was designed to track the receipt of material from the schools. It was sorted by school number and displayed the following information: participation status, scheduled administration date and the shipment receipt date. The comment field in this report showed any school for which a shipment had not been received within three days of the scheduled test date.

The Processing Status Report was divided into two sections. The first was sorted by school and grade within each assessment. The following preliminary data for each were entered from the administration schedule as the shipment was opened by the receiving department: school number, session code, test date, preliminary count date, preliminary to-be-assessed counts, preliminary total-assessed counts, and completeness flags. The actual to-be-assessed count, actual total-assessed count, actual withdrawn ineligible count, actual count date, actual number excluded, and actual absent count were entered programmatically following the completion of processing. The second section of the Processing Status Report sorted and totaled the various documents by form within each grade and assessment.

The PCS Exceptions Report listed all schools and sessions with discrepancies, that is, materials not returned within three days, school or session given a completeness flag. Once all discrepancies were resolved for a school, the school would be removed from the report.

NCS transmitted electronic files containing the above data to Westat weekly. Hard copy of the PCS Exception Report, Alerts, and Documents Processed Report were also sent to ETS and Westat weekly.

6.5 DATA TRANSMISSION BEFORE SCORING

Delivery of data to the scoring center was accomplished via T1 transmission lines that linked the mainframe computers and the NAEP servers at the document-scanning site in the NCS main facility with the scoring servers that were dedicated to distributing work to the professional readers at the scoring center. The actual task of scheduling items for downloading was accomplished using a code written by the Image Software Development team. This code enabled the person scheduling the download to choose a team of readers and select the scheduled items from a list of all items that the team would be scoring throughout the scoring project. This process was repeated for all teams of readers until all anticipated work was scheduled.

6.6 CLASSROOM-BASED WRITING STUDY

Approximately 200 schools participating in the national writing assessments also conducted the Classroom-Based Writing Study. This study involved collecting two examples of student writing from an intact classroom at the selected schools. Participating students were also asked to complete a brief questionnaire of the assignment for which the writing samples were written. Teachers of participating classes were interviewed and an audiotape of the interview was shipped to NCS for transcription. Details of this study will be published in a forthcoming NAEP report.

Chapter 7

PROFESSIONAL SCORING[1]

Connie Smith, Charles Brungardt, and Timothy Robinson
National Computer Systems

7.1 OVERVIEW

The 1998 NAEP assessment required the scoring of constructed responses in reading, writing, and civics at grades 4, 8, and 12 on the national level. At the state level, constructed responses were scored at grades 4 and 8 for reading and grade 8 for writing. All preparations were completed and scoring accomplished on a schedule that allowed faster reporting and delivery of data than in previous years. Also, to measure longitudinal trends in reading, the project required National Computer Systems (NCS) to replicate scoring from the 1994 NAEP reading assessment for most of the reading items and to demonstrate that scoring of this subject was statistically comparable across years.

To accomplish the task of scoring the constructed responses, NCS's Performance Assessment Scoring Center (PSC) employed more than 300 professional and 82 clerical scorers on a two-shift schedule. The professional scorer is required to have, at a minimum, a baccalaureate degree from a four-year college or university; an advanced degree, scoring experience, and/or teaching experience is preferred. The clerical scorers, who coded the pre-writing exercise, have at least a high school diploma. NCS worked with Educational Testing Service (ETS) to prepare training materials and carry out the training of the scoring teams. Table 7-1 lists the processing and scoring totals for each subject and grade.

Table 7-1
Processing and Scoring Totals for the 1998 NAEP Assessment

	Booklets Processed	Number of Constructed Responses*	Number of Discrete Response Items†	Number of Scorers and Team Leaders‡	Dates of Training and Scoring
Total	447,961	3,770,952	335	—	—
National & State Grade 4 Reading	125,517	1,535,479	46	160 / 16	3/23/98 – 4/24/98
National & State Grade 8 Reading	110,746	1,470,932	69	100 / 10	3/23/98 – 4/24/98
National Grade 12 Reading	13,431	195,444	76**	40 / 4	3/23/98 – 4/24/98
National Grade 4 Writing	19,937	49,347	20	30 / 3	4/28/98 – 7/1/98
National & State Grade 8 Writing	124,346	268,238	23	129 / 12	4/28/98 – 7/1/98
National Grade 12 Writing	25,433	55,695	23	30 / 3	4/28/98 – 7/1/98
National Grade 4 Civics	8,087	52,454	21	27 / 3	4/27/98 – 5/11/98
National Grade 8 Civics	10,337	72,450	28	27 / 3	4/27/98 – 5/11/98
National Grade 12 Civics	10,031	70,913	29	36 / 4	4/27/98 – 5/11/98

* This is the number of student responses to the constructed-response items. These scored responses include those that were rescored for reliability estimation.
† This is the number of discrete constructed-response items in assessment booklets.
‡ Because readers scored items from all grades and all types of booklets, it is not possible to break the numbers down by how many scored each classification of items.
** This included 75 image and 1 paper.

[1] Connie R. Smith was the NCS project manager for 1998 NAEP, Charles Brungardt was the NCS project director for 1998 NAEP scoring, and Timothy Robinson was the NCS senior processing coordinator for 1998 NAEP.

Figures 7-1 and 7-2 provide flowcharts for image scoring (see Section 7.4) and paper scoring (see Section 7.5). Further detail is provided in NCS's *1998 NAEP Assessment Report of Processing and Professional Scoring Activities* (National Computer Systems, 1998).

