ED 393 094 CS 012 419

AUTHOR Mazzeo, John; And Others
TITLE Technical Report of the NAEP 1994 Trial State Assessment Program in Reading.
INSTITUTION Educational Testing Service, Princeton, N.J.; National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
REPORT NO NCES-96-116
PUB DATE Dec 95
NOTE 545p.; For related documents, see ED 388 962-963.
PUB TYPE Collected Works - General (020) -- Reports - Research/Technical (143)

EDRS PRICE MF02/PC22 Plus Postage.
DESCRIPTORS *Data Analysis; Grade 4; Intermediate Grades; *Program Design; Program Implementation; *Reading Achievement; *Reading Research; Test Items
IDENTIFIERS *National Assessment of Educational Progress; *Trial State Assessment (NAEP)

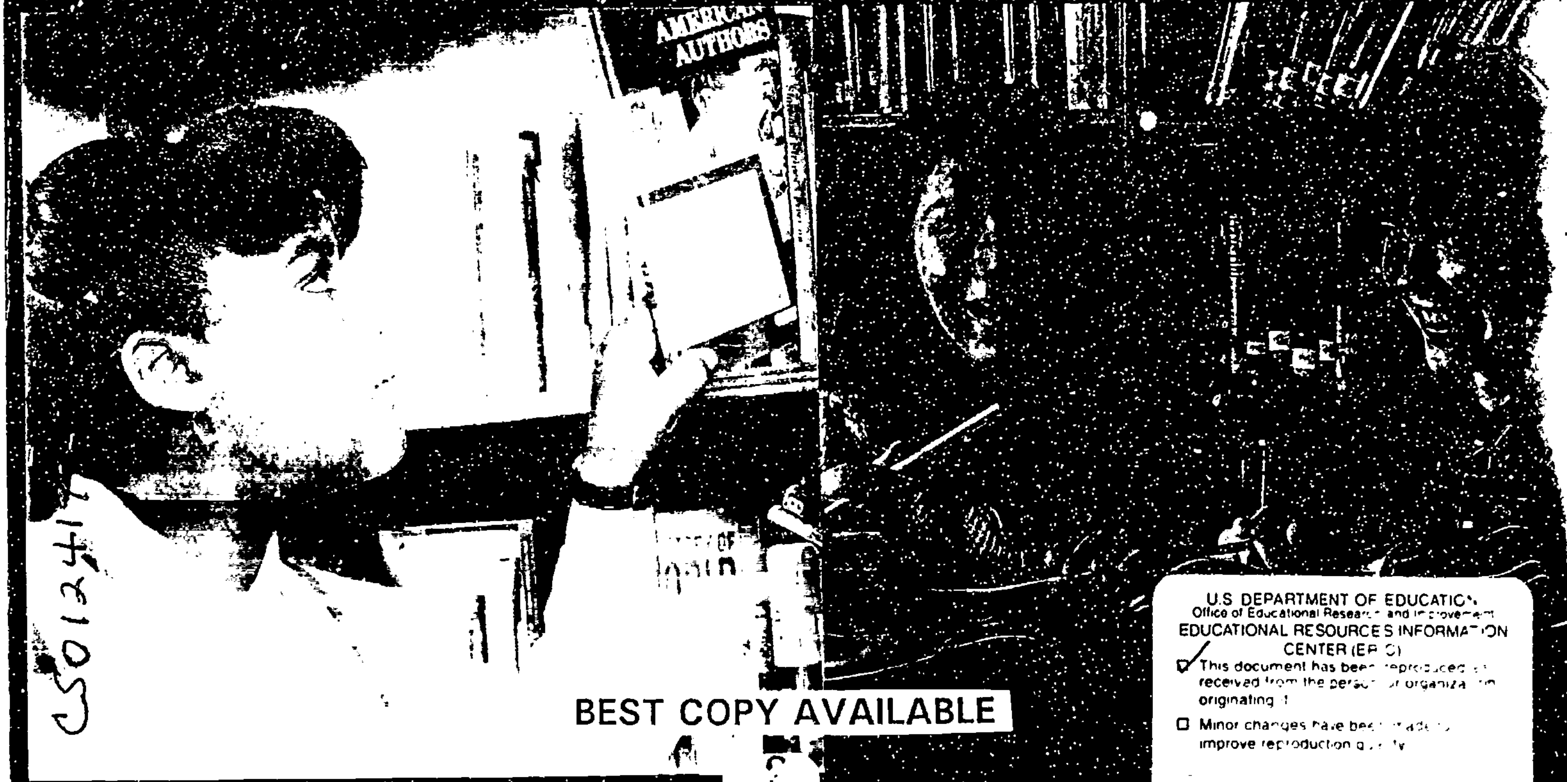
ABSTRACT

This report provides a description of the design for the National Assessment of Educational Progress (NAEP) 1994 Trial State Assessment in Reading and gives an overview of the steps involved in the implementation of the program from the planning stages through to the analysis and reporting of data. The report does not provide the results of the assessment--rather, it provides information on how those results were derived. Chapters in the report are (1) "Overview: The Design, Implementation, and Analysis of the 1994 Trial State Assessment Program in Reading" (John Mazzeo and Nancy L. Allen); (2) "Developing the Objectives, Cognitive Items, Background Questions, and Assessment Instruments" (Jay R. Campbell and Patricia L. Donahue); (3) "Sample Design and Selection" (James L. Green and others); (4) "State and School Cooperation and Field Administration" (Nancy Caldwell and Mark Waksberg); (5) "Processing and Scoring Assessment Materials" (Patrick Bourgeacq and others); (6) "Creation of the Database, Quality Control of Data Entry, and Creation of the Database Products" (John J. Ferris and others); (7) "Weighting Procedures and Variance Estimation" (Mansour Fahimi and others); (8) "Theoretical Background and Philosophy of NAEP Scaling Procedures" (Eugene G. Johnson and others); (9) "Data Analysis and Scaling for the 1994 Trial State Assessment in Reading" (Nancy L. Allen and others); and (10) "Conventions Used in Reporting the Results of the 1994 Trial State Assessment in Reading" (John Mazzeo and Clyde M. Reese). Contains 48 tables and 25 figures of data. Appendixes provide a list of participants in the objectives and item development process, a summary of participation rates, conditioning variables and contrast codings, IRT parameters for reading items, information on Trial State Assessment reporting subgroups, and discussions of setting achievement levels, the effect of monitoring an assessment sessions in nonpublic schools, correction of the NAEP program documentation error, and the information weighting error. (RS)

NATIONAL CENTER FOR EDUCATION STATISTICS

TECHNICAL REPORT OF THE

[illegible] PROGRAM IN READING



U.S. DEPARTMENT OF EDUCATION
OFFICE OF EDUCATIONAL RESEARCH AND IMPROVEMENT

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

Technical Report of the NAEP 1994 Trial State Assessment Program in Reading

John Mazzeo
Nancy L. Allen
Debra L. Kline

in collaboration with

Patrick B. Bourgeacq
Mary Lyn Bourque
Charles L. Brungardt
John Burke
Nancy Caldwell
Jay R. Campbell
Patricia L. Donahue
Mansour Fahimi
John J. Ferris
David S. Freund
James L. Green
Eddie H.S. Ip
Steven P. Isham
Eugene G. Johnson

Tillie Kennel
Robert J. Mislevy
Clyde M. Reese
Linda L. Reynolds
Timothy Robinson
Alfred M. Rogers
Keith F. Rust
Patricia M. Garcia Stearns
Brent W. Studer
Spencer Swinton
Bradley J. Thayer
Neal Thomas
Mark M. Waksberg
Lois H. Worthington

with a Foreword by Gary W. Phillips

December 1995

THE NATION'S REPORT CARD naep

Prepared by Educational Testing Service under contract
with the National Center for Education Statistics

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

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

Office of Educational Research and Improvement
Sharon P. Robinson
Assistant Secretary

National Center for Education Statistics
Jeanne E. Griffith
Acting Commissioner

Education Assessment Division
Gary W. Phillips
Associate Commissioner

December 1995

FOR MORE INFORMATION:

For ordering information on this report, write:

National Library of Education
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, D.C. 20208-5641

or call 1-800-424-1616 (in the Washington, D.C. metropolitan area call 202-219-1651).

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.

Educational Testing Service is an equal opportunity, affirmative action employer.

TECHNICAL REPORT
OF THE NAEP 1994 TRIAL STATE ASSESSMENT PROGRAM
IN READING

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

ACKNOWLEDGMENTS

The design, development, analysis, and reporting of the Trial State Assessment Program was truly a collaborative effort among staff from State Education Agencies, the National Center for Education Statistics (NCES), Educational Testing Service (ETS), Westat, and National Computer Systems (NCS). The program benefitted from the contributions of hundreds of individuals at the state and local levels—Governors, Chief State School Officers, State and District Test Directors, State Coordinators, and district administrators—who tirelessly provided their wisdom, experience, and hard work. Finally, and most importantly, NAEP is grateful to the students and school staff who participated in the Trial State Assessment.

This report documents the design and data analysis procedures behind the 1994 Trial State Assessment in reading. It also provides insight into the rationale behind the technical decisions made about the program. The development of this Technical Report, and especially of the Trial State Assessment Program, is the culmination of effort by many individuals who contributed their considerable knowledge, experience, and creativity to the 1994 Trial State Assessment Program in reading.

The 1994 Trial State Assessment was funded through the National Center of Education Statistics in the Office of Educational Research and Improvement of the U.S. Department of Education. Emerson Elliott, past NCES Commissioner, provided consistent support and guidance. NCES staff—particularly Jeanne Griffith, Gary Phillips, Steve Gorman, Peggy Carr, Sharif Shakrani, Susan Ahmed, Larry Ogle, Shi-Cheng Wu, Maureen Treacy, and Sheida White—worked closely and collegially with ETS, Westat, and NCS staff and played a crucial role in all aspects of the program. The technical report was reviewed by Jamal Abedi of UCLA and Ralph Lee, Marilyn Binkley, Sue Ahmed, and Shi-Chang Wu of NCES.

The members of the National Assessment Governing Board (NAGB) and NAGB staff provided advice and guidance throughout, and their contractor, American College Testing, worked with various panels in setting the achievement levels, and carried out a variety of analyses related to the levels. NAGB's contractor for the reading consensus project, the Council of Chief State School Officers, worked diligently under tight time constraints to create the forward-thinking framework underlying the assessment.

NAEP owes a great deal to the numerous panelists and consultants who worked so diligently on developing the assessment and providing a frame for interpreting the results, including those who helped create the objectives, develop the assessment instruments, set the achievement levels, and provide the anchoring descriptions.

Under the NAEP contract to ETS, Archie Lapointe served as executive director and Paul Williams as project director. Steve Lazer managed test development activities, and Jay Campbell worked with the Reading Item Development committee to develop the assessment instruments. Jules Goodison managed the operational aspects together with John Olson. Eugene Johnson led the measurement and research efforts; John Barone directed data analysis activities. ETS management has been very supportive of NAEP's technical work. Special thanks go to Nancy Cole and Ernie Anastasio as well as to Henry Braun and Charles Davis of ETS research management.

The guidance of the NAEP Design and Analysis Committee on the technical aspects of NAEP has been outstanding. The members are Sylvia Johnson (chair), Albert Beaton, Jeri Benson, William Cooley, Jeremy Finn, Huynh Huynh, Gaea Leinhardt, David Lohman, Bengt Muthén, Anthony Nitko, Ingram Olkin, Tej Pandey, and Juliet Shaffer. We were saddened by the untimely death of Clifford Clogg, whose service to NAEP will not be forgotten.

Statistical and psychometric activities were led by Nancy Allen, Spencer Swinton, and Eddie Ip under the direction of Eugene Johnson, Jim Carlson, and John Mazzeo. Major contributions were also made by Huahua Chang, John Donoghue, Frank Jenkins, Jo-lin Liang, Eiji Muraki, and Neal Thomas. Robert Mislevy provided valuable statistical and psychometric advice.

Under the leadership of John Barone, the division of Data Analysis and Technicology Research developed the operating systems and carried out the data analyses. David Freund and Alfred Rogers developed and maintained the large and complex NAEP data management systems; Kate Pashley managed database activities and Patricia O'Reilly was responsible for the restricted-use version of the data. Alfred Rogers developed the production versions of key analysis and scaling systems. Special thanks go to Steven Isham, who performed the reading data analyses, assisted by Lois Worthington. Laura Jerry led the computer-based development and production of the state reading reports and Jennifer Nelson produced the data compendium. They were assisted by Phillip Leung, Inge Novatkoski, Steven Isham, and David Freund. Alfred Rogers developed the report maps. Other members of this division who made important contributions to NAEP data analyses were Laura Jenkins, Michael Narcowich, Craig Pizzuti, and Minhwei Wang.

The staff of Westat, Inc. contributed their exceptional talents in all areas of sample design and data collection. Field administration and data collection were carried out under the direction of Renee Slobasky and Nancy Caldwell. Keith Rust developed and supervised the sampling design. Mark Waksberg, Leslie Wallace, Debra Vivari, Dianne Walsh, Leyla Mohadjer, Adam Chu, Valerija Smith, and Jacqueline Severynse undertook major roles in these activities.

Critical to the program was the contribution of National Computer Systems, Inc. Printing, distribution, scoring, and processing of the assessment materials were carried out under the leadership of Judy Moyer and Brad Thayer, with additional contributions by Patrick Bourgeacq, Charles Brungardt, Patricia Garcia Stearns, Tillie Kennel, Linda L. Reynolds, Timothy Robinson, and Brent Studer.

Jay Campbell, John Mazzeo, and Clyde Reese collaborated on the text for the assessment reports. Karen Miller coordinated the quality control and checking of the reports, much of which was carried out by Alice Kass, Yvette Hillard, and Christy Schwager. Thanks go also to the many individuals who reviewed the reports, especially the editors who improved the text and the data analysts who checked the accuracy of the data.

Mary Michaels and Sharon David-Johnson coordinated the cover design and final production of this technical report. Debra Kline was responsible for organizing, scheduling, editing, and ensuring the accuracy of the final report.

FOREWORD

This technical report summarizes some of the most complex statistical methodology used in any survey or testing program in the United States. In its 25-year history, the National Assessment of Educational Progress (NAEP) has pioneered such state-of-the-art techniques as matrix sampling and item response theory models. Today it is the leading survey using the advanced plausible values methodology, which uses a multiple imputation procedure in a psychometric context.

The 1994 Trial State Assessment in reading followed the same basic design as that used for the 1990 and 1992 Trial State Assessments in mathematics and reading. Properties of the 1994 reading assessment common to the 1990 and 1992 assessments include: 1) continuing the use of focused-BIB spiraling, item response theory models, and plausible values; 2) keeping the national and Trial State Assessment samples and scales separate; 3) doing separate stratifications and conditioning in each of the state samples; 4) making each state sample have power similar to the regional samples from the national assessment (this is how the sample sizes for the states were determined); 5) equating state and national scales using the aggregate of the state samples and a national subsample that was representative of the aggregate of the states; and 6) using power rules and other statistical considerations to determine which subgroup comparisons were supported by sufficient school and student sample sizes. One new activity in the 1994 assessment was the inclusion of nonpublic schools at the state level. The goal was to make the state estimates more representative of the total student population and, where possible, provide state estimates for the nonpublic school subgroups.

The 1994 Trial State Assessment provided many opportunities to test the limits of statistical theory and thereby advance the state of the art. Some examples include: 1) conditioning on a smaller set of principal components rather than a larger set of background variables and 2) the use of the two-parameter polytomous item response theory model for scaling constructed-response and extended constructed-response items. It is expected that in the future the conditioning models may be expanded in ways that will help secondary analysts who want to use hierarchical linear models as part of their statistical analysis procedures.

The Trial State Assessment has many statistical challenges ahead that must be dealt with. As the NAEP project plans for the 1996 assessment, it must find ways to: 1) accurately report results for nonpublic schools (which have less well developed sampling frames); 2) provide accommodations and adaptations for students with disabilities and limited English proficiency; and, 3) provide reports to the States within a six-month period. The project can and will meet these challenges.

The NAEP project is not only characterized by elegant statistical procedures, but it is also noted for the dedicated professionalism of its staff. It is the stubborn insistence that surveys are scientific activities and relentless quest for improved methodology that have made NAEP credible for over two decades.

Gary W. Phillips
Associate Commissioner
National Center for Education
Statistics

Chapter 1

OVERVIEW:

THE DESIGN, IMPLEMENTATION, AND ANALYSIS OF THE 1994 TRIAL STATE ASSESSMENT PROGRAM IN READING

John Mazzeo and Nancy L. Allen

Educational Testing Service

The National Assessment shall conduct a 1994. . .trial reading assessment for the 4th grade, in states that wish to participate, with the purpose of determining whether such assessments yield valid and reliable State representative data. (Section 406 (i)(2)(C)(i) of the General Education Provisions Act, as amended by Pub. L. 103-33 (US.C. 1221e-1(a(2)(B)(iii)))

The National Assessment shall include in each sample assessment. . .students in public and private schools in a manner that ensures comparability with the national sample. (Section 406(i)(2)(C)(i) of the General Education Provisions Act, as amended by Pub. L. 103-33 (U.S.C. 1221e-1(a)(2)(B)(iii)))

1.1 OVERVIEW

In April 1988, Congress reauthorized the National Assessment of Educational Progress (NAEP) and added a new dimension to the program—voluntary state-by-state assessments on a trial basis in 1990 and 1992, in addition to continuing the national assessments that NAEP had conducted since its inception. In this report, we will refer to the voluntary state-by-state assessment program as the Trial State Assessment Program. This program, which is designed to provide representative data on achievement for participating jurisdictions, is distinct from the assessment designed to provide nationally representative data, referred to in this report as the national assessment. (This terminology is also used in all other reports of the 1990, 1992, and 1994 assessments.) It should be noted that the word trial in Trial State Assessment refers to the Congressionally mandated trial to determine whether such assessments can yield valid, reliable state representative data. All instruments and procedures used in the 1990, 1992, and 1994 Trial State and national assessments were previously piloted in field tests conducted in the year prior to each assessment.

The 1990 Trial State Assessment Program collected information on the mathematics knowledge, skills, and understanding of a representative sample of eighth-grade students in public schools in 37 states, the District of Columbia, and two territories. The second phase of the Trial State Assessment Program, conducted in 1992, collected information on the mathematics knowledge, skills, and understanding of a representative sample of fourth- and eighth-grade students and the reading skills and understanding of a representative sample of fourth-grade students in public schools in 41 states, the District of Columbia, and two territories.

The 1994 Trial State Assessment Program, described in this technical report, once again assessed the reading skills and understanding of representative samples of fourth-grade students in participating jurisdictions. The participation of jurisdictions in the Trial State Assessment has been, and continues to be, voluntary. The 1994 program broke new ground in two ways. The 1994 NAEP authorization called for the assessment of samples of both public and private school students. Thus, for the first time in NAEP, jurisdiction-level samples of students from Catholic schools, other religious schools and private schools, Domestic Department of Defense Education Activity schools, and Bureau of Indian Affairs schools were added to the Trial State program. Second, samples of students from the Department of Defense Education Activity overseas schools participated as a jurisdiction, along with the states and territories that have traditionally had the opportunity to participate in Trial State Assessment Program.

Table 1-1 lists the jurisdictions that participated in the 1994 Trial State Assessment Program. More than 120,000 students at grade 4 participated in the reading assessment in those jurisdictions. Students were administered the same assessment booklets that were used in NAEP's 1994 national grade 4 reading assessment.

The reading framework that guided both the 1994 Trial State Assessment and the 1994 national assessment is the same framework used for the 1992 NAEP assessments. The framework was developed for NAEP through a consensus project of the Council of Chief State School Officers, funded by the National Assessment Governing Board. Hence, 1994 provides the first opportunity to report jurisdiction-level trend data for a NAEP reading instrument for those states and territories that participated in both the 1992 and 1994 Trial State Assessment programs. In addition, questionnaires completed by the students, their reading teachers, and principals or other school administrators provided an abundance of contextual data within which to interpret the reading results.

The purpose of this report is to provide technical information about the 1994 Trial State Assessment in reading. It provides a description of the design for the Trial State Assessment and gives an overview of the steps involved in the implementation of the program from the planning stages through to the analysis and reporting of the data. The report describes in detail the development of the cognitive and background questions, the field procedures, the creation of the database and data products for analysis, and the methods and procedures used for sampling, analysis, and reporting. It does not provide the results of the assessment—rather, it provides information on how those results were derived.

This report is one of several documents that provide technical information about the 1994 Trial State Assessment. For those interested in performing their own analyses of the data, this report and the user guide for the secondary-use data should be used as primary sources of information about NAEP. Information for lay audiences is provided in the procedural appendices to the reading subject-area reports; theoretical information about the models and procedures used in NAEP can be found in the special NAEP-related issue of the *Journal of Educational Statistics* (Summer 1992/Volume 17, Number 2).

Table 1-1
Jurisdictions Participating in the
1994 Trial State Assessment Program

Jurisdictions			
Alabama	Hawaii	Mississippi	Pennsylvania
Arizona	Idaho	Missouri	Rhode Island
Arkansas	Indiana	Montana*	South Carolina
California	Iowa	Nebraska	Tennessee
Colorado	Kentucky	New Hampshire	Texas
Connecticut	Louisiana	New Jersey	Utah
Delaware	Maine	New Mexico	Virginia
DoDEA Overseas*	Maryland	New York	Washington*
District of Columbia**	Massachusetts	North Carolina	West Virginia
Florida	Michigan	North Dakota	Wisconsin
Georgia	Minnesota		Wyoming
Guam			

* Washington, Montana, and DoDEA (Department of Defense Education Activity) overseas schools participated in the 1994 program but did not participate in the 1992 program.

** The District of Columbia participated in the testing portion of the 1994 Trial State Assessment Program. However, in accordance with the legislation providing for participants to review and give permission for release of their results, the District of Columbia chose not to publish their results in the reports.

Educational Testing Service (ETS) was the contractor for the 1994 NAEP programs, including the Trial State Assessment. ETS was responsible for overall management of the programs as well as for development of the overall design, the items and questionnaires, data analysis, and reporting. National Computer Systems (NCS) was a subcontractor to ETS on both the national and Trial State NAEP programs. NCS was responsible for printing, distribution, and receipt of all assessment materials, and for scanning and professional scoring. All aspects of sampling and of field operations for both the national and Trial State Assessments were the responsibility of Westat, Inc. The National Center for Education Statistics contracted directly with Westat for these services for the national assessment. Westat was a subcontractor to ETS in providing sampling and field operations services for the Trial State Assessment.

This technical report provides information about the technical bases for a series of reports that have been prepared for the 1994 Trial State Assessment Program in reading, including:

- A *State Report* for each participating jurisdiction that describes the reading proficiency of the fourth-grade public- and nonpublic-school students in that jurisdiction and relates their proficiency to contextual information about reading policies and instruction.

- The report *NAEP 1994 Reading: A First Look*, which provides overall public-school results and results for major NAEP reporting subgroups for all of the jurisdictions that participated in the Trial State Assessment Program, as well as selected results from the 1994 national reading assessment.

- The *NAEP 1994 Reading Report Card for the Nation and the States*, which provides both public- and nonpublic-school data for all of the jurisdictions that participated in the Trial State Assessment Program along with a more complete report of the results from the 1994 national reading assessment.

- The *Executive Summary of the NAEP 1994 Reading Report Card for the Nation and the States*, providing the highlights of the *Reading Report Card*.

- The *Cross-State Data Compendium from the NAEP 1994 Reading Assessment*, which includes jurisdiction-level results for all the demographic, instructional and experiential background variables included in the *Reading Report Card* and *State Report*.

- *Data Almanacs* for each jurisdiction that contain a detailed breakdown of the reading proficiency data according to the responses to the student, teacher, and school questionnaires for the public-school, nonpublic-school, and combined populations as a whole and for important subgroups of the public-school population. There are six sections to each almanac:

    - *The Distribution Data Section* provides information about the percentages of students at or above the three composite-scale achievement levels (and below basic). For the composite scale and each reading scale, this almanac also provides selected percentiles for the public-school, nonpublic-school, and combined populations and for the standard demographic subgroups of the public-school population.

    - *The Student Questionnaire Section* provides a breakdown of the composite scale proficiency data according to the students' responses to questions in the three student questionnaires included in the assessment booklets.

    - *The Teacher Questionnaire Section* provides a breakdown of the composite-scale proficiency data according to the teachers' responses to questions in the reading teacher questionnaire.

    - *The School Questionnaire Section* provides a breakdown of the composite-scale proficiency data according to the principals' (or other administrators') responses to questions in the school characteristics and policies questionnaire.

    - *The Scale Section* provides a breakdown of the proficiency data for the two reading scales (Reading for Literary Experience, Reading to Gain Information) according to selected items from the questionnaires.

- ▲ *The Reading Item Section* provides the response data for each reading item in the assessment.

Organization of the Technical Report

This chapter provides a description of the design for the Trial State Assessment in reading and gives an overview of the steps involved in implementing the program from the planning stages to the analysis and reporting of the data. The chapter summarizes the major components of the program, with references to later chapters for more details. The organization of this chapter, and of the report, is as follows:

- Section 1.2 provides an overview of the design of the 1994 Trial State Assessment Program in reading.

- Section 1.3 summarizes the development of the reading objectives and the development and review of the items written to measure those objectives. Details are provided in Chapter 2.

- Section 1.4 discusses the assignment of the cognitive questions to assessment booklets. An initial discussion is provided of the partially balanced incomplete block (PBIB) spiral design that was used to assign cognitive questions to assessment booklets and assessment booklets to individuals. A more complete description is provided in Chapter 2.

- Section 1.5 outlines the sampling design used for the 1994 Trial State Assessment Program. A fuller description is provided in Chapter 3.

- Section 1.6 summarizes Westat's field administration procedures, including securing school cooperation, training administrators, administering the assessment, and conducting quality control. Further details appear in Chapter 4.

- Section 1.7 describes the flow of the data from their receipt at National Computer Systems through data entry, professional scoring, and entry into the ETS/NAEP database for analysis, and the creation of data products for secondary users. Chapters 5 and 6 provide a detailed description of the process.

- Section 1.8 provides an overview of the data obtained from the 1994 Trial State Assessment in reading.

- Section 1.9 summarizes the procedures used to weight the assessment data and to obtain estimates of the sampling variability of subpopulation estimates. Chapter 7 provides a full description of the weighting and variance estimation procedures.

- Section 1.10 describes the initial analyses performed to verify the quality of the data in preparation for more refined analyses, with details given in Chapter 9.

- Section 1.11 describes the item response theory subscales and the overall reading composite that were created for the primary analysis of the Trial State Assessment data. Further discussion of the theory and philosophy of the scaling technology appears in Chapter 8, with details of the scaling process in Chapter 9.

- Section 1.12 provides an overview of the linking of the scaled results from the Trial State Assessment to those from the national reading assessment. Details of the linking process appear in Chapter 9.

- Section 1.13 describes the reporting of the assessment results, with further details supplied in Chapter 10.

- Appendices A through G include a list of the participants in the objectives and item development process, a summary of the participation rates, a list of the conditioning variables, the IRT parameters for the reading items, the reporting subgroups, composite and derived common background and reporting variables, a description of the process used to define achievement levels, and a description of analyses comparing the performance of monitored and unmonitored schools for the nonpublic-school samples.

1.2 DESIGN OF THE TRIAL STATE ASSESSMENT IN READING

The major aspects of the design for the Trial State Assessment in reading included the following:

- Participation at the jurisdiction level was voluntary.

- Students from public and nonpublic schools were assessed. Nonpublic schools included Catholic schools, other religious schools, private schools, Domestic Department of Defense Education Activity schools, and Bureau of Indian Affairs schools. Separate representative samples of public and nonpublic schools were selected in each participating jurisdiction and students were randomly sampled within schools. The size of a jurisdiction's nonpublic-school samples was proportional to the percentage of students in that jurisdiction attending such schools.

- The fourth-grade reading assessment used for the 1994 NAEP Trial State Assessment, and included in the 1994 national NAEP instrument, consisted of eight 25-minute blocks of exercises. Six of these blocks were previously administered as part of the 1992 national and Trial State Assessments. Each block contained one reading passage and a combination of constructed-response and multiple-choice items. Passages selected for the assessment were drawn from texts that might be found and used by students in real, everyday reading. Entire stories, articles, or sections of textbooks were used, rather than excerpts or abridgments. The type of items—constructed-response or multiple-choice—was determined by the nature of the task. In addition, the constructed-response items were of two types: *Short constructed-response* items required students to respond to a question with a few words or a few sentences, while *extended constructed-response* items required students to respond to a question with a few paragraphs. Each student was given two of the eight blocks of items.

- A complex form of matrix sampling called a partially balanced incomplete block (PBIB) spiraling design was used. With PBIB spiraling, students in an assessment session received different booklets, which provides for greater reading content coverage than would have been possible had every student been administered the identical set of items, without imposing an undue testing burden on the student.

- Background questionnaires given to the students, the students' reading teachers, and the principals or other administrators provided a variety of contextual information. The background questionnaires for the Trial State Assessment were identical to those used in the national fourth-grade assessment.

- The assessment time for each student was approximately 63 minutes. Each a sessed student was assigned a reading booklet that contained a 5-minute background questionnaire, followed by two of the eight 25-minute blocks of reading items, a 5-minute reading background questionnaire, and a 3-minute motivation questionnaire. Sixteen different booklets were assembled.

- The assessments took place in the five-week period between January 31 and March 4, 1994. One-fourth of the schools in each state were assessed each week throughout the first four weeks; the fifth week was reserved for makeup sessions.

- Data collection was, by law, the responsibility of each participating jurisdiction. Security and uniform assessment administration were high priorities. Extensive training was conducted to assure that the assessment would be administered under standard, uniform procedures. For jurisdictions that had participated in the 1992 Trial State Assessment, 25 percent of the public-school assessment sessions and 50 percent of the nonpublic-school assessment sessions were monitored by the contractor's staff. For the remaining jurisdictions, 50 percent of both public- and nonpublic-school sessions were monitored.

1.3 DEVELOPMENT OF READING OBJECTIVES, ITEMS, AND BACKGROUND QUESTIONS

The 1994 Trial State Assessment and national NAEP program in reading were based on a reading framework[1] developed through a national consensus process, set forth by law, that calls for "active participation of teachers, curriculum specialists, subject matter specialists, local school administrators, parents, and members of the general public" (Public Law 100-297, Part C, 1988). This same framework was used for the 1992 Trial State Assessment in reading.

[1] *Reading Framework for the 1992 National Assessment of Educational Progress* (Washington, DC: National Assessment Governing Board, U.S. Department of Education, 1992). In addition, questionnaires completed by the students, their reading teacher, and principal or other school administrator provided an abundance of contextual data within which to interpret the reading results.

The process of developing the framework was carried out in late 1989 and early 1990 under the direction of the National Assessment Governing Board (NAGB), which is responsible for formulating policy for NAEP, including developing assessment objectives and test specifications. To prepare the 1992 reading framework, NAGB awarded a contract to the Council of Chief State School Officers (CCSSO). As the framework was being developed, the project staff continually sought guidance and reaction from a wide range of people in the fields of reading and assessment, from school teachers and administrators, and from state coordinators of reading and reading assessment. After thorough discussion and some amendment, the recommended framework was adopted by NAGB in March 1990.

The 1992 and 1994 NAEP reading assessments measured three general types of text and purposes for reading, the first two of which were measured at the fourth grade:

> **Reading for Literary Experience** usually involves the reading of novels, short stories, poems, plays, and essays. In these reading situations, readers explore the human condition and consider relationships among events, emotions, and possibilities. In reading for literary experience, readers are guided by what and how an author might write in a specific genre and by their expectations of how the text will be organized. The readers' orientation when reading for literary experience usually involves looking for how the author explores or uncovers experiences and engaging in vicarious experiences through the text.
>
> **Reading to Gain Information** usually involves the reading of articles in magazines and newspapers, chapters in textbooks, entries in encyclopedias and catalogues, and entire books on particular topics. The type of prose found in such texts has its own features. To understand it, readers need to be aware of those features. For example, depending upon what they are reading, readers need to know the rules of literary criticism, or historical sequences of cause and effect, or scientific taxonomies. In addition, readers read to gain information for different purposes—for example, to find specific pieces of information when preparing a research project, or to get some general information when glancing through a magazine article. These purposes call for different orientations to text from those in reading for a literary experience because readers are focused specifically on acquiring information.
>
> **Reading to Perform a Task** usually involves the reading of documents such as bus or train schedules; directions for games, repairs, and classroom or laboratory procedures; tax or insurance forms; recipes; voter registration materials; maps; referenda; consumer warranties; and office memos. When they read to perform tasks, readers must use their expectations of the purposes of the documents and the structure of documents to guide how they select, understand, and apply such information. The readers' orientation in these tasks involves looking for specific information so as to do something. Readers need to be able to apply the information, not simply understand it, as is usually the case in reading to be informed. Furthermore, readers engaging in this type of reading are not likely to savor the style or thought in these texts, as they might in reading for literary experience. Reading to Perform a Task was not measured at grade 4.

All items underwent extensive reviews by specialists in reading, measurement, and bias/sensitivity, as well as reviews by representatives from State Education Agencies. The items repeated from the 1992 NAEP assessment were originally field tested in 1991. Additional items for the 1994 assessment were field tested in 1993 on a representative group of approximately 6,800 students across 27 jurisdictions; about 500 responses were obtained to each item in the field test. Based on field test results, items that had not been used previously in a NAEP assessment were revised or modified as necessary and then again reviewed for sensitivity, content, and editorial concerns. With the assistance of ETS/NAEP staff and outside reviewers, the Reading Item Development Committee selected the items to include in the 1994 assessment.

Chapter 2 includes specific details about developing the objectives and items for the Trial State Assessment.

1.4 ASSESSMENT INSTRUMENTS

The assembly of cognitive items into booklets and their subsequent assignment to assessed students was determined by a PBIB design with spiraled administration. Details of this design, identical to the design used in 1992, are provided in Chapter 2. In addition to the student assessment booklets, three other instruments provided data relating to the assessment—a reading teacher questionnaire, a school characteristics and policies questionnaire, and an IEP/LEP student questionnaire.

The *student assessment booklets* contained five sections and included both cognitive and noncognitive items. In addition to two 25-minute sections of cognitive questions, each booklet included two 5-minute sets of general and reading background questions designed to gather contextual information about students, their experiences in reading, and their attitudes toward the subject, and one 3-minute section of motivation questions designed to gather information about the students' levels of motivation for taking the assessment.

The *teacher questionnaire* was administered to the reading teachers of the fourth-grade students participating in the assessment. The questionnaire consisted of three sections and took approximately 20 minutes to complete. The first section focused on teachers' general background and experience; the second, on teachers' background related to reading; and the third, on classroom information about reading.

The *school characteristics and policies questionnaire* was given to the principal or other administrator in each participating school and took about 15 minutes 'o complete. The questions asked about the principal's background and experience, school policies, programs, facilities, and the demographic composition and background of the students and teachers.

The *IEP/LEP student questionnaire* was completed by the teachers of those students who were selected to participate in the Trial State Assessment sample but who were determined by the school to be ineligible to be assessed because they either had an Individualized Education Plan (IEP) and were not mainstreamed at least 50 percent of the time, or were categorized as Limited English Proficient (LEP). Each questionnaire took approximately three minutes to complete and asked about the nature of the student's exclusion and the special programs in which the student participated.

1.5 THE SAMPLING DESIGN

The target populations for the Trial State Assessment Program in reading consisted of fourth-grade students enrolled in public schools and nonpublic schools. The representative sample of public-school fourth graders assessed in the Trial State Assessment came from about 100 schools in each jurisdiction, unless a jurisdiction had fewer than 100 schools with a fourth grade, in which case all or almost all schools were asked to participate. The nonpublic-school samples differed in size across the jurisdictions, with the number of schools selected proportional to the nonpublic-school enrollment within each jurisdiction. On average, about 15 nonpublic schools were included for each jurisdiction. The school sample in each state was designed to produce aggregate estimates for the state and for selected subpopulations (depending upon the size and distribution of the various subpopulations within the state), and also to enable comparisons to be made, at the state level, between administration of assessment tasks with monitoring and without monitoring. The public schools were stratified by urbanization, percentage of Black and Hispanic students enrolled, and median household income. The nonpublic schools were stratified by type of control (Catholic, private/other religious, other nonpublic), metro status, and enrollment size per grade.

In most states, up to 30 students were selected from each school, with the aim of providing an initial target sample size of approximately 3,000 public-school students per state. The student sample size of 30 for each school was chosen to ensure that at least 2,000 public-school students participated from each state allowing for school nonresponse, exclusion of students, inaccuracies in the measures of enrollment, and student absenteeism from the assessment. In states with fewer schools, larger numbers of students per school were often required to ensure target samples of roughly 3,000 students. In certain jurisdictions, all eligible fourth graders were targeted for assessment.

Students within a school were sampled from lists of fourth-grade students. The decisions to exclude students from the assessment were made by school personnel, as in the national assessment, and were based on the same criteria for exclusion (described in section 1.4) used for the national assessment. Each excluded student was carefully accounted for to estimate the percentage of the state population deemed unassessable and the reasons for exclusion.

Chapter 3 describes the various aspects of selecting the sample for the 1994 Trial State Assessment—the construction of the public- and nonpublic-school frames, the stratification processes, the updating of the school frames with new schools, the actual sample selection, and the sample selection for the field test.

1.6 FIELD ADMINISTRATION

The administration of the 1994 program and the 1993 field test required collaboration between staff in the participating states and schools and the NAEP contractors, especially Westat, the field administration contractor. The purpose of the field test conducted in 1993 was to try out blocks of items that were to be used as replacements for the 1992 assessment blocks released to the public.

Each jurisdiction volunteering to participate in the 1993 field test and in the 1994 Trial State Assessment was asked to appoint a state coordinator as liaison between NAEP staff and the participating schools. In addition, Westat hired and trained a supervisor for each state and six field managers, each of which was assigned to work with groups of states. The state supervisors were responsible for working with the state coordinators, overseeing assessment activities, training school district personnel to administer the assessment, and coordinating the quality-control monitoring efforts. Each field manager was responsible for working with the state coordinators of 7-8 states and the supervision of the state supervisors assigned to those states. An assessment administrator was responsible for preparing for and conducting the assessment session in one or more schools. These individuals were usually school or district staff and were trained by Westat. Westat also hired and trained three to five quality control monitors in each state. For states that had previously participated in the state assessment program, 25 percent of the public-school sessions and 50 percent of the nonpublic-school sessions were monitored. For states new to the program, 50 percent of all sessions were monitored. During the field test, the state supervisors monitored all sessions.

Chapter 4 describes the procedures for obtaining cooperation from states and provides details about the field activities for both the field test and 1994 program. Chapter 4 also describes the planning and preparations for the actual administration of the assessment, the training and monitoring of the assessment sessions, and the responsibilities of the state coordinators, state supervisors, assessment administrators, and quality control monitors.

1.7 MATERIALS PROCESSING AND DATABASE CREATION

Upon completion of each assessment session, school personnel shipped the assessment booklets and forms to NAEP subcontractor National Computer Systems for professional scoring, entry into computer files, and checking. The files were then sent to Educational Testing Service for creation of the database. Careful checking assured that all data from the field were received. Chapter 5 describes the printing, distribution, receipt, processing, and final disposition of the 1994 Trial State Assessment materials.

The volume of collected data and the complexity of the Trial State Assessment processing design, with its spiraled distribution of booklets, as well as the concurrent administration of this assessment and the national assessments, required the development and implementation of flexible, innovatively designed processing programs, and a sophisticated Process Control System. This system, described in Chapter 5, allowed an integration of data entry and workflow management systems that included carefully planned and delineated editing, quality control, and auditing procedures.

Chapter 5 also describes the data transcription and editing procedures used to generate the disk and tape files containing various assessment information, including the sampling weights required to make valid statistical inferences about the population from which the Trial State Assessment sample was drawn. Before any analysis could begin, the data from these files underwent a quality control check at ETS. The files were then merged into a comprehensive, integrated database. Chapter 6 describes the transcribed data files, the procedure of merging them to create the Trial State Assessment database, and the results of the quality control process, and the procedures used to create data products for use in secondary research.

1.8 THE TRIAL STATE ASSESSMENT DATA

The basic information collected from the Trial State Assessment in reading consisted of the responses of the assessed students to 85 reading exercises organized around eight distinct reading passages. To limit the assessment time for each student to about one hour, a partially balanced incomplete block (PBIB) spiral design was used to assign a subset of the full exercise pool to each student. The partially balanced design differed slightly from the fully balanced incomplete block (BIB) spiral design used for the 1990 and 1992 Trial State Assessments in mathematics. Both the PBIB and BIB designs are variants of matrix sampling designs.

The full set of reading items was divided into eight unique blocks, each requiring 25 minutes for completion. Four of the blocks contained literary passages; the items accompanying these blocks were designed to assess student abilities in Reading for Literary Experience. The other four blocks were based on informational prose passages (e.g., magazine articles, newspaper articles, sections of textbook chapters, etc.); the items accompanying these passages were designed to assess student abilities in Reading to Gain Information. Each assessed student received a booklet containing two of the eight blocks according to a design that ensured that each block was administered to a representative sample of students within each jurisdiction. The design also ensured that each Reading for Literary Experience block was paired in exactly one booklet with every other Reading for Literary Experience block. Similarly, each Reading to Gain Information block was paired in exactly one booklet with every other Reading to Gain Information block. Furthermore, each Reading for Literary Experience block was paired in exactly one booklet with one of the Reading to Gain Information blocks. The data also included responses to the background questionnaires (described in section 1.4). Further details on the assembly of cognitive instruments and the data collection design can be found in Chapter 2.

The national data to which the Trial State Assessment results were compared were taken from nationally representative samples of public- and nonpublic-school students in the fourth grade. These samples were a part of the full 1994 national reading assessment, in which nationally representative samples of students in public and private schools from three age cohorts were assessed: students who were either in the fourth grade or 9 years old; students who were either in the eighth grade or 13 years old; and students who were either in the twelfth grade or 17 years old.

The assessment instruments used in the Trial State Assessment were also used in the fourth-grade national assessments and were administered using the identical procedures in both assessments. The time of testing for the state assessments (January 31 to February 25, 1994) occurred within the time of testing of the national assessment (January 3 to April 1, 1994). The state assessments differed from the national assessment, however, in one important regard: Westat staff collected the data for the national assessment while, in accordance with the NAEP legislation, data collection activities for the Trial State Assessment were the responsibility of each participating jurisdiction. The data collection activities included ensuring the participation of selected schools and students, assessing students according to standardized procedures, and observing procedures for test security.

1.9 WEIGHTING AND VARIANCE ESTIMATION

A complex sample design was used to select the students to be assessed in each of the participating jurisdictions. The properties of a sample from a complex design are very different from those of a simple random sample in which every student in the target population has an equal chance of selection and in which the observations from different sampled students can be considered to be statistically independent of one another. The properties of the sample from the complex Trial State Assessment design were taken into account in the analysis of the assessment data.

One way that the properties of the sample design were addressed was by using sampling weights to account for the fact that the probabilities of selection were not identical for all students. These weights included adjustments for school and student nonresponse. All population and subpopulation characteristics based on the Trial State Assessment data used sampling weights in their estimation. Chapter 7 provides details on the computation of these weights.

In addition to deriving appropriate estimates of population characteristics, it is essential to obtain appropriate measures of the degree of uncertainty of those statistics. One component of uncertainty is a result of sampling variability, which measures the dependence of the results on the particular sample of students actually assessed. Because of the effects of cluster selection (schools are selected first, then students are selected within those schools), observations made on different students cannot be assumed to be independent of each other (and, in fact, are generally positively correlated). As a result, classical variance estimation formulas will produce incorrect results. Instead, a variance estimation procedure that takes the characteristics of the sample into account was used for all analyses. This procedure, called jackknife variance estimation, is discussed in Chapter 7.

Jackknife variance estimation provides a reasonable measure of uncertainty for any statistic based on values observed without error. Statistics such as the average proportion of students correctly answering a given question meet this requirement, but other statistics based on estimates of student reading proficiency, such as the average reading proficiency of a subpopulation, do not. Because each student typically responds to relatively few items within a particular purpose of reading (i.e., for literary experience or to gain information), there exists a nontrivial amount of imprecision in the measurement of the proficiency of a given student. This imprecision adds an additional component of variability to statistics based on estimates of individual proficiencies. The estimation of this component of variability is discussed in Chapter 8.

1.10 PRELIMINARY DATA ANALYSIS

After the computer files of student responses were received from NCS, all cognitive and noncognitive items were subjected to an extensive item analysis. Each block of cognitive items was subjected to item analysis routines, which yielded for each item the number of respondents, the percentage of responses in each category (100 x item score), the percentage who omitted the item, the percentage who did not reach the item, and the correlation between the item score and the item block score (r-polyserial). In addition, the item analysis program provided summary statistics for each block, including reliability (internal consistency) coefficient. These analyses were used to check on the scoring of the items, to verify the appropriateness of the difficulty level of the items, and to check for speededness. The results also were reviewed by knowledgeable project staff in search of aberrations that might signal unusual results or errors in the database.

Tables of the weighted percentages of students with responses in each category of each cognitive and background item were created and distributed to each state and jurisdiction. Additional analyses comparing the data from the monitored sessions with those from the unmonitored sessions were conducted to determine the comparability of the assessment data from the two types of administrations. Differential item functioning (DIF) analyses were carried out to identify items that were differentially difficult for various subgroups and to reexamine such items with respect to their fairness and their appropriateness for inclusion in the scaling process. Further details of the preliminary analyses appear in Chapter 9.

1.11 SCALING THE ASSESSMENT ITEMS

The primary analysis and reporting of the results from the Trial State Assessment used item response theory (IRT) scale-score models. Scaling models quantify a respondent's tendency to provide correct answers to the domain of items contributing to a scale as a function of a parameter called proficiency. Proficiency can be viewed as a summary measure of performance across the domain of items that make up the scale. Three distinct IRT models were used for scaling: 1) 3-parameter logistic models for multiple-choice items; 2) 2-parameter logistic models for short constructed-response items that were scored correct or incorrect; and 3) generalized partial credit models for short and extended constructed-response items that were scored on a multipoint scale. Chapter 8 provides an overview of the scaling models used. Further details on the application of these models are provided in Chapter 9.

Two distinct scales were created for the Trial State Assessment to summarize fourth-grade students' reading abilities according to two purposes for reading: Reading for Literary Experience and Reading to Gain Information. For reasons discussed in Chapter 9, these scales were defined identically to, but separately from, those used for the scaling of the national NAEP fourth-grade reading data. Although the items comprising each scale were identical to those used for the national program, the item parameters for the Trial State Assessment scales were estimated from combined data from all jurisdictions participating in the Trial State Assessment. Item parameter estimation was based on an item calibration sample consisting of an approximately 25 percent sample of all available data. To ensure equal representation in the scaling process, each jurisdiction was equally represented in the item calibration sample, as were monitored and unmonitored administrations from each jurisdiction. Chapter 9 provides further details about item parameter estimation.

The fit of the IRT model to the observed data was examined within each scale by comparing the estimates of the empirical item characteristic functions with the theoretic curves. For binary-scored items, nonmodel-based estimates of the expected proportions of correct responses to each item for students with various levels of scale proficiency were compared with the fitted item response curve; for the short and extended partial-credit constructed-response items, the comparisons were based on the expected proportions of students with various levels of scale proficiency who achieved each score level. In general, the item level results were well fit by the scaling models.

Using the item parameter estimates, estimates of various population statistics were obtained for each jurisdiction. The NAEP methods use random draws ("plausible values") from estimated proficiency distributions for each student to compute population statistics. Plausible values are not optimal estimates of individual student proficiencies; instead, they serve as intermediate values to be used in estimating population characteristics. Under the assumptions of the scaling models, these population estimates will be consistent, in the sense that the estimates approach the model-based population values as the sample size increases, which would not be the case for subpopulation estimates obtained by aggregating optimal estimates of individual proficiency. Chapter 8 provides further details on the computation and use of plausible values.

In addition to the plausible values for each scale, a composite of the two reading scales was created as a measure of overall reading proficiency. This composite was a weighted average of the plausible values for the two reading scales, in which the weights were proportional to the relative importance assigned to each purpose in the reading objectives. The definition of the composite for the Trial State Assessment program was identical to that used for the national fourth-grade reading assessment. More details about composite scores are given in Chapter 9.

1.12 LINKING THE TRIAL STATE RESULTS TO THE NATIONAL RESULTS

A major purpose of the Trial State Assessment Program was to allow each participating jurisdiction to compare its 1994 results with the nation as a whole and with the region of the country in which that jurisdiction is located. For meaningful comparisons to be made between each of the Trial State Assessment jurisdictions and the relevant national sample, results from these two assessments had to be expressed in terms of a similar system of scale units.

The results from the Trial State Assessment were linked to those from the national assessment through linking functions determined by comparing the results for the aggregate of all students assessed in the Trial State Assessment with the results for fourth-grade students in the State Aggregate Comparison (SAC) subsample of the national NAEP. The SAC subsample of the national NAEP is a representative sample of the population of all grade-eligible public-school students within the aggregate of 41 participating states and the District of Columbia (Guam and the Department of Defense Overseas Education Activity schools were not included in the aggregate). Specifically, the grade 4 SAC subsample consists of all fourth-grade students in public schools in the states and the District of Columbia who were assessed in the national cross-sectional reading assessment.

A linear transformation within each scale was used to link the results of the Trial State Assessment to the national assessment. The adequacy of linear linking was evaluated by comparing, for each scale (Reading for Literary Experience and Reading to Gain Information), the distribution of reading proficiency based on the aggregation of all assessed students at each grade from the participating states and the District of Columbia with the equivalent distribution based on the students in the SAC subsample. In the estimation of these distributions, the students were weighted to represent the target population of public-school students in the specified grade in the aggregation of the states and the District of Columbia. If a linear linking were adequate, the distribution for the aggregate of states and the District of Columbia and that for the SAC subsample would have, to a close approximation, the same shape in terms of the skewness, kurtosis, and higher moments of the distributions. The only differences in the distributions allowed by linear linking would be in the means and variances. Generally, this was found to be the case.

Each reading scale was linked by matching the mean and standard deviation of the scale proficiencies across all students in the Trial State Assessment (excluding Guam and the Department of Defense Overseas Activity Schools) to the corresponding scale mean and standard deviation across all students in the SAC subsample. Further details of the linking are given in Chapter 9.

1.13 REPORTING THE TRIAL STATE ASSESSMENT RESULTS

Each jurisdiction in the Trial State Assessment received a summary report providing the state's results with accompanying text and tables, national and regional comparisons, and (for states that had participated in the 1992 state program) trend comparisons to the previous assessment. These reports were generated by a computerized report-generation system for which graphic designers, statisticians, data analysts, and report writers collaborated to develop shells of the reports in advance of the analysis. These prototype reports were provided to State Education Agency personnel for their reviews and comments. The results of the data analysis were then automatically incorporated into the reports, which displayed tables and graphs of the results and interpretations of those results, including indications of subpopulation comparisons of statistical and substantive significance.

Each report contained state-level estimates of mean proficiencies, both for the state as a whole and for categories of the key reporting variables: gender, race/ethnicity, level of parental education, and type of location. Results were presented for each reading proficiency scale, for the overall reading composite, and by achievement levels. Results were also reported for a variety of other subpopulations based on variables derived from the student, teacher, and school questionnaires. Standard errors are included for all statistics.

A second report, *1994 NAEP Reading: A First Look*, was released in April of 1995 (several months prior to the release of the state reports and the other documents described below), presenting selected national and state public-school findings for the composite reading proficiency scale. The report compared 1994 results to 1992 results and included findings with respect to the reading achievement levels established by the National Assessment Governing Board.

A third report, the *NAEP 1994 Reading Report Card for the Nation and the States*, highlighted key assessment results for the nation and summarized results across the states and territories participating in the assessment. This report contained composite scale results (proficiency means, proportions at or above achievement levels, etc.) for the nation, for each of the four regions of the country, and for each jurisdiction in the Trial State Assessment, both overall, by the primary demographic reporting variables, and for both public and nonpublic schools. In addition, results were reported for each of the reading scales.

The fourth report, entitled *Cross-state Data Compendium for the NAEP 1994 Reading Assessment*, contains state-by-state results for all variables reported on in the *Report Card* and *State Report* .

The fifth report is a six-section almanac. The first section, or "distribution" section, provides results for the achievement levels and percentiles. Three of the sections of the almanac (referred to as proficiency sections) present summary tables based on responses to each of the questionnaires (student, reading teacher, and school) administered as part of the Trial State Assessment. The fifth section of the almanac, the scale section, reports proficiency means and associated standard errors for the two purpose-for-reading scales. Results in this section are reported for the total group in each state, as well as for select subgroups of interest. The final section of the almanac, the "p-value" section, provides the total-group proportion for each response alternative for each cognitive item included in the assessment.

The production of the state reports, *Reading Report Card*, *Cross-State Data Compendium*, and the almanacs required a large number of decisions about a variety of data analysis and statistical issues. Chapter 10 documents the major conventions and statistical procedures used in generating the reports and almanacs. The chapter describes the rules, based on effect size and school and student sample size considerations, that were used to establish whether a particular category contained data sufficient to report reliable results for a particular state. Chapter 10 also describes the multiple comparison and effect-size-based inferential rules that were used for evaluating the statistical and substantive significance of subpopulation comparisons.

To provide information about the generalizability of the results, a variety of information about participation rates was reported for each jurisdiction. This information included school participation rates, both in terms of the initially selected samples of schools and in terms of the finally achieved samples, including replacement schools. The student participation rates, the rates of students excluded due to Limited English Proficiency (LEP) and Individualized Education Plan (IEP) status, and the estimated proportions of assessed students who are classified as IEP or LEP were also reported for each state. These rates are described and reported in Appendix B.

Chapter 2

DEVELOPING THE OBJECTIVES, COGNITIVE ITEMS, BACKGROUND QUESTIONS, AND ASSESSMENT INSTRUMENTS

Jay R. Campbell and Patricia L. Donahue

Educational Testing Service

2.1 OVERVIEW

The framework that was developed for the 1992 NAEP Trial State Assessment in reading also served as the framework for the 1994 assessment. Similar to all previous NAEP assessments, the objectives in reading were developed through a broad-based consensus process. To prepare the framework and objectives, initially for the 1992 reading assessment, the National Assessment Governing Board (NAGB) contracted with the Council of Chief State School Officers (CCSSO). The development process involved a steering committee, a planning committee, and CCSSO project staff. Educators, scholars, and citizens, representative of many diverse constituencies and points of view, participated in the national consensus process to design objectives for the reading assessment.

The instrument used in the 1994 reading assessment was composed of a combination of reading passages and questions from the 1992 assessment and a set of passages and questions newly developed for 1994. Those passages and questions carried over from the 1992 instrument comprised two-thirds of the 1994 instrument. The remaining third was made up of new passages and questions developed according to the same framework that was used for the 1992 assessment. Maintaining two-thirds of the instrument across the two assessment years allowed for the reporting of trends in reading performance. At the same time, developing a new set of passages and questions made it possible to release one-third of the 1992 assessment for public use.

In developing the new portion of the 1994 NAEP reading assessment, the same framework, objectives, and procedures used in 1992 were followed. After careful reviews of the objectives, reading materials were selected and questions were developed that were appropriate to the objectives. All questions underwent extensive reviews by specialists in reading, measurement, and bias/sensitivity, as well as reviews by state representatives.

The objectives and question development effort were governed by four major considerations:

- The objectives for the reading assessment had to be developed through a consensus process, involving subject-matter experts, school administrators, teachers, and parents.

- As outlined in the ETS proposal for the administration of the NAEP contract (ETS, 1992), the development of the items had to be guided by a Reading Instrument Development Panel and receive further review by state representatives and classroom teachers from across the country. In addition, the items had to be carefully reviewed for potential bias.

- As described in the ETS Standards of Quality and Fairness (ETS, 1987), all materials developed at ETS had to be in compliance with specified procedures.

- As per federal regulations, all NAEP cognitive and background items had to be submitted to a federal clearance process.

This chapter includes details about developing the objectives and items for the Trial State Assessment in reading. The chapter also describes the instruments, the student assessment booklets, reading teacher questionnaire, school characteristics and policies questionnaire, and IEP/LEP student questionnaire. Various committees worked on the development of the framework, objectives, and items for the reading assessment. The list of committee members and consultants who participated in the 1994 development process is provided in Appendix A.

2.2 FRAMEWORK AND ASSESSMENT DESIGN PRINCIPLES

The reading framework for the 1992 and 1994 assessments was developed according to guidelines established by the Steering Committee. These guidelines determined that the design of the framework be performance-oriented with a focus on reading processes. The framework would embody a broad view of reading that addressed the high levels of literacy needed for employability, personal development, and citizenship. Also, the framework would take into account contemporary research on reading and literacy and would expand the range of assessment tools to include formats that more closely resembled desired classroom activities.

The objectives development was guided by the consideration that the assessment should reflect many of the states' curricular emphases and objectives in addition to what various scholars, practitioners, and interested citizens believed should be included in the curriculum. Accordingly, the committee gave attention to several frames of reference:

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

- The term "reading literacy" should be used in the broad sense of knowing when to read, how to read, and how to reflect on what has been read. It represents a complex, interactive process that goes beyond basic or functional literacy.

- The reading assessment should use authentic passages and tasks that are both broad and complete in their coverage of important reading behaviors so that the assessment tool will demonstrate a close link to desired classroom instruction and students' reading experiences.

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

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

2.3 FRAMEWORK DEVELOPMENT PROCESS

The National Assessment Governing Board is responsible for guiding NAEP, including the development of the reading assessment objectives and test specifications. Appointed by the Secretary of Education from lists of nominees proposed by the board itself in various statutory categories, the 24-member board is compoced of state, local, and federal officials, as well as educators and members of the public.

NAGB began the development process for the 1992 reading objectives (that also served as objectives for the 1994 assessment) by conducting a widespread mail review of the objectives for the 1990 reading assessment and by holding a series of public hearings throughout the country. The contract for managing the remainder of the consensus process was awarded to the Council of Chief State School Officers. The development process included the following activities:

- A Steering Committee consisting of members recommended by each of 15 national organizations (see Appendix A) was established to provide guidance for the consensus process. The committee responded to the progress of the project and offered advice. Drafts of each version of the document were sent to members of the committee for review and reaction.

- A Planning Committee (see Appendix A) was established to identify the objectives to be assessed in reading in 1992, and subsequently in 1994, and to prepare the framework document. The members of this committee consisted of experts in reading, including college professors, an academic dean, a classroom teacher, a school administrator, state-level assessment and reading specialists, and a representative of the business community. This committee met with the Steering Committee and as a separate group. A subgroup also met to develop item specifications. Between meetings, members of the committee provided information and reactions to drafts of the framework.

- The project staff at the Council of Chief State School Officers met regularly with staff from the National Assessment Governing Board and the National Center for Education Statistics to discuss progress made by the Steering and Planning committees.

During this development process, input and reactions were continually sought from a wide ıange of members of the reading field, experts in assessment, school administrators, and state staff in reading assessment. In particular, the process was informed by innovative state assessment efforts and work being done by the Center for the Learning and Teaching of Literature (Langer, 1989, 1990).

2.4 FRAMEWORK FOR THE ASSESSMENT

The framework adopted for the 1992 reading assessment and used for developing new portions of the 1994 instrument is organized according to a four-by-three matrix of reading *stances* by reading *purposes*. These stances included:

- Initial Understanding,
- Developing an Interpretation,
- Personal Reflection and Response, and
- Demonstrating a Critical Stance.

These stances were assessed across three global purposes defined as:

- Reading for Literary Experience,
- Reading to Gain Information, and
- Reading to Perform a Task.

Different types of texts were used to assess the various purposes for reading. Students' reading abilities were evaluated in terms of a single purpose for each type of text. At grade 4, only reading for literary experience and reading to gain information were assessed, while all three global purposes were assessed at grades 8 and 12. Figures 2-1 and 2-2 describe the four reading stances and three reading purposes that guided the development of the 1992 and 1994 Trial State Assessments in reading. The interactions among the aspects of reading assessed are presented in Figure 2-3.

2.5 DISTRIBUTION OF ASSESSMENT ITEMS

In recognition that the demands made of readers change as readers move from grade to grade, the Planning Committee recommended that the proportion of items related to each of the reading purposes vary according to grade level. The relative contribution of each reading purpose to the overall proficiency score is presented in Table 2-1. The weighting of each reading purpose scale changes from grade to grade to reflect the changing demands made of students as they mature.

Figure 2-1
Description of Reading Stances

Readers interact with text in various ways as they use background knowledge and understanding of text to construct, extend, and examine meaning. The NAEP reading assessment framewo:k specified four reading stances to be assessed that represent various interactions between readers and texts. These stances are not meant to describe a hierarchy of skills or abilities. Rather, they are intended to describe behaviors that readers at all developmental levels should exhibit.

Initial Understanding

Initial understanding requires a broad, preliminary construction of an understanding of the text. Questions testing this aspect ask the reader to provide an initial impression or unreflected understanding of what was read. In the 1992 and 1994 NAEP reading assessments, the first question following a passage was usually one testing initial understanding.

Developing an Interpretation

Developing an interpretation requires the reader to go beyond the initial impression to develop a more complete understanding of what was read. Questions testing this aspect require a more specific understanding of the text and involve linking information across parts of the text as well as focusing on specific information.

Personal Reflection and Response

Personal response requires the reader to connect knowledge from the text more extensively with his or her own personal background knowledge and experience. The focus is on how the text relates to personal experience; questions on this aspect ask the readers to reflect and respond from a personal perspective. For the 1992 and 1994 NAEP reading assessments, personal response questions were typically formatted as constructed-response items to allow for individual interpretations and varied responses.

Demonstrating a Critical Stance

Demonstrating a critical stance requires the reader to stand apart from the text, consider it, and judge it objectively. Questions on this aspect require the reader to perform a variety of tasks such as critical evaluation, comparing and contrasting, application to practical tasks, and understanding the impact of such text features as irony, humor, and organization. These questions focus on the reader as interpreter/critic and require reflection and judgments.

Figure 2-2
Description of Purposes for Reading

Reading involves an interaction between a specific type of text or written material and a reader, who typically has a purpose for reading that is related to the type of text and the context of the reading situation. The 1992 and 1994 NAEP reading assessments presented three types of text to students representing each of three reading purposes: literary text for literary experience, informational text to gain information, and documents to perform a task. Students' reading abilities were evaluated in terms of a single purpose for each type of text.

Reading for Literary Experience

Reading for literary experience involves reading literary text to explore the human condition, to relate narrative events with personal experiences, and to consider the interplay in the selection among emotions, events, and possibilities. Students in the NAEP reading assessment were provided with a wide variety of literary text, such as short stories, poems, fables, historical fiction, science fiction, and mysteries.

Reading to Gain Information

Reading to gain information involves reading informative passages in order to obtain some general or specific information. This often requires a more utilitarian approach to reading that requires the use of certain reading/thinking strategies different from those used for other purposes. In addition, reading to gain information often involves reading and interpreting adjunct aids such as charts, graphs, maps, and tables that provide supplemental or tangential data. Informational passages in the NAEP reading assessment included biographies, science articles, encyclopedia entries, primary and secondary historical accounts, and newspaper editorials.

Reading to Perform a Task

Reading to perform a task involves reading various types of materials for the purpose of applying the information or directions in completing a specific task. The reader's purpose for gaining meaning extends beyond understanding the text to include the accomplishment of a certain activity. Documents requiring students in the NAEP reading assessment to perform a task included directions for creating a time capsule, a bus schedule, a tax form, and instructions on how to write a letter to a senator. Reading to perform a task was assessed only at grades 8 and 12.

Figure 2-3
1992 and 1994 NAEP Framework — Aspects of Reading Literacy

Purposes for Reading	Reading Stances			
	Initial Understanding Requires the reader to provide an initial impression or unreflected understanding of what was read.	**Developing an Interpretation** Requires the reader to go beyond the initial impression to develop a more complete understanding of what was read.	**Personal Reflection and Response** Requires the reader to connect knowledge from the text with his/her own personal background knowledge.	**Demonstrating a Critical Stance** Requires the reader to stand apart from the text and consider it.
Reading for Literary Experience	What is the story/plot about? How would you describe the main character?	How did the plot develop? How did this character change from the beginning to the end of the story?	How did this character change your idea of ________? Is this story similar to or different from your own experiences?	Rewrite this story with _______ as a setting or _______ as a character. How does this author's use of _______ (irony, personification, humor) contribute to _______?
Reading to Gain Information	What does this article tell you about ________? What does the author think about this topic?	What caused this event? In what ways are these ideas important to the topic or theme?	What current event does this remind you of? Does this description fit what you know about _______? Why?	How useful would this article be for _______ ? Explain. What could be added to improve the author's argument?
Reading to Perform a Task	What is this supposed to help you do? What time can you get a non-stop flight to X?	What will be the result of this step in the directions? What must you do before this step?	In order to _______, what information would you need to find that you don't know right now? Describe a situation where you could leave out step X.	Why is this information needed? What would happen if you omitted this?

Table 2-1
Weighting of the Reading Purpose Scales on the Composite Reading Scale

Grade	Purposes for Reading		
	For Literary Experience	To Gain Information	To Perform a Task
4	55%	45%	(No Scale)
8	40%	40%	20%
12	35%	45%	20%

Table 2-2
Percentage Distribution of Items by Reading Stance for All Grades
as Specified by the Reading Framework

Initial Understanding/ Developing an Interpretation	Personal Reflection and Response	Demonstrating a Critical Stance
33%	33%	33%

Readers use a range of cognitive abilities and assume various stances as they engage in various reading experiences. In the 1992 and 1994 NAEP reading assessments, four stances were assessed within each of the reading purposes. While reading, students form an initial understanding of the text and connect ideas within the text to generate interpretations. In addition, they extend and elaborate their understanding by responding to the text personally and critically and by relating ideas in the text to prior experiences or knowledge. In accordance with development specifications, items were developed to fulfill the reading stance requirements. Table 2-2 shows the distribution of items by reading stances for all three grade levels, as specified in the NAEP Reading Framework. The distribution requirements for the exercise specifications combined the stances of initial understanding and developing an interpretation.