Figure 7-1
Image Scoring Flow Chart

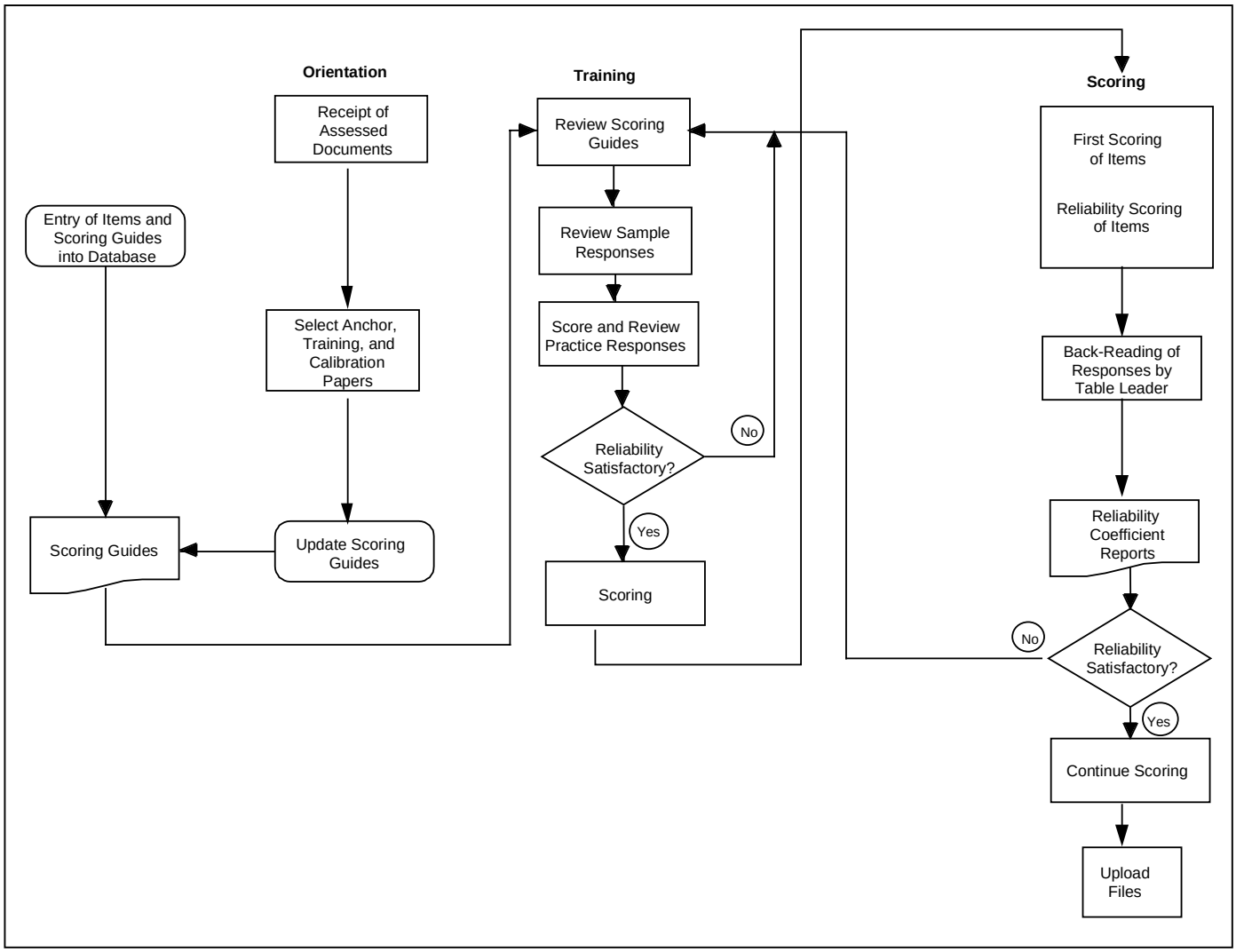
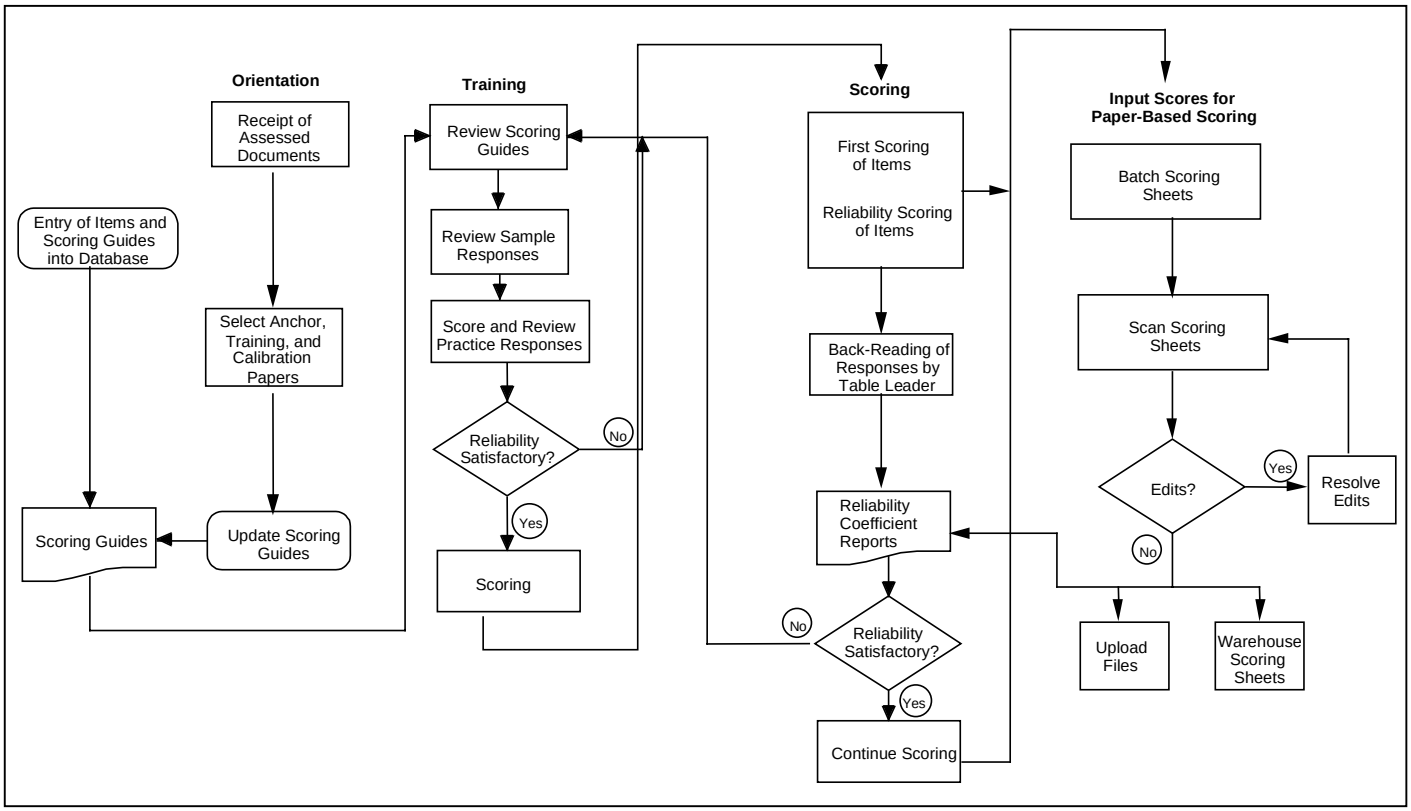


Figure 7-2
Paper Scoring Flow Chart



7.2 SELECTION OF TRAINING PAPERS

Clerical staff began the process of copying all responses for rangefinding and creation of anchor and training sets in November of 1997 by copying all the responses (approximately 400 per prompt) for the writing prompts that did not change wording or format between the field test and operational assessment. In January and February of 1998, the clerical staff copied more sample responses, including approximately 300 responses for each writing item that had undergone changes in wording or format, 200 responses for each writing item that remained the same since the field test, 200 responses for each new reading item, and 150 responses for each civics item. NCS clerical staff wrote the booklet identification numbers on each page of each response so that the training samples could be linked back to the identification numbers of the booklet they came from. They then sorted the papers by item and sent the samples to ETS for the rangefinding, while keeping the samples in Iowa City for those items to be reviewed at NCS.

Rangefinding[2] and creation of training sets took place at ETS for the three new reading blocks, all the writing prompts, and those civics blocks assigned to ETS staff for training. The process took place

[2] *Rangefinding* is the process of interpreting the scoring guide onto student responses. These scored responses are then used in the various training sets (i.e., anchor, practice, calibration, and qualification papers.)

in Iowa City for civics blocks assigned to NCS trainers. After review by each subject's coordinator, ETS returned the training sets to NCS staff, who reproduced them for scoring. Correct scores were written on all the anchor papers, while only the table leaders and trainers had keys for the practice, calibration (see Section 7.4.3), and qualification sets. Trainers also kept annotations, explaining the thought process behind each score assigned. If any of these changed during training for scoring, the table leaders kept notes explaining the reason.

7.3 CALIBRATION POLICIES

When scoring was expected to last longer than a few hours (for example, items with a state sample), a calibration set was created to refresh the training and avoid scorer drift. Responses were chosen from the current sample (see Section 7.4.3). The table leader invoked the calibration tool in the backreading tool (see Section 7.4.2) to create calibration sets. In general, each team scored calibration sets whenever they took a break longer than 15 minutes, such as when returning from lunch.

7.4 IMAGE SCORING

During processing, images of the student responses to each of the constructed-response items were digitized, placed in an image archive, and grouped according to scoring purpose (e.g., grade 4 reading, grade 4 writing, and validity). Two of the significant advantages of the image-scoring system were the ease of regulating the flow of work to scorers and the ease of monitoring scoring. The image system provided table leaders with tools to determine scorer qualification, to backread scores, to determine scorer calibration, to monitor interrater reliability, and to gauge the rate at which scoring was being completed. These tools are described in Sections 7.4.1 through 7.4.10.

7.4.1 Reader Qualification

Teams used copies of paper sets to determine whether each individual scorer was sufficiently prepared to score. All extended items in reading and civics and all items in writing required scorers to qualify. Short items in reading and civics did not require special qualification. Once scorers demonstrated readiness for scoring, either through the trainer's perception during the training of short constructed-response items or the formal 80 percent correct on the qualification set for extended constructed-response items, the table leader used the qualification tool to route work to the team. To make sure that all scorers had a common understanding of the training, the teams usually gathered around one terminal at the beginning of scoring, read several papers aloud, and scored them as a group. Then the teams broke into pairs for scoring, followed by individual scoring.