2.6 DEVELOPING THE COGNITIVE ITEMS

The development of cognitive items began with a careful selection of grade-appropriate passages for the assessment. Passages were selected from a pool of reading selections contributed by teachers from across the country. The framework stated that the assessment passages should represent authentic, naturally occurring reading material that students may encounter in or out of school. These passages were reproduced in test booklets as they had appeared in their original publications. Final passage selections were made by the Reading Instrument Development Panel. Lastly, in order to guide the development of items, passages were outlined or mapped to identify essential elements of the text.

The Trial State Assessment included constructed-response (short and extended) and multiple-choice items. The decision to use a specific item type was based on a consideration of the most appropriate format for assessing the particular objective. Both types of constructed-response items were designed to provide an in-depth view of students' ability to read thoughtfully and generate their own responses to reading. Short constructed-response questions (scored with either a 2- or 3- level scoring rubric) were used when students needed to respond in only one or two sentences in order to demonstrate full comprehension. Extended constructed-response questions (scored with a 4-level scoring rubric) were used when the task required more thoughtful consideration of the text and engagement in more complex reading processes. Multiple-choice items were used when a straightforward, single correct answer was all that was required. Guided by the NAEP reading framework, the Instrument Development Panel monitored the development of all three types of items to assess objectives in the framework. For more information about item scoring, see Chapter 5.

The Trial State Assessment at grade 4 included eight different 25-minute "blocks," each consisting of one reading passage and a set of multiple-choice and constructed-response items to assess students' comprehension of the written material. Students were asked to respond to two 25-minute blocks within one booklet. Four blocks assessed reading for literary experience and four assessed reading to gain information. Even though the number of items varied within each block, the amount of assessment time was the same for each block and each reading purpose.

As with the 1992 instrument development effort, a detailed series of steps was used to create the new assessment items for 1994 that reflected the objectives.

1. Item specifications and prototype items were provided in the *1992 and 1994 Reading Framework.*

2. The Reading Instrument Development Panel 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 Panel made recommendations about priorities for the assessment and types of items to be developed.

3. Passages were chosen for the assessment through an extensive selection process that involved the input of teachers from across the country as well as the Reading Instrument Development Panel.

4. Item writers from both inside and outside ETS were selected based on their knowledge about reading theory and practices and experience in creating items according to specifications.

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

6. Passages and items were reviewed by grade-appropriate teachers across the country for developmental appropriateness.

7. Representatives from the State Education Agencies met and reviewed all items and background questionnaires (see section 2.8 for a discussion of the background questionnaires).

8. Language editing and sensitivity reviews were conducted according to ETS quality control procedures.

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

10. The field test was conducted in 23 states, the District of Columbia, and three territories.

11. Representatives from State Education Agencies met and reviewed the field test results.

12. Based on the field test analyses, new items for the 1994 assessment were revised, modified, and re-edited, where necessary. The items once again under went ETS sensitivity review.

13. The Reading Instrument Development Panel selected the blocks to include in the 1994 assessment.

14. 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. In total, the items that appeared in the Trial State Assessment underwent 86 separate reviews, including reviews by NAEP/ETS staff, external reviewers, State Education Agency representatives, and federal officials.

The overall pool of items for the Trial State Assessment consisted of 84 items, including 37 short constructed-response items, 8 extended constructed-response items, and 39 multiple-choice items. Table 2-3 provides the percentage of assessment time (based on field observations and basic assumptions made in item development) devoted to each reading stance within the two purposes for reading.

Table 2-3
Percentage of Assessment Time Devoted to the Reading Stances
Within Each Purpose for Reading for the 1994 Reading Trial State Assessment

		Initial Understanding/ Developing an Interpretation		**Personal Reflection and Response**		**Demonstrating a Critical Stance**	
Grade	**Purpose for Reading**	**Target**	**Actual***	**Target**	**Actual***	**Target**	**Actual***
4	For Literary Experience	33%	45%	33%	22%	33%	33%
	To Gain Information	33%	52%	33%	27%	33%	20%
	Overall	33%	49%	33%	25%	33%	27%

*Actual percentages are based on the classifications agreed upon by NAEP's Instrument Development Panel.

2.7 STUDENT ASSESSMENT BOOKLETS

Each student assessment booklet included two sections of cognitive reading items and three sections of background questions. The assembly of reading blocks into booklets and their subsequent assignment to sampled students was determined by a *partially balanced incomplete block* (PBIB) design with *spiraled* administration.

The first step in implementing PBIB spiraling for the grade 4 reading assessment required constructing blocks of passages and items that required 25 minutes to complete. These blocks were then assembled into booklets containing two 5-minute background sections, one 3-minute motivation questionnaire, and two 25-minute blocks of reading passages and items according to a partially balanced incomplete block design. The overall assessment time for each student was approximately 63 minutes.

At the fourth-grade level, the blocks measured two purposes for reading—reading for literary experience and reading to gain information. The reading blocks were assigned to booklets in such a way that every block within a given purpose for reading was paired with every other block measuring the same purpose but was only paired with one block measuring the other purpose for reading. Every block appears in four booklets, three times within booklets measuring the same purpose and once in a booklet measuring both purposes. This is the *partially balanced* part of the balanced incomplete block design.

The PBIB design for the both the 1992 and 1994 national reading assessments (and also for the trial state assessments) was *focused*, in that each block was paired with every other reading block assessing the same purpose for reading but not with all the blocks assessing the other purpose for reading. The *focused*-PBIB design also balances the order of presentation of the blocks of items—every block appears as the first cognitive block in two booklets and as the second cognitive block in two other booklets. This design allows for some control of context effects (see Chapter 9).

The design required that eight blocks of grade 4 reading items be assembled into sixteen booklets. The assessment booklets were then *spiraled* and bundled. Spiraling involves interweaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in each position in a bundle.

The final step in the PBIB-spiraling procedure was the assigning of the booklets to the assessed students. The students within an assessment session were assigned booklets in the order in which the booklets were bundled. Thus, students in an assessment session received different booklets, and only a few students in a session received the same booklet. Across all jurisdictions in the Trial State Assessment, representative and randomly equivalent samples of about 25,625 students responded to each item.

Table 2-4 provides the composition of each block of items administered in the Trial State Assessment Program in reading. Table 2-5 shows the order of the blocks in each booklet and how the 8 cognitive blocks were arranged across the 16 booklets to achieve the PBIB-spiral design. The 1994 design was identical to that used in 1992. The two new blocks that were developed for the 1994 assessment at grade 4 (R8 and R9) were arranged within the booklet design in the same manner as were the 1992 blocks that they replaced.

2.8 QUESTIONNAIRES

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

Table 2-4
Cognitive and Noncognitive Block Information

Block	Type	Total Number of Items	Number of Multiple-choice Items	Number of Constructed-response Items	Booklets Containing Block
B1	Common Background	22	22	0	30 - 45
R2	Reading Background	15	15	0	30 - 45
RB	Reading Motivation	5	5	0	30 - 45
R3	Reading for Literary Experience	11	6	5	30, 31, 35, 43
R4	Reading for Literary Experience	12	5	7	30, 33, 34, 42
R5	Reading for Literary Experience	11	7	4	31, 32, 34, 44
R6	Reading to Gain Information	10	5	5	36, 39, 40, 44
R7	Reading to Gain Information	10	4	6	37, 38, 40, 42
R8*	Reading to Gain Information	9	3	6	38, 39, 41, 43
R9*	Reading for Literary Experience	9	3	6	32, 33, 35, 45
R10	Reading to Gain Information	12	6	6	36, 37, 41, 45

* New blocks for the 1994 assessment.

Table 2-5
Booklet Contents

Booklet Number	Common Background Block	Cognitive Blocks		Reading Background Block	Reading Motivation Block
		1st	2nd		
R1	B1	R4	R3	R2	RB
R2	B1	R3	R5	R2	RB
R3	B1	R5	R9	R2	RB
R4	B1	R9	R4	R2	RB
R5	B1	R4	R5	R2	RB
R6	B1	R3	R9	R2	RB
R7	B1	R6	R10	R2	RB
R8	B1	R10	R7	R2	RB
R9	B1	R7	R8	R2	RB
R10	B1	R8	R6	R2	RB
R11	B1	R6	R7	R2	RB
R12	B1	R10	R8	R2	RB
R13	B1	R7	R4	R2	RB
R14	B1	R8	R3	R2	RB
R15	B1	R5	R6	R2	RB
R16	B1	R9	R10	R2	RB

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

2.8.1 Student Questionnaires

In addition to the cognitive questions, the 1994 Trial State Assessment included three student questionnaires. Two of these were five-minute sets of general and reading background questions designed to gather contextual information about students, their instructional and recreational experiences in reading, and their attitudes toward reading. The third, a three-minute questionnaire, was given to students at the end of each booklet to determine students' motivation in completing the assessment and their familiarity with assessment tasks. In order to ensure that all fourth-grade students understood the questions and had every opportunity to respond to them, the three questionnaires were read aloud by administrators as students read along and responded in their booklets.

The **student demographics (common core) questionnaire** (22 questions) included questions about race/ethnicity, language spoken in the home, mother's and father's level of education, reading materials in the home, homework, attendance, which parents live at home, and which parents work. This questionnaire was the first section in every booklet. In many cases the questions used were continued from prior assessments, so as to document changes in contextual factors that occur over time.

Three categories of information were represented in the second five-minute section of reading background questions called the **student reading questionnaire** (14 questions):

> *Time Spent Studying Reading*: Students were asked to describe both the amount of instruction they received in reading and the time spent on reading homework.
>
> *Instructional Practices*: Students were asked to report their instructional experiences related to reading in the classroom, including group work, special projects, and writing in response to reading. In addition, they were asked about the instructional practices of their reading teachers and the extent to which the students themselves discussed what they read in class and demonstrated use of skills and strategies.
>
> *Attitudes Towards Reading*: Students were asked a series of questions about their attitudes and perceptions about reading, such as whether they enjoyed reading and whether they were good in reading.

The **student motivation questionnaire** (5 questions) asked students to describe how hard they tried on the NAEP reading assessment, how difficult they found the assessment, how many questions they thought they got right, how important it was for them to do well, and how familiar they were with the assessment format.

2.8.2 Teacher, School, and IEP/LEP Student Questionnaires

To supplement the information on instruction reported by students, the reading teachers of the fourth graders participating in the Trial State Assessment were asked to complete a questionnaire about their instructional practices, teaching backgrounds, and characteristics. The teacher questionnaire contained two parts. The first part pertained to the teachers' background and general training. The second part pertained to specific training in teaching reading and the procedures the teacher uses for *each class* containing an assessed student.

The Teacher Questionnaire, Part I: Background and General Training (25 questions) included questions pertaining to gender, race/ethnicity, years of teaching experience, certification, degrees, major and minor fields of study, course work in education, course work in specific subject areas, amount of in-service training, extent of control over instructional issues, and availability of resources for their classroom.

The **Teacher Questionnaire, Part II: Training in Reading and Classroom Instructional Information** (46 questions) included questions on the teacher's exposure to various issues related to reading and teaching reading through pre- and in-service training, ability level of students in the class, whether students were assigned to the class by ability level, time on task, homework assignments, frequency of instructional activities used in class, methods of assessing student progress in reading, instructional emphasis given to the reading abilities covered in the assessment, and use of particular resources.

A **School Characteristics and Policies Questionnaire** was given to the principal or other administrator of each school that participated in the trial state assessment program. This information provided an even broader picture of the instructional context for students' reading achievement. This questionnaire (64 questions) included questions about background and characteristics of school principals, length of school day and year, school enrollment, absenteeism, drop-out rates, size and composition of teaching staff, policies about grouping students, curriculum, testing practices and uses, special priorities and school-wide programs, availability of resources, special services, community services, policies for parental involvement, and school-wide problems.

The **IEP/LEP Student Questionnaire** was completed by the teachers of those students who were selected to participate in the trial state assessment sample but who were determined by the school to be ineligible to be assessed. In order to be excluded from the assessment, students must have had an Individualized Education Plan (IEP) and had not mainstreamed at least 50 percent of the time or were categorized as Limited English Proficient (LEP). In addition, the school staff would have needed to determine that it was inappropriate to include these students in the assessment. This questionnaire asked about the nature of the student's exclusion and the special programs in which the student participated.

2.9 DEVELOPMENT OF FINAL FORMS

The field tests of new items for the 1994 assessment were conducted in February and March 1993 and involved 6,800 students in 233 schools in 23 states, the District of Columbia, and three territories. The intent of the field test was to try out the items and procedures and to give the states and the contractors practice and experience with the proposed materials and procedures. About 500 responses were obtained to each item in the field test.

The field test data were collected, scored, and analyzed in preparation for meetings with the Reading Instrument Development Panel. Four objectives guided these reviews: to determine which items were most suitable for assessing reading comprehension in accordance with the framework; to determine the need for revisions of items that lacked clarity, or had ineffective item formats; to prioritize items to be included in the Trial State Assessment; and to determine appropriate timing for assessment items. Committee members, ETS test development staff, and NAEP/ETS staff reviewed the materials. Item analyses (which provided the mean percentage of correct responses, the r-biserial correlations, and the difficulty level for each item) were used as a guide in identifying and flagging for further review those test items that were not measuring the intended objective well. In addition, another meeting of representatives from state education agencies was convened to review the field test results.

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

Chapter 3

SAMPLE DESIGN AND SELECTION

James L. Green, John Burke, and Keith F. Rust

Westat, Inc.

3.1 OVERVIEW

For the 1994 Trial State Assessment in reading, a combined sample of approximately 2,800 fourth-grade public- and nonpublic-school students was assessed in most jurisdictions. Each sample was designed to produce aggregate estimates as well as estimates for various subpopulations of interest with approximately equal precision for the participating jurisdictions. In most of the jurisdictions, about 2,500 students from approximately 100 public schools were assessed. The nonpublic-school sample sizes were more varied, usually from about 100 to 500 students in up to 22 nonpublic schools. The tables in Appendix B provide more detailed information about participation rates for schools and students.

The target population for the 1994 Trial State Assessment Program included students in public and nonpublic schools who were enrolled in the fourth grade at the time of assessment. The sampling frame included public and nonpublic schools having the relevant grade 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 the fourth-grade enrollment in the schools. Special procedures were used for jurisdictions with many small schools, and for jurisdictions having small numbers of grade-eligible schools.

The sampling frame for each jurisdiction was first stratified by urbanization status of the area in which the school was located. The urbanization classes 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, 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, schools were sorted by median household income of the ZIP Code area where the school was located.

A systematic random sample of about 100 fourth-grade schools was drawn with probability proportional to the fourth-grade enrollment of the school from the stratified frame of schools within each jurisdiction. Each selected school provided a list of eligible enrolled students, from which a systematic sample of students was drawn. One session of 30 students was sampled within each school, except in Delaware, where as many as three sessions were sampled within a given school. The number of sessions (i.e., multiples of 30 students) selected in each Delaware school was proportional to the fourth-grade enrollment of the school. Overlap between the 1994 state and national samples was minimized.

For jurisdictions that had participated in the 1992 Trial State Assessment, 25 percent of their selected public schools were designated at random to be monitored during the assessment so that reliable comparisons could be made between sessions administered with and without monitoring. For jurisdictions that had not participated in the previous assessment, 50 percent of their selected public schools were designated to be monitored. Fifty percent of all nonpublic schools were designated to be monitored, regardless of whether or not the jurisdiction had previously participated.

The 1994 assessment was preceded in 1993 by a field test. The principal goals of the field test were to test procedures and new items contemplated for 1994. Furthermore, three states and one territory used the field test to observe and react to proposed strategies. Twenty-four states participated in the field test. Schools that participated in the field test were given a chance of selection in the 1994 assessment. Section 3.2 documents the procedures used to select the schools for the field test.

Section 3.3 describes the construction of the sampling frames, including the sources of school data, missing data problems, and definition of in-scope schools. Section 3.4 includes a description of the various steps in stratification of schools within participating jurisdictions. School sample selection procedures (including new and substitute schools) are described in section 3.5. Section 3.6 includes the steps involved in selection of students within participating schools.

3.2 SAMPLE SELECTION FOR THE 1993 FIELD TEST

The Trial State Assessment 1993 field test was conducted together with the field test for the national portion of the assessment. Twenty-four states participated in the field test, which was conducted for grades 4, 8, and 12. Pairs of schools were identified, with one of each pair to be included in the test. This allowed state participation in the selection of the test schools and also facilitated replacement of schools that declined to participate in the assessment. Sampling weights were not computed for the field test samples.

3.2.1 Primary Sampling Units

The sampling frame for the field test primary sampling units (PSUs) was derived from the national frame of NAEP PSUs[1]. The 60 national frame PSUs that were noncertainty selections for the 1992 national NAEP were excluded from the field test sampling frame. National frame PSUs in the District of Columbia, Delaware, Hawaii, New Hampshire, Rhode

[1] The frame of NAEP PSUs was the frame used to draw the national NAEP samples for 1986 to 1992. Refer to the 1990 national technical report (Johnson & Allen, 1992) for more information.

Island, and Wyoming were excluded due to the heavy burden these states experience in the national and state assessments. National frame PSUs in Alaska were excluded to control field test costs.

One hundred PSUs were selected from the resulting field test frame. Forty PSUs were selected with certainty and 60 noncertainty PSUs were selected, one per noncertainty stratum. The PSUs were selected systematically and with probability proportional to the 1980 FSU general population using a starting point that was selected to avoid overlap with PSUs selected for the national assessment studies from 1986 to 1992.

3.2.2 Selection of Schools and Students

Public, private, and Catholic schools with fourth-, eighth- or twelfth-grade students were in scope for the field test assessment. Schools with fewer than 25 fourth graders were eliminated from the frame to avoid the relatively high per student cost of conducting assessments in small schools. For the same reason, schools with fewer than 40 eighth or twelfth graders were eliminated from the frame. Schools that were selected in the 1992 national and state NAEP samples were eliminated from the frame to avoid undue burden.

Three hundred pairs of schools were selected for each grade from the resulting frame by selecting two to eight pairs of schools within each of the 100 PSUs. In each of the 60 noncertainty PSUs two pairs of schools were selected. In the 40 certainty PSUs, from two to eight pairs were selected in proportion to the size of the PSU. The first member of each pair was selected systematically and with probability proportional to grade enrollment. The twelfth-grade sample was drawn first, followed by the eighth- and fourth-grade samples. Each school selected for a grade was removed from the frame before the next grade's schools were drawn. In this way, no school was selected for more than one grade.

The second member of each pair was selected in such a way that the "distance" from the primary selection was the smallest across all schools in the sampling frame that were not selected for the fourth-, eighth- and twelfth-grade samples. The distance measure was a function of the percent of Black students, percent of Hispanic students, grade enrollment, and percent of students living below poverty.

3.2.3 Assignment to Sessions for Different Subjects

Six to eight different session types were assigned in a given state. The particular number of session types varied by grade and no individual school held more than three sessions. Table 3-1 gives the overall number of sessions assigned by grade and session type.

Table 3-1
Number of Sessions by Grade and Session Type

Session Type	Grade 4	Grade 8	Grade 12
Reading	44	106	103
Mathematics	87	85	82
Mathematics Estimation	22	21	21
Science	154	148	145
U.S. History	124	190	168
Geography	122	147	147
Advanced Mathematics	--	53	40
Advanced Science	--	--	60
Totals	553	750	766

The number of sessions assigned to an individual school depended on the size of the school and was determined as follows:

Grade 4:	3 sessions for the 50 largest schools 2 sessions for all other schools with 55 or more students 1 session for all other schools
Grade 8:	2 sessions for the 90 smallest schools 3 sessions for all other schools
Grade 12:	3 sessions for the 180 largest schools 2 sessions for all others.

3.3 SAMPLING FRAME FOR THE 1994 ASSESSMENT

3.3.1 Choice of School Sampling Frame

In order to draw the school samples for the 1994 Trial State Assessment, it was necessary to obtain 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 experience with the 1992 Trial State Assessment and national assessments from 1984 to 1992, the file made available by Quality Education Data, Inc. (QED) was elected as the sampling frame. The National Center for Education Statistics' Common Core of Data (CCD) school file was used to check the completeness of the QED file. This approach was the same as that used to develop frames for the 1992 Trial State Assessment.

The QED list covers all jurisdictions except Puerto Rico. The version of the QED file used was released in late 1992, in time for selection of the school sample in early 1993. The file was missing minority and urbanization data for a sizable minority of schools (due to the inability of QED to match these schools with the corresponding CCD file). Considerable efforts were undertaken to obtain these variables for all schools in jurisdictions where these variables were to be used for stratification. These efforts are described in the next section.

Table 3-2 shows the distribution of fourth-grade schools and enrollment within schools as reported in the 1992 QED file. Enrollment was estimated for each grade as the ratio of total school enrollment divided by the number of grades in the school.

3.3.2 Missing Stratification

As stated earlier, the sampling frame for the 1994 Trial State Assessment was the most recent version of the QED file as of January 1993. The CCD file was used to extract information on urbanization (type of location) and minority enrollment in the cases where these variables were missing on the QED file. For public schools, missing values remaining in urbanization or minority enrollment data were imputed.

Schools with missing values in urbanization data were assigned the urbanization of other school records within the same jurisdiction, county, and city when urbanization did not vary within the given city. Any schools still missing urbanization were assigned values from the CCD file, where possible, or were assigned the modal value of urbanization within their city. Any remaining missing values were assigned individually based on city and 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 codes and Census data. The minority data were extracted only for those schools in jurisdictions in which minority stratification was performed.

Metro status was assigned to each nonpublic school based on Census definitions as of December 31, 1992 and FIPS county code. The QED school type was used to assign Catholic school status to nonpublic schools. Values for metro status and Catholic school status were found for all schools in the frame.

Median 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 income were assigned the mean value of median income for the three-digit ZIP code prefix or county within which they were located.

Table 3-2
Distribution of Fourth-grade Schools and Enrollment as Reported in QED 1992

Jurisdiction	Public Schools		Nonpublic Schools	
	Total Schools	Total Enrollment	Total Schools	Total Enrollment
Alabama	770	59833	211	4639
Alaska	352	9666	52	476
Arizona	663	55734	200	3581
Arkansas	542	35222	111	1864
California	4727	422095	1962	45740
Colorado	759	49556	183	3144
Connecticut	565	39660	205	4500
Delaware	53	7207	66	1762
District of Columbia	116	6234	41	1215
Florida	1413	164279	740	17588
Georgia	1015	98414	327	7506
Hawaii	173	14891	82	2740
Idaho	316	18995	55	671
Illinois	2334	144400	936	25159
Indiana	1152	75809	474	8575
Iowa	781	38813	202	4474
Kansas	822	36986	165	3193
Kentucky	819	50820	201	5448
Louisiana	789	64458	319	10883
Maine	405	17438	78	904
Maryland	771	56587	308	8236
Massachusetts	1031	68314	333	8438
Michigan	1873	126676	749	16049
Minnesota	844	62300	420	7791
Mississippi	458	40967	138	3805
Missouri	1092	66583	424	10027
Montana	468	12885	62	615
Nebraska	948	22132	171	3267
Nevada	234	18140	34	700
New Hampshire	265	15156	68	1131
New Jersey	1319	88171	567	15005
New Mexico	384	26206	128	1809
New York	2249	197261	1281	37925
North Carolina	1113	87415	232	5436
North Dakota	343	9875	58	840
Ohio	2016	139722	721	20593
Oklahoma	953	49375	92	1794
Oregon	753	40374	174	2395
Pennsylvania	1870	131024	1220	28292
Rhode Island	179	11466	68	1704
South Carolina	553	50842	198	4058
South Dakota	395	11245	103	1247
Tennessee	926	69647	223	4973
Texas	3124	282576	666	15363
Utah	433	37681	31	582
Vermont	251	7926	39	397
Virginia	1034	83093	301	6072
Washington	1034	71984	336	5289
West Virginia	593	24688	98	1160
Wisconsin	1152	63161	747	13412
Wyoming	233	8345	24	232
Total	47457	3392327	16624	382699

3.3.3 In-scope Schools

The target population for the 1994 fourth-grade Trial State Assessment in reading included students in regular public and nonpublic schools who were enrolled in the fourth grade. Nonpublic schools include parochial schools, private schools, Bureau of Indian Affairs schools and Domestic Department of Defense Education Activity schools. Special education schools were not included.

3.4 STRATIFICATION WITHIN JURISDICTIONS

3.4.1 Stratification Variables

Selection of schools within participating jurisdictions involved two stages of explicit stratification and one stage of implicit stratification. The two explicit stages for public schools were urbanization and minority enrollment. The two explicit stages for nonpublic schools were metro status and school type. The final stage for both public and nonpublic schools was median income.

3.4.2 Urbanization Classification

The NCES "type of location" variable was used to stratify fourth-grade schools into seven different urbanization levels:

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;

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

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

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

Large Town: a place not within an MSA, but with a population greater than or equal to 25,000 but greater than 50,000 and defined as urban by the U.S. Bureau of Census;

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

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

The urbanization strata were created by collapsing type of location categories. The nature of the collapsing varied across jurisdictions and grades. Each urbanization stratum included a minimum of 10 percent of eligible students in the participating jurisdiction. Table 3-3 provides the urbanization .categories (created by collapsing type of location) used within each jurisdiction.

3.4.3 Minority Classification

The second stage of stratification was minority enrollment. Minority enrollment strata were formed within urbanization strata, based on percentages of Black and Hispanic students. The three cases that occur are described in the following paragraphs.

> *Case 1:* Urbanization strata with less than 10 percent Black students and 7 percent Hispanic students were not stratified by minority enrollment.
>
> *Case 2:* Urbanization strata with greater than or equal to 10 percent Black students or 7 percent Hispanic students, but not more than twenty percent of each, were stratified by ordering percent minority enrollment within the urbanization classes and dividing the schools into three groups with about equal numbers of students per minority group.
>
> *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) with the higher concentration. The stratification was performed as follows. The minority group with the higher percentage gave the primary stratification variable; the remaining group gave the secondary stratification variable. Within urbanization class, the schools were sorted based on the primary stratification variable and divided into two groups of schools containing approximately equal numbers of students. 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 groups (e.g., low Black/low Hispanic, low Black/high Hispanic, high Black/low Hispanic, and high Black/high Hispanic).

The cutpoints in minority enrollment used to classify urbanization strata into these three cases were developed empirically. They ensure that there is good opportunity to stratify by race and ethnicity, without creating very small strata that would lead to sampling inefficiency.

The minority groups were formed solely for the purpose of creating efficient stratification design at this stage of sampling. These classifications were not used directly in analysis and reporting of the data, but acted to reduce sampling errors for achievement-level estimates. Table 3-3 provides information on minority stratification for the participating jurisdictions.

Table 3-3
Distribution of the Selected Public Schools by Sampling Strata

Urbanization	Minority	Originally Selected Schools
Alabama		
Midsize Central City	Low Percent Minority	10
Midsize Central City	Medium Percent Minority	10
Midsize Central City	High Percent Minority	10
Urban Fringe of Midsize Central City	Low Percent Minority	8
Urban Fringe of Midsize Central City	Medium Percent Minority	9
Urban Fringe of Midsize Central City	High Percent Minority	8
Large/Small Town	Low Percent Minority	9
Large/Small Town	Medium Percent Minority	9
Large/Small Town	High Percent Minority	9
Rural	Low Percent Minority	8
Rural	Medium Percent Minority	9
Rural	High Percent Minority	8
		107
Arizona		
Large Central City	Low Percent Minority	9
Large Central City	Medium Percent Minority	9
Large Central City	High Percent Minority	9
Midsize Central City	Low Percent Minority	11
Midsize Central City	Medium Percent Minority	11
Midsize Central City	High Percent Minority	11
Urban Fringe of Large/Midsize Central City	Low Percent Minority	5
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	6
Urban Fringe of Large/Midsize Central City	High Percent Minority	5
Large/Small Town and Rural	Low Percent Minority	9
Large/Small Town and Rural	Medium Percent Minority	10
Large/Small Town and Rural	High Percent Minority	10
		105
Arkansas		
Midsize Central City + Urban Fringe	Low Percent Minority	9
Midsize Central City + Urban Fringe	Medium Percent Minority	10
Midsize Central City + Urban Fringe	High Percent Minority	10
Large/Small Town	Low Percent Minority	16
Large/Small Town	Medium Percent Minority	15
Large/Small Town	High Percent Minority	15
Rural	Low Percent Minority	11
Rural	Medium Percent Minority	10
Rural	High Percent Minority	11
		107

Table 3-3 (continued)
Distribution of the Selected Public Schools by Sampling Strata

Urbanization	Minority	Originally Selected Schools
California		
Large Central City	Low Percent Minority	7
Large Central City	Medium Percent Minority	8
Large Central City	High Percent Minority	7
Midsize Central City	Low Percent Minority	7
Midsize Central City	Medium Percent Minority	7
Midsize Central City	High Percent Minority	7
Urban Fringe of Large Central City	Low Percent Minority	10
Urban Fringe of Large Central City	Medium Percent Minority	11
Urban Fringe of Large Central City	High Percent Minority	11
Urban Fringe of Midsize Central City	Low Percent Minority	3
Urban Fringe of Midsize Central City	Medium Percent Minority	4
Urban Fringe of Midsize Central City	High Percent Minority	3
Large/Small Town and Rural	Low Percent Minority	7
Large/Small Town and Rural	Medium Percent Minority	7
Large/Small Town and Rural	High Percent Minority	7
		106
Colorado		
Large/Midsize Central City	Low Percent Minority	12
Large/Midsize Central City	Medium Percent Minority	11
Large/Midsize Central City	High Percent Minority	11
Urban Fringe of Large/Midsize Central City	Low Percent Minority	13
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	13
Urban Fringe of Large/Midsize Central City	High Percent Minority	12
Large/Small Town	Low Percent Minority	6
Large/Small Town	Medium Percent Minority	7
Large/Small Town	High Percent Minority	6
Rural	Low Percent Minority	6
Rural	Medium Percent Minority	5
Rural	High Percent Minority	6
		108
Connecticut		
Large Central City	Low Black/Low Hispanic	4
Large Central City	Low Black/High Hispanic	5
Large Central City	High Black/Low Hispanic	4
Large Central City	High Black/High Hispanic	5
Midsize Central City	Low Percent Minority	6
Midsize Central City	Medium Percent Minority	8
Midsize Central City	High Percent Minority	6
Urban Fringe of Large Central City	None	17
Urban Fringe of Midsize Central City	None	13
Large/Small Town and Rural	None	37
		105

Table 3-3 (continued)
Distribution of the Selected Public Schools by Sampling Strata

Urbanization	Minority	Originally Selected Schools
Delaware		
Midsize Central City	Low Percent Minority	5
Midsize Central City	Medium Percent Minority	6
Midsize Central City	High Percent Minority	5
Urban Fringe of Midsize Central City	Low Percent Minority	3
Urban Fringe of Midsize Central City	Medium Percent Minority	1
Urban Fringe of Midsize Central City	High Percent Minority	2
Small Town	Low Percent Minority	4
Small Town	Medium Percent Minority	2
Small Town	High Percent Minority	1
Rural	Low Percent Minority	9
Rural	Medium Percent Minority	7
Rural	High Percent Minority	8
		53
District Of Columbia		
Large Central City	Low Percent Minority	39
Large Central City	Medium Percent Minority	38
Large Central City	High Percent Minority	39
		116
Florida		
Large Central City	Low Black/Low Hispanic	4
Large Central City	Low Black/High Hispanic	4
Large Central City	High Black/Low Hispanic	4
Large Central City	High Black/High Hispanic	4
Midsize Central City	Low Percent Minority	12
Midsize Central City	Medium Percent Minority	12
Midsize Central City	High Percent Minority	12
Urban Fringe of Large/Midsize Central City	Low Percent Minority	12
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	11
Urban Fringe of Large/Midsize Central City	High Percent Minority	11
Large/Small Town and Rural	Low Percent Minority	7
Large/Small Town and Rural	Medium Percent Minority	6
Large/Small Town and Rural	High Percent Minority	7
		106
Georgia		
Large/Midsize Central City	Low Percent Minority	8
Large/Midsize Central City	Medium Percent Minority	9
Large/Midsize Central City	High Percent Minority	8
Urban Fringe of Large/Midsize Central City	Low Percent Minority	10
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	9
Urban Fringe of Large/Midsize Central City	High Percent Minority	10
Small Town	Low Percent Minority	11
Small Town	Medium Percent Minority	10
Small Town	High Percent Minority	10
Rural	Low Percent Minority	6
Rural	Medium Percent Minority	7
Rural	High Percent Minority	7
		105
Guam		
None	None	21

Table 3-3 (continued)
Distribution of the Selected Public Schools by Sampling Strata