7.4.2 Backreading Process

After scoring began, NCS table leaders reviewed each scorer's progress using a backreading utility that allowed the table leader to review papers scored by each scorer on the team. Typically, a table leader reviewed approximately 10 percent of all responses scored by each scorer. Table leaders made certain to note the score the scorer awarded each response as well as the score a second scorer gave that same paper. This was done as an interrater reliability check. Alternatively, a table leader could choose to review all responses given a particular score to determine if the team as a whole was scoring consistently. Both of these review methods used the same display screen and showed the identification number of the scorer and the scores awarded. If the table leader disagreed with the score given an item, he or she discussed it with the scorer for possible correction. This discussion was used as a training tool to ensure

that all scorers assigned the same score to similar responses. Whether or not the table leader agreed with the score, he or she assigned a table-leader score in backreading. If this score agreed with the first score, the score was recorded only for statistical purposes. If the scores disagreed, then the table-leader score overrode the first score as the reported score.

7.4.3 Calibration Process

During backreading, the table leader had a pool of 300 responses for each item, which were available to use in the calibration process. The table leader viewed samples of these responses together with the scores assigned by the first and, if applicable, second scorer. From this pool, the table leader chose which responses to put into the pool that would be available to scorers during calibration, choosing responses that were scored correctly and were a good measure to keep scoring on track. From this pool, the table leader built sets with the desired number of responses, usually between 5 and 10, to be displayed to scorers for calibration. When the scorers invoked the calibration window, all scorers received the same responses and scored them. After scorers had finished scoring this pool, the table leader could look at reliability reports, which included only the data from the calibration set just run. Thus, this type of calibration served to refresh training and avoid drift in scoring. Because paper calibration sets from 1994 reading still existed, some reading teams used hard copies to calibrate scorers.

7.4.4 Short-Term Trend Rescoring

To measure comparability of this year's reading scoring to the scoring of the same items done in 1994, a minimum of 600 on-task responses per item from 1994 were scanned and loaded into the system with their scores from 1994 as the first score.

"On-task" responses generate scores of 1, 2, 3, 4, 5, 6, or 7. "Off-task" scores are received when the response

- is blank,
- is "I don't know,"
- is totally erased,
- contains only comments for the test developer or scorer, or
- contains other unelicited remarks, drawings, or both.

These responses were loaded into a separate computer application to keep the data separate from regular scoring. At staggered intervals during the scoring process, the table leader released items from the 1994 cycle for scorers to read and score. Since the 1994 scores were preloaded as first scores, this year's teams in effect scored 100 percent of them a second time. Typically, the table leaders released 100 responses after training was finished but before beginning the scoring of current-year responses. The table leader and trainers then looked at reliability reports and *t*-tests and performed backreading to gauge consistency with 1994 scoring and make adjustments in scoring where appropriate. The remainder of the responses were released in equal amounts when scoring was one-third finished, two-thirds finished, and 90 percent finished. Note that the time intervals between rescored sessions varied with the number of responses to be scored per item.

Cross-year reliability results for each constructed-response item used in both 1998 and 1994 are provided in Tables C-7 through C-12 in Appendix C.

7.4.5 Validity Sets Tool

In order to score a validity set, the table leader updated the scorers' qualification to the same item they were regularly scoring for the validity application. Then, when scorers opened the scoring window, they received the validity papers. Validity papers, student responses prescored by the trainer during rangefinding, were used to prevent reader drift over the course of scoring. All scorers were in effect second scoring against the preloaded first scores. Unlike calibration sets, where all scorers read the same responses, with the validity sets, each scorer received different responses. Since the validity papers were under a separate application, the reliability reports and *t*-tests and backreading were available independently of the regular scoring. Before the next time the validity sets were used, the table leader used a tool to reset the items to make them available for scoring again, and also reset the reliability statistics. They accomplished this by executing a command in the report menu that then prompted them for a topic name. When the system carried out this command, it reset scoring and statistics only for the batch involved in the validity process.

7.4.6 *t*-Tests

To perform a *t*-test, the table leader executed a command in the report window that prompted the table leader for the item, the application, and the cubicle to which the item was assigned. The system then displayed an analysis of the data, which could be printed. The test results were based only on responses for which both scores were on-task. The display showed number of scores compared, number of scores with exact agreement, percent of scores with exact agreement, mean of the preloaded scores, mean of the currently assigned scores, mean difference, variance of the mean difference, standard error of the mean difference, and the t value.

7.4.7 Procedure for Monitoring Interrater Reliability

During the scoring of an item or the scoring of a calibration set, table leaders monitored progress using interrater reliability. This was done using a computer display that functioned in either of two modes: (1) to display information of all first readings versus all second readings, or (2) to display all readings of an individual that were also scored by other scorers versus the scores assigned by those other scorers. The information was displayed as a matrix, with scores awarded during first readings displayed in rows and scores awarded during second readings displayed in columns for mode one and the individual's scores in rows and all other scorers in columns for mode two. In this format, instances of exact agreement fell along the diagonal of the matrix. For completeness, data in each cell of the matrix contained the number and percentage of cases of agreement (or disagreement). The display also contained information on the total number of second readings and the overall percentage of reliability on the item. Also, the computer program provided on demand a separate calculation for exact and adjacent agreement rates for each writing item. Since the interrater reliability reports were cumulative, a printed copy of the reliability of each item was made periodically and compared to previously generated reports. Scoring staff saved printed copies of all final reliability reports and archived them with the training sets.

7.4.8 Process for Monitoring Frequency Distribution of Scores

For each topic, table leaders could run a report that showed the frequency distribution of scores. The report displayed separate frequencies for first and second scores. For each score level, the report showed the number of responses as an integer and as a percentage of the total. The report could be updated and printed on demand.

7.4.9 Process for Monitoring the Rate of Scoring

The table leaders were able to monitor work flow for each item using a status tool that displayed the number of responses scored, the number of responses first-scored that still needed to be second-scored, the number of responses remaining to be first-scored, and the total number of responses remaining to be scored. This allowed the team leaders and performance assessment specialists to accurately monitor the rate of scoring and to estimate the time needed for completion of the various phases of scoring.

7.4.10 Scoring Buttons

To assign a score, scorers clicked the mouse over a button displayed in the scoring window. Since buttons included only valid score values, there was no editing for out-of-range scores.

7.5 PAPER SCORING

The 1998 NAEP assessment used paper scoring only for one item, the "tax form" item in grade 12 reading. The tax form items were packaged into sets of 20. The development staff printed score sheets with the identification numbers for the 20 books contained in each packet on a score sheet. Separate score sheets were printed for the responses selected for second scoring. As soon as the last student response on any score sheet was completed, the score sheets were collected and taken to a central clerical support area to be scanned on the NCS paper-based scoring system using OpScan 7 scanners. As each sheet was processed, the scanning system edited the incoming data against tables to ensure that all responses were scored with one and only one valid score, and that only raters who were qualified to score an item scored it. Any discrepancies (e.g., no score assigned, double gridding, out-of-range scores, or invalid scorer identification numbers) were flagged and resolved before the data from that sheet were accepted into the scoring system database. Interrater agreement reports were generated on demand.

All score data were stored on personal computers at NCS as the responses were scanned. When scoring was completed, the scanner operator ran a query to make sure that all score sheets were accounted for. Once all edits were corrected, the PC file was renamed and put into an export file, which automatically created the mainframe file. This file was then uploaded to the mainframe to be merged with the mainframe student files.

7.6 LARGE-PRINT BOOKS AND OTHER SPECIAL ACCOMMODATIONS

NCS's Performance Assessment Scoring Center (PSC) scored responses for a number of students whose special accommodations made the books nonscannable. These included large-print books as well as responses typed on a separate sheet of paper outside the booklet. Altogether, there were 37 such books for reading, 3 for civics, and 61 for writing.

Since the books were nonscannable, they were transported to the scoring center after processing. Clerical staff created a log to account for all the special accommodations books and a score sheet for each book listing the constructed-response items in that book. The books were routed to the table leaders in charge of the different items in each book. As the team scored an item, the table leader marked the score for that response, his or her scorer identification number, and the date scored. Once all items in each book for a given subject were scored, the scoring sheets were returned to development staff to enter those scores manually into the records for those books.