Urbanization	Minority	Originally Selected Schools
Hawaii		
Midsize Central City	None	32
Urban Fringe of Midsize Central City	None	50
Small Town and Rural	None	23
		105
Idaho		
Midsize Central City and Urban Fringe	None	21
Large Town	None	18
Small Town	None	34
Rural	None	36
		109
Indiana		
Midsize Central City	Low Percent Minority	9
Midsize Central City	Medium Percent Minority	9
Midsize Central City	High Percent Minority	9
Urban Fringe of Large/Midsize Central City	Low Percent Minority	9
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	8
Urban Fringe of Large/Midsize Central City	High Percent Minority	9
Large/Small Town	None	32
Rural	None	21
		106
Iowa		
Midsize Central City and Urban Fringe	None	36
Large/Small Town	None	36
Rural	None	37
		109
Kentucky		
Midsize Central City and Urban Fringe	Low Percent Minority	11
Midsize Central City and Urban Fringe	Medium Percent Minority	11
Midsize Central City and Urban Fringe	High Percent Minority	11
Large/Small Town	None	37
Rural	None	37
		107
Louisiana		
Large Central City	Low Percent Minority	3
Large Central City	Medium Percent Minority	4
Large Central City	High Percent Minority	4
Midsize Central City	Low Percent Minority	9
Midsize Central City	Medium Percent Minority	8
Midsize Central City	High Percent Minority	8
Urban Fringe of Large/Midsize Central City	Low Percent Minority	6
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	6
Urban Fringe of Large/Midsize Central City	High Percent Minority	6
Large/Small Town	Low Percent Minority	11
Large/Small Town	Medium Percent Minority	11
Large/Small Town	High Percent Minority	10
Rural	Low Percent Minority	6
Rural	Medium Percent Minority	7
Rural	High Percent Minority	6
		105

Table 3-3 (continued)
Distribution of the Selected Public Schools by Sampling Strata

Urbanization	Minority	Originally Selected Schools
Maine		
Midsize Central City and Urban Fringe	None	20
Small Town	None	56
Rural	None	39
		115
Maryland		
Large/Midsize Central City	Low Percent Minority	6
Large/Midsize Central City	Medium Percent Minority	6
Large/Midsize Central City	High Percent Minority	7
Urban Fringe of Large/Midsize Central City	Low Percent Minority	20
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	21
Urban Fringe of Large/Midsize Central City	High Percent Minority	19
Large/Small Town and Rural	Low Percent Minority	8
Large/Small Town and Rural	Medium Percent Minority	9
Large/Small Town and Rural	High Percent Minority	9
		105
Massachusetts		
Large/Midsize Central City	Low Percent Minority	11
Large/Midsize Central City	Medium Percent Minority	12
Large/Midsize Central City	High Percent Minority	11
Urban Fringe of Large/Midsize Central City	None	33
Large/Small Town and Rural	None	38
		105
Michigan		
Large/Midsize Central City	Low Percent Minority	9
Large/Midsize Central City	Medium Percent Minority	8
Large/Midsize Central City	High Percent Minority	8
Urban Fringe of Large Central City	None	24
Urban Fringe of Midsize Central City	Low Percent Minority	4
Urban Fringe of Midsize Central City	Medium Percent Minority	4
Urban Fringe of Midsize Central City	High Percent Minority	4
Large/Small Town	None	29
Rural	None	16
		106
Minnesota		
Large/Midsize Central City	Low Percent Minority	5
Large/Midsize Central City	Medium Percent Minority	4
Large/Midsize Central City	High Percent Minority	5
Urban Fringe of Large/Midsize Central City	None	36
Large/Small Town	None	25
Rural	None	32
		107

Table 3-3 (continued)
Distribution of the Selected Public Schools by Sampling Strata

Urbanization	Minority	Originally Selected Schools
Mississippi		
Midsize Central City	Low Percent Minority	5
Midsize Central City	Medium Percent Minority	4
Midsize Central City	High Percent Minority	5
Urban Fringe of Large/Midsize Central City	Low Percent Minority	4
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	3
Urban Fringe of Large/Midsize Central City	High Percent Minority	3
Large/Small Town	Low Percent Minority	16
Large/Small Town	Medium Percent Minority	17
Large/Small Town	High Percent Minority	15
Rural	Low Percent Minority	11
Rural	Medium Percent Minority	11
Rural	High Percent Minority	11
		105
Missouri		
Large/Midsize Central City	Low Percent Minority	5
Large/Midsize Central City	Medium Percent Minority	5
Large/Midsize Central City	High Percent Minority	6
Urban Fringe of Large/Midsize Central City	Low Percent Minority	12
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	12
Urban Fringe of Large/Midsize Central City	High Percent Minority	11
Large/Small Town	None	28
Rural	None	30
		109
Montana		
Midsize Central City and Urban Fringe	None	24
Large Town	None	12
Small Town	None	41
Rural	None	58
		135
Nebraska		
Midsize Central City and Urban Fringe	Low Percent Minority	15
Midsize Central City and Urban Fringe	Medium Percent Minority	14
Midsize Central City and Urban Fringe	High Percent Minority	14
Large/Small Town	None	40
Rural	None	61
		144
New Hampshire		
Midsize Central City and Urban Fringe	None	26
Large/Small Town	None	57
Rural	None	26
		109

Table 3-3 (continued)
Distribution of the Selected Public Schools by Sampling Strata

		Originally
Urbanization	Minority	Selected Schools
New Jersey		
Large/Midsize Central City	Low Black/Low Hispanic	6
Large/Midsize Central City	Low Black/High Hispanic	5
Large/Midsize Central City	High Black/Low Hispanic	5
Large/Midsize Central City	High Black/High Hispanic	6
Urban Fringe of Large Central City	Low Percent Minority	13
Urban Fringe of Large Central City	Medium Percent Minority	13
Urban Fringe of Large Central City	High Percent Minority	13
Urban Fringe of Midsize Central City	Low Percent Minority	7
Urban Fringe of Midsize Central City	Medium Percent Minority	6
Urban Fringe of Midsize Central City	High Percent Minority	7
Large/Small Town and Rural	None	25
		106
New Mexico		
Midsize Central City and Urban Fringe	Low Percent Minority	14
Midsize Central City and Urban Fringe	Medium Percent Minority	14
Midsize Central City and Urban Fringe	High Percent Minority	14
Large Town	Low Percent Minority	5
Large Town	Medium Percent Minority	6
Large Town	High Percent Minority	5
Small Town	Low Percent Minority	11
Small Town	Medium Percent Minority	11
Small Town	High Percent Minority	11
Rural	Low Percent Minority	5
Rural	Medium Percent Minority	6
Rural	High Percent Minority	6
		108
New York		
Large/Midsize Central City	Low Black/Low Hispanic	12
Large/Midsize Central City	Low Black/High Hispanic	11
Large/Midsize Central City	High Black/Low Hispanic	12
Large/Midsize Central City	High Black/High Hispanic	12
Urban Fringe of Large/Mid size Central City	Low Percent Minority	10
Urban Fringe of Large/Mid size Central City	Medium Percent Minority	10
Urban Fringe of Large/Mid size Central City	High Percent Minority	9
Large/Small Town and Rural	None	29
		105

Table 3-3 (continued)
Distribution of the Selected Public Schools by Sampling Strata

Urbanization	Minority	Originally Selected Schools
North Carolina		
Midsize Central City	Low Percent Minority	10
Midsize Central City	Medium Percent Minority	11
Midsize Central City	High Percent Minority	10
Urban Fringe of Midsize Central City	Low Percent Minority	3
Urban Fringe of Midsize Central City	Medium Percent Minority	4
Urban Fringe of Midsize Central City	High Percent Minority	3
Large/Small Town	Low Percent Minority	11
Large/Small Town	Medium Percent Minority	11
Large/Small Town	High Percent Minority	11
Rural	Low Percent Minority	11
Rural	Medium Percent Minority	11
Rural	High Percent Minority	10
		106
North Dakota		
Midsize Central City and Urban Fringe	None	36
Large/Small Town	None	30
Rural	None	63
		129
Pennsylvania		
Large/Midsize Central City	Low Percent Minority	7
Large/Midsize Central City	Medium Percent Minority	7
Large/Midsize Central City	High Percent Minority	7
Urban Fringe of Large/Midsize Central City	Low Percent Minority	11
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	12
Urban Fringe of Large/Midsize Central City	High Percent Minority	11
Large/Small Town	None	34
Rural	None	18
		107
Rhode Island		
Large Central City	Low Hispanic/Low Black	5
Large Central City	Low Hispanic/High Black	4
Large Central City	High Hispanic/Low Black	5
Large Central City	High Hispanic/High Black	5
Midsize Central City	Low Percent Minority	4
Midsize Central City	Medium Percent Minority	5
Midsize Central City	High Percent Minority	4
Urban Fringe of Large/Midsize Central City	None	45
Large/Small Town and Rural	None	29
		106

Table 3-3 (continued)
Distribution of the Selected Public Schools by Sampling Strata

Urbanization	Minority	Originally Selected Schools
South Carolina		
Midsize Central City	Low Percent Minority	5
Midsize Central City	Medium Percent Minority	5
Midsize Central City	High Percent Minority	6
Urban Fringe of Midsize Central City	Low Percent Minority	9
Urban Fringe of Midsize Central City	Medium Percent Minority	10
Urban Fringe of Midsize Central City	High Percent Minority	9
Small Town	Low Percent Minority	12
Small Town	Medium Percent Minority	12
Small Town	High Percent Minority	12
Rural	Low Percent Minority	9
Rural	Medium Percent Minority	8
Rural	High Percent Minority	8
		105
Tennessee		
Large Central City	Low Percent Minority	8
Large Central City	Medium Percent Minority	8
Large Central City	High Percent Minority	9
Midsize Central City	Low Percent Minority	5
Midsize Central City	Medium Percent Minority	5
Midsize Central City	High Percent Minority	4
Urban Fringe of Large/Midsize Central City	Low Percent Minority	5
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	5
Urban Fringe of Large/Midsize Central City	High Percent Minority	5
Large/Small Town	Low Percent Minority	10
Large/Small Town	Medium Percent Minority	10
Large/Small Town	High Percent Minority	10
Rural	None	22
		106
Texas		
Large Central City	Low Hispanic/Low Black	7
Large Central City	Low Hispanic/High Black	7
Large Central City	High Hispanic/Low Black	8
Large Central City	High Hispanic/High Black	7
Midsize Central City	Low Percent Minority	9
Midsize Central City	Medium Percent Minority	9
Midsize Central City	High Percent Minority	9
Urban Fringe of Large/Midsize Central City	Low Percent Minority	5
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	4
Urban Fringe of Large/Midsize Central City	High Percent Minority	5
Small Town	Low Percent Minority	7
Small Town	Medium Percent Minority	8
Small Town	High Percent Minority	8
Rural	Low Percent Minority	5
Rural	Medium Percent Minority	4
Rural	High Percent Minority	5
		107

Table 3-3 (continued)
Distribution of the Selected Public Schools by Sampling Strata

Urbanization	Minority	Originally Selected Schools
Utah		
Midsize Central City	None	36
Urban Fringe of Midsize Central City	None	32
Large/Small Town	None	16
Rural	None	22
		106
Virginia		
Midsize Central City	Low Percent Minority	13
Midsize Central City	Medium Percent Minority	12
Midsize Central City	High Percent Minority	12
Urban Fringe of Large/Midsize Central City	Low Percent Minority	10
Urban Fringe of Large/Midsize Central City	Medium Percent Minority	9
Urban Fringe of Large/Midsize Central City	High Percent Minority	10
Large/Small Town	Low Percent Minority	5
Large/Small Town	Medium Percent Minority	6
Large/Small Town	High Percent Minority	5
Rural	Low Percent Minority	8
Rural	Medium Percent Minority	8
Rural	High Percent Minority	8
		106
Washington		
Large/Midsize Central City	None	35
Urban Fringe of Large/Midsize Central City	None	31
Large/Small Town	Low Percent Minority	7
Large/Small Town	Medium Percent Minority	8
Large/Small Town	High Percent Minority	7
Rural	None	18
		106
West Virginia		
Midsize Central City	None	16
Urban Fringe of Midsize Central City	None	12
Large/Small Town	None	32
Rural	None	52
		112
Wisconsin		
Large Central City	Low Percent Minority	5
Large Central City	Medium Percent Minority	5
Large Central City	High Percent Minority	5
Midsize Central City	None	23
Urban Fringe of Large/Midsize Central City	None	14
Large/Small Town	None	29
Rural	None	27
		108
Wyoming		
Midsize Central City	None	13
Urban Fringe of Midsize Central City	Low Percent Minority	6
Urban Fringe of Midsize Central City	Medium Percent Minority	5
Urban Fringe of Midsize Central City	High Percent Minority	5
Small Town	None	61
Rural	None	31
		121

3.4.4 Metro Status

All schools in the sampling frame were assigned metro status based on their FIPS county code and Census Bureau Metropolitan Statistical Area Definitions as of December 31, 1992. The field indicated if the school was located within a metropolitan area or not. This field was used as the first stage stratification variable for nonpublic schools. Table 3-4 provides information on metro status stratification for the participating jurisdictions.

3.4.5 School Type

All nonpublic schools in the sampling frame were assigned a school type (Catholic or other nonpublic) based on their QED school type variable. This field was used as the second stage stratification variable for nonpublic schools. Table 3-4 provides information on school type stratification for the participating jurisdictions.

3.4.6 Median Household Income

Prior to the selection of the school samples, the schools were sorted by their primary and secondary stratification variables in a serpentine order. 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 income. The data on median household income were related to the ZIP code area in which the school is located. These data, derived from the 1990 Census, were obtained from Donnelly Marketing Information Services.

3.5 SCHOOL SAMPLE SELECTION FOR THE 1994 TRIAL STATE ASSESSMENT

3.5.1 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 undue burden on individual schools, NAEP developed a policy for 1994 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 remainder of this section. Three fourth-grade schools, two in Delaware and one in the District of Columbia, were selected for both the national and state assessments.

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

Table 3-4
Distribution of the Selected Nonpublic Schools by Sampling Strata

Metro Status	School Type	Originally Selected Schools
Alabama		
In metro area	Catholic	2
In metro area	Other nonpublic	7
Not in metro area	Other nonpublic	2
Arizona		
In metro area	Catholic	2
In metro area	Other nonpublic	5
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	3
Arkansas		
In metro area	Catholic	2
In metro area	Other nonpublic	4
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	2
California		
In metro area	Catholic	5
In metro area	Other nonpublic	10
Colorado		
In metro area	Catholic	4
In metro area	Other nonpublic	5
Not in metro area	Other nonpublic	2
Connecticut		
In metro area	Catholic	11
In metro area	Other nonpublic	5
Not in metro area	Other nonpublic	1
Delaware		
In metro area	Catholic	16
In metro area	Other nonpublic	15
Not in metro area	Other nonpublic	3
District of Columbia		
In metro area	Catholic	14
In metro area	Other nonpublic	12
Florida		
In metro area	Catholic	4
In metro area	Other nonpublic	11
Not in metro area	Other nonpublic	1
Georgia		
In metro area	Catholic	1
In metro area	Other nonpublic	8
Not in metro area	Other nonpublic	3

Table 3-4 (continued)
Distribution of the Selected Nonpublic Schools by Sampling Strata

Metro Status	School Type	Originally Selected Schools
Guam		
*	Catholic	6
*	Other nonpublic	5
Hawaii		
In metro area	Catholic	7
In metro area	Other nonpublic	12
Not in metro area	Catholic	2
Not in metro area	Other nonpublic	3
Idaho		
In metro area	Catholic	1
In metro area	Other nonpublic	1
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	5
Indiana		
In metro area	Catholic	8
In metro area	Other nonpublic	7
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	2
Iowa		
In metro area	Catholic	6
In metro area	Other nonpublic	1
Not in metro area	Catholic	6
Not in metro area	Other nonpublic	4
Kentucky		
In metro area	Catholic	8
In metro area	Other nonpublic	4
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	1
Louisiana		
In metro area	Catholic	11
In metro area	Other nonpublic	6
Not in metro area	Catholic	2
Not in metro area	Other nonpublic	2
Maine		
In metro area	Catholic	2
In metro area	Other nonpublic	3
Not in metro area	Catholic	2
Not in metro area	Other nonpublic	5

* Metro status did not apply to Guam.

Table 3-4 (continued)
Distribution of the Selected Nonpublic Schools by Sampling Strata

Metro Status	School Type	Originally Selected Schools
Maryland		
In metro area	Catholic	9
In metro area	Other nonpublic	10
Massachusetts		
In metro area	Catholic	11
In metro area	Other nonpublic	5
Not in metro area	Catholic	1
Michigan		
In metro area	Catholic	8
In metro area	Other nonpublic	8
Not in metro area	Catholic	2
Not in metro area	Other nonpublic	2
Minnesota		
In metro area	Catholic	9
In metro area	Other nonpublic	5
Not in metro area	Catholic	4
Not in metro area	Other nonpublic	3
Mississippi		
In metro area	Catholic	1
In metro area	Other nonpublic	3
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	7
Missouri		
In metro area	Catholic	12
In metro area	Other nonpublic	5
Not in metro area	Catholic	2
Not in metro area	Other nonpublic	2
Montana		
In metro area	Catholic	1
In metro area	Other nonpublic	2
Not in metro area	Catholic	4
Not in metro area	Other nonpublic	7
Nebraska		
In metro area	Catholic	9
In metro area	Other nonpublic	3
Not in metro area	Catholic	6
Not in metro area	Other nonpublic	6
New Hampshire		
In metro area	Catholic	6
In metro area	Other nonpublic	3
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	3

Table 3-4 (continued)
Distribution of the Selected Nonpublic Schools by Sampling Strata

Metro Status	School Type	Originally Selected Schools
New Jersey		
In metro area	Catholic	17
In metro area	Other nonpublic	6
New Mexico		
In metro area	Catholic	3
In metro area	Other nonpublic	5
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	5
New York		
In metro area	Catholic	15
In metro area	Other nonpublic	9
Not in metro area	Catholic	1
North Carolina		
In metro area	Catholic	1
In metro area	Other nonpublic	6
Not in metro area	Other nonpublic	2
North Dakota		
In metro area	Catholic	4
In metro area	Other nonpublic	1
Not in metro area	Catholic	6
Not in metro area	Other nonpublic	6
Pennsylvania		
In metro area	Catholic	19
In metro area	Other nonpublic	9
Not in metro area	Catholic	2
Not in metro area	Other nonpublic	1
Rhode Island		
In metro area	Catholic	14
In metro area	Other nonpublic	4
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	1
South Carolina		
In metro area	Catholic	1
In metro area	Other nonpublic	8
Not in metro area	Other nonpublic	3
Tennessee		
In metro area	Catholic	2
In metro area	Other nonpublic	7
Not in metro area	Other nonpublic	2

Table 3-4 (continued)
Distribution of the Selected Nonpublic Schools by Sampling Strata

Metro Status	School Type	Originally Selected Schools
Texas		
In metro area	Catholic	3
In metro area	Other nonpublic	4
Not in metro area	Other nonpublic	1
Utah		
In metro area	Catholic	2
In metro area	Other nonpublic	4
Not in metro area	Other nonpublic	1
Virginia		
In metro area	Catholic	3
In metro area	Other nonpublic	6
Not in metro area	Other nonpublic	2
Washington		
In metro area	Catholic	3
In metro area	Other nonpublic	8
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	2
West Virginia		
In metro area	Catholic	4
In metro area	Other nonpublic	3
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	3
Wisconsin		
In metro area	Catholic	14
In metro area	Other nonpublic	10
Not in metro area	Catholic	6
Not in metro area	Other nonpublic	6
Wyoming		
In metro area	Catholic	1
In metro area	Other nonpublic	2
Not in metro area	Catholic	1
Not in metro area	Other nonpublic	4

overlap between state and national samples. The procedure also recognizes the impact of the heavy within-PSU sampling in noncertainty PSUs in some jurisdictions, even though the unconditional probabilities of selection for such schools in the national samples were quite low. The procedure worked as follows:

Let $N = 1$ if the school is selected in the national sample; let $N = 0$ otherwise. Let $P_N = P(N = 1)$. Thus, $P_N = 0$ for schools not located within a selected national sample PSU. Let π_S denote the probability that a school is to be selected for the state fourth-grade sample. Schools to be included with certainty in the state sample ($\pi_S = 1$) 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 undue variance to the state estimates. For actually drawing the state samples, a conditional probability of selection, π_S^* ,was derived as follows for each school in the frame having fourth-grade enrollment:

$$\pi_S^* = 1 \qquad \text{if } \pi_S = 1$$

$$\pi_S^* = \max\left[0, \frac{\pi_S - 1 + P_N}{P_N}\right] \qquad \text{if } \pi_S < 1,\ P_N > 0, \text{ and } N = 1$$

$$\pi_S^* = \min\left[1, \frac{\pi_S}{1 - P_N}\right] \qquad \text{if } \pi_S < 1,\ P_N \geq 0, \text{ and } (N = 0)$$

$$\pi_S^* = \min(1, \pi_S) \qquad \text{if } \pi_S < 1,\ P_N = 0$$

The values of π_S^* are 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 selection probabilities for schools selected for the national sample. If P_N and π_S are relatively small, then $\frac{\pi_s - 1 + P_N}{P_N} < 0$, so that there was no chance of selecting the school for the state sample if it was in the national sample. The probability that a school was selected in the state sample, conditional on the national PSU sample but unconditional on the national school sample selection within PSUs is equal to π_S, as desired. This follows from the above formulation of π_S^* and the fact that $P(N = 1) = P_N$. The quantity π_s is used as the basis for weighting the schools, and hence the students, in the state samples.

To illustrate the implementation of these expressions for drawing the state sample, consider the following example. Suppose that $\pi_S = 0.3$ and $P_N = 0.25$. Then $\pi_S^* = 0.4$ if the school is not selected for the national sample. Thus in this case the school is selected with probability 0.4. If the school is selected for the national sample, $\pi_S^* = 0$. Thus there is no chance that this school will be selected for both the national and state samples. Integrating over the national sampling process gives the required unconditional state selection probability of 0.3 (=0.4 * 1(1 - 0.25) + 0(0.25)).

3.5.2 Selection of Schools in Small Jurisdictions

For jurisdictions with small numbers of public schools—specifically, Delaware, the District of Columbia, and Guam—all of the eligible fourth-grade public schools were included in the sample with certainty. This did not occur in any of the nonpublic school samples.

3.5.3 New School Selection

A district-level file was constructed from the fourth-grade school frame. The file was divided into a small districts file, consisting 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. The remainder of districts were denoted as "large" districts.

A sample of large 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 large districts (eligible for sampling) were divided into two files within each jurisdiction. 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 jurisdiction's enrollment. 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 except Alaska and Nevada, where four and seven districts were selected with equal probability and six and three districts were selected with probability proportional to enrollment, respectively. The 10 selected districts in each jurisdiction were then sent a listing of all their schools that appeared on the QED sampling frame, and were asked to provide information about new schools not included in the QED frame. These listings, provided by selected districts, were used as sampling frames for selection of new schools.

The eligibility of a school was determined based on its grade span. A school was classified as "new " if the changes of grade span were 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 frame.

Each fourth-grade school was assigned the measure of size:

$$\begin{cases} 60 & \text{if enrollment } \leq 70 \\ \textit{enrollment} & \text{if enrollment } > 70 \end{cases}$$

The probability of selecting a school was $\min\left[\frac{\text{sampling rate} \; * \; \text{measure of size}}{\text{P(district)}}, 1\right]$,

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.

In each jurisdiction, the sampling rate used for the main sample of fourth-grade schools was used to select the new schools for "large" districts. Additionally, all new eligible schools coming from "small" districts (those with at most one fourth-grade and one eighth-grade school) that had a school selected in the regular sample for the fourth grade were included in the sample with certainty.

Table 3-5 shows the number of new schools coming from the large and small districts for the fourth-grade samples.

3.5.4 Designating Monitor Status

One-fourth of the selected public schools were designated at random to be monitored during the assessment field period in all jurisdictions that had also participated in the 1992 Trial State Assessment. One-half of the selected public schools were designated to be monitored in jurisdictions that had not participated in the 1992 assessment—specifically Montana, Washington, and Department of Defense Education Activity Overseas. One-half of all nonpublic schools in every jurisdiction (regardless of 1992 participation) were designated to be monitored. The details of the implementation of the monitoring process in the field are given in Chapter 4. The purpose of monitoring a random quarter or half of the schools was to ensure that the procedures were being followed throughout each jurisdiction by the school and district personnel administering the assessments, and to provide data adequate for assessing whether there was a significant difference in assessment results between monitored and unmonitored schools within each jurisdiction.

The following procedure was used to determine the sample of schools to be monitored. The initially selected schools were sorted in the order in which they were systematically selected. New schools from "large' districts were added to the sample at the end of the list in random order. The sorted schools were then paired, and one member of every other pair was assigned at random, with probability 0.5, to be monitored. One member of each pair was assigned to be monitored in jurisdictions requiring 50 percent monitoring of public schools as well as for all nonpublic school samples. If there was an odd number of schools, the last school was assigned monitor status as if it were part of a pair.

3.5.5 School Substitution and Participation

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. Two passes were made at the substitution, one

Table 3-5
Distribution of New Schools Coming From Large and Small Districts in the Fourth-grade Sample

Jurisdiction	Number of New Schools	
	Large Districts	Small Districts
Alabama	0	0
Arizona	0	0
Arkansas	1	1
California	0	0
Colorado	0	1
Connecticut	0	0
Delaware	1	0
DoDEA Overseas	0	0
District of Columbia	1	0
Florida	1	0
Georgia	1	1
Guam	0	0
Hawaii	1	0
Idaho	0	0
Indiana	1	0
Iowa	1	0
Kentucky	0	0
Louisiana	0	0
Maine	0	2
Maryland	1	0
Massachusetts	0	0
Michigan	0	0
Minnesota	0	0
Mississippi	0	0
Missouri	0	0
Montana	0	0
Nebraska	0	0
New Hampshire	0	0
New Jersey	1	0
New Mexico	0	0
New York	1	0
North Carolina	0	2
North Dakota	0	0
Pennsylvania	0	0
Rhode Island	3	0
South Carolina	1	0
Tennessee	0	0
Texas	1	0
Utah	0	0
Virginia	1	0
Washington	0	0
West Virginia	0	0
Wisconsin	0	0
Wyoming	0	0
Total	17	7

assigning substitutes from outside the sampled school's district and a second pass lifting this constraint. This strategy was instigated by the fact that most school nonresponse is really at the 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 fourth-grade schools and all 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 0.60. 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 potcntial substitutes was less than the number of sampled schools or the distance measure for all remaining potential substitutes outside of the same district was greater than 0.60.

In the second pass, the different district constraint was lifted and the maximum distance allowed was raised to 0.75. This generally brought in a small number of additional assigned substitutes. Although the selected cut-off points of 0.60 and 0.75 on the distance measure were somewhat arbitrary, they were decided upon by reviewing a large number of listings befnrehand and finding a consensus on the distance measures at which substitutes began to appear unacceptable.

Table 3-6 includes information about the number of substitutes provided in each jurisdiction. Of the 44 jurisdictions participating, 34 were provided with at least one substitute. Among jurisdictions receiving no substitutes, the majority had 100 percent participation from the original sample. In a few cases, however, refusals did occur after the November 1 deadline. The number of substitutes provided to a jurisdiction ranged from zero to 24 in the fourth-grade sample. A total of 243 substitutes were selected. Some jurisdictions did not attempt to solicit participation from the substitute schoo. provided, as they considered the timing too late to seek cooperation from schools not previously notified about the assessmert.

Tables 3-7 and 3-8 shows the number of schools in the fourth grade reading samples, together with school response rates observed within participating jurisdictions. The table also shows the number of substitutes in each jurisdiction that were associated with a nonparticipating original school selection, and the number of those that participated. The numbers of participating schools differ slightly from those given in Chapter 4. The numbers of participating schools in Tables 3-7 and 3-8 indicate the numbers of schools from which useable assessment data were received. In a few instances, assessments were conducted but the data were never received.

3.6 STUDENT SAMPLE SELECTION

Schools initially sent a complete list of students to a central location in November 1993. Schools were not asked to list students in any particular order, but were asked to implement

Table 3-6
Substitute School Counts for Fourth-grade Schools

Jurisdiction	Substitutes
Alabama	8
Arizona	0
Arkansas	9
California	12
Colorado	1
Connecticut	2
Delaware	0
DoDEA Overseas	0
District of Columbia	0
Florida	3
Georgia	1
Guam	0
Hawaii	2
Idaho	24
Indiana	10
Iowa	15
Kentucky	10
Louisiana	2
Maine	4
Maryland	3
Massachusetts	1
Michigan	17
Minnesota	12
Mississippi	4
Missouri	2
Montana	6
Nebraska	8
New Hampshire	9
New Jersey	7
New Mexico	0
New York	22
North Carolina	0
North Dakota	18
Pennsylvania	4
Rhode Island	6
South Carolina	4
Tennessee	2
Texas	3
Utah	0
Virginia	1
Washington	0
West Virginia	1
Wisconsin	7
Wyoming	0
Total	243

Table 3-7
Distribution of the Fourth-grade Public-school Sample by Jurisdiction

Jurisdiction	Weighted Percent School Participation		Number of Schools in the Original Sample			Number of Substitute Schools for Nonparticipating Originals		Number of Participating Schools
	Before Substitution	After Substitution	Total	Not Eligible	Participated	Provided	Participated	
Alabama	86.78	93.39	107	1	92	14	7	99
Arizona	99.04	99.04	107	2	104	1	0	104
Arkansas	86.20	94.09	109	6	89	9	8	97
California	80.09	90.52	106	0	85	21	11	96
Colorado	100.00	100.00	108	0	108	0	0	108
Connecticut	96.47	96.47	105	1	100	3	0	100
Delaware	100.00	100.00	54	3	51	0	0	51
DoDEA Overseas	99.25	99.25	83	1	81	1	0	81
Dist. of Columbia	100.00	100.00	117	10	107	0	0	107
Florida	100.00	100.00	107	0	107	0	0	107
Georgia	99.05	99.05	107	2	104	1	0	104
Guam	100.00	100.00	21	0	21	0	0	21
Hawaii	99.07	99.07	106	1	104	0	0	104
Idaho	69.23	91.45	109	1	74	27	24	98
Indiana	83.07	92.48	107	0	89	18	10	99
Iowa	85.29	99.05	110	2	92	16	15	107
Kentucky	88.48	96.16	107	2	93	11	8	101
Louisiana	100.00	100.00	105	2	103	0	0	103
Maine	94.41	96.99	116	9	101	4	3	104
Maryland	94.23	96.15	106	2	98	6	2	100
Massachusetts	97.02	97.02	105	3	99	3	0	99
Michigan	63.19	79.77	106	3	65	38	17	82
Minnesota	85.67	95.22	107	2	90	14	10	100
Mississippi	95.20	99.04	105	1	99	4	4	103
Missouri	96.48	98.40	109	2	103	4	2	105
Montana	85.10	88.58	135	6	105	23	6	111
Nebraska	70.59	77.35	144	2	101	38	8	109
New Hampshire	71.17	[illegible]	109	0	77	25	9	86
New Jersey	85.15	91.29	107	2	89	15	7	96
New Mexico	100.00	100.00	108	3	105	0	0	105
New York	74.54	90.57	106	0	79	26	17	96
North Carolina	99.05	99.05	108	2	105	1	0	105
North Dakota	79.62	91 19	129	1	101	22	16	117
Pennsylvania	79.85	83.69	107	2	84	20	4	88
Rhode Island	80.22	85.54	109	2	86	14	6	92
South Carolina	95.26	97.15	106	1	100	4	2	102
Tennessee	71.85	73.79	106	3	74	27	2	76
Texas	91.26	93.20	108	4	95	9	2	97
Utah	100.00	100.00	106	1	105	0	0	105
Virginia	98.10	99.05	107	2	103	2	1	104
Washington	100.00	100.00	106	2	104	0	0	104
West Virginia	99.07	100.00	112	1	110	1	1	111
Wisconsin	79.15	85.56	108	3	83	21	7	90
Wyoming	98.38	98.38	121	5	112	4	0	112
Total			4673	98	4077	447	209	4286

Table 3-8
Distribution of the Fourth-grade Nonpublic-school Sample by Jurisdiction

Jurisdiction	Weighted Percent School Participation		Number of Schools in the Original Sample			Number of Substitute Schools for Nonparticipating Originals		Number of Partici-pating Schools
	Before Substitution	After Substitution	Total	Not Eligible	Participated	Provided	Participated	
Alabama	91.67	95.83	11	0	8	3	1	9
Arizona	34.93	34.93	11	0	3	8	0	3
Arkansas	80.83	93.78	9	0	6	3	1	7
California	41.96	51.42	15	4	5	6	1	6
Colorado	71.36	85.00	11	1	7	3	1	8
Connecticu'	72.70	81.89	17	1	11	4	2	13
Delaware	72.66	72.66	34	3	22	5	0	22
Dist. of Columbia	41.74	41.74	26	0	12	3	0	12
Florida	52.21	73.28	16	1	8	7	3	11
Georgia	74.03	83.77	12	1	8	3	1	9
Guam	96.02	96.02	11	0	9	0	0	9
Hawaii	80.40	87.54	24	2	17	5	2	19
Idaho	89.29	89.29	8	0	7	1	0	7
Indiana	85.09	85.09	18	4	10	4	0	10
Iowa	100.00	100.00	17	1	16	0	0	16
Kentucky	69.71	85.12	14	0	10	4	2	12
Louisiana	81.91	91.30	21	0	17	4	2	19
Maine	79.45	100.00	12	4	7	1	1	8
Maryland	62.74	69.81	19	2	10	7	1	11
Massachusetts	94.52	100.00	17	2	14	1	1	15
Michigan	00.00	00.00	20	3	0	17	0	0
Minnesota	90.63	98.78	21	0	18	3	2	20
Mississippi	64.40	64.40	12	1	7	4	0	7
Missouri	90.35	90.35	21	0	19	2	0	19
Montana	64.78	64.78	14	2	7	5	0	7
Nebraska	48.05	48.05	24	0	11	11	0	11
New Hampshire	53.85	53.85	13	2	5	6	0	5
New Jersey	76.19	76.19	23	1	17	5	0	17
New Mexico	100.00	100.00	14	5	9	0	0	9
New York	40.34	61.52	25	0	10	15	5	15
North Carolina	32.26	32.26	9	2	2	5	0	2
North Dakota	76.60	90.88	17	2	12	2	2	14
Pennsylvania	72.42	72.42	31	5	17	9	0	17
Rhode Island	92.98	92.98	20	1	17	2	0	17
South Carolina	69.12	85.71	12	3	5	4	2	7
Tennessee	41.06	41.06	11	1	4	6	0	4
Texas	24.24	39.39	8	1	2	5	1	3
Utah	22.88	22.88	7	1	1	5	0	1
Virginia	80.75	80.75	11	1	8	1	0	8
Washington	00.00	00.00	14	0	0	14	0	0
West Virginia	86.13	86.13	11	2	7	2	0	7
Wisconsin	65.71	65.71	36	4	20	12	0	20
Wyoming	00.00	00.00	8	0	0	8	0	0
Total			705	63	405	215	31	436

checks to ensure that all fourth-grade students were listed. Based on the total number of students on this list, called 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 calculator that had been programmed with the sampling algorithm. The calculator generated a random start that was used to systematically select the student line numbers (30 per session). Delaware was the only jurisdiction for which more than one session was conducted in a school. Up to three sessions were conducted in Delaware public schools, with the exact number of sessions being determined by the fourth-grade enrollment of each school. 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 had an Individualized Education Plan or because they were Limited English Proficient. More details on the procedures for student exclusion are presented in the report on field procedures for the Trial State Assessment Program.

When the assessment was conducted in a school, a count was made of the number of nonexcluded students who did not attend the session. If this number exceeded three students, the school was instructed to conduct a make-up session to which were invited all students who had been absent from the initial session.

Tables 3-9 and 3-10 provide the distribution of the student samples and response rates by jurisdiction.

Table 3-9
Distribution of the Fourth-grade Public-school Student Sample and Response Rates by Jurisdiction

Jurisdiction	Weighted Student Response Rate (Percent)	Number of Students			
		In Original Sample	Excluded from Sample	To Be Assessed	Actually Assessed
Alabama	96.07	2911	162	2749	2646
Arizona	94.27	3010	204	2806	2651
Arkansas	95.96	2808	169	2639	2535
California	93.86	2801	404	2397	2252
Colorado	94.25	3120	221	2899	2730
Connecticut	95.63	2944	248	2696	2578
Delaware	95.58	2496	153	2343	2239
DoDEA Overseas	94.58	2666	117	2549	2413
District of Columbia	94.52	3072	271	2801	2646
Florida	93.96	3167	326	2841	2666
Georgia	95.45	3072	171	2901	2766
Guam	95.91	2517	220	2297	2203
Hawaii	95.45	3020	154	2866	2732
Idaho	96.12	2847	145	2702	2598
Indiana	95.86	2919	153	2766	2655
Iowa	95.54	3024	140	2884	2759
Kentucky	96.68	2963	114	2849	2758
Louisiana	96.08	3007	181	2826	2713
Maine	94.32	2854	275	2579	2436
Maryland	95.20	2897	216	2681	2555
Massachusetts	95.43	2874	236	2638	2517
Michigan	94.87	2397	135	2262	2142
Minnesota	95.49	2915	133	2782	2655
Mississippi	96.68	3022	169	2853	2762
Missouri	95.04	2972	156	2816	2670
Montana	95.72	2711	93	2618	2501
Nebraska	94.85	2634	114	2520	2395
New Hampshire	95.58	2441	145	2296	2197
New Jersey	95.30	2799	162	2637	2509
New Mexico	94.68	3022	241	2781	2635
New York	95.34	2847	227	2620	2495
North Carolina	95.83	3127	173	2954	2833
North Dakota	96.63	2690	59	2631	2544
Pennsylvania	94.13	2569	143	2426	2290
Rhode Island	94.70	2614	140	2474	2342
South Carolina	96.39	2999	190	2809	2707
Tennessee	95.63	2217	127	2090	1998
Texas	96.45	2862	317	2545	2454
Utah	94.82	3034	153	2881	2733
Virginia	94.65	3089	216	2873	2719
Washington	94.45	3054	158	2896	2737
West Virginia	95.88	3087	213	2874	2757
Wisconsin	96.34	2609	189	2420	2331
Wyoming	95.92	2943	130	2813	2699
Total		125643	8063	117580	112153

Table 3-10
Distribution of the Fourth-grade Nonpublic-school Student Sample and Response Rates by Jurisdiction

Jurisdiction	Weighted Student Response Rate (Percent)	Number of Students			
		In Original Sample	Excluded from Sample	To Be Assessed	Actually Assessed
Alabama	95.00	212	4	208	199
Arkansas	94.67	164	1	163	154
California	97.11	153	0	153	149
Colorado	93.93	139	0	139	130
Connecticut	95.11	310	5	305	290
Delaware	97.53	558	0	558	544
District of Columbia	96.67	281	4	277	267
Florida	98.12	273	1	272	267
Georgia	96.59	225	0	225	217
Guam	97.90	380	0	380	372
Hawaii	96.23	430	2	428	415
Idaho	96.02	98	0	98	94
Indiana	95.03	231	2	229	219
Iowa	98.83	334	3	331	327
Kentucky	96.61	287	0	287	278
Louisiana	96.81	474	2	472	457
Maine	94.63	90	0	90	85
Maryland	96.96	286	3	283	275
Massachusetts	96.15	321	7	314	302
Minnesota	96.06	415	8	407	390
Mississippi	95.70	169	6	163	156
Missouri	95.38	392	1	391	372
Montana	94.49	157	0	157	148
Nebraska	97.17	218	0	218	211
New Jersey	95.86	399	3	396	379
New Mexico	92.30	229	22	207	191
New York	96.35	389	7	382	369
North Dakota	93.22	277	7	270	253
Pennsylvania	94.43	456	2	454	427
Rhode Island	96.17	369	1	368	354
South Carolina	98.07	160	0	160	156
Virginia	95.95	159	1	158	151
West Virginia	97.17	135	1	134	130
Wisconsin	95.01	407	1	406	388
Total		9577	94	9483	9116

Chapter 4

STATE AND SCHOOL COOPERATION AND FIELD ADMINISTRATION

Nancy Caldwell and Mark M. Waksberg

Westat, Inc.

4.1 OVERVIEW

By volunteering to participate in the Trial State Assessment and in the field test that preceded it, each jurisdiction assumed responsibility for securing the cooperation of the schools sampled by NAEP. The participating jurisdictions were responsible for the actual administration of the 1994 Trial State Assessment at the school level. The 1993 field test, however, operated within the framework of the national (rather than Trial State) model. Therefore, for the field test, NAEP field staff were responsible for securing cooperation, scheduling, and conducting the assessments. This chapter describes state and school cooperation and field administration procedures for both the 1993 field test and the 1994 program. Section 4.2 presents information on the field test, while section 4.3 focuses on the 1994 Trial State Assessment.

4.2 THE FIELD TEST

4.2.1 Conduct of the Field Test

In preparation for the 1994 state and national assessment programs, a field test of the forms, procedures, and booklet items was helo in 1993. In the 1993 field test, assessments were piloted in five subject areas: reading, mathematics, science, U.S. history, and world geography. The field test design focused on instructionally relevant approaches to assessment such as performance-based science tasks, the use of calculators, protractors, rulers and other manipulatives in mathematics, and the introduction of a world atlas and a retail catalog as resource tools in the geography and reading assessments.

In August 1993, letters were sent from the U.S. Department of Education to all Chief State School Officers inviting them to participate in the field test of materials and procedures for 1994. In an effort to secure the participation of more schools and to lessen the burden of participation on jurisdictions, ETS and Westat offered to perform all of the work involved, including sampling, communicating with school staff, and administering the assessment.

The school sample for the field test included both public and nonpublic schools and was designed to involve as many jurisdictions as possible, thus limiting the burden on each jurisdiction. However, small jurisdictions in which all of the schools had been involved in the

1992 NAEP were excluded from the 1993 field test sample, which had the effect of eliminating the small jurisdictions entirely. As a result, the field test sample was spread very roughly in proportion to the population across 38 jurisdictions. Each participating jurisdiction was asked to appoint a state coordinator to serve as the liaison between NAEP/Westat staff and the participating schools. State coordinators were asked only to notify districts of their inclusion and to support the schools' participation in the field test.

The original school sample comprised 905 schools. For each originally sampled school, up to three substitutes or "alternate" schools were named by Westat. The three levels of alternate schools included specified substitutes within the same district that were demographically comparable to the originally selected schools, an option that allowed district superintendents to choose their own alternate schools. In the event that a district was not able to participate, an "out-of-district" alternate school was offered. The type and number of sessions scheduled for an originally selected school remained constant across alternates.

From October to December 1992, all districts and schools in the 1993 field test sample were contacted, cooperation obtained, and assessment schedules set. To accomplish this, 11 of the most experienced NAEP supervisors were each responsible for gaining cooperation in districts and schools in several jurisdictions. In January 1993, the NAEP field staff expanded to 51 supervisors. Each supervisor, including those in the original group, was responsible for sampling and conducting assessments in a single region of approximately 20 schools.

4.2.2 Results of the Field Test

A total of 844 originally selected schools and alternates actually participated in the field test. The final assessed sample of schools included 300 schools at grade 4, 273 schools at grade 8, and 272 schools at grade 12.

A total of 46,849 students participated in the field test: 13,962 students at grade 4, 17,439 students at grade 8, and 15,448 students at grade 12. The overall student participation rate was 86.8 percent: 93 percent at grade 4, 89.4 percent at grade 8, and 79.3 percent at grade 12. A total of 811 students (1.7%) who were sampled for the assessment were excluded from participation by their schools.

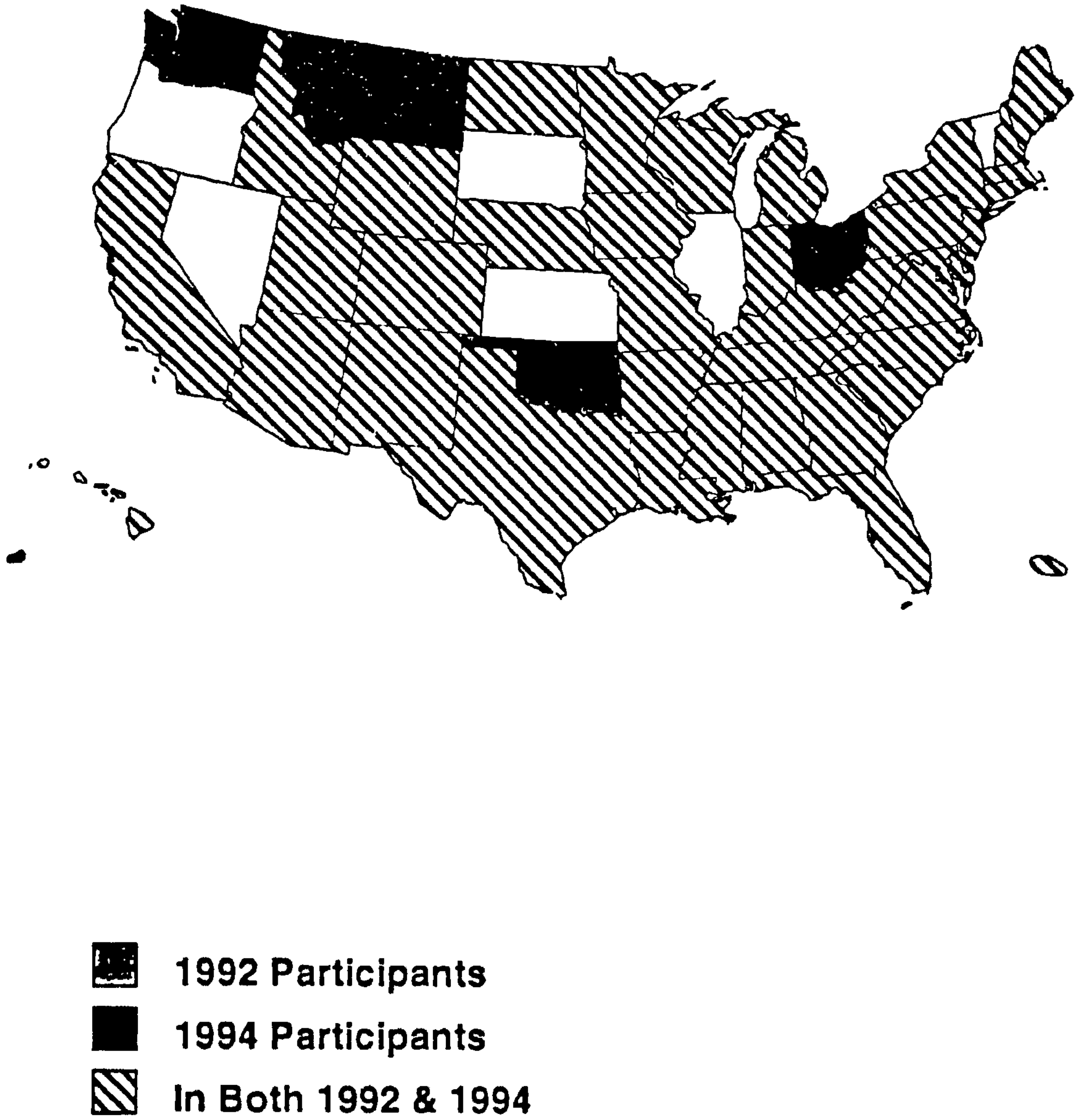
Depending on the size of the school, a school's sample numbered approximately 30 to 60 students, who were assessed in either one or two sessions. The desired number of student responses to the assessment items being tested was achieved.

4.3 THE 1994 TRIAL STATE ASSESSMENT

Forty-one states, the District of Columbia, and Guam volunteered for the 1994 Trial State Assessment, as did the Department of Defense Education Activity (DoDEA) overseas schools. Figure 4-1 identifies the jurisdictions participating in the last two assessment

Figure 4-1

Map of Participating Jurisdictions, 1992 and 1994 Trial State Assessments



years (similar information is presented in table form in Chapter 1). As was the case for the 1992 Trial State assessment, each jurisdiction designated its own coordinator to oversee all assessment activities in their jurisdiction.

4.3.1 Overview of Responsibilities

Data collection for the 1994 Trial 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 personnel to conduct the assessments; and
- conducting an extensive quality assurance program.

Each jurisdiction volunteering to participate in the 1994 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;
- 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 state supervisor to use when selecting the sample of students;
- notify assessment administrators about training and send them their manuals; and
- send the lists of sampled students to the schools

At the school level, an assessment administrator 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 sessions(s);
- completing assessment forms; and
- preparing the assessment materials for shipment.

Westat hired and trained six field managers and 46 state supervisors, one for each jurisdiction (two supervisors were hired for DoDEA overseas schools, one working in Europe and the other in the Far East). 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 about cooperation and scheduling;
- make sure the arrangements for the assessments were set and assessment administrators identified; and
- schedule the assessment administrators training sessions.

The primary tasks of the state supervisors were to

- select the sample of students to be assessed;
- recruit and hire the quality control monitors throughout their jurisdiction;
- conduct in-person assessment administrator training sessions; and
- coordinate the monitoring of the assessment sessions and makeup sessions.

Westat also hired and trained an average of four quality control monitors in each jurisdiction to monitor the assessment sessions.

4.3.2 Schedule of Data Collection Activities

August 1993

Westat sent the samples of schools selected for the national and Trial State Assessment to the state coordinators. At the time of this mailing, a final decision had not been made as to which grades would be included in the Trial State Assessment, so the lists included fourth-, eighth-, and twelfth-grade schools. Some state coordinators chose to inform all selected districts and schools immediately, while others waited until the final authorization was received from Congress and then informed only the schools with fourth grade.

October 1993

Westat field managers visited each jurisdiction to explain the computerized state coordinator system, which could be used to keep track of assessment-related activities.

Westat distributed Student Listing Forms, Principal Questionnaires, and the list of the schools selected for the Trial State Assessment, updated with a suggested week of assessment and number of sessions.

September-November 1993

State coordinators obtained cooperation from districts and schools and reported participation status to Westat field managers via printed lists or computer files.

State coordinators sent Student Listing Forms, Supplemental Student Listing Forms, and Principal Questionnaires to participating schools.

November 3-6, 1993

State supervisors were trained.

November 15, 1993	Suggested cutoff for decisions on participating schools and submission of list of grade-eligible students to state coordinators for sampling purposes.
November 29-December 10, 1993	NAEP state supervisors visited state coordinators to select student samples and prepare Administration Schedules listing the students selected for each session. Westat provided the schedule of training sessions and copies of the *Manual for Assessment Administrators* to state coordinators for distribution.
December 7, 1993-January 7, 1994	State coordinators notified assessment administrators of the date and time of training and sent each a copy of the *Manual for Assessment Administrators*.
January 6-8, 1994	Quality control monitors were trained.
January 10-28, 1994	Assessment administrators were trained.
January 31-February 25, 1994	Assessments were conducted. Unannounced visits were made by quality control monitors to a predetermined subset of the sessions.
February 23-March 4, 1994	Make-up sessions were held as necessary.

4.3.3 Preparations for the Trial State Assessment

The focal point of the schedule for the Trial State Assessment was the period between January 31 and February 25, 1994 when the assessments were conducted in the schools. However, as with any undertaking of this magnitude, the project required many months of planning and preparation.

Westat selected the samples of schools according to the procedures described in Chapter 3. On August 18, 1993, lists of the selected schools and other materials describing the Trial State Assessment Program were sent to state coordinators. Most state coordinators also preferred that NAEP provide a suggested assessment date for each school. School listings were updated with this information and were sent to the state coordinators, along with other descriptive materials and forms, in October.

State coordinators were also given the option of receiving the school information in the form of a computer database with accompanying management information software. This system enabled state coordinators to keep track of the cooperating schools, the assessment schedule, the training schedule, and the assessment administrators. Coordinators could choose to receive a laptop computer and printer or to have the system installed on their own computer.

Westat field managers traveled to the state offices to explain the computer system to the state coordinators and their staff. All but two state coordinators chose to use the computerized system.

Six of the most experienced NAEP supervisors were chosen to be field managers, the primary link between NAEP and the state coordinators. In October, the field managers visited offices of the state coordinators to explain the computer system to state staff. The field managers kept in frequent contact with the state coordinators as the state coordinators secured the cooperation of the selected schools and established the assessment schedule.

The field managers used the same computer system as the state coordinators to keep track of the schools and the schedule. The state coordinators sent updates via computer disks, telephone, or print to their field manager, who then entered the information into the system. Weekly transmissions were made from the field manager to Westat.

By the first of November, Westat hired one state supervisor for each participating jurisdiction. The state supervisors attended a training session held November 3-6, 1993. 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.

The state supervisors' first task after training 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 15. 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 preprogrammed calculator, the supervisors generally selected a sample of 30 students per session type per school. The exceptions to this were in small schools and jurisdictions with fewer than the necessary 100 fourth-grade public schools. In the jurisdictions with fewer schools, larger student samples were required from schools that participated. In the 1994 Trial State Assessment, this was only necessary in the schools in Delaware and DoDEA overseas.

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.

During the period from mid-November through December, 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. Since studies have shown no measurable difference between the performance in monitored and unmonitored sessions, the ratio of monitored schools was lowered to reduce the costs of the field work. In any jurisdiction in which the fourth grade had previously participated in the Trial State Assessment, the percentage of public schools to be monitored was reduced to 25 percent. Because nonpublic schools were included for the first time, their monitor rate was 50 percent. Also, any jurisdiction that had not previously participated at the fourth-grade level was monitored at 50 percent as well. The schools to be monitored were known only to contractor staff; it was not on any of the listings provided to state staff.

Approximately 200 quality control monitors were trained in a session held January 6-8, 1994. 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.

Almost immediately following the quality control monitor training, supervisors began conducting training for assessment administrators. Each quality control monitor attended several of these training sessions, to assist the state supervisor and to become thoroughly familiar with the assessment administrator's responsibilities. Almost 4,700 assessment administrators were trained in about 450 training sessions across the nation.

To ensure uniformity in the training sessions, Westat developed a highly structured program involving a script for trainers, a videotape, and an example to be completed by the trainees. 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.

The videotape, similar to the one used in the 1992 Trial State Assessment, was developed by Westat to provide background for the study and to simulate the various steps of the assessment that would be repeated by the administrators. The portions of the videotape depicting an assessment had been taped in a classroom with students in attendance to closely simulate an actual assessment session. The videotape was divided into sections with breaks for review by the trainer and practice for the trainees.

The final component of the presentation was a training example consisting of a set of exercises keyed to each part of the training package. A portion of the videotape was shown and then reviewed by the trainer. Related exercises were then completed by the trainees before the next subject was discussed.

The entire training session generally ran for about three and one-half hours. Sessions usually began in the morning and ended with lunch. This reduction in time (from about five hours in 1990) for the training session, initiated in 1992, was appreciated by the trainees.

All of the information presented in the training session was included in the *Manual for Assessment Administrators*, developed by Westat. Copies of the manual were sent by Westat to the state coordinators at the beginning of December so that they could be distributed to the assessment administrators before the training sessions.

4.3.4 Monitoring of Assessment Activities

Two weeks prior to the scheduled assessment date, the 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 Observation Form. Most of the questions asked in the pre-assessment call were designed to gauge whether the assessment administrator had received all materials needed and had completed the preparations for the assessment. The 40-page Quality Control Monitor Observation Form is included in the *Report on Data Collection Activities for the 1994 National Assessment of Educational Progress* (Westat, Inc, 1995).

Pre-assessment calls were made to all schools regardless of whether they were to be monitored. If the sessions in a school were not observed, the quality control monitor called the assessment administrator three days after the assessment to find out how the session went, to obtain the assessment administrator's impressions of the manual, training, and materials and to ensure that all post-assessment activities had been completed.

If the sessions in a school were 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 Observation 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.

4.3.5 School and Student Participation

Table 4-1 shows the results of the state coordinators' efforts to gain the cooperation of the selected schools. Overall, 4,295 public schools and 437 nonpublic schools participated in the 1994 Trial State Assessment. This is about 86 percent (unweighted) of the eligible schools in the original sample at each grade and about 91 percent (unweighted) of the sample after substitution.

Table 4-1
School Participation, 1994 Trial State Assessment

Status	Public	Nonpublic	Total
Schools in original sample	4671	705	5376
Schools not eligible (e.g. closed, no grade 4)	98	63	161
Eligible schools in original sample	4573	642	5215
Noncooperating (e.g. school, district, or jurisdiction refusal)	489	236	725
Participating	4084	406	4490
Substitutes provided for noncooperating schools	441	214	655
Participating substitutes	211	31	242
Total schools participating after substitution	4295	437	4732

Participation results for students in the 1994 Trial State Assessment in reading are given in Table 4-2. Approximately 140,000 students were sampled. As can be seen from the table, the original sample, which was selected by the NAEP state supervisors, comprised about 136,000 (or 97%) of this number. The original sample size was increased somewhat after the supplemental samples had been drawn (from students newly enrolled since the creation of the original lists).

Table 4-2
Student Participation, 1994 Trial State Assessment

Status	Public	Nonpublic	Total
Sampled	130452	10176	140628
Original sample	126596	10013	136609
Supplemental sample	3856	136609	4019
Withdrawn	4805	127	4932
Excluded	8068	121	8189
To be assessed	117579	9928	127507
Assessed	112150	9544	121694
Initial sessions	111187	9503	120690
Make-up sessions	963	41	1004

Assessment administrators removed some students from the total sample according to NAEP criteria: first, those students who had left their schools since the time that they were sampled (withdrawn); then, those judged incapable of participating meaningfully in the assessment by school staff (excluded). Any student who had an Individualized Education Plan (IEP) for reasons other than being gifted and talented or who was classified as Limited English Proficient (LEP) could be considered for exclusion. To be excluded, an IEP student had to be "mainstreamed less than 50 percent of the time in academic subjects and/or judged incapable of participating meaningfully in the assessment." For an LEP student to be excluded, he or she had to be "a native speaker of a language other than English, and enrolled in an English-speaking school (not including a bilingual education program) for less than two years and judged incapable of taking part in the assessment."

These exclusions left 127,507 fourth graders to be assessed in reading. Of these, 121,694 were actually assessed, yielding an unweighted student participation rate of 95.4 percent.

4.3.6 Results of the Observations

During the assessment sessions, the quality control monitors noted instances when the assessment administrators deviated from the prescribed procedures and whether any of these deviations were serious enough to warrant their intervention. Quality control monitors reported no instances where there were serious breaches of the procedures or major problems that would question the validity of the assessment.

Deviation from prescribed procedures occurred most often in the administrator's reading of the script that introduced the assessment and provided the directions. Even so, in at least 92 percent of the observed sessions in the public and nonpublic schools, the assessment administrator read the script verbatim or with only slight deviations. Examples of major deviations included skipping sections of the script, adding substantially to the script, and forgetting to pass out materials at the appropriate times. The quality control monitor intervened in these instances.

Most of the other procedures that could have had some bearing on the validity of the results were adhered to very well by the assessment administrators. In 99 percent of the observed public-school sessions and 98 percent of the observed nonpublic school sessions, the assessment administrators opened the bundles of booklets at the appropriate time and handled questions from the students correctly. Ninety-eight percent of the public-school sessions and 100 percent of the nonpublic-school sessions were timed correctly.

After the assessment session was over, assessment administrators were asked how they thought the assessment went and whether they had any comments or suggestions. Overall, assessment administrators stated that they thought 99 percent of the sessions went either very well or satisfactorily. This figure was consistent across the public and nonpublic schools, as well as for both monitored and unmonitored sessions. The percentage of assessment administrators who thought their session had gone "very well" was about three percentage points higher in the monitored sessions than in the unmonitored.

Comments about the assessment materials and procedures were generally favorable. Criticisms or suggestions included that there were too many forms and too much paperwork; coding the booklet covers was tedious and problematic for students; and schools needed more information about NAEP and assessment results.

In addition to these interviews, Westat sent a debriefing form to all of the NAEP state supervisors and met in person with half of them. This meeting produced suggestions for future assessments, especially many minor changes in the procedures, materials, and training plans. In addition, the state supervisors recommended that district and particularly school staff receive more information describing the background and objectives of NAEP and the Trial State Assessments. They also stated that many school staff were very interested in results for their students, or at least summary results for their jurisdiction.

State coordinators were also sent a questionnaire about their experiences, suggestions, and comments, to which 39 coordinators responded. Generally, the state coordinators felt that the assessments went more smoothly than in the past. They also commented favorably on the training package and other materials. Like the assessment administrators, the state coordinators criticized the amount of work required to prepare for the assessments. They made many other suggestions about the computerized data system, sampling procedures, training program, and design of the assessment. All of these suggestions will be reviewed as future assessments are planned.

The results of the assessment and comments from assessment administrators and state coordinators were summarized in a report presented to the NAEP Network in October 1994. At that time, each participating jurisdiction received a summary of its participation data, data collection activities, results of the assessment, and assessment administrators' comments.

Chapter 5

PROCESSING AND SCORING ASSESSMENT MATERIALS

Patrick B. Bourgeacq, Charles L. Brungardt, Patricia M. Garcia Stearns, Tillie Kennel, Linda L. Reynolds, Timothy Robinson, Brent W. Studer, and Bradley J. Thayer

National Computer Systems

5.1 OVERVIEW

This portion of the report reviews the activities conducted by National Computer Systems (NCS) for the 1994 NAEP Trial State Assessment. The 1994 assessment was an exciting one for NAEP and NCS because of the introduction of image scoring to the assessment. The advent of image scoring eliminated almost all paper handling during scoring and improved monitoring and reliability scoring. A short-term trend study was added to the assessment to compare the scoring of paper and scoring of images of student responses from both 1992 and 1994.

In the early 1990s, NCS developed and implemented flexible, innovatively designed processing programs and a sophisticated process control system that allowed the integration of data entry and work flow management systems. The planning, preparation, and quality-conscious application of these systems in 1992 and 1994 has made the NAEP project an exercise in coordinated teamwork and excellence.

This chapter begins with a description of the various tasks performed by NCS, detailing printing, distribution, receipt control, scoring, and processing activities. It also discusses specific activities involved in processing the assessment materials, and presents an analysis of several of those activities. The chapter provides documentation for the professional scoring effort—scoring guides, training papers, papers illustrating sample score points, calibration papers, calibration bridges, and interreader reliability reports. The detailed processing specifications and documentation of the NAEP process control system are presented later.

5.1.1 Innovations for 1994

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

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

The success of this new method of transferring data has moved NAEP closer to achieving another goal—the simultaneous scoring of constructed-response items at multiple locations. This process enhanced the reliability and monitoring of scoring and allowed both NCS and ETS to focus attention on the intellectual process of scoring student responses.

Tables 5-1 and 5-2 give an overview of the processing volume and the schedule for the 1994 NAEP Trial State Assessment.

Table 5-1
1994 NAEP Trial State Assessment Processing Totals

Document/Category	Totals
Number of sessions	4,842
Assessed student booklets	122,052
Absent student booklets	5,810
Excluded student booklets	8,189
IEP/LEP questionnaires	17,118
School questionnaires	4,690
Teacher questionnaires	17,231
Scanned documents	62,058
Scanned sheets	2,544,434
Key-entered documents	0

Table 5-2
1994 NAEP Trial State Assessment
NCS Schedule

Activity	Planned Start Date	Planned Finish Date	Actual Start Date	Actual Finish Date
Subcontractor meeting	11/08/93	11/09/93	11/08/93	11/09/93
Network meeting to review items	06/18/93	06/18/93	06/18/93	06/18/93
Printing	09/02/93	10/15/93	09/02/93	11/18/93
NCS submits receipt-control specifications plan	10/01/93	10/01/93	10/01/93	10/01/93
All reading materials at NCS	10/15/93	10/15/93	11/18/93	11/18/93
Initial packaging begins	10/15/93	01/03/94	10/15/93	01/03/94
Weekly status reports on receipt control and procedures	10/18/93	05/31/94	10/18/93	06/10/94
State address file from Westat	11/17/93	11/17/93	11/17/93	11/17/93
95% of public schools to NCS from Westat	11/19/93	11/19/93	11/19/93	11/19/93
Print nonpublic-school Administration Schedule	11/24/93	11/24/93	11/24/93	11/24/93
Print public-school Administration Schedule	11/24/93	11/24/93	11/24/93	11/24/93
NCS ships Administration Schedules to public-school supervisors	11/29/93	11/29/93	11/29/93	11/29/93
Ship public-school Administration Schedule	11/29/93	11/29/93	11/29/93	11/29/93
Ship nonpublic-school Administration Schedule	11/29/93	11/29/93	11/29/93	11/29/93
NCS ships Administration Schedules to nonpublic-school supervisors	12/13/93	12/13/93	12/13/93	12/13/93
Materials due in districts	01/02/94	01/02/94	01/02/94	01/02/94
Final packaging	01/03/94	02/18/94	01/03/94	02/18/94
Distribution	01/14/94	02/18/94	01/14/94	02/18/94
Public and nonpublic test administration	01/31/94	03/04/94	01/31/94	04/11/94
Receiving	02/01/94	03/11/94	02/01/94	03/18/94
Processing	02/02/94	04/25/94	02/02/94	05/02/94
Scoring training preparation	02/21/94	03/11/94	02/21/94	03/11/94
Project through clean post	03/25/94	04/08/94	03/25/94	04/15/94
Constructed-response scoring/training	03/28/94	05/13/94	03/28/94	05/27/93
Ship weights data tape to Westat	04/19/94	04/19/94	04/20/94	04/20/94
State questionnaires data tape delivered	06/13/94	06/13/94	06/13/94	06/13/94
State reading data tape delivered	06/16/94	06/17/94	06/17/94	06/17/94

5.2 PRINTING

5.2.1 Overview

For the 1994 NAEP Trial State Assessment, 16 discrete documents were designed. More than 112,500 booklets and forms, totaling over 5.4 million pages, were printed. A list of these materials and key dates for their production is found in Table 5-3.