7.7 TRAINING

The training on each item was conducted by subject-area specialists from ETS and NCS. Dates for training and scoring can be found in Table 7-1. All of the assessments were scored item-by-item so that each scorer worked with only one set of rubrics at a time. After scoring all available responses, a team then proceeded with training and scoring the next item.

Training involved explaining the item and its scoring rubric to the team and discussing types of student responses that represented the various score points in the guide. Typically, two or three student responses were chosen to anchor each score point. When review of the anchor packet was completed, the scorers scored 10 to 20 "practice papers," previously scored by subject-area specifications that represented the entire range of score points the item could receive. The trainer then led the team in a discussion of the practice papers to focus the scorers on how the scoring rubrics should be interpreted. After the trainer and table leader determined that the team had reached consensus, the table leader then released work on the image-scoring system to the scorers. The scorers initially took turns reading aloud their first "live" responses to the team or worked in pairs as a final check before beginning work individually. Once the practice session was completed, the formal scoring process began.

During training, scorers and the table leader kept notes of scoring decisions. The table leader was then responsible for compiling those notes and ensuring that all scorers were in alignment with the decisions. Teams varied greatly in the amount of time spent scoring as a group before breaking into individual scoring. This time ranged from five minutes to five hours.

7.8 SCORING

All scoring for each item was conducted via computer image except for the grade 12 reading "tax form" item. During scoring, the table leaders continued to compile notes on scoring decisions for the scorers' reference and guidance. Additionally, table leaders closely monitored interrater reliability using both team and individual statistics as a reference. Consistently throughout the scoring of each item, the table leaders also performed backreading duties in which they reviewed a sample of the responses scored by each scorer on the team. The table leaders and performance assessment specialists continuously monitored the progress of each team and noted all scoring-related decisions to ensure that training and scoring progressed smoothly and in a timely manner.

7.9 INTERRATER RELIABILITY

A subsample of the reading, writing, and civics responses for each item were scored by a second scorer to obtain statistics on interrater reliability. In general, items administered only to the national sample received 25 percent second scoring, while those given in both the national and state samples received less. Thus, all civics items received 25 percent second scoring; all grade 12 reading received 25 percent second scoring; grades 4 and 8 reading items received 6 percent second scoring; grades 4 and 12 writing received 25 percent second scoring, and grade 8 writing items received 10 percent second scoring, except for the three 50-minute prompts, which received 25 percent second scoring because they were administered only in the national sample. The reliability information obtained from second scoring was also used by the team leaders to monitor the capabilities of all scorers and maintain uniformity of scoring across scorers. Reliability reports were generated on demand by the table leader, team leader, or performance assessment specialist as needed. They were displayed at a computer workstation. Printed copies were reviewed daily by both NCS and ETS lead scoring staff. In addition to the immediate feedback provided by the on-line reliability reports, each table leader could also review the actual responses scored by a scorer by using the backreading tool (see Section 7.4.2). In this way, the table

leader was able to monitor each scorer carefully and correct difficulties in scoring almost immediately with a high degree of efficiency. Table 7-2 provides the interrater reliability ranges.

Table 7-2
Interrater Reliability Ranges for the NAEP 1998 Assessment

		Total Number of Unique Items	Number and Percentage of Items in Percentage Exact Agreement Range							
			60–69%		70–79%		80–89%		Above 90%	
	Grade		Number	Percent	Number	Percent	Number	Percent	Number	Percent
Reading										
	4	46	—	—	3	6.5	16	34.7	27	58.6
	8	69	1	1.4	4	5.8	28	40.6	36	52.2
	12	76	1	1.3	4	5.2	36	47.4	35	46.1
Writing										
	4	20	4	20.0	16	80.0	—	—	—	—
	8	23	18	78.3	4	17.4	—	—	—	—
	12	23	10	43.5	9	39.1	3	13.0	—	—
Civics										
	4	21	—	—	3	14.3	11	52.4	7	33.3
	8	28	1	3.6	6	21.4	17	60.7	4	14.3
	12	29	—	—	8	27.6	20	70.0	1	3.4

Detailed results of interrater scoring reliability for the reading, writing, and civics constructed-response items are provided in Appendix C.

7.9.1 Scoring of Reading

The reading portion of the 1998 NAEP assessment included a total of 154 discrete constructed-response items. Four items were scored on an accelerated schedule between March 23 and 27. Scoring for the rest of the items took place between March 30 and April 24. The items scored included short-answer constructed responses and extended constructed responses. Each constructed-response item had a unique scoring rubric that identified the range of possible scores for the item and defined the criteria to be used in evaluating student responses. Note that these numerical values were for scoring only; they do not reflect the IRT-based scores used in analysis of the data. Chapter 15 describes the IRT values used in the data analysis.

During the course of the project, each team scored constructed-response items using a 2-, 3-, or 4-point scale as outlined below:

Dichotomous Items

1	=	unacceptable response
2, 3, or 4	=	acceptable response

(Items that originated in the 1992 NAEP used 1 and 4 for dichotomously scored items; items from the 1994 NAEP used 1 and 3; items developed in the 1997 field test used 1 and 2.)

Short Three-Point Items

1 = evidence of little or no comprehension
2 = evidence of partial or surface comprehension
3 = evidence of full comprehension

Extended Items

1 = unsatisfactory
2 = partial
3 = essential
4 = extensive

Table 7-3 lists the number of reading constructed-response items by item type and score-point level.

Table 7-3
Number of Constructed-Response Items by Score-Point Levels for the 1998 NAEP Reading Assessment

Item Type	Grade	2-Category	3-Category	4-Category	Total
Reading Items – Total					
	4	19	11	6	36
	4/8	8	—	2	10
	8	11	16	5	32
	8/12	13	9	5	27
	12	22	19	8	49
Reading Items – New in 1998					
	4	3	2	1	6
	4/8	—	—	—	—
	8	1	4	1	6
	8/12	2	4	1	7
	12	—	—	1*	1
Reading Items – Trend from 1994					
	4	16	9	5	30
	4/8	8	—	2	10
	8	10	12	4	26
	8/12	11	5	4	20
	12	22	19	7	48

* Even though the grade 12 tax form stimulus had been used in previous assessments, it is counted here as a new item, because no rescoring was done and it was not used to measure trend.

Note: "—" indicates that this category was not applicable.

7.9.2 Scoring of Writing

The writing portion of the 1998 NAEP assessment included a total of 66 discrete constructed-response items. Scoring was conducted from April 28 to July 1. The amount of space given students to respond ranged from four pages for the 25-minute prompts to eight pages for the 50-minute prompts. Trainers used generic holistic scoring guides for each grade that identified the range of possible scores for the item and defined the criteria to be used in evaluating student responses. Note that these numerical values were for scoring only; they do not reflect the IRT-based scores used in analysis of the data. Chapter 19 describes the IRT values used in the data analysis.

All writing scoring rubrics used a six-point scale as follows:

6	=	excellent response
5	=	skillful response
4	=	sufficient response
3	=	uneven response
2	=	insufficient response
1	=	inappropriate (grade 4) or unsatisfactory (grade 8 and 12) response

The IRT numerical values used in analysis of the data are described in Chapter 19. Table 7-4 lists the number of writing constructed-response items by item type and score-point level.

Table 7-4
Number of Constructed-Response Items by Score-Point Levels for the 1998 NAEP Writing Assessment

Item Type	Grade	6-Category	Total
Writing Items			
	4	20	20
	8	23	23
	12	23	23
Prewriting Items			
	4	20	20
	8	23	23
	12	23	23

7.9.2.1 Selective Rescoring

To address problems of low reliability at the upper-score levels, the ETS staff chose 13 prompts at grade 4, 9 at grade 8, and 8 at grade 12 to conduct a selective rescoring of responses. For each prompt involved in the selective rescoring, all responses that received either a first or second score of 5 or 6 were downloaded again to the scoring center. Specially selected trainers prepared additional training material focusing on the upper-level scores. One trainer did all of the grade 4 selectively rescored items with the team that the trainer had worked with throughout the project. Three trainers, each with a specially selected team of 10 scorers, prepared and carried out the rescoring for the grade 8 responses. One team rescored responses to narrative prompts, another rescored responses to informative prompts, and the third worked exclusively on persuasive prompts. At grade 12, one trainer and team rescored responses to six of the prompts, while another trainer and group rescored two. Scores of 5 and 6 from the original scoring were deleted from the active files, though copies were maintained to provide an audit trail. All frequency

distributions and interrater agreement reports attached to this report show the status of the items after the selective rescoring was finished.

7.9.2.2 Prewriting Coding

All students were given a blank page to use for prewriting planning. Codes were developed for the type of prewriting planning students did during the assessment. Prewriting coding took place during the evening shift from May 11 through 26, working 41/2 hours from 6:00 p.m. to 10:30 p.m. The first evening, the ETS writing coordinator trained the table leaders, who in turn trained their teams of clerical scorers the following evening.

The coders classified the prewriting strategies for all items using the same coding guide, anchor set, and practice papers. All coding was completed by May 26.

The codes used to classify prewriting were as follows:

1	=	rough draft
2	=	list
3	=	outline
4	=	diagram
6	=	picture
7	=	multiple

Note that when a response showed multiple prewriting strategies the different, specific strategies used by a student were not recorded by the coders. Also note that the code value of "5" was originally planned to indicate that the student used a table as a prewriting strategy. However, that category was eliminated before training began.

7.9.3 Scoring of Civics

The civics portion of the 1998 NAEP assessment included a total of 78 discrete constructed-response items. It was scored from April 27 to May 11 on an evening shift that ran from 6:00 p.m. to 10:30 p.m. The items scored included short-answer constructed responses and extended constructed responses. Each constructed-response item had a unique scoring rubric that identified the range of possible scores for the item and defined the criteria to be used in evaluating student responses.

During the course of the scoring, each team scored constructed-response items using a 3- or 4-point scale as outlined below:

Short Item

1	=	unacceptable
2	=	partial
3	=	acceptable

Extended Items

1	=	unacceptable
2	=	partial
3	=	acceptable
4	=	complete

The IRT numerical values used in analysis of the data are described in Chapter 23. Table 7-5 lists the number of constructed-response items by item type and score-point level.

Table 7-5
Number of Constructed-Response Items by Score-Point Levels for the 1998 NAEP Civics Assessment

Item Type	Grade	3-Category	4-Category	Total
Civics Items				
	4	15	6	21
	8	22	6	28
	12	23	6	29

7.10 PREPARATION FOR TAPE CREATION

The 1998 NAEP assessment data collection resulted in several classes of data files—student, school, teacher, SD/LEP student, student/teacher match, and student-response information. Student-response information included response data from all assessed students in 1998. Data resolution activities occurred prior to the submission of data files to ETS and Westat to resolve any irregularities that existed.

7.11 UPLOADING OF SCORES TO THE NAEP DATABASE

An important quality control component of the image-scoring system was the inclusion, for purposes of file identification, of an exact copy of the student edit record, including the student booklet identification number, with every image of a student's response to a constructed-response item. When all the responses for an individual item had been scored, the system automatically submitted all item scores assigned during the scoring, along with their edit records, to a queue to be transmitted to the mainframe. A custom edit program matched the edit records of the scoring files to those of the original edit records on the mainframe. As matches were confirmed, the scores were applied to those individual files.

7.12 SD/LEP STUDENT QUESTIONNAIRES

SD/LEP questionnaires were completed for those students who were selected to participate in the assessment sample and were classified as students with disabilities (SD), or were categorized as limited English proficient (LEP) students. This questionnaire, which was completed by someone at the school knowledgeable about the student, asked about the student's background and the special programs in which the student participated. NCS processed the SD/LEP student questionnaires via optical mark recognition (OMR) scanning. Edits performed on the questionnaires assured that responses to questions fell within the valid range for that question. SD/LEP questionnaires were then matched to a student record. SD/LEP questionnaires that were not matched to a student document were cross-referenced with

the corresponding administration schedule, roster of questionnaires, and student data files to correct, if necessary, the information needed to result in a match.

7.13 SCHOOL QUESTIONNAIRES

In 1998, NCS continued to use intelligent character recognition (ICR) technology to capture percentage figures written by school personnel directly in boxes on the school questionnaire, rather than requiring the school official to grid ovals in a matrix. The data were then verified by an edit operator.

7.14 TEACHER QUESTIONNAIRE MATCH

The same processes that were followed in previous cycles were used in 1998 to achieve the best possible student/teacher match rate. Student identification numbers that were not matched to a teacher questionnaire were cross-referenced with the corresponding administration schedule and roster of questionnaires to verify (and change, if necessary) the teacher number, teacher period, and questionnaire number recorded on these control documents. The NAEP school identification numbers listed on the roster of questionnaires and teacher questionnaire were verified and corrected, if necessary. Once these changes were made, any duplicate teacher numbers existing within a school were, if possible, cross-referenced for resolution with the roster(s) of questionnaires. Since this information was located together on a single, central control document, the ability to match and resolve discrepant or missing fields was simplified.

7.15 DELIVERY

After all data-processing activities were completed, data cartridges, or diskettes were created and shipped via overnight delivery to ETS or Westat. NCS maintains a duplicate archive file for security and back-up purposes.

7.16 STORAGE OF DOCUMENTS

After batches of processed documents had successfully passed the editing process, they were sent to the NCS warehouse for storage. Due to the large number of rescoring projects done with NAEP material, the documents were unspiraled and sequenced by grade and book type after all of the processing and scoring was completed. This allows for efficient document retrieval to fill requests for specific booklets or book types for future projects. Unspiraled and sequenced booklets were then assigned a new inventory number by grade and book type and were sent back to the warehouse for storage. The storage locations of all documents were recorded on the inventory control system.

7.17 QUALITY CONTROL DOCUMENTS

ETS required that a random sample of books be pulled for an additional quality control check. The 1998 NAEP assessment of reading, writing, and civics documents to be scored were all image scanned (aside from the exception noted previously). For image-scanned documents, a scoring sheet was not used, so ETS used scores sent to them on a data tape to verify the accuracy of applied scores. All of these documents were selected prior to sending the booklets to storage and were then sent to ETS to verify the accuracy and completeness of the data. A random sample of all the questionnaires used in the 1998 NAEP assessment was also sent to ETS along with the quality assurance booklets used for processing and scoring. The quality control analyses of these booklets are discussed in Chapter 8.

Chapter 8

CREATION OF THE DATABASE, QUALITY CONTROL OF DATA ENTRY, AND CREATION OF THE DATABASE PRODUCTS[1]

John J. Ferris, Katharine E. Pashley, David S. Freund, and Alfred M. Rogers
Educational Testing Service

8.1 INTRODUCTION

The data-processing, scoring, and editing procedures described in Chapters 6 and 7 resulted in the generation of disk and tape files containing various data for students (assessed and excluded), teachers, schools, and SD/LEP (students with disabilities and students with limited English proficiency) information. The weighting procedures described in Chapters 10 and 11 resulted in the generation of data files that included the sampling weights required to make valid statistical inferences about the population from which the 1998 fourth-, eighth- and twelfth-grade NAEP samples were drawn. These files were merged into a comprehensive, integrated database. The creation of the database is described in Section 8.2.

Section 8.2.2 describes a central repository or master catalog of this information. The master catalog is accessible by all analysis and reporting programs and provides correct parameters for processing the data fields and consistent labeling for identifying the results of the analyses.

To evaluate the effectiveness of the quality control of the data-entry process, the corresponding portion of the final integrated database was verified in detail against a sample of the original instruments received from the field. The results of this procedure are given in Section 8.3.

The integrated database was the source for the creation of the NAEP item information database and the NAEP secondary-use data files. These are described in Section 8.4.

8.2 CREATION OF THE DATABASE

The data processing conducted by National Computer Systems (NCS) resulted in the transmittal to ETS of four data files for each of fourth, eighth and twelfth grade: one file for the student background and item-response data and one file for each of the three questionnaires—teacher, school characteristics and policies, and SD/LEP. The sampling weights, derived by Westat, comprised additional files for each grade. (See Chapters 10 and 11 for a discussion of the sampling weights.) These files at each grade were the foundation for the analysis of the 1998 NAEP data. Before data analyses could be performed, these data files had to be integrated into a coherent and comprehensive database.