The printing effort began in June 1993, with the design of the booklet covers and the administration schedule. This was a collaborative effort involving staff from ETS, Westat, and NCS. The covers were designed to facilitate the use of intelligent character recognition (ICR) to gather data. The administration schedule, which was designed to use both ICR and OMR (optical mark recognition), was the primary source of demographic data and also served as the session header for booklets when processed. Spaces for the same information were included on the student booklet cover as a backup source. For elements not individualized on the administration schedule (school number, Zip code, ILSQ number, and a "do not use" field), both handwritten information and OMR ovals were used on the booklet cover to assure complete, accurate data collection.

5.2.2 Trial State Assessment Printing

For the Trial State Assessments booklets, ETS provided one camera-ready copy of each unique cognitive block as well as of each set of directions and background sections.

The printing effort for the Trial State Assessment materials began in June 1993, with the receipt of short-term trend reading blocks. The same camera-ready copy was used for these blocks as was used in the 1992 assessment; only the block designation on each page was changed. Camera-ready copy of the other reading blocks and all directions and background sections followed in August.

Because a large number of documents had to be printed in a relatively short period of time, preparatory work was started before all parts of the test booklets were received. Upon receipt of camera-ready materials from ETS, NCS made duplicate copies of each unique block and booklet component. These were then checked for consistency in design. An attempt was also made to proofread text and check response foils. Any problems or questions were referred to ETS personnel. Whenever possible, corrections or changes were made by NCS; other times replacement copy was supplied by ETS. During this time, the number of pages for each assessment booklet was calculated to ensure that no booklets would exceed size limitations.

As each block was received and as many issues as possible resolved, camera-ready materials were sent to the NCS forms division along with a guide indicating the number of times each cognitive block and booklet component would be repeated in the assessment battery. Preliminary work such as adding gridding ovals for response options began and the required numbers of negatives for each block and booklet component were made. Performance of these preliminary tasks was crucial to meeting the delivery schedule.

Table 5-3

Documents Printed for the 1994 NAEP Trial State Assessment

Sample	Grade/age	Document	Subject	Type	No. Pages	Final Copy from ETS	Approval to Print	Printed Documents Received	Quantity Printed
Main/state	4/9	Booklet R1	Reading	Image scan	36	08/25/93	10/28/93	11/16/93	13094
Main/state	4/9	Booklet R2	Reading	Image scan	36	08/25/93	10/28/93	11/16/93	13130
Main/state	4/9	Booklet R3	Reading	Image scan	36	08/25/93	10/28/93	11/16/93	13095
Main/state	4/9	Booklet R4	Reading	Image scan	40	08/25/93	10/28/93	11/15/93	13225
Main/state	4/9	Booklet R5	Reading	Image scan	36	08/25/93	10/29/93	11/15/93	13275
Main/state	4/9	Booklet R6	Reading	Image scan	36	08/25/93	10/28/93	11/15/93	13275
Main/state	4/9	Booklet R7	Reading	Image scan	36	08/25/93	10/28/93	11/16/93	13280
Main/state	4/9	Booklet R8	Reading	Image scan	36	08/25/93	10/29/93	11/16/93	13291
Main/state	4/9	Booklet R9	Reading	Image scan	36	08/25/93	10/29/93	11/16/93	13330
Main/state	4/9	Booklet R10	Reading	Image scan	36	08/25/93	10/29/93	11/16/93	13255
Main/state	4/9	Booklet R11	Reading	Image scan	32	08/25/93	10/29/93	11/18/93	13055
Main/state	4/9	Booklet R12	Reading	Image scan	36	08/25/93	10/29/93	11/18/93	13166
Main/state	4/9	Booklet R13	Reading	Image scan	40	08/25/93	10/29/93	11/16/93	13280
Main/state	4/9	Booklet R14	Reading	Image scan	36	08/25/93	11/01/93	11/16/93	13280
Main/state	4/9	Booklet R15	Reading	Image scan	36	08/25/93	11/01/93	11/16/93	13155
Main/state	All	Administration Schedule	—	ICR & OMR	2	N/A	09/02/93	09/20/93	70705
State	4	Reading Teacher Questionnaire	—	W201	12	09/02/93	11/09/93	11/19/93	30300
Main/state	All	IEP/LEP Questionnaire	—	W201	4	09/16/93	10/11/93	12/01/93	76581
Main/state/bridge	4	School Questionnaire	—	ICR & OMR	12	09/02/93	10/20/93	11/18/93	12320
Main/state	All	Roster/Questionnaire IEP/LEP and Excluded Student Questionnaire	—	ICR & OMR	1	N/A	08/05/93	08/27/93	40050
State	4	Teacher Questionnaire Roster	—	ICR & OMR	2	N/A	09/16/93	09/30/93	14050

The actual assembly of booklets began after all components for a particular booklet were received and the Office of Management and Budget had given its approval. Using mock-ups of booklets and "booklet maps" as guides, the NCS printer assembled prepared negatives into complete booklets.

Rosters for teacher questionnaires, school questionnaires and IEP/LEP student questionnaires were designed by NCS and reviewed by ETS. After approval, NCS produced camera-ready copy and mounted it on layout sheets for printing.

School and teacher questionnaires were the last materials to be printed. NCS mounted camera-ready pages of the questionnaires received from ETS on NCS Mark Reflex layout sheets. In some cases spacing of text and answer foils had to be adjusted so that the gridding ovals would appear in scannable positions. Portions of questionnaire pages requiring redesign were revised by NCS to include shaded boxes to make use of ICR technology and were submitted to ETS for approval.

The printer forwarded proofs for each unique booklet for review by NCS and ETS personnel. Clean-up work, where necessary, was indicated on the proofs. A content change in several blocks required multiple camera-ready copies that could be stripped into each affected booklet. ETS approved the proofs, and NCS reported this, along with any necessary changes, to the printer. Once approved, the booklets were printed in the colors agreed upon by NCS and ETS. Because reading booklets contained short-term trend items, the same colors were used as in the 1992 assessmen'. NCS and ETS personnel checked sample copies to check for color accuracy. Any booklets that did not meet color specifications were reprinted.

As the booklets and forms were printed by vendors, pallets of documents were received and entered into NCS's inventory control system. Sample booklets were selected and quality-checked for printing and collating errors. All printing for the 1994 NAEP Trial State Assessment was completed by November 30, 1993.

5.3 PACKAGING AND SHIPPING

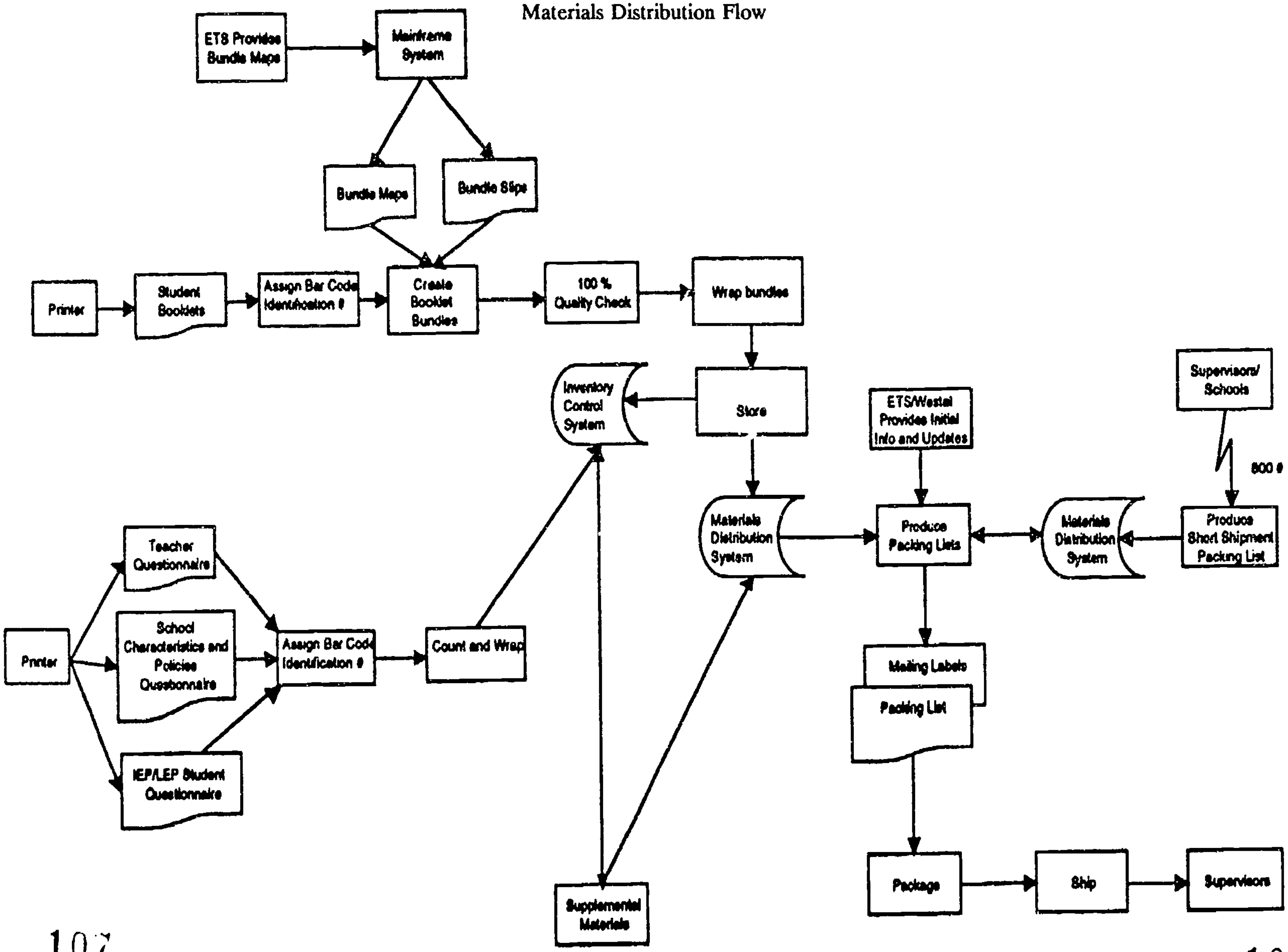
5.3.1 Distribution

The distribution effort for the 1994 NAEP Trial State Assessment involved packaging and mailing documents and associated forms and materials to individual schools. The NAEP materials distribution system, initially developed by NCS in 1990 to control shipments to the schools and supervisors, was enhanced and utilized. Files in this system contained the names and addresses for shipment of materials, scheduled assessment dates, and a listing of all materials available for use by a participant. Changes to any of this information were made directly in the distribution system file either manually by NCS staff or via file updates provided by Westat. The complex packaging effort, booklet accountability system, and on-line bundle assignment and distribution system is illustrated in Figure 5-1.

Bar code technology, introduced by NCS in the 1990 assessment, continued to be utilized in document control. To identify each document, a unique ten-digit numbering system was

Figure 5-1

1994 NAEP Trial State Assessment
Materials Distribution Flow



devised, consistıng of the :hree-digit booklet number or form type, a six-digit sequential number, and a check digit. Each form was assigned a range of ID numbers. Bar codes reflecting this ID number were applied to the front cover of each document through NCS bar code technology using an ink jet printer. After administration of the assessment, as bar codes were read during the scanning process, the document ID number was incorporated into each student record.

The booklets were then spiraled into bundles, according to the design specified by ETS. Bundles of 11 booklets were created in the pattern dictated by the bundle maps. The booklets were arranged in such a manner that each booklet appeared first in a bundle approximately the same number of times and the booklets were evenly distributed across the bundles. This assured that sample sizes of individual booklets would not be jeopardized if entire bundles were not used. Since all Administration Schedules for each scheduled session were preprinted with the booklet IDs designated for that session, only bundles of 11 booklets were created. Three bundles of booklets were preassigned to each session, giving each 33 booklets. This number most closely approximated the average projected session size of 30 students and allowed extra booklets either for additional students or to replace defective booklets. There were 16 unique spiral bundle types for the 1994 NAEP Trial State Assessment.

Each group of 11 booklets had a bundle slip/header sheet that indicated the subject area, bundle type, bundle number, and a list of the booklet types to be included in the bundle, along with a list of any other essential materials to be used with the session. All booklets had to be arranged in the exact order listed on the bundle header sheet. To ensure the security of the NAEP assessments, the following plan was used to account for all booklets: All bundles were taken to a bar code reader/document transport machine where they were scanned to interpret each bundle's bar codes. The file of scanned bar codes was then transferred from the personal computer connected to the scanner to a mainframe data set.

The unique bundle number on the header sheet informs the system program as to what type of bundle should follow. A computer job was run to compare the bundle type expected to the sequence of booklets that was actually scanned after the header. This job also verified that the appropriate number of booklets was included in each bundle. Any discrepancies were printed on an error listing and forwarded to the packaging department. The error was corrected and the bundle was again read into the system. This process was repeated until all bundles were correct. As a bundle cleared the process, it was flagged on the system as ready for distribution. All bundles were shrink-wrapped in clear plastic, bound with plastic strips, and labeled "Do not open until 45 minutes before assessment." The bundles were then ready for distribution.

Using sampling files provided by Westat, NCS assigned bundles to schools and customized the bundle slips and packing lists. File data from Westat was coupled with the file of bundle numbers and the corresponding booklet numbers. This file was then used to preprint all booklet identification numbers, school name, school number and session type, directly onto the scannable Administration Schedule. As a result, every session had specific bundles assigned to it in advance. This increased the quality of the booklet accountability system by enabling NCS to identify where any booklet should be at any time during the assessment.

Distribution of materials was accomplished in waves according to the assessment date. Booklets were boxed by session, with the appropriate additional nonreusable materials included with each session. If the quantities of materials received were insufficient to conduct the assessment, additional materials could be requested by school supervisors via the NAEP toll-free line.

Initially, a total of 5,182 sets of session materials were shipped for the 1994 NAEP Trial State Assessments. Approximately 143 additional shipments of booklets and miscellaneous materials were sent. All outbound shipments were recorded in the NCS outbound mail management system. A bar code containing the school number on each address label was read into the system, which determined the routing of the shipment and the charges. Information was recorded in a file on the system, which, at the end of the day, was transferred to the mainframe from a personal computer. A computer program could then access information to produce reports on shipments sent, regardless of the carrier used.

5.3.2 Short Shipment and Phones

A toll-free telephone line was maintained for administrators to request additional materials for the Trial State Assessments. To process a shipment, a clerk asked the caller information such as primary sampling unit, school ID, assessment type, city, jurisdiction, and Zip code. This information was then entered into the online short shipment system and a particular school and mailing address was displayed on the screen to be verified with the caller. The system allowed NCS staff to change the shipping address for individual requests. The clerk proceeded to the next screen, which displayed the materials to be selected. After the clerk entered the requested items, the due date, and the method of shipment, the system produced a packing list and mailing labels. Approximately 650 such calls were received regarding the Trial State Assessment. The number and types of calls are summarized in Table 5-4.

Table 5-4
1994 NAEP Trial State Assessment
Phone Request Summary

Number of Calls	Request
117	Excluded Student Questionnaires/IEP/LEP
318	Teacher Questionnaires
50	Additional bundles (some due to increasing sessions or replenishing supervisor's supply)
11	School Characteristics and Policies Questionnaire
92	Additional miscellaneous materials (some missing in original shipment, some due to increasing sessions or sample)
34	Change in administration date, disposition, session information, tracing unreceived shipments, general questions

5.4 PROCESSING

5.4.1 Overview

The following describes the stages of work involved in receiving and processing the documents used in the 1994 Trial State Assessment, as illustrated in Figure 5-2. NCS staff created a set of predetermined rules and specifications to be followed by the processing departments within NCS. Project staff performed a variety of procedures on materials received from the assessment supervisors before releasing these materials into the processing system. Control systems were used to monitor and route all NAEP materials returned from the field. The NAEP process control system contained the status of all sampled schools for all sessions and their scheduled assessment dates. As materials were returned, the process control system 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 the processed counts. NCS report programs allowed ETS, Westat, and NCS staff to monitor the progress and the receipt control operations. The processing flow is illustrated in Figure 5-2.

An "alerts" process was utilized to record, monitor, and categorize all discrepant or problematic situations. Throughout the processing cycle, alert situations were identified based upon the processing specifications. These situations were either flagged by computer programs or identified using clerical procedures. All situations that could not be directly resolved by the staff involved in the given process were documented. A form describing the problem was completed and the information was forwarded to project personnel for resolution.

NCS's work flow management system was used to track batches of student booklets, school questionnaires, teacher questionnaires, IEP/LEP student questionnaires, and rosters through each processing step, allowing project staff to monitor the status of all work in progress. The work flow management system was also used by NCS to analyze the current work load, by project, across all work stations. By routinely monitoring these data, NCS's management staff was able to assign priorities to various components of the work and monitor all phases of the data receipt and processing.

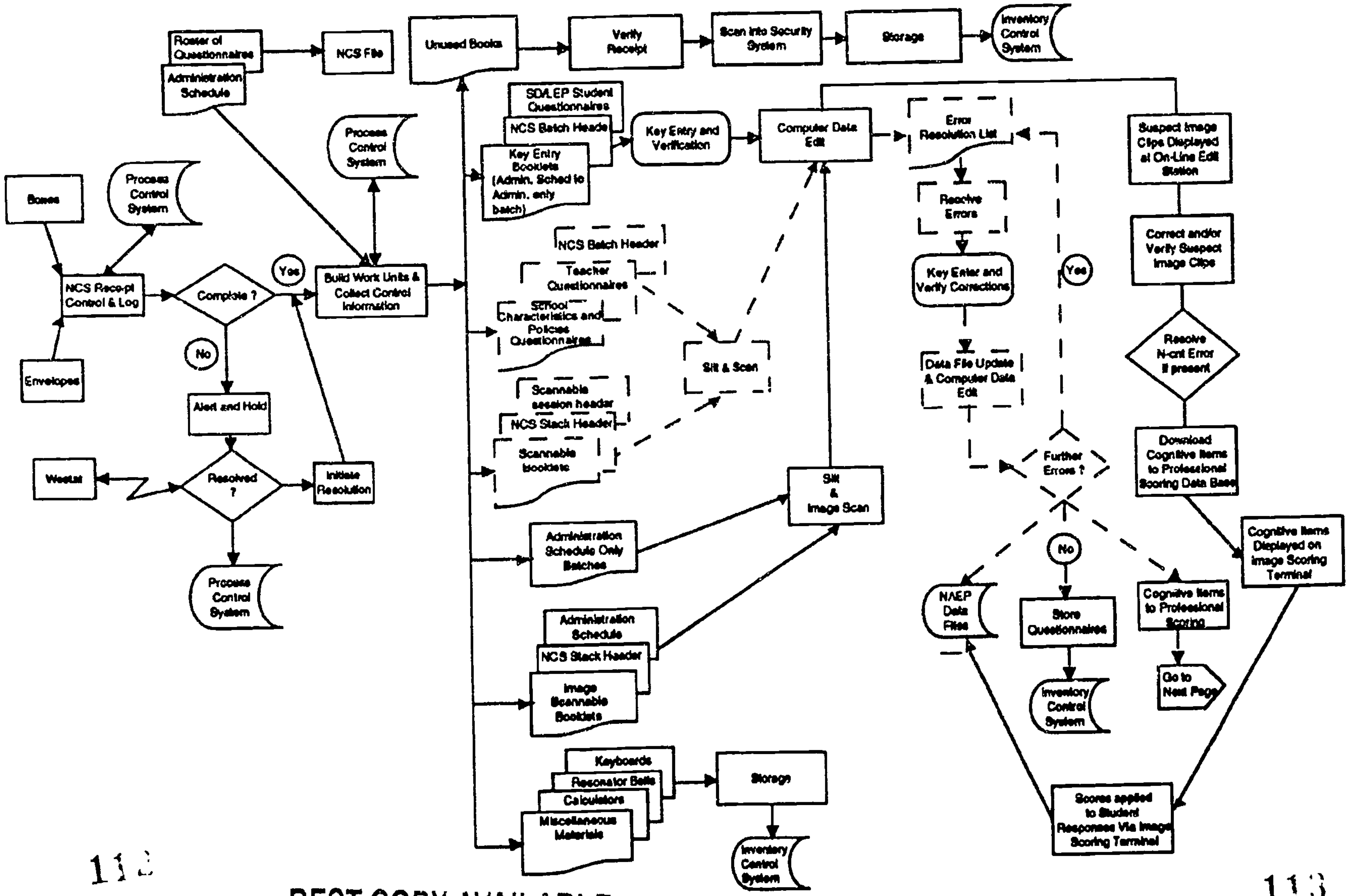
NCS used a team approach to facilitate the flow of materials through all data processing steps. The image processing team checked in the materials from the field, created the batches to be scanned, scanned the booklets, edited the information when the program found errors or inconsistencies, selected quality control samples, and sent the completed batches to the warehouse for storage. Advantages to the team environment included less duplication of effort and improved quality control measures.

5.4.2 Document Receipt and Tracking

All shipments were to be returned to NCS packaged in the original boxes. As mentioned earlier, NCS packaging staff applied a bar code label to each box that indicated the NAEP school ID number. When the shipment arrived at the NCS dock area, this bar code was scanned to a personal computer file and sorted by assessment type. The shipment was then

Figure 5-2

1994 NAEP Trial State Assessment
Materials Processing Flow Chart



11.

113

BEST COPY AVAILABLE

forwarded to the receiving area. The personal computer file was then transferred to the mainframe and the shipment receipt date was applied to the appropriate school within the process control system. This provided the current status of receipts regardless of any processing delays. The receipt was reflected on the control system status report provided to the receiving department and was also supplied to Westat via electronic data file transfer.

The process control system could be updated manually to reflect changes. Receiving personnel also checked the shipment to verify that the contents of the box matched the school and session indicated on the label. Each shipment was checked for completeness and accuracy. If it was discovered that a shipment had not been received within seven days of the scheduled assessment date, project staff were alerted. Project staff would then check the administration status of the session and, in some cases, initiated a trace on the shipment.

If multiple sessions were returned in one box, the contents of the package were removed and separated by session. The shipment was checked to verify that all booklets preprinted or hand-written on the Administration Schedule were returned with the shipment and that all administration codes matched from the booklet covers to the Administration Schedule. If discrepancies were discovered at any step in this process, the receiving staff issued an alert and held the session for resolution by the NAEP project staff.

If a make-up session had been scheduled, receiving staff issued an information alert to facilitate tracking, and the documents were placed on holding shelves until the make-up session documents arrived.

Once all booklets listed on the Administration Schedule for sessions containing scannable documents were verified as being present, the entire set of session materials, including the Administration Schedule and booklets, was forwarded to the batching area and a batch created on the work flow management system using the scannable Administration Schedule as a session header. The booklets were batched by grade level and assessment type. Each batch was assigned a unique batch number. The batch number, created on the image capture environment system and automatically uploaded to the work flow management system, facilitated the internal tracking of the batches and allowed departmental resource planning. All other scannable documents, questionnaires, and rosters were batched by document type in the same manner.

The batched documents were then forwarded to the scanning area, where all information on the Administration Schedule and booklets were scanned via a W201 image scanner. All information from the Administration Schedules was read by the intelligent character recognition engine and verified by online editing staff. Information gathered throughout this process, which included the school number, session code, counts of the students in original sample, supplemental sample, and total sample; numbers of students withdrawn, excluded, to be assessed, absent, assessed in original, and assessed in makeup; and total number of assessed students was transferred electronically to Westat on a weekly basis to produce participation statistics.

Two rosters were used to account for all questionnaires. The Roster of Questionnaires recorded the distribution and return of the school questionnaire and the IEP/LEP student questionnaire. The Roster of Teacher Questionnaires recorded teacher questionnaires distributed and returned for their respective students. Some questionnaires may not have been available for return with the shipment. These were returned to NCS at a later date in an envelope provided for that purpose. The questionnaires were submitted for scanning as sufficient quantities became available for batching.

Receipt of the questionnaires was entered into the system using the same process used for the Administration Schedules. The rosters were grouped with other rosters of the same type from other sessions, and a batch was created on the image capture environment system. The batch was then forwarded to scanning where all information on the rosters was scanned into the system.

A sophisticated booklet accountability system was used to track all booklets distributed. Prior to the distribution of materials, unique booklet numbers were read into a file by bundle. This file was then used to control distribution by assigning specific bundles to supervisors or schools. This assignment was recorded in the materials distribution system.

When shipments were received, the used booklets were submitted to processing. Unused booklets were batched and their booklet ID bar codes were read into a file by the bar code scanner. This file and the processed documents file were later compared to the original bundle security file. A list of unmatched booklet IDs was printed in a report that was used to confirm nonreceipt of individual booklets. At the end of the assessment period, the supervisors returned all unused materials. When these materials were returned, the booklet IDs were read into a file by the bar code scanner. Any major discrepancies were directed to Westat for follow-up. The unused materials were then inventoried and sent to the NCS warehouse for storage.

The Receipt Control Status Report displayed the current status of all schools. This report could be sorted by school number or by scheduled administration date. As the receiving status of a school was updated through the receiving, opening, and batching processes, the data collected were added to this report. Data represented on this report included participation status, shipment receipt date, and receipt of the Roster of Questionnaires. The comment field in this report showed any school for which a shipment had not been received within seven days of the completion of the assessment administration.

5.4.3 Data Entry

The transcription of the student response data into machine-readable form was achicved through the use of three separate systems: 1) data entry, which included optical mark recognition scanning, image scanning, intelligent character recognition), and key entry; 2) validation (edit); and 3) resolution.

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 two methods was used to transcribe NAEP data to a computerized form. The data on scannable documents were collected using NCS optical scanning equipment and also captured images of the constructed response items. Nonscannable materials were keyed through an interactive online system. In both of these cases, the data were edited and suspect cases were resolved before further processing.

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

During the scanning process (shown in Figure 5-3), each scannable NAEP document was uniquely identified using a print-after-scan number consisting of the scan batch number and the sequential number within the batch. The number was assigned to and printed on one side of each sheet of each document as it exited the scanner. This permitted the data editors to quickly and accurately locate specific documents during the editing phase. The print-after-scan number remained with the data record and provided a method for easy identification and quick retrieval of any document.

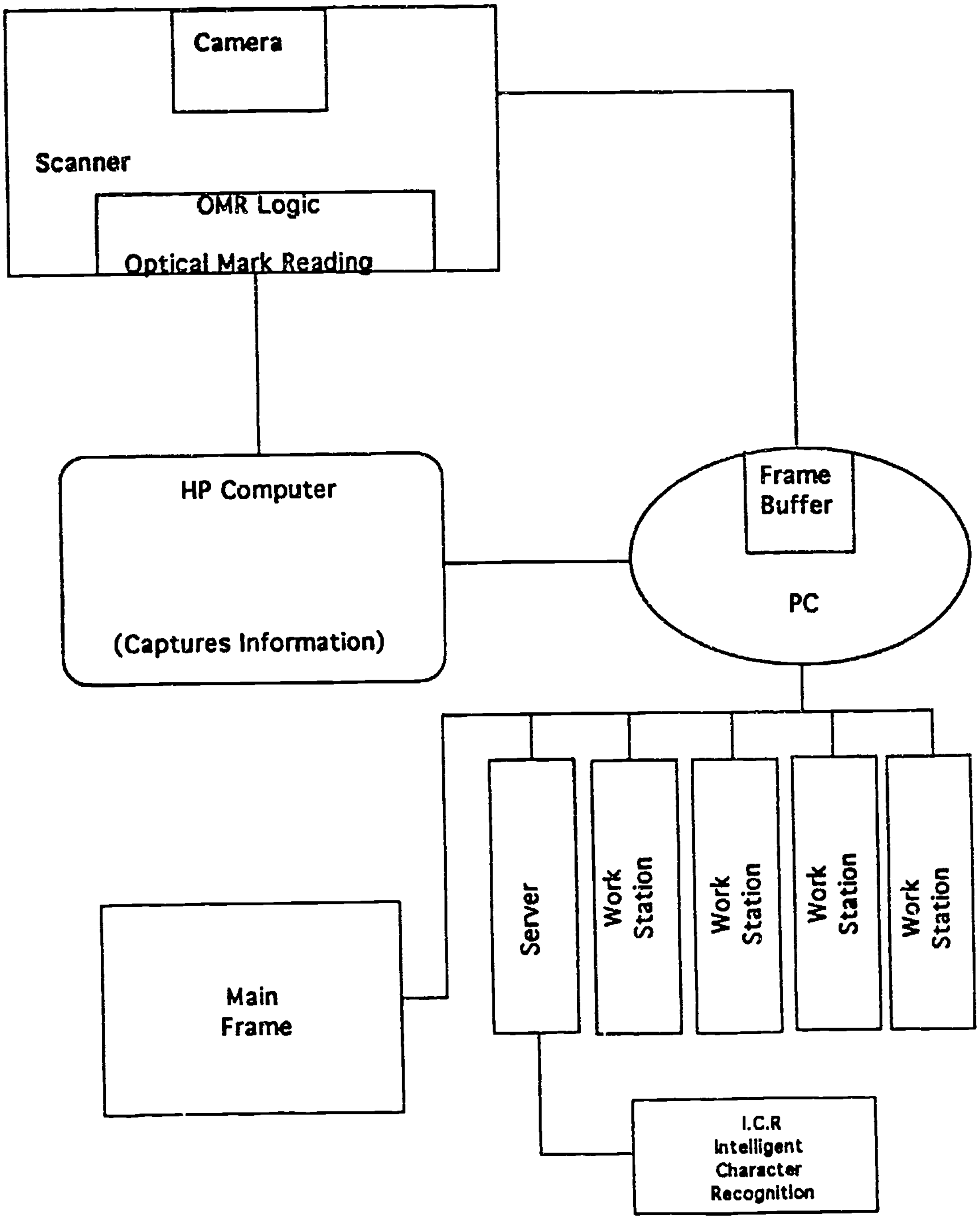
The data values were captured from the booklet covers and Administration Schedules and were coded as numeric data. Unmarked fields were coded as blanks and processing staff were alerted to missing or uncoded critical data. Fields that had multiple marks were coded as asterisks. The data values for the item responses and scores were returned as numeric codes. The multiple-choice, single response format items were assigned codes depending on the position of the response alternative; that is, the first choice was assigned the code "1," the second was assigned "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" and the unmarked choices as blanks. The images of constructed response items were saved as a digitized computer file. The area of the page that needed to be clipped was defined prior to scanning through the document definition process. The fields from unreadable pages were coded "X" as a flag for resolution staff to correct.

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

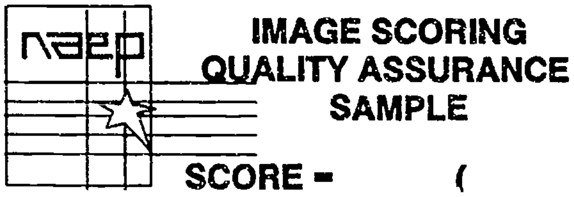
An intelligent character recognition engine was used to read various hand and machine print on the front cover of the assessment and supervisor documents. Information from the Administration Schedule, the Rosters of Questionnaires, and some questions in the school questionnaire were read by the engine and verified by a key entry operator. Analysis by NCS development staff of the accuracy of characters read via intelligent character recognition determined that the recognition engine read as well as two people processing information using a key entry and 100 percent verify method of data input. In all, the intelligent character recognition engine read nearly 6 million characters for the 1994 NAEP Trial State Assessment. This saved NAEP field staff and school personnel a significant amount of time since they no longer had to enter this data by gridding rows and columns of data.

Figure 5-3

1994 NAEP Trial State Assessment
Image Scanning Flow Chart



To provide yet another quality check on the image scanning and scoring system, NCS staff implemented a quality check process by creating labels with a valid score designated on them. Each unique item scored via the image system had two quality control labels per valid score. These labels were attached to blank, unused booklets by clerical staff and sent through the scanning process. An example of the label used is given below.



naep

IMAGE SCORING
QUALITY ASSURANCE
SAMPLE

SCORE = ( )

Although the quality control booklets were batched and processed separately from assessed student booklets, they were sent through the same process as the student document" Since all of a specific item are batched together for transmission to the scoring facility, the labeled responses were integrated with and transmitted simultaneously to the scoring facility with the student responses. During the scoring process, both student responses and the quality control items were randomly displayed so scores could be applied.

When a reader saw the quality control label on the monitor, he or she notified the team leader to watch and confirm the score while the reader assigned the score given on the label. The quality control booklets were included in the pool of all items to be drawn from for the 25 percent reliability rescore. Analysis of the data captured from this quality assurance process showed 100 percent accuracy in the system software design for capturing scores assigned to constructed-response items and linking them back to the original student document.

A key entry and verification process was used to make corrections to the teacher questionnaires and the IEP/LEP student questionnaires. The Falcon system that was used to enter this data is an online data entry system designed to replace most methods of data input such as keypunch, key-to-disk, and many of the microcomputer data entry systems. The terminal screens were uniquely designed for NAEP to facilitate operator speed and convenience. The fields to be entered were titled to reflect the actual source document.