The database ultimately comprised four files per cohort: three student files (reading, writing, and civics) and a single school file. The student files were separated by subject area to improve maintenance and efficiency of the databases and data analyses. Each record on the student file contained a student's responses to the particular assessment booklet the student was administered (in the case of excluded

[1] John J. Ferris was responsible for the evaluation of the quality of the database and the data-entry process; Katharine E. Pashley was responsible for database generation under the supervision of David S. Freund; Alfred M. Rogers created the secondary-use data files.

students, a booklet was assigned, but the student-response fields contain a special code indicating no response), and the information from the questionnaire that the student's teacher completed. Additionally, for a student (assessed or excluded) who was identified as a student with a disability (SD) or of limited English proficiency (LEP), the data from the SD/LEP questionnaire are included. This questionnaire is filled out for all students both assessed and excluded, identified as SD, LEP, or both. (See Chapter 2 for information regarding assessment instruments.) Also added to the student files were variables with school-level information supplied by Quality Education Department, Inc. (QED), including demographic information about schools such as distributions of student populations by race/ethnicity. Since the teacher data are not from a representative sample of teachers and since the focus of NAEP is to report student-level results, the teacher-response data were added to the student records in cases where the student's teacher responded to a teacher questionnaire. The school data were on separate files that could be analyzed on their own and could also be linked to the student files through the unique school identification code.

The creation of the student data files for fourth, eighth, and twelfth grade began with the reorganization of the data files received from NCS. This involved two major tasks:

1. The files were restructured, eliminating unused (blank) areas to reduce the size of the files.

2. In cases where students had chosen not to respond to an item, the missing responses were recoded as either "omit" or "not reached," as discussed in Chapter 12 of this report.

8.2.1 Merging Files

Following the reorganization of data files, the student-response data were merged with the student-weights files. The resulting file was then merged with the SD/LEP and teacher data. In all merging steps, the 10-digit booklet identification (the 3-digit booklet number common to every booklet with the same block of items, a 6-digit serial number unique to the booklet a student was given, and a single check digit) was used as the matching criterion. The teacher data can be linked to the student data through four data variables: primary sampling unit (PSU), school code, teacher ID, and classroom period.

The school file for each grade was created by merging the school characteristics and policies questionnaire file with the file of school weights and school variables, supplied by Westat. The PSU and school codes were used as the matching criteria. Since some schools did not return a questionnaire, some of the records in the school file contained only school-identifying information and sampling-weight information. The school data can be linked to the student data through the PSU and school code variables.

When the student and school files for each grade had been created, the database was ready for analysis. In addition, whenever new data values (such as composite background variables or plausible values) were derived, they were added to the appropriate database files using the same matching procedures described above.

For archival purposes and to provide data for outside users, restricted-use data files and codebooks for each jurisdiction in the state assessment were generated from this database. The restricted-use data files contain all responses and response-related data from the assessment, including responses from the student booklets, teacher questionnaires, and school characteristics and policies questionnaires, scale scores, sampling weights, and variables used to compute standard errors.

8.2.2 Creating the Master Catalog

A critical part of any database is its processing control and descriptive information. Having a central repository for this information, which may be accessed by all analysis and reporting programs, will provide correct parameters for processing the data fields and consistent labeling for identifying the results of the analyses. The NAEP master catalog file was designed and constructed to serve these purposes for the NAEP database.

Each record of the master catalog contains the processing (e.g., response options), labeling, classification (e.g., content), and location information for each assessment exercise and other data variables in the NAEP database. The control parameters are used by the access routines in the analysis programs to define the manner in which the data values are to be transformed and processed.

Each data variable has a 50-character label in the master catalog describing the contents of the variable and, where applicable, the source of the variable. The variables with discrete or categorical response values (e.g., multiple-choice items and professionally scored items, but not weight variables) have additional label fields in the catalog containing 8- and 20-character labels for those response values. These short labels can be used for reporting purposes as a concise description of the responses for these discrete items.

The classification area of the master catalog record contains distinct fields corresponding to predefined classification categories (e.g., reading purpose and reading stance) for the data variables. For a particular classification variable, a nonblank value indicates the code of the subcategory within the classification category for the data variable. This classification area permits the grouping of identically classified items or other variables by performing a selection process on one or more classification fields in the master catalog.

According to NAEP design, it is possible for assessment exercises to appear in more than one student sample and in more than one block of exercises within each sample. The location fields of the catalog record contain age cohort, block, and, where applicable, the order within the block for each appearance of the assessment exercise.

The master catalog file was constructed concurrently with the collection and transcription of the national and state assessment data so that it would be ready for use by analysis programs when the database was created. As new data fields were derived and added to the database, their corresponding descriptive and control information were entered into the master catalog.

8.3 QUALITY CONTROL OF NAEP DATA ENTRY FOR 1998

This section describes the evaluation of the data-entry process for the 1998 national assessment. As in past years, the NAEP database was found to be more than accurate enough to support the analyses that were done. Overall, the observed error rates were comparable to those of past assessments, including those of the teacher questionnaires, which returned to more typical levels after displaying a somewhat elevated error rate in 1996. Derived error rate limits were around one error per thousand responses except for the school questionnaire data, which was nearly five per thousand (see discussion below).

The purpose of the analysis reported in this section is to assess the quality of the data resulting from the complete data-entry system, beginning with the actual instruments collected in the field and ending with the final machine-readable database used in the analyses. The process involved the selection of instruments at random from among those returned from the field and the comparison of each entire

instrument, character by character, with its representation in the final database. In this way, we were able to measure the error rates in the data as well as the success of the data-entry system.

Of course the observed error rate cannot be taken at face value. For example, the sample of school questionnaires that happened to be selected for close inspection contained two errors out of a total of 2,251 characters. To conclude that the entire school questionnaire database has an error rate of $\frac{2}{2,251}$, or .0009, would be too optimistic; we may simply have been lucky (or unlucky) with this particular random sample. What is needed is an indication of how bad the true error rate might be, given what we observed. Such an indication is provided by confidence limits. Confidence limits indicate how likely it is that a value falls inside a specified range in a specified context or distribution. In our analysis, the specified range is an error rate between zero and some maximum value beyond which we are confident at a specified level (traditionally 99.8%) that the true error rate does not lie (for the school questionnaires, this error rate is .0046). The specified context or distribution turns out to be the cumulative binomial probability distribution. An example will demonstrate this technique:

> Let us say that 1,000 booklets were processed, each with 100 characters of data transcribed for a total of 100,000 characters. Let us say further that 5 of these characters were discovered to be in error in a random sample of 50 booklets that were completely checked; in other words, five errors were found in a sample of 5,000 characters. The following expression may be used to establish the probability that the true error rate is .0025 or less, rather than the single-value estimate of the observed rate, one in a thousand (.001):
>
> $$\sum_{j=0}^{5} \binom{5000}{j} \times .0025^{j} \times (1-.0025)^{(5000-j)} = .0147$$
>
> This is the sum of the probability of finding five errors plus the probability of finding four errors plus . . . etc. . . . plus the probability of finding zero errors in a sample of 5,000 with a true error rate of .0025; that is, the probability of finding five or fewer errors by chance when the true error rate is .0025. Notice that we did not use the size of the database in this expression. Actually, the assumption here is that our sample of 5,000 was drawn from a database that is infinite. The smaller the actual database is, the more confidence we can have in the observed error rate; for example, had there been only 5,000 in the total database, our sample would have included all the data, and the observed error rate would have been the true error rate. The result of the above computation allows us to say, conservatively, that .0025 is an upper limit on the true error rate with 98.53 percent (i.e., 1 - .0147) confidence; that is, we can be quite sure that our true error rate is no larger than .0025. As noted above, in NAEP quality control we use a more stringent confidence limit of 99.8%, which yields an even more conservative upper bound on the true error rate; with 99.8% confidence, we would state that the true error rate in this example is no larger than .0031, rather than .0025.