5.4.4 Data Validation

NCS used the same format used in the 1992 assessment and the 1993 field test to set up the document definition files for the large numbers of unique documents used in the 1994 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 used it in many documents. The procedure used was a *document* file that pointed to the appropriate blocks on a *block* file that pointed to appropriate items on an *item* file. With this method of definition, a document was made up of blocks, which were made up of items.

Each dataset produced by the scanning system contained data for a particular batch. These data had to be edited for type and range of response. The data entry and resolution system used was able to process a variety of materials from all age groups, subject areas, control documents, and questionnaires simultaneously, 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 pre-edit data file or an edit log 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 pre-edit data file and/or recorded all errors on the edit log. The data fields on an edit log record identified each data problem by the batch sequence number, booklet serial number, section or block code, field name or item number, and data value. After each batch had been processed, the program generated a listing or online edit file of the data problems and resolution guidelines. An edit log listing was printed at the termination of the program for all non-image documents and image "clips" were routed to online editing stations for those documents that were image-scanned.

As the program processed each data record, it first read the booklet number and checked it against the session code for appropriate session type. Any mismatch was recorded on the error log and processing continued. The booklet number was then compared against the first three digits of the student identification number. If they disagreed, a message was written to the error log. The remaining booklet cover fields were read and validated for the correct range of values. The school codes had to be identical to those on the process control system record. All data values that were out of range were read "as is" but flagged as suspect. All data fields that were read as asterisks were recorded on the edit log or online edit file.

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

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

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

Moving the translated and edited data field into the output buffer was the last task performed in this phase of processing.

When the entire document had been processed, the completed string of data was written to the data file. When the program encountered the end of a file, it closed the dataset and generated an edit listing for non-image and key-entered documents. Image scanned items which required correction were displayed on an online editing terminal.

Accuracy checks were performed on each non-image batch processed. The record of every 500th document of each booklet/document type was printed in its entirety, with a minimum of one document type per batch. This record was checked, item by item, against the source document. If inconsistencies were found, project personnel were contacted and processing stopped.

5.4.5 Editing for Non-image and Key-entered Documents

Throughout the system, quality procedures and software ensured that the NAEP data were correct. The machine edits performed during data capture verified that each sheet of each document was present and that each field had an appropriate value. All batches entered into the system, whether key entered or machine scanned, were checked for completeness.

Data editing took place after these checks. This consisted of a computerized edit review of each respondent's record and the clerical edits necessary to make corrections based upon the computer edit. This data editing step was repeated until all data fell within a valid range.

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

During data entry, counts of processed documents were generated by type. These counts were balanced against the information captured from the administration schedules. The number of assessed and absent students processed had to match the numbers indicated on the process control system.

In the second phase of data editing, an experienced editing staff used a predetermined set of specifications to review the field errors and record any necessary correction to the student data file. The same computerized edit list used in the first phase was used to perform this function. The process was as follows:

The editing staff reviewed the edit log prepared by the computer and the area of the source document that was noted as being "suspect" or containing possible errors. The current composition of the field was shown in the edit box. The editing staff checked this piece of information against the NAEP source document. At that point, one of the following took place:

> *Correctable error*: If the error was correctable by the editing staff according to the editing specifications, the corrections were noted on the edit log.
>
> *Alert*: If an error was not correctable according to the specifications, an alert was issued to the operations coordinator for resolution. Once the correct information was obtained, the correction was noted on the edit log.
>
> *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, the programs were tailored to allow this information to be accepted into the data record and no corrective action was taken.

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

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

5.4.6 Data Validation and Editing of Image-processed Documents

The paper edit log was replaced by online viewing of suspect data for all image-processed documents. The edit criteria for each item or items in question also appeared on the screen at the same time the suspect item was displayed for rapid resolution. Corrections were made at this time. The system employed an edit/verify system which ultimately enabled two different online-edit operators to view the same suspect data and work on it separately. The "verifier" must make sure that the two responses (one from either the "entry" operator or the intelligent character recognition engine) were the same before the system would accept that item as being corrected. The verifier was able to overrule or agree with the original correction made if the two were discrepant. If the editor was unable to determine the appropriate response, he or she escalated the suspect situation to a supervisor.

When an entire batch was through the edit phase, it was then eligible for the count verification phase. The administration schedule data were examined systematically for booklet IDs that should have been processed (assessed, absent, and excluded administration codes). The documents under an individual administration schedule were then inspected to ensure that all of the booklet IDs listed on it were present.

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 then produced, and, if errors remained, the paper edit log was forwarded to another editor. When this edit was satisfied, the appropriate tracking mechanisms (the process control and work flow management systems) were updated.

5.4.7 Data Transmission

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

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

5.5 PROFESSIONAL SCORING

5.5.1 Overview

Scoring of the 1994 NAEP Trial State Assessment constructed-response items was conducted using NCS's image technology. All 1994 responses were scored online by readers working at image stations. The logistical problems associated with handling large quantities of student booklets were removed for those items scored on the image system.

One of the greatest advantages image technology presented for NAEP scoring was in the area of sorting and distributing work to scorers. All student responses for a particular item, regardless of where spiraling had placed that item in the various booklet forms, were grouped together for presentation to a team of readers. This allowed training to be conducted one item at a time, rather than in blocks of related items, thus focusing readers' attention on the complexities of a single item.

A number of tools built into the system allowed table leaders and trainers to closely and continuously monitor reader performance. A detailed discussion of these tools can be found later in this chapter.

The system automatically routed 25 percent of student responses to other members of the team for second scoring. Readers were given no indication of whether the response had been scored by another reader, thereby making the second scoring truly blind. On-demand, real-time reports on interreader reliability (drawn from those items that were second-scored) presented extremely valuable information on team and individual scoring. Information on adjacent and perfect agreement, score distribution, and quantity of responses scored were continuously available for consultation. Similarly, back-reading of student responses could be accomplished in an efficient and timely manner. Table leaders were able to read a large percentage of responses, evaluating the appropriateness and accuracy of the scores assigned by readers on their teams.

Project management tools assisted table leaders in making well-informed decisions. For example, knowledge of the precise number of responses remaining to be scored for a particular item allowed table leaders to determine the least disruptive times for lunch breaks.

Concerns about possible reader fatigue or other problems that might result from working continuously at a computer terminal proved unfounded. Both readers and table leaders responded with enthusiasm to the system, remarking on the ease with which student responses could be read and on the increased sense of professionalism they felt in working in this technological environment. Readers took periodic breaks, in addition to their lunch break, to reduce the degree of visual fatigue. Readers were grouped in teams of 6 to 10 readers per team. Individual rooms were set up with each room containing teams for a single subject area.

5.5.2 Training Paper Selection

A pool of papers to be used during training for the national main assessment was selected by NCS staff in February 1994. During the interview process, NCS scoring specialists identified those candidates with team leader potential. Individuals recruited to be team leaders during the actual scoring were asked to select student responses to send to ETS test development specialists, who created the master training set. Team leaders were used for this task because it gave them the advantages of working on specific items, learning the make-up of the various booklets, learning the terminology, and understanding the processing of the booklets at NCS. This was especially important in 1994, because most scoring activities occurred via the image processing system.

The training set for each short (two- or three-point) item included 40 papers:

- 10 anchor papers
- 20 practice papers
- 5 papers in calibration set #1
- 5 papers in calibration set #2

The training set for each extended (four-point) item included 85 papers:

- 15 anchor papers
- 40 practice papers
- 10 papers in each of two qualification sets
- 5 papers in calibration set #1
- 5 papers in calibration set #2

To ensure that the ETS test development specialist would have a wide range of student responses to encompass all score points, NCS personnel copied approximately 100 papers for each two- or three-point item and 200 papers for each four-point item. To ensure that training papers represented the range of responses obtained from the sample population, NCS personnel selected papers randomly from across the sample. The student identifier (barcode) was written on the copy. The responses were numbered sequentially, copied, and sent via overnight delivery to ETS. When the training packet was compiled, the ETS test development specialist faxed the composition of the packets back using the sequential numbers. ETS staff kept its copy of the training sets. A total of 4,100 student responses were forwarded to ETS to be used in the creation of training packets.

From the faxed sheets, packets were created for each item using the first generation copy. These packets were then forwarded to the NCS communication center for copying, and stored for the team's use in training. ETS also sent the most up-to-date version of the scoring guide for each item to be included in the scoring guide.

5.5.3 General Training Guidelines

ETS personnel conducted training for the constructed-response items on an item-by-item basis, so that each item could be scored immediately after training. Reading items tied to a common stimulus were trained and scored sequentially, finishing one block before proceeding to the next.

In all, 13 team leaders and 120 readers worked from March 28 to May 27, 1994 to complete scoring for the 1994 NAEP Trial State Assessment. Each member of a team received a copy of the stimulus and training materials for the items which his or her team would be scoring. Before training, each team member read the stimulus and discussed it under the guidance of the trainer where applicable. Next, ETS staff conducted training sessions to explain the anchor papers, exemplifying the various score point levels. The team proceeded with each member scoring the practice papers, and then discussing those papers as a group while the trainer clarified issues and answered questions. The papers selected for each training set were chosen to illustrate a range from easily classifiable responses to borderline responses for each score point.

When the trainer was confident the readers were ready to begin scoring short constructed responses, the table leader signaled the system to release the responses to the team members who had successfully completed training. For extended constructed-response items, each team member was given a qualifying set which had been prescored by the trainer in conjunction with the table leader. Readers were required to score an exact match on 80 percent of the items in order to qualify for scoring. If a reader failed on the first attempt, the trainer discussed the discrepant scores with the reader and administered a second qualifying set. Again, 80 percent exact agreement was required to score the item. During the beginning stages of scoring, the team members discussed student responses with the trainer and table leader to ensure that issues not addressed in training were handled in the same manner by all team members.

After the initial training, readers scored the items, addressing questions to the table leader and/or trainer when appropriate. Depending upon n-counts, length of responses and complexity of the rubric, scoring of an individual item ranged anywhere from one-half hour to two weeks. Whenever a break longer than 15 minutes occurred in scoring, each team member received a set of calibration papers which had been prescored by the trainer and table leader. Each team member scored the calibration set individually, and then the team discussed the papers to ensure against scorer drift.

5.5.4 Table Leader Utilities and Reliability Reports

Among the many advantages of the image scoring system is the ease with which work flow to readers can be regulated and scoring can be monitored. One of the utilities at a table leader's disposal was a qualification algorithm executed upon completion of training on an extended constructed-response item. At that time, a table leader passed out a qualification packet of 10 papers whose scores had been entered as a master key on the table leader's workstation. Upon completion of the packet, the table leader entered each reader's scores into the computer for tabulation and the computer calculated each reader's percent of exact, adjacent, and nonadjacent agreement with the master key. If a reader had a percent of exact agreement above a predetermined threshold, the reader was authorized to begin scoring that item. Readers not reaching the predetermined threshold were handled on a case-by-case basis, typically receiving individual training by the ETS trainer or the NCS table leader before being allowed to begin scoring. A table leader also had the authority to cancel a reader's qualification to score an item if review of a reader's work indicated the reader was scoring inaccurately.

After scoring commenced, review of each reader's progress was conducted using a back-reading utility that allowed a table leader to review every paper scored by each reader on the table. Typically a table leader would choose the ID number of a reader and review a minimum of 10 percent of the responses scored by that reader, making certain to note the score the reader awarded each response as well as the score a second reader gave that same paper as an interreader reliability check. Alternately, a table leader could select to review all responses receiving a specific score in order to determine if the whole team was scoring consistently. Both review methods utilized the same display screen and revealed the ID number of the reader and the score awarded. If the table leader disagreed with the score given a response, the table leader would discuss the discrepancy with the reader and possibly replace the score of the questionable response. Scores were replaced by the table leader only when the scorer had made an obvious error. The main purpose of this monitoring was to provide early identification of problems and opportunities to retrain scorers when needed.

A minimum of 25 percent of the 1994 reading responses were scored twice. The image system presented all responses in the same manner, so the reader could not discern which responses were being first-scored and which were designated for a second scoring. The table leader and the ETS trainer were able to monitor these figures on demand. The system showed the overall reliability for the group scoring the item and individual reliability of the qualified readers.

During the scoring of an item, the table leader could monitor progress using an interreader reliability tool. This display tool could be used in either of two modes—to display information of first readings versus second readings, or to display first reading of an individual versus second readings of that individual.

The table leaders were able to monitor work flow using a status tool that displayed the number of items completed, the number of items that still needed second scoring, and the number of items that had not been scored up to that time.

Table 5-5 shows the number of constructed-response items falling into each range of percentages of exact agreement. Tables 9-2 and 9-3 in Chapter 9 show more reliability information about the constructed-response items used in the NAEP scale.

Table 5-5
1994 NAEP Trial State Assessment
Number of Constructed-response Items
in Each Range of Percentages of Exact Agreement Between Readers

Grade 4 Reading Items	Number of Unique Items	60-69%	70-79%	80-89%	90-100%
Short constructed-response items	37	0	0	8	29
Extended constructed-response items	8	0	1	6	1

5.5.5 Main and Trial State Reading Assessment

It is important to note that the student responses in the fourth-grade reading assessments were scored concurrently for the national and the state samples. Another advantage of image-based item-by-item scoring is that the comparability of the scoring of the two samples is ensured since all responses are scored simultaneously and in a manner which makes is impossible for the scorers to know from which sample any individual response is. Because of this, the following discussion addresses both national (main) and state reading.

5.5.6 Training for the Main and State Reading Assessment

The reading assessment followed the basic training procedures outlined in section 5.5.4. One trainer provided all the training for the fourth-grade items scored. One trainer followed the fourth-grade items through from beginning to end.

5.5.7 Scoring the Main and State Reading Assessment

Each constructed-response item had a unique scoring standard that identified the range of possible scores for the item and defined the criteria to be used in evaluating the students' responses. Point values were assigned with the following meanings:

Dichotomous items from the 1992 assessment

- 1 = Unacceptable
- 4 = Acceptable

Dichotomous items developed during the 1993 field test

- 1 = Evidence of little or no comprehension
- 3 = Evidence of full comprehension

Three-point items developed during the 1993 field test

- 1 = Evidence of little or no comprehension
- 2 = Evidence of partial or surface comprehension
- 3 = Evidence of full comprehension

All four-point items

- 1 = Evidence of unsatisfactory comprehension
- 2 = Evidence of partial comprehension
- 3 = Evidence of essential comprehension
- 4 = Evidence of extensive comprehension

The scores for these items also included a 0 for no response, 8 for an erased or crossed-out response, and a 9 for any response found to be unratable (i.e., illegible, off-task, responses written in a language other than English, or responses of "I don't know").

During scoring, the table leaders compiled notes on various responses for the readers' reference and guidance and for the permanent record. In addition, trainers were accessible for consultation in interpreting the guides for unusual or unanticipated responses. The table leaders conducted constant online back-reading of all team members' work throughout the scoring process, bringing to the attention of each reader any problems relating to scoring. When deemed appropriate, scoring issues were discussed among the team as a whole. Table leaders also monitored n-counts of responses scored and individual and team reliability figures throughout the course of scoring.

Each item was scored by a single team immediately after training for that item. Team sizes averaged 10 scorers.

Grade 4 items came from both a national and a state-by-state sample. Responses were delivered by image in such a way that the student demographics were unknown to the reader. Thus, readers did not know from which sample any given item came when it appeared on the screen. In the case of overlap items, all readers scored responses at both grade levels.

5.5.8 1992 Short-Term Trend and Image/Paper Special Study

Sixteen blocks from the 1992 reading assessment were re-used in the 1994 assessment to provide data with which to study trends over time. To accomplish this, a random sample of responses from the 1992 assessment were pulled from the warehouse for rescoring to determine whether or not the scoring performed in 1994 was comparable to the scoring performed in 1992. For the national sample, ETS measurement personnel identified three booklets at grade 4, four booklets at grade 8, and five booklets at grade 12 which contained all of the blocks needed for the study. The entire sample of those booklets was used for the rescore study. Since each block appears in four booklets, rescoring the entire sample of one booklet resulted in a 25 percent rescore of the responses from 1992. For the state sample, 12 booklets were pulled for each unique booklet type (R30 through R45) for each of the 41 jurisdictions which participated in both the 1992 and 1994 Trial State assessments. Since each block appeared in four different booklet types, 48 responses to each item were rescored for each jurisdiction. These booklets were scanned to capture the same clip areas used for the 1994 responses. Thus they appeared identical to the reader when viewing them on the monitor and were presented at the same time as the 1994 responses.

After scanning was completed, the national sample of the rescore booklets was transported to the scoring facilities to be scored on paper. Paper scoring took place at the same time as image scoring. This process yielded data to compare the paper-based scoring done in 1992, the paper-based scoring done in 1994, and the image-based scoring done in 1994. Analyses performed on these data will be documented in *The NAEP 1994 Technical Report.*

5.5.9 Calibration Bridges

Unanticipated delays in receipt and processing of student booklets resulted in a situation in which scoring for some constructed-response items began before all or most of the student responses for those items were available for scoring. The result was that the responses for most of the 1994 constructed-response items were scored in two different scoring sessions ("sweeps"). To maintain the highest standards of scoring and measurement precision and to ensure that calibration error was not introduced as a result of the split scoring sessions and the time elapsed between them, a plan was devised to calibrate the scoring of sweeps 1 and 2. In some instances, it was determined that scoring could resume with a review of training and a regular calibration set to ensure consistency and reliability. In other instances, a calibration bridge was constructed to provide statistical linkage between the two scoring sessions.

It was determined that scoring could continue without the calibration bridge in those instances in which completed scoring had met two criteria: 50 percent scored on the first sweep and interreader reliability equal to or greater than 95 percent.

For those items not meeting these criteria, a set of papers was scored to provide a reliability link or calibration bridge between completed scoring ("first sweep") and subsequent scoring ("second sweep"). The procedures followed for completing the calibration bridge were as follows:

1. Approximately 12,500 processed booklets were pulled from inventory and, from them, samples of student responses were constructed for each item designated for the calibration bridge scoring.

2. A file of all pulled booklet ID numbers was created along with all scores assigned in the first sweep of scoring. This allowed for matching scores assigned in the first sweep to those given in the paper-based calibration bridge rescoring.

3. For each designated item, each scorer read and scored at least 10 student responses drawn from this sample (10 different papers for each scorer). No 0 score papers were included, and 20 percent of the responses were scored twice for interreader reliability.

4. The clerical support staff entered the scores in a spreadsheet program which produced data on reader agreement, score distributions, mean scores, and standard deviations of the mean scores.

5. The data from the calibration bridge scoring was compared to the data for the first sweep scoring on the same item.

6. After reviewing these data, items meeting the following criteria were determined to be ones for which second sweep scoring could then proceed:

   - items for which the Diff T test was not significant, >.05, e.g. the null hypothesis cannot be rejected.

   - items for which the bridge/sweep percent agreement was higher than the designated threshold of 90 percent reliability for two-point items, 80 percent reliability for three-point items, and 75 percent reliability for four-point items.

   - items for which the bridge interreader reliability was no more than six percentage points lower than first sweep interreader reliability.

7. For those items not meeting the criteria, readers were retrained. Following the retraining, five different papers from the sample were read by each reader and results evaluated.

8. Following analysis of the results of scoring following this retraining, a decision was made to continue scoring or, alternately, to rescore all the previously scored responses along the remaining responses for the item under consideration. A total of 95 calibration bridges were conducted in reading.

5.5.10 The Performance Scoring Center

The performance scoring center uses a desktop scanner interfaced with a PC for collecting score data. The software, scanner, and performance center scoring sheet used for NAEP were all developed by NCS. This scoring system is designed to add efficiency and portability to traditional paper-based scoring projects. The scoring system software is customized to NAEP's needs including all items and valid score ranges. The demographic information, batch, sequence, and barcode numbers are pre-slugged onto the performance center scoring sheet obtained from the clean-post file after the editing process. These score sheets are then delivered to the scoring center for use when scoring the student documents

The performance scoring center system offers unique attributes that are ideal for paper-based scoring projects. One advantage the system offers is the capability of scanning scoring sheets in random order. This provides the means for continuous scanning if a scoring sheet is rejected with an error (e.g., score out of range). Another advantage is the ability to produce inter-reader reliability information upon request. The reliability reports produced record the total occurrences of second score for each item. It also reflects the total for agree and disagree and calculates the percentages of agreement. Reports can be produced, on request, on an item basis for a particular team or an individual reader. This enables us to ensure the validity of scoring by item, reader and team. Additionally, the performance scoring center system has the unique ability to produce reports that indicate [illegible] of sheets left to scan by project, batch, and sheet. This guarantees scoring sheet accountability and assures that a score is assigned to every student response.

The 1994 national and state assessments had some components that were not conducive to image scoring but were ideal for scoring using the performance scoring center system, including the NAEP Packet (1992 rescore). The NAEP 1992 rescore items did not require second scoring; therefore, there is no interreader reliability information to report on that component.

5.6 DATA DELIVERY

The 1994 NAEP data collection resulted in several classes of data files—student, school, teacher, excluded student, IEP/LEP student, sampling weight, student/teacher match and item information. Item information included item data from all assessed students in 1994, item data for the short-term reading trend, and item data from the special study comparing image-based and paper-based scoring. Data resolution activities occurred prior to the submission of data files to ETS and Westat to resolve any irregularities that existed. This section details additional steps performed before creating of the final data files to ensure the most complete and accurate information was captured.

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

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

Some of the assessed students were determined to be ineligible for the assessment because they did not match the particular age/grade being sampled or because of unusual circumstances. At the conclusion of each assessment, it was necessary to delete the records of these students from the NAEP database. Deleting this information required compiling a list of all student records that had been processed with administration codes other than those for assessed students. To do this, the process control system and the Administration Schedule data were referenced. If the system showed a discrepancy, project personnel pulled the Administration Schedules and other documentation (e.g., alerts, student booklets, etc.) to verify and resolve the discrepancy.

The edits and data verification performed on the IEP/LEP student questionnaires assured that information regarding the IEP/LEP status of the students was not left blank. If there was no indication as to IEP or LEP on the questionnaire cover, the edit clerk cross-checked the administration schedule(s) and student booklet cover to confirm the IEP/LEP status of the student. If this information was not available from the questionnaire cover, booklet cover, or the administration schedule, the edit clerk viewed the information indicated in question #1 (which asked why the student was classified IEP or LEP) to see whether responses written there might yield useful information. Then the determination was made as to how the student should be classified.

The school questionnaires were revised for 1994 so that some items that had required school staff to provide a percentage figure by gridding ovals in a matrix were changed to allow the respondent to simply write the percentage in a box. These data was then captured via ICR technology and verified by an edit operator.

To obtain the best possible match of teacher questionnaires to student records, the same processes that were followed in 1992 were refined in 1994. The first step in matching was to identify teacher questionnaires that had not been returned to NCS for processing, so as not to include the students of these teachers from the matching process. Student identification numbers that were not matched to a teacher questionnaire were then crossreferenced with the corresponding Administration Schedule and Roster of Teacher Questionnaires to verify the teacher number, teacher period, and questionnaire number recorded on these control documents. If a change could be made that would result in a match, the correction was applied to the student record. The NAEP school numbers listed on the Roster of Questionnaires, Administration Schedule, and teacher questionnaire were verified and corrected, if necessary.

Once these resolutions were made, any duplicate teacher numbers that existed within a school were crossreferenced with the Rosters of Questionnaires for resolution, if possible. In one jurisdiction that had multiple sessions in many schools, a number of the schools used a single Roster of Questionnaires for each session. This resulted in a larger than expected number of duplicate teacher numbers that could not be resolved. The overall quality of the matching process improved in 1994 as a result of the inclusion of the teacher number and period on the Administration Schedule. Since this information was located together on a single, central control document, the ability to match and resolve discrepant or missing fields was simplified.

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

5.7 MISCELLANEOUS

5.7.1 Storage of Documents

After the batches of image-scanned documents had successfully passed the editing process, they were sent to the warehouse for storage. Batches of 1992 rescore booklets were sent to the scoring area after passing the edit phase of processing, because they were also to be scored on paper. Once paper scoring was completed, 1992 rescore booklets were also sent to the warehouse for storage. The storage locations of all documents were recorded on the inventory control system. Unused materials were sent to temporary storage to await completion of the entire assessment. After the data tape was accepted, extra inventory was destroyed and a nominal supply of materials was stored permanently.

5.7.2 Quality Control Documents

ETS requested that a random sample of booklets and the corresponding scores/scoring sheets be pulled for an additional quality control check. Because no scoring sheet was available for image-scanned documents, ETS used scores sent to them on a data tape to verify the accuracy of applied scores. For nonscannable trend booklets and for the 1992 rescore booklets that were scored on paper, both the booklet and its corresponding score sheet were sent to ETS. An average of 20 of each booklet and scores/scoring sheets for each document type were selected at random by NCS. 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.

5.7.3 Alert Analysis

Even though Receiving Department personnel were trained in the resolution of many problematic situations, some problems required resolution by NAEP staff. These are listed in Table 5-6. The types of problems were categorized and codes ("N" for national and "S" for Trial State) were assigned. For any unusual situations, Westat was called so that the Assessment Supervisors could be notified immediately to avoid further problems in test administration.

Many discrepancies were found in the receiving process that did not require an alert to be issued, but did require a great deal of effort to resolve in order to provide the most complete and accurate information. These included blank fields on covers of booklets as well as discrepancies between the booklet covers and the administration schedule. There were a total of 311 alerts for the Trial State Assessment.

Table 5-6
Alerts for 1994 National and Trial State Assessments

Code	Description
N1/S1	Booklet covers not fully completed or bubbled
N2/S3	Information on covers does not match Administration Schedule
N3/S3	Handwritten or photocopied Administration Schedule
N4/S4*	Student Listing Form returned
N5/S5	Questionnaires discrepant with roster
N8/S8*	School shipments returned unused
N10/S10	Booklets missing or unaccounted for (i.e., make-up sessions)
N11/S11	Administration Code questionable
N17/S17	Roster/Administration Schedule not received
N25/S25	Transcribing document
N26	Excluded Student Questionnaire not assigned/ # not recorded on booklet
N27/S27	IEP/LEP not assigned/ # not recorded on booklet
N28/S28*	Booklets with an administration code of 14, 19, or 27
N29/S29*	Names returned on Administration Schedule/Roster
N30/S30	Other

* Alerts requiring only an information code.

Chapter 6

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

John J. Ferris, David S. Freund, and Alfred M. Rogers

Educational Testing Service

6.1 OVERVIEW

The data transcription and editing procedures described in Chapter 5 resulted in the generation of disk and tape files containing various data for assessed students, excluded students, teachers, and schools. The weighting procedures described in Chapter 7 resulted in the generation of data files that included the sampling weights required to make valid statistical inferences about the population from which the 1994 fourth-grade Trial State Reading Assessment samples were drawn. These files were merged into a comprehensive, integrated database. The creation of the database is described in section 6.2.

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 the sample of original instruments received from the field. The results of this procedure are given in section 6.3.

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 6.4.

6.2 CREATION OF THE DATABASE

6.2.1 Merging Files into the Trial State Assessment Database

The transcription process conducted by National Computer Systems resulted in the transmittal to ETS of four data files for fourth grade: one file for each of the three questionnaires (teacher, school, and IEP/LEP student) and one file for the student response data. The sampling weights, derived by Westat, Inc., comprised an additional three files—one for students, one for schools, and one for excluded students. (See Chapter 7 for a discussion of the sampling weights.) These seven files were the foundation for the analysis of the 1994 Trial State Assessment data. Before data analyses could be performed, these data files had to be integrated into a coherent and comprehensive database.

The 1994 Trial State Reading Assessment database for fourth grade consisted of three files—student, school, and excluded student. Each record on the student file contained a student's responses to the particular assessment booklet the student was administered (booklets R1 to R16) and the information from the questionnaire that the student's reading teacher completed. Additionally, for those assessed students who were identified as having an Individualized Education Plan (IEP) or Limited English Proficiency (LEP), data from the IEP/LEP Questionnaire is included. (Note that beginning with the 1994 assessment, the IEP/LEP questionnaire replaces the excluded student questionnaire. This questionnaire is filled out for all students identified as IEP and/or LEP, both assessed and excluded. See Chapter 2 for information regarding assessment instruments.) Since teacher response data can be reported only at the student level, it was not necessary to have separate teacher files. The school files and student files (both assessed and excluded) were separate files and could be linked via the state, school, and school type codes.

The creation of the student data files began with the reorganization of the data files received from National Computer Systems. This involved two major tasks: 1) the files were restructured, eliminating unused (blank) areas to reduce the size of the files; and 2) in cases where students had chosen not to respond to an item, the missing responses were recoded as either "omitted" or "not reached," as appropriate. Next, the student response data were merged with the student weights file. The resulting file was then merged with the teacher response data. In both merging steps, the booklet ID (th: three-digit booklet number and the six-digit serial number) was used as the matching criterion.

The school file was created by merging the school questionnaire file with the school weights file and a file of school-level variables, supplied by Westat and Quality Education Department, Inc. (QED), that included demographic information about the schools such as Race/Ethnicity percentages. The state, school, and school type codes were used as the matching criteria. Since some schools did not return a questionnaire and/or were missing QED data, some of the records in the school file contained only school-identifying information and sampling weight information.

The excluded student file was created by merging the IEP/LEP student questionnaire file with the excluded student weights file. The assessment booklet serial number was used as the matching criterion.

When the student, school, and excluded student files 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 as described above.

For archiving purposes, restricted-use data files and codebooks for eacl. jurisdiction were generated from this database. The restricted-use data files contain all responses and response-related data from the assessment, including responses from the student booklets and teacher and school questionnaires, proficiency scores, sampling weights, and variables used to compute standard errors.

6.2.2 Creating the Master Catalog

A critical part of any database is its processing control and descriptive information. Having a central repository of this information, which may be accessed by all analysis and repor‑ting programs, will provide correct parameters for processing the data fields and consistent label ng for identifying the results of the analyses. The Trial State Assessment master catalog file was designed and constructed to serve these purposes for the Trial State Assessment database.

Each record of the master catalog contains the processing, labeling, classification, and location information for a data field in the Trial State Assessment database. The control parameters are used by the access routines in the analysis programs to define the manner in which the data values are to be transformed and processed.

Each data field has a 50-character label in the master catalog describing the contents of the field and, where applicable, the source of the field. The data fields with discrete or categorical values (e.g., multiple-choice and constructed-response items, but not weight fields) have additional label fields in the catalog containing 8- and 20-character labels for those values.

The classification area of the master catalog record contains distinct fields corresponding to predefined classification categories (e.g., reading content area) for the data fields. For a particular classification field, a nonblank value indicates the code of the subcategory within the classification categories for the data field. This classification area permits the grouping of identically classified items or data fields by performing a selection process on one or more classification fields in the master catalog.

The master catalog file was constructed concurrently with the collection and transcription of the Trial State Assessment data so that it would be ready for use by analysis programs when the database was created. As new data fields were derived and added to the database, their corresponding descriptive and control information were entered into the master catalog. The machine-readable catalog files are available as part of the secondary-use data files package for use in analyzing the data with programming languages other than SAS and SPSS-X (see the *NAEP 1994 Trial State Assessment in Reading Secondary-use Data Files User Guide*).

6.3 QUALITY CONTROL EVALUATION

The purpose of the data entry quality control procedure is to gauge the overall accuracy of the process that transforms responses into machine-readable data. The procedure involves examining the actual responses made in a random sample of booklets and comparing them, mark by mark and character by character, with the responses recorded in the final database, which is used for analysis and reporting.

In the present assessment, the selection of booklets for this comparison took place at the point of first entry into the recording process for data from the field. In past assessments, this selection took place only after data had reached the final database, in order to assure that only relevant booklets were involved in the quality control evaluation. While the new method of selection did result in some irrelevant booklets—due to absentee students or other problems—sufficient numbers of booklets were ultimately selected that did appear in the final database. The earlier availability of booklets for quality control evaluation and the improved efficiency of this new selection process were adequate compensation for the loss of control over which booklets were involved in quality control evaluation.

6.3.1 Student Data

Sixteen assessment booklets, R1 through R16, were administered as part of the Trial State Assessment in reading. Table 6-1 provides the numbers of each booklet for which data were scanned into data files. The variation in these numbers is trivial, indicating very good control of the distribution process.

The number of students assessed in each of the 44 participating jurisdictions varied from a low of 2,081 to a high of 3,147. All but two jurisdictions met or exceeded the target participation rate for public schools. The average number of students assessed in each jurisdiction was 2,766. This was somewhat higher than the average in 1992.

For the first time, the data entry process relied on image processing technology for recording the scores assigned by professional readers to the students' constructed responses. The scanned image of a student's response to one of these items was presented on the computer screen of a reader's work station. After determining the score for the item, the reader then entered this score using the keyboard at the work station.