Calculations of true probabilities based on a combinatorial analysis have been done (e.g., Grant, 1964). Even when the sample was as much as 10% of a population of 50, the estimate of the probability based on the binomial theorem was not much different from the correct probability. NAEP does not sample at a rate greater than about 2%. Thus, the computations of the upper limits on the true error rates based on the binomial theorem are likely to be highly accurate approximations.

The individual instruments are briefly discussed in the following sections and a summary table (Table 8-1) gives the upper 99.8 percent confidence limit for the error rate for each of the instruments as well as the sampling information. The 99.8 percent confidence limit and the selection rates indicated were chosen to make these results comparable to those of administrations since 1983, all of which used the same parameters.

Table 8-1
Summary of Quality Control Error Analysis for NAEP 1998 Data Entry

Instrument/ Sample	Selection Rate	Different Booklets	Number of Booklets Sampled	Number of Characters Sampled	Number of Errors	Observed Error Rate	Upper 99.8% Confidence Limit
Student Booklets – Nat'l. Main	1/278	266	509	29,802	16	.0005	.0011
SD/LEP Student Questionnaires	1/77	3	217	19,964	8	.0004	.0010
Teacher Questionnaires	1/68	4	131	14,811	6	.0004	.0012
School Characteristics and Policies Questionnaires	1/53	3	40	2,251	2	.0009	.0046

8.3.1 Student Booklet Data

Data from about 140,000 students were processed across all samples in this assessment. Roughly one booklet in 278 was selected for close examination, which is a somewhat higher rate than that used in past assessments, when a rate of approximately one in 350 was used. The higher selection rate improves the chance of drawing sufficient numbers of each booklet when there is a large number of different books. The student data error rates were consistently low in all subject areas and across all three grades, typically involving an occasional multiple response taken as a single one. The overall quality of the data was very high.

8.3.2 SD/LEP Student Questionnaire Data

In this assessment, 16,703 SD/LEP student questionnaires were scanned. The quality control sampling rate was 1 in 77, a somewhat higher rate than that used in previous assessments. The data showed about the same error rate as that in the previous assessment—comparable to the rate for the student data. The few problems encountered involved the scanner's mistaking an erasure for a genuine response or failing to identify a multiple response as such.

8.3.3 Teacher Questionnaire Data

In this assessment, 8,959 teacher questionnaires were collected and scanned. About 1.5 percent of these questionnaires was sampled for the quality control procedure. The error rates for these questionnaires were about the same as for the student categories of data, and much improved over the 1996 error rates. Since there has been no significant change in the format of these questionnaires, the improved error rates may be attributable to improved administration procedures.

8.3.4 School Characteristics and Policies Questionnaire Data

In this assessment, 2,102 school characteristics and policies questionnaires were collected. They were sampled at a rate of about 1 in 53. Only two scanning errors were found in these questionnaires, both of which involved the scanner's failing to pick up a valid response. In spite of this apparently good error rate of less than one in a thousand, the application of the binomial theorem yields an upper bound on the true error rate of .0046 (at the same confidence level). While this may seem surprisingly high, an error rate limit derived from an application of the binomial theorem is appropriate here, since the sample population is large, as noted in the above discussion of the application of this technique.

8.4 NAEP DATABASE PRODUCTS

The NAEP database described to this point serves primarily to support analysis and reporting activities that are directly related to the NAEP contract. This database has a singular structure and access methodology that is integrated with the NAEP analysis and reporting programs. One of the directives of the NAEP contract is to provide secondary researchers with a nonproprietary version of the database that is portable to any computer system. In the event of transfer of NAEP to another client, the contract further requires ETS to provide a full copy of the internal database in a format that may be installed on a different computer system.

The secondary-use data files are designed to enable any researcher with an interest in the NAEP database to perform secondary analysis on the same data as those used at ETS. The data, documentation, and supporting files are distributed on CD-ROM media. For each sample in the assessment, the following files are provided: the response data file, a printable data file layout and codebook file, a file of control statements that will generate an SPSS system file, a file of control statements that will generate a SAS system file, and a machine-readable catalog file. Each codebook is in portable document file (PDF) format, which may be browsed, excerpted, and printed using the Adobe Acrobat Reader program on a variety of platforms. Each machine-readable catalog file contains sufficient control and descriptive information to permit the user who does not have either SAS or SPSS to set up and perform data analysis.

The remainder of this section summarizes the procedures used in generating the data files and related materials.

8.4.1 File Definition

The design of the 1998 assessment perpetuates two features of the 1990, 1992, 1994, and 1996 assessment design: the focused BIB or PBIB booklet design and the direct matching of teacher questionnaires to student assessment instruments. In addition, the sample of students who were excluded from the assessment is now incorporated into the appropriate assessed student subject-area sample.

The focused BIB or PBIB design within the main assessment isolates the primary subject areas to separate groups of booklets. This permits the division of the main sample into subject-specific subsamples. The data files generated from these subsamples need only contain the data that are relevant to their corresponding subject areas and are therefore smaller and more manageable than their counterparts in previous assessments.

The intent of the 1998 assessment design was to collect data from the reading, writing, or civics teachers of fourth-grade and eighth-grade students who participated in the assessments of, respectively, reading, writing, or civics. A portion of the teacher questionnaire contained questions that were directly related to each matched student. This change in the design afforded a very high matching rate between

student and teacher data. Therefore, for those subject areas in each grade cohort for which teacher data were collected, the teacher responses were appended to each student record in the secondary-use data files.

8.4.2 Definition of the Variables

The initial step in the variable definition process was the generation of a labels file of descriptors of the variables for each data file to be created. Each record in a labels file contains, for a single data field, the variable name, a short description of the variable, and processing control information to be used by later steps in the data-generation process. This file could be edited for deletion of variables, modification of control parameters, or reordering of the variables within the file. The labels file is an intermediate file only; it is not included on the released data files.

The variables on all data files are grouped and arranged in the following order: identification information, weights, derived variables, scale scores (where applicable), and response data. On the student data files, these fields are followed by the teacher-response data and the SD/LEP student questionnaire data, where applicable. The identification information is taken from the front covers of the instruments. The weight data include sample descriptors, selection probabilities, nonresponse adjustments, and replicate weights for the estimation of sampling error. The derived data include sample descriptions from other sources and variables that are derived from the response data for use in analysis or reporting.

For each subject area of the 1998 assessment, the item-response data within each block of questions (see Section 1.5) were left in their order of presentation. The responses to cognitive blocks that were not present in a given booklet were left blank, signifying a condition of “missing by design.”

In order to process and analyze the spiral sample data effectively, the user must also be able to determine, from a given booklet record, which blocks of item response data were present and their relative order in the instrument. This problem was remedied by the creation of a set of control variables, one for each block, which indicated not only the presence or absence of the block but its order in the instrument. These control variables are included with the derived variables.

8.4.3 Data Definition

To enable the data files to be processed on any computer system using any procedural or programming language, it was desirable that the data be expressed in numeric format. This was possible, but not without the adoption of certain conventions for re-expressing the data values.

During creation of the NAEP database, the responses to all multiple-choice items were transcribed and stored in the database using the letter codes printed in the instruments. This scheme afforded the advantage of saving storage space for items with 10 or more response options, but at the expense of translating these codes into their numeric equivalents for analysis purposes. The response data fields for most of these items would require a simple alphabetic-to-numeric conversion. However, the data fields for items with 10 or more response choices would require “expansion” before the conversion, since the numeric value would require two column positions. One of the processing control parameters on the labels file indicates whether or not the data field is to be expanded before conversion and output.

The ETS database contained special codes to indicate certain response conditions: “I don’t know” responses, multiple responses, omitted responses, not-reached responses, and unresolvable responses, which include out-of-range responses and responses that were missing due to errors in printing

or processing. The scoring guides for the reading, writing, and civics constructed-response items included additional special codes for ratings of “illegible,” “off task,” or nonrateable by the scorers. All of these codes had to be re-expressed in a consistent numeric format.

The following convention was adopted and used in the designation of these codes: The “illegible” response codes were converted to 5, the “off task” response codes were converted to 6, the “I don’t know” and nonrateable response codes were converted to 7, the “omitted” response codes were converted to 8, the “not reached” response codes were converted to 9, and the multiple-response codes were converted to 0, and the out-of-range and missing responses were coded as blank fields, corresponding to the “missing by design” designation.

This coding scheme created conflicts for those multiple-choice items that had seven or more valid response options as well as the “I don’t know” response and for those constructed-response items whose scoring guide had five or more categories. These data fields were also expanded to accommodate the valid response values and the special codes. In these cases, the special codes were “extended” to fill the output data field: The “I don’t know” and nonrateable codes were extended from 7 to 77, the omitted response codes were extended from 8 to 88, and so on.

Each numeric variable on the secondary-use files was classified as either continuous or discrete. The continuous variables include the weights, scale scores, identification codes, and questionnaire responses where counts or percentages were requested. The discrete variables include those items for which each numeric value corresponds to a response category. The designation of “discrete” also includes those derived variables to which numeric classification categories have been assigned. The constructed-response items were treated as a special subset of the discrete variables and were assigned to a separate category to facilitate their identification in the documentation.

8.4.4 Data File Catalogs

The catalog file is created by the GENCAT program from the labels file and the 1998 master catalog file. Each record on the labels file generates a catalog record by first retrieving the master catalog record corresponding to the field name. The master catalog record contains usage, classification, and response code information, along with positional information from the labels file, field sequence number, output column position, and field width. Like the labels file, the catalog file is an intermediate file and is not included on the released data files.

The information for the response codes consists of the valid data values for the discrete numeric fields, and a 20-character description of each. The GENCAT program uses additional control information from the labels file to determine if extra response codes should be generated and saved with each catalog record. The first flag controls generation of the “I don’t know” or nonrateable response code; the second flag regulates omitted or not-reached code generation; and the third flag denotes the possibility of multiple responses for that field and sets up an appropriate response code. All of these control parameters, including the expansion flag, may be altered in the labels file by use of a text editor, in order to control the generation of data or descriptive information for any given field.

The catalog file supplies control and descriptive information for many of the subsequent secondary-use data-processing steps.

8.4.5 Data File Layouts

The data file layouts are the first user product to be generated in the secondary-use data files process. The generation program, GENLYT, uses a catalog file as input and produced a printable file. The layout file is little more than a formatted listing of the catalog file.

Each line of the layout file contains the following information for a single data field: sequence number, field name, output column position, field width, number of decimal places, data type, value range, key or correct response value, and a short description of the field. The sequence number of each field is implied from its order on the labels file. The field name is an 8-character label for the field that is to be used consistently by all secondary-use data files materials to refer to that field on that file. The output column position is the relative location of the beginning of that field on each record for that file, using bytes or characters as the unit of measure. The field width indicates the number of columns used in representing the data values for a field. If the field contains continuous numeric data, the value under the number of decimal places entry indicates how many places to shift the decimal point before processing data values.

The data type category uses five codes to designate the nature of the data in the field: Continuous numeric data are coded “C”; discrete numeric data are coded “D”; constructed-response item data are coded either “OS” (if the item was dichotomized for scaling) or “OE” (if it was scaled under a polytomous response model). Additionally, the discrete numeric fields that include “I don’t know” response codes are coded “DI.” If the field type is discrete numeric, the value range is listed as the minimum and maximum permitted values separated by a hyphen to indicate range. If the field is a response to a scorable item, the correct option value, or key, is printed. If the field is an assigned score that was scaled as a dichotomous item using cut-point scoring, the range of correct scores is printed. Each variable is further identified by a 50-character descriptor.

8.4.6 Data Codebooks

The data codebook is a printed document containing complete descriptive information for each data field. Most of this information originates from the catalog file, while the remaining data comes from the counts file and the IRT parameters file.

Each data field receives at least one line of descriptive information in the codebook. If the data type is continuous numeric, no more information is given. If the variable is discrete numeric, the codebook lists the response codes, response-code labels, and frequencies of each value in the data file. Additionally, if the field represents an item used in IRT scaling, the codebook lists the parameters used by the scaling program.

Certain blocks of cognitive items in the 1998 assessment that are to be used again in later assessments for trend comparisons have been designated as nonreleased. In order to maintain their confidentiality, generic labels have been substituted for the response category descriptions of these items in the data codebooks and the secondary-use files.

The frequency counts are not available on the catalog file, but must be generated from the data. The GENFREQ program creates the counts file using the field name to locate the variable in the database, and the response code values to validate the range of data values for each field. This program also serves as a check on the completeness of the response codes in the catalog file, as it flags any data values not represented by a value and label.

The IRT parameter file is linked to the catalog file through the field name. Printing of the IRT parameters is governed by a control flag in the classification section of the catalog record. If an item has been scaled for use in deriving the scale score estimates, the IRT parameters are listed to the right of the values and labels, and the score value for each response code is printed to the immediate right of the corresponding frequency.

The layout and codebook files are written by their respective generation programs to print-image disk data files. Draft copies are printed and distributed for review before the production copy is generated. The production copy combines the layout and codebook files for each sample in a portable document file (PDF) format. This file may be browsed, excerpted and printed using the Adobe Acrobat Reader program on a variety of platforms and operating systems.

8.4.7 Control Statement Files for Statistical Packages

An additional requirement of the NAEP cooperative agreement is to provide, for each secondary-use data file, a file of control statements each for the SAS and SPSS statistical systems that will convert the raw data file into the system data file for that package. Two separate programs, GENSAS and GENSPX, generate these control files using the catalog file as input.

Each of the control files contains separate sections for variable definition, variable labeling, missing value declaration, value labeling, and creation of scored variables from the cognitive items. The variable definition section describes the locations of the fields, by name, in the file, and, if applicable, the number of decimal places or type of data. The variable label identifies each field with a 50-character description. The missing value section identifies values of those variables that are to be treated as missing and excluded from analyses. The value labels correspond to the response codes in the catalog file. The code values and their descriptors are listed for each discrete numeric variable. The scoring section is provided to permit the user to generate item score variables instead of the item response variables.

Each of the code generation programs combines three steps into one complex procedure. As each catalog file record is read, it is broken into several component records according to the information to be used in each of the resultant sections. These record fragments are tagged with the field sequence number and a section sequence code. They are then organized by section code and sequence number. Finally, the reorganized information is output in a structured format dictated by the syntax of the processing language.

The generation of the system files accomplishes the testing of these control statement files. The system files are saved for use in special analyses by NAEP staff. These control statement files are included on the distributed data files to permit users with access to SAS and/or SPSS to create their own system files.

8.4.8 Machine-Readable Catalog Files

For those NAEP data users who have neither SAS nor SPSS capabilities, yet require processing control information in a computer-readable format, the distribution files also contain machine-readable catalog files. Each machine-readable catalog record contains processing control information, IRT parameters, and response codes and labels. The machine-readable catalog files are described in and are available as part of the secondary-use data files package for use in analyzing the data with programming languages such as SAS and SPSS (see the *NAEP 1998 Reading Data Companion*, [Rogers, Kokolis,

Stoeckel, & Kline, 2000], the *NAEP 1998 Writing Data Companion*, [Rogers, Kokolis, Stoeckel, & Kline, 2000], and the *NAEP 1998 Civics Data Companion*, [Rogers, Kokolis, Stoeckel, & Kline, 2000]).

8.4.9 NAEP Data on Disk

The complete set of secondary-use data files described above are available on CD-ROM as part of the NAEP Data on Disk product suite. This medium is ideal for researchers and policy makers operating in a personal computing environment.

The NAEP Data on Disk product suite includes two other components that facilitate the analysis of NAEP secondary-use data. The PC-based NAEP data extraction software, NAEPEX, enables users to create customized extracts of NAEP data and to generate SAS or SPSS control statements for preparing analyses or generating customized system files. The NAEP analysis modules, which currently run under SPSS® for Windows™, use output files from the extraction software to perform analyses that incorporate statistical procedures appropriate for the NAEP design (e.g., minimum sample size requirements, appropriate row-wise and column-wise t-tests, and automatic calculation of correct and consistent standard errors and degrees of freedom).