This new process raised the question of what to verify or check in a quality control operation. The usual issue—whether the response that ended up in the final database is the same as the original intended response—could not be raised here, since the reader's intention, which defines the data, was entered ***directly into the database*** without any intermediate steps. The question of whether readers consistently and accurately applied agreed-upon scoring rubrics was not at issue here; that question falls into the province of reader reliability studies. In short, the data for these items existed in only one form, the database itself, and could not be verified against any earlier or preliminary form.

Rather than abdicate all quality control responsibility for these items, we chose to verify the process itself. Two important questions were examined:

1. Was the identity of the respondent maintained? Did a respondent's scores end up in his or her data record and not someone else's?

2. Was the identity of the item maintained? Did the score for each constructed-response item in this booklet end up correctly identified in the database, or was it transposed with another item response or perhaps left out?

Four different booklets in this assessment contained some number of constructed-response items requiring professional scoring. To verify that the system was functioning correctly, four sets of artificial data were carefully constructed, one set for the constructed-response items in each of these booklets. Each set consisted of two booklets, representing a total of eight "respondents". These booklets were filled in with pre-assigned scores and processed in the usual way, the only difference being that the readers were presented with the score to assign, rather than with a passage to be evaluated.

To assure correct identification of a booklet (question #1 above), the score pattern of each booklet was made unique, even under the assumption that the scorers made one recording error in every booklet. Such an error would not be relevant to the question of whether the

Table 6-1

Number of Reading Booklets Scanned and Selected for Quality Control Evaluation

Booklet Number	Total Booklets Scanned	Total Booklets Selected
R1	7,604	20
R2	7,562	19
R3	7,591	20
R4	7,630	19
R5	7,639	17
R6	7,616	18
R7	7,562	21
R8	7,525	22
R9	7,583	19
R10	7,627	21
R11	7,614	21
R12	7,637	20
R13	7,656	21
R14	7,646	21
R15	7,611	19
R16	7,615	22
Total	121,718	320

correct respondent was being scored. As noted above, the question of whether the correct score is being assigned needs to be addressed through reader reliability studies.

To assure that item identity was maintained within a booklet (question #2 above), different responses were used across the constructed-response items for each booklet. Since the number of different responses was almost never adequate to allow making each response unique within a booklet, a second sample of each booklet was needed. Any item response which had to be duplicated within the first booklet of such a pair was designed to be different in the second booklet, and vice versa.

We are pleased to report reassurance for both of the above questions. Both item and respondent integrity were maintained in these booklets of artificial data.

Student booklets were sampled in adequate numbers and the average rate of selection was about one out of 380, a selection rate comparable to that used in past assessments at both the state and national levels. The few errors found during this quality control examination did not cluster by booklet number, so there is no reason to believe that the variation in numbers of booklets selected had a significant effect on the estimates of overall error rate confidence limits reported below.

The quality control evaluation detected 14 errors in these student booklet samples, about evenly divided between multiple responses that were not identified as such by the scanner and erasures that were recorded instead of ignored. As usual, there was some indication that the error rate could be improved with further tuning of the scanner procedures; the erroneously scanned responses would not have challenged human judgment—indeed, that was the criterion used to determine whether a mis-scanning had occurred. Not to lose sight of the final goal, however, the process as it stands can still be described as adequate for the support of conclusions about educational progress in America. A very large volume of data was scanned with consistently usable results. The usual quality control analysis based on the binomial theorem permits the inferences described in Table 6-2.

Table 6-2
Inference from the Quality Control Evaluation of Grade 4 Data

Subsample	Selection Rate	Different Booklets Sampled	Number of Booklets Sampled	Characters Sampled	Number of Errors	Observed Rate	Upper 99.8% Confidence Limit
Student	1/380	16	320	19,792	14	.0007	.0015
Teacher	1/104	1	154	14,168	11	.0008	.0017
School	1/77	1	61	6,588	3	.0005	.0019
IEP/LEP	1/215	1	75	5,850	12	.0021	.0044

6.3.2 Teacher Questionnaires

A total of 16,011 questionnaires from reading teachers were associated with student data in the final database. These questionnaires were sampled at the rate of 1 in 104, roughly double the rate used in previous years. The 154 selected questionnaires contained a total of eleven errors in eleven different booklets, usually involving the scanner's mistaking an erasure for a response, but occasionally involving the failure of the scanner to pick up a multiple response. In every case, the respondent's intention was clear to the human eye, but the scanner seemed unprepared to exercise the same judgment that a careful observer would. The resulting error rate for the teacher questionnaire data was about the same as that for the student data. The quality of the teacher data is more than adequate for the purposes to which it was put.

6.3.3 School Questionnaires

A total of 4,704 questionnaires were collected from school administrators. These questionnaires were sampled for quality control evaluation at the rate of 1 in 77, resulting in the selection of 61 questionnaires. The three errors that were found represent an error rate about the same as that for the teacher questionnaire data, and about the same as that for school questionnaires in past years.

6.3.4 IEP/LEP Student Questionnaires

A total of 16,149 IEP/LEP questionnaires were scanned. About half of these questionnaires appear in the main student database, representing students who were included in the assessment. In the past, all students given this kind of questionnaire were excluded, and the instrument was referred to as the excluded student questionnaire. The overall selection rate was about 1 in 215, comparable to that used in earlier assessments for this questionnaire. Seventy-five questionnaires were selected in all. Both the selection rates and the resulting error rates were about the same in the two pools of students. Nearly all of the 12 errors found were due to the scanner's mistaking an erasure for an intended response. The quality of these data appears to be about as high as the other questionnaires—that is to say, adequate for the purposes to which it was put.

The results of the evaluation of all questionnaire data, as well as the student data, are summarized in Table 6-2.

6.4 NAEP DATABASE PRODUCTS

The NAEP database described to this point serves primarily to support analysis and reporting activities that are directly related to the NAEP contract. This database has a singular structure and access methodology that is integrated with the NAEP analysis and reporting programs. One of the directives of the NAEP contract is to provide secondary researchers with a nonproprietary version of the database that is portable to any computer system. In the event of transfer of NAEP to another client, the contract further requires ETS to provide a full copy of the internal database in a format that may be installed on a different computer system.

In fulfillment of these requirements, ETS provides two sets of database products: the item information database and the secondary-use data files. The contents, format and usage of these products are documented in the publications listed under the appropriate sections below.

6.4.1 The Item Information Database

The NAEP item information database contains all of the descriptive, processing, and usage information for every assessment item developed and used for NAEP since 1970. The primary unit of this database is the item. Each NAEP item is associated with different levels of information, including usage across years and age cohorts, subject area classifications, response category descriptors, and locations of response data on secondary-use data files.

The item information database is used for a variety of essential NAEP tasks: providing statistical information to aid in test construction, determining the usage of items across assessment years and ages for trend and cross-sectional analyses, labeling summary analyses and reports, and organizing items by subject area classifications for scaling analysis.

The creation, structure, and use of the NAEP item information database for all items used up to and including the 1994 assessment are fully documented in the NAEP publications *A Guide to the NAEP Item Information Database* (Rogers, Barone, & Kline, 1995) and *A Primer for the NAEP Item Information Database* (Rogers, Kline, Barone, Mychajlowycz, & Forer, 1989).

The procedures used to create the 1994 version of the item information database are the same as those documented in the guide. The updated version of the guide also contains the subject area classification categories for the cognitive items.

6.4.2 The Secondary-use Data Files

The secondary-use data files are designed to enable any researcher with an interest in the NAEP database to perform secondary analysis on the same data as those used at ETS. The three elements of the distribution package are the data files, the printed documentation, and copies of the questionnaires and released item blocks. A set of files for each sample or instrument contains the response data 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 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 printed documentation consists of two volumes: a guide to the use of the data files, and a set of data file layouts and codebooks for each of the participants in the assessment.

The remainder of this section summarizes the procedures used in generating the data files and related materials.

6.4.2.1 File Definition

There are essentially five samples for analysis in the 1994 Trial State Reading Assessment: the assessed students, the excluded students, and the schools in the state-by-state component, and the assessed students and the schools in a matched comparison sample drawn from the national reading assessment. Each state sample is divided into separate files by each jurisdiction, resulting in a total of over 130 files, but the same file formats, linking conventions, and analysis considerations apply to each file within a given sample. For example, the analysis specification that links school and assessed student data for California would apply identically to New York, Illinois, or any other participant or group of participants.

Each participant data file still requires its own data codebook, detailing the frequencies of data values within that jurisdiction. The file layouts, SPSS and SAS syntax and machine-readable catalog files, however, need only be generated for each sample, since the individual jurisdiction data files within a state sample are identical in format and data code definition.

6.4.2.2 Definition of the Variables

The lifting of the restraint on confidential data simplified the variable definition process as it permitted the transfer of *all* variables from the database to the secondary-use files.

The initial step in this process was the generation of a LABELS file of descriptors of the variables for each data sample 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 next program in the processing stream, GENLYT, produced a printed layout for each file from the information in its corresponding LABELS file. These layouts were initially reviewed for the ordering of the variables.

The variables on all data files were grouped and arranged in the following order: identification information, weights, derived variables, proficiency scale scores (where applicable), and response data. On the student data files, these fields were followed by the teacher response data and the IEP/LEP student questionnaire data, where applicable. The identification information is taken from the front covers of the instruments. The weight data include sample descriptors, selection probabilities, and replicate weights for the estimation of sampling error. The derived data include sample descriptions from other sources and variables that are derived from the response data for use in analysis or reporting.

For each assessed student sample in the state component and national comparison sample, the item response data within each block were left in their order of presentation. The blocks, however, were arranged according to the following scheme: common background, subject-related background, the cognitive blocks in ascending numerical order, and student motivation. The responses to cognitive blocks that were not present in a given booklet were left blank, signifying a condition of "missing by design."

In order to process and analyze the spiral sample data effectively, the user must also be able to determine, from a given booklet record, which blocks of item response data were present and their relative order in the instrument. This problem was remedied by the creation of a set of control variables, one for each block, which indicated not only the presence or absence of the block but its order in the instrument. These control variables were included with the derived variables.

6.4.2.3 Data Definition

To enable the data files to be processed on any computer system using any procedural or programming language, it was desirable that the data be expressed in numeric format. This was possible, but not without the adoption of certain conventions for reexpressing the data values.

As mentioned in section 6.1, the responses to all multiple-choice itums 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 included out-of-range responses and responses that were missing due to errors in printing or processing. The scoring guides for the reading constructed-response items included additional special codes for ratings of "illegible," "off task," and non-rateable by the scorers. All of these codes had to be reexpressed in a consistent numeric format.

The following convention was adopted and used in the designation of these codes: The "I don't know" and non-rateable response codes were always converted to 7; the "omitted" response codes were converted to 8; the "not-reached" response codes were converted to 9; the multiple response codes were converted to 0; the "illegible" codes were converted to 5; and the "off task" codes were converted to 6. The out-of-range and missing responses were coded as blank fields, corresponding to the "missing by design" designation.

This coding scheme created conflicts for those multiple-choice items that had seven or more valid response options as well as the "I don't know" response and for those constructed-response items whose scoring guide had five or more categories. These data fields were also expanded to accommodate the valid response values and the special codes. In these cases, the special codes were "extended" to fill the output data field: The "I don't know" and non-rateable codes were extended from 7 to 77, omitted response codes from 8 to 88, etc.

Each numeric variable on the secondary-use files was classified as either continuous or discrete. The continuous variables include the weights, proficiency values, identification codes, and item responses where counts or percentages were requested. The discrete variables include those items for which each numeric value corresponds to a response category. The designation of "discrete" also includes those derived variables to which numeric classification categories have been assigned. The constructed-response items were treated as a special subset of the discrete variables and were assigned to a separate category to facilitate their identification in the documentation.

6.4.2.4 Data File Layouts

The data file layouts, as mentioned above, were the first user product to be generated in the secondary-use data files process. The generation program, GENLYT, used a LABELS file and a CATALOG file as input and produced a printable file. The LAYOUT file is little more than a formatted listing of the LABELS file.

Each line of the LAYOUT file contains the following information for a single data field: sequence number, field name, output column position, field width, number of decimal places, data type, value range, key or correct response value, and a short description of the field. The sequence number of each field is implied from its order on the LABELS file. The field name is an 8-character label for the field that is to be used consistently by all secondary-use data files materials to refer to that field on that file. The output column position is the relative location of the beginning of that field on each record for that file, using bytes or characters as the unit of measure. The field width indicates the number of columns used in representing the data values for a field. If the field contains continuous numeric data, the value under the number of decimal places entry indicates how many places to shift the decimal point before processing data values.

The data type category uses five codes to designate the nature of the data in the field: Continuous numeric data are coded "C"; discrete numeric data are coded "D"; constructed-response item data are coded "OS" if the item was dichotomized for scaling and "OE" if it was scaled under a polytomous response model. Additionally, the discrete numeric fields that include "I don't know" response codes are coded "DI." If the field type is discrete numeric, the value range is listed as the minimum and maximum permitted values separated by a hyphen to indicate range. If the field is a response to a scorable item, the correct option value, or key, is printed; if the field is an assigned score that was scaled as a dichotomous item using cut point scoring, the range of correct scores is printed. Each variable is further identified by a 50-character descriptor.

6.4.2.5 Data File Catalogs

The LABELS file contains sufficient descriptive information for generating a brief layout of the data file. However, to generate a complete codebook document, substantially more information about the data is required. The CATALOG file provides most of this information.

The CATALOG file is created by the GENCAT program from the LABELS file and the 1994 master catalog file. Each record on the LABELS file generates a CATALOG record by first retrieving the master catalog record corresponding to the field name. The master catalog record contains usage, classification, and response code information, along with positional information from the LABELS file: field sequence number, output column position, and field width. Like the LABELS file, the CATALOG file is an intermediate file and is not included on the released data files.

The information for the response codes, also referred to as "foils," consists of the valid data values for the discrete numeric fields, and a 20-character description of each. The GENCAT program uses additional control information from the LABELS file to determine if extra foils should be generated and saved with each CATALOG record. The first flag controls generation of the "I don't know" or non-rateable foil; the second flag regulates omitted or not-reached foil generation; and the third flag denotes the possibility of multiple responses for that field and sets up an appropriate foil. All of these control parameters, including the expansion flag, may be altered in the LABELS file by use of a text editor, in order to control the generation of data or descriptive information for any given field.

The LABELS file supplies control information for many of the subsequent secondary-use data processing steps. The CATALOG file provides detailed information for those and other steps.

6.4.2.6 Data Codebooks

The data codebook is a printed document containing complete descriptive information for each data field. Most of this information originates from the CATALOG file; the remaining data came from the COUNTS file and the IRT parameters file.

Each data field receives at least one line of descriptive information in the codebook. If the data type is continuous numeric, no more information is given. If the variable is discrete numeric, the codebook lists the foil codes, foil labels, and frequencies of each value in the data file. Additionally, if the field represents an item used in IRT scaling, the codebook lists the parameters used by the scaling program.

Certain blocks of cognitive items in the 1994 assessment that are to be used again in later assessments for trend comparisons have been designated as nonreleased. In order to maintain their confidentiality, generic labels have been substituted for the response category descriptions of these items in the data codebooks and the secondary-use files.

The frequency counts are not available on the catalog file, but must be generated from the data. The GENFREQ program creates the COUNTS file using the field name to locate the variable in the database, and the foil values to validate the range of data values for each field. This program also serves as a check on the completeness of the foils in the CATALOG file, as it flags any data values not represented by a foil value and label.

The IRT parameter file is linked to the CATALOG file through the field name. Printing of the IRT parameters is governed by a control flag in the classification sectioi of the

CATALOG record. If an item has been scaled for use in deriving the proficiency estimates, the IRT parameters are listed to the right of the foil values and labels, and the score value for each response code is printed to the immediate right of the corresponding frequency.

The LAYOUT and CODEBOOK files are written by their respective generation programs to print-image disk data files. Draft copies are printed and distributed for review before the production copy is generated. The production copy is printed on an IBM 3800 printer that uses laser-imaging technology to produce high-quality, reproducible documentation.

6.4.2.7 Control Statement Files for Statistical Packages

An additional requirement of the NAEP contract is to provide, for each secondary-use data file, a file of control statements each for the SAS and SPSS statistical systems that will convert the raw data file into the system data file for that package. Two separate programs, GENSAS and GENSPX, generate these control files using the CATALOG file as input.

Each of the control files contains separate sections for variable definition, variable labeling, missing value declaration, value labeling, and creation of scored variables from the cognitive items. The variable definition section describes the locations of the fields, by name, in the file, and, if applicable, the number of decimal places or type of data. The variable label identifies each field with a 50-character description. The missing value section identifies values of those variables that are to be treated as missing and excluded from analyses. The value labels correspond to the foils in the CATALOG file. The code values and their descriptors are listed for each discrete numeric variable. The scoring section is provided to permit the user to generate item score variables in addition to the item response variables.

Each of the code generation programs combines three steps into one complex procedure. As each CATALOG file record is read, it is broken into several component records according to the information to be used in each of the resultant sections. These record fragments are tagged with the field sequence number and a section sequence code. They are then sorted 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.

6.4.2.8 Machine-readable Catalog Files

For those NAEP data users who have neither SAS nor SPSS capabilities, yet require processing control information in a computer-readable format, the distribution files also contain machine-readable catalog files. Each machine-readable catalog record contains processing control information, IRT parameters, and foil codes and labels.

6.4.2.9 NAEP Data on Disk

The complete set of secondary-use data files described above are available on CD-ROM as part of the NAEP Data on Disk product suite. This medium can be ideal for researchers and policy makers operating in a personal computing environment.

The NAEP Data on Disk product suite includes two other components 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.

Chapter 7

WEIGHTING PROCEDURES AND VARIANCE ESTIMATION

Mansour Fahimi, Keith F. Rust, and John Burke
Westat, Inc.

7.1 OVERVIEW

Following the collection of assessment and background data from and about assessed and excluded students, the processes of deriving sampling weights and associated sets of replicate weights were carried out. The sampling weights are needed to make valid inferences from the student samples to the respective populations from which they were drawn. Replicate weights are used in the estimation of sampling variance, through the procedure known as *jackknife repeated replication*.

Each student was assigned a weight to be used for making inferences about the state's students. This weight is known as the *full-sample* or *overall* sample weight. The full-sample weight contains three components. First, a base weight is established which is the inverse of the overall probability of selection of the sampled student. The base weight incorporates the probability of selecting a school and the student within a school. This weight is then adjusted for two sources of nonparticipation—school level and student level. These weighting adjustments seek to reduce the potential for bias from such nonparticipation by increasing the weights of students from schools similar to those schools not participating, and increasing the weights of students similar to those students from within participating schools who did not attend the assessment session as scheduled. The details of how these weighting steps were implemented are given in sections 7.2 and 7.3.

Section 7.4 addresses the effectiveness of the adjustments made to the weights using the procedures described in section 7.3. The section examines characteristics of nonresponding schools and students, and investigates the extent that nonrespondents differ from respondents in ways not accounted for in the weight adjustment procedures. Section 7.5 considers the distributions of the final student weights in each state, and whether there were outliers that called for further adjustment.

In addition to the full-sample weights, a set of replicate weights was provided for each student. These replicate weights are used in calculating the sampling errors of estimates obtained from the data, using the jackknife repeated replication method. Full details of the method of using these replicate weights to estimate sampling errors are contained in the NAEP technical reports from the 1992 assessment (Johnson & Carlson, 1994) and earlier. Section 7.6 of this report describes how the sets of replicate weights were generated for the 1994 Trial State Assessment data. The methods of deriving these weights were aimed at reflecting the features of the sample design appropriately in each state, so that when the jackknife variance estimation procedure is implemented, approximately unbiased estimates of sampling variance result.

7.2 CALCULATION OF BASE WEIGHTS

The base weight assigned to a school was the reciprocal of the probability of selection of that school. The school base weight reflected the actual probability used to select the school from the frame, including the impact of avoiding schools selected for the national sample. For "new" schools selected using the supplemental new school sampling procedures (see section 3.5.3), the school base weight reflected the combined probability of selection of the district, and school within district.

The student base weight was obtained by multiplying the school base weight by the within-school student weight, where the within-school student weight reflected the probability of selecting students within the school. Additional details about the weighting process are given in the sections below.

7.2.1 Calculation of School Base Weights

The base weight for sample school *i* was computed as:

$$W_i = \frac{E}{mE_i}$$

where

E_i = the enrollment in the given school;

E = $\sum_{s=1}^{S} E_s$ the state-wide enrollment obtained by summing E_S across all schools in the state frame; and

m = the number of schools selected from the state.

In each state, all schools included in the sample with certainty were assigned school base weights of unity.

Schools sampled with certainty were sometimes selected more than once in the systematic sampling process. If a large school was selected more than once, each selection was treated separately in the selection of students within a school. For example, a school that was selected twice was allocated twice the usual numbers of students for the assessments; a school that was selected three times was allocated three times the usual numbers of students for the assessments.

7.2.2 Weighting New Schools

New public schools were identified and sampled through a two-stage sampling process, involving the selection of districts, and then of new schools within selected districts. This process is described in Chapter 3. There were two distinct processes used depending upon the size of the district.

Within each state, public school districts were partitioned into those having (at most) one school with grade 4, one with grade 8, and one with grade 12, versus all other districts. For the first set of small districts, the selection of the grade 4 school from the frame, in the initial sample of schools for the state, triggered an inquiry of the district as to whether there were in fact any additional schools with grade 4 (not contained on the school sampling frame). Any school so identified was automatically added to the sample for the assessment. Thus the selection probability of such a school was equal to that of the grade 4 school from the district that was included on the school frame, and the school base weight was calculated accordingly.

For the larger districts (those having multiple schools at least one of grades 4, 8, and 12), a sample of districts was selected in each state. Districts in the sample were asked to identify schools having grade 4 that were not included on the school frame. A sample of these newly identified schools was then selected. The base weight for these schools reflected both the probability that the district was selected for this updating process, and that the school was included in the NAEP sample, having been identified as a new by the district.

7.2.3 Treatment of New and Substitute Schools

Schools that replaced a refusing school (i.e., substitute schools) were assigned the weight of the refusing school, unless the substitute school also refused. Thus the substitute school was treated as if it were the original school that it replaced, for purposes of obtaining school base weights.

7.2.4 Calculation of Student Base Weights

Within the sampled schools, eligible students were sampled. The within-school probability of selection therefore depended on the number of eligible students in the school and the number of students selected for the assessment (usually 30). The within-school weights for the substitute schools were further adjusted to compensate for differences in the sizes of the substitute and the originally sampled (replaced) schools. Thus, in general, the within-school student weight for the jth student in school i was equal to:

$$W_{ij}^{within} = \frac{N_i}{n_i} \times K_i$$

where

N_i = the number of eligible students enrolled in the school, as reported in the sampling worksheets;

n_i = the number of students selected; and

K_i = $\frac{E_i}{E_i^s}$

with

E_i = the QED grade enrollment of the originally sampled (replaced) school; and

E_i^s = the QED grade enrollment of the substitute school.

The factor K_i in the above formula for the within-school student weight applies to only a few schools in each state. This factor adjusts the count of eligible students in a substitute school to be consistent with corresponding count of the originally sampled (replaced) school. For nonsubstitute schools K_i was set to 1.

7.3 ADJUSTMENTS FOR NONRESPONSE

As mentioned earlier, the base weight for a student was adjusted by two factors: one to adjust fo- nonparticipating schools for which no substitute participated, and one to adjust for students who were invited to the assessment but did not appear in the scheduled sessions.

7.3.1 Defining Initial School-Level Nonresponse Adjustment Classes

School-level nonresponse adjustment classes were created separately for public and nonpublic schools within each state. For each set these classes were defined as a function of their sampling strata, as detailed next.

Public Schools. For each state, the initial school nonresponse adjustment classes were formed by crossclassifying the level of urbanization and minority status (see Chapter 3 for definitions of these characteristics). Where there were no minority strata within a particular level of urbanization, a categorized version of median income was used. For this purpose within each level of urbanization, public schools were sorted by the median income, and then divided into three groups of about equal size, representing low, middle, and high income areas.

Nonpublic Schools. For each state (excluding District of Columbia and Guam nonpublic schools), nonresponse adjustment classes were formed by crossclassifying school type (Catholic and nonCatholic) and metropolitan status (metro/nonmetro) area. For District of Columbia nonpublic schools these classes were defined by crossclassifying school type and two levels of estimated grade enrollment (25 or fewer students, versus 26 or more students). For Guam, initial nonresponse classes for nonpublic schools were defined by school type only. The District of Columbia is entirely metropolitan, and Guam is entirely nonmetropolitan, so alternatives were needed for these two jurisdictions.

Department of Defense Educational Activity (DoDEA) Overseas Schools. For the jurisdiction comprised of DoDEA Overseas schools, there was only one nonresponding school. This school, along with the remaining schools in the Atlantic region, formed the first nonresponse class while all remaining DoDEA Overseas schools were assigned to the second nonresponse class.

7.3.2 Constructing the Final Nonresponse Adjustment Classes

The objective in forming the nonresponse adjustment classes is to create as many classes as possible which are internally as homogeneous as possible, but such that the resulting nonresponse adjustment factors are not subject to large random variation. Consequently, all initial nonresponse adjustment classes deemed unstable were collapsed with suitable neighboring classes so that: (1) the combined class contained at least 6 schools, and (2) the resulting nonresponse adjustment factor did not exceed 1.35 (in a few cases a factor slightly in excess of 1.35 was permitted). These limits had been used for the 1992 Trial State Assessment.

Public Schools. For these schools, inadequate nonresponse adjustment classes were reinforced by collapsing adjacent levels of minority status (or median income level if minority information was missing). In doing so, different categories of urbanization were not mixed (to the extent possible).

Nonpublic Schools. For nonpublic schools, excluding schools in District of Columbia and Guam, inadequate classes were reinforced by collapsing adjacent levels of metropolitan area status. Catholic and nonCatholic schools were kept apart to the extent possible, particularly when the only requirement to combine such schools was as a means of reducing the adjustment factors below 1.35. For schools in the District of Columbia, inadequate classes were collapsed over similar values of estimated grade enrollment. Catholic and nonCatholic schools were kept apart to the extent possible. For nonpublic schools in Guam, Catholic and nonCatholic schools were collapsed together in order to form a stable nonresponse adjustment class.

7.3.3 School Nonresponse Adjustment Factors

The school-level nonresponse adjustment factor for the *i*th school in the *h*th class was computed as:

$$F_h^{(1)} = \frac{\sum_{i \in C_h} W_{hi}^{sch} E_{hi}}{\sum_{i \in C_h} W_{hi}^{sch} E_{hi} \delta_{hi}}$$

where

C_h = the subset of school records in class *h*;

W_{hi}^{sch} = the base weight of the *i*th school in class *h*;

E_{hi} = the QED grade enrollment for the *i*th school in class *h*;

$$\delta_{hi} = \begin{cases} 1 & \text{if the } i\text{th school in adjustment class } h \text{ participated in the assessments; and} \\ 0 & \text{otherwise.} \end{cases}$$

In the calculation of the above nonresponse adjustment factors, a school was said to have participated if:

- It was selected for the sample from the QED frame or from the lists of new schools provided by participating school districts, and student assessment data were obtained from the school; or

- The school participated as a substitute school and student assessment data were obtained (so that the substitute participated in place of the originally selected school).

Both the numerator and denominator of the nonresponse adjustment factor contained only in-scope schools.

The nonresponse-adjusted weight for the *i*th school in class *h* was computed as:

$$W_{hi}^{adj} = F_h^{(1)} W_{hi}^{sch} \quad .$$

7.3.4 Student Nonresponse Adjustment Classes

The initial student nonresponse classes were formed using the final school nonresponse classes, crossclassified by the quality control monitoring status (see section 3.5.4) and student age. Age was used to classify students into two groups (those born in September 1983 or earlier and those born in October 1983 or later). Following creation of the initial student nonresponse adjustment classes, all weak classes were identified for possible reinforcements. A class was considered to be unstable when any of the following conditions was true for the given class:

- Number of responding eligible students was fewer than 20;
- Nonresponse adjustment factor exceeded 2.0; and
- Number of responding eligible students was fewer than 31 and nonresponse adjustment factor exceeded 1.5.

All classes deemed unstable in the previous step were collapsed with other classes using the following rules:

- Collapsed across monitoring status within all other classes;
- If a resulting class still needed to be collapsed, then the previous collapsing was undone, and now collapsed across minority/income categories; and
- If a resulting class still needed to be collapsed, it was further collapsed across the three fields—monitor status, urbanization level, and age category—in that order.

7.3.5 Student Nonresponse Adjustments

As described above, the student-level nonresponse adjustments for the assessed students were made within classes defined by the final school-level nonresponse adjustment classes, monitoring status of the school, and age group of the students. Subsequently, in each state, the final student weight for the *j*th student of the *i*th school in class *k* was then computed as:

$$W_{kij}^{final} = W_i^{adj} \times W_{ij}^{within} \times F_k \times \delta_{kj}$$

where

W_i^{adj} = the nonresponse-adjusted school weight for school *i*;

W_{ij}^{within} = the within-school weight for the *j*th student in school *i*;

and

$$F_k = \frac{\sum_j W_{ij}}{\sum_j W_{ij}\delta_{kj}} \quad .$$

In the above formulation, the summation included all students, *j*, in the *k*th final (collapsed) nonresponse class. The indicator variable δ_{kj} had a value of 1 when the *j*th student in adjustment class *k* participated in the assessment; otherwise, $\delta_{kj} = 0$.

For excluded students the same basic procedures as described above for assessed students were used, except that the numerator and denominator contained excluded rather than assessed students, and monitoring status and student age group were not used to form the adjustment classes. Weights are provided for excluded students so as to estimate the size of this group and its population characteristics. Table 7-1 summarizes the unweighted and final weighted counts of assessed and excluded students for each state.

7.4 CHARACTERISTICS OF NONRESPONDING SCHOOLS AND STUDENTS

In the previous section procedures were described for adjusting the survey weights so as to reduce the potential bias of nonparticipation of sampled schools and students. To the extent that a nonresponding school or student is different from those respondents in the same nonresponse adjustment class, potential for nonresponse bias remains.

In this section, we examine the potential for remaining nonresponse bias in two related ways. First we examine the weighted distributions, within each grade and state, of certain characteristics of schools and students, both for the full sample and for respondents only. This analysis is of necessity limited to those characteristics that are known for both respondents and nonrespondents, and hence cannot directly address the question of nonresponse bias. The approach taken does reflect the reduction in bias obtained through the use of nonresponse weighting adjustments. As such, it is more appropriate than a simple comparison of the characteristics of nonrespondents with those of respondents for each state.

The second approach involves modeling the probability that a school is a nonrespondent, as a function of the nonresponse adjustment class within which the school is located, together with other school characteristics. This has been achieved using linear logistic regression models, with school response status as the dependent variable. By examining how much better one can predict school nonresponse using school characteristics, over and above using the membership of the nonresponse adjustment class to make this prediction, we can obtain some insight into the remaining potential for nonresponse bias. If these factors are substantially marginally predictive, there is a danger that significant nonresponse bias remains. These models have been developed for public schools in each of the seven states having public school participation (after substitution) of below 90 percent (with a participation rate prior to substitution in excess of 70 percent).

Table 7-1
Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction

Jurisdiction	Assessed		Excluded		Assessed and Excluded	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Alabama	2,845	57,099	163	3,131	3,008	60,230
Arizona	2,651	52,297	191	3,899	2,842	56,196
Arkansas	2,689	32,550	167	1,978	2,856	34,528
California	2,401	370,558	358	48,031	2,759	418,589
Colorado	2,860	51,259	204	3,792	3,064	55,052
Connecticut	2,868	38,888	237	3,250	3,105	42,138
Delaware	2,783	9,239	146	503	2,929	9,742
DoDEA Overseas	2,413	8,350	108	399	2,521	8,749
District Of Columbia	2,913	6,241	262	507	3,175	6,748
Florida	2,933	168,380	302	17,846	3,235	186,225
Georgia	2,983	102,798	164	5,548	3,147	108,346
Guam	2,575	2,693	192	220	2,767	2,913
Hawaii	3,147	15,474	141	714	3,288	16,188
Idaho	2,692	17,922	140	947	2,832	18,869
Indiana	2,874	75,590	153	3,787	3,027	79,377
Iowa	3,086	39,125	133	1,752	3,219	40,877
Kentucky	3,036	50,820	108	1,850	3,144	52,670
Louisiana	3,170	64,663	165	3,572	3,335	68,236
Maine	2,521	15,837	257	1,682	2,778	17,519
Maryland	2,830	61,712	205	4,408	3,035	66,120
Massachusetts	2,819	65,486	237	5,118	3,056	70,604
Michigan	2,142	112,908	122	6,954	2,264	119,862
Minnesota	3,045	67,251	115	2,661	3,160	69,912
Mississippi	2,918	39,288	169	2,340	3,087	41,628
Missouri	3,042	68,884	152	3,211	3,194	72,094
Montana	2,649	12,660	86	430	2,735	13,089
Nebraska	2,606	26,458	103	1,063	2,709	27,521
New Hampshire	2,197	14,296	132	896	2,329	15,192
New Jersey	2,888	93,268	155	5,208	3,043	98,476
New Mexico	2,826	24,972	239	2,219	3,065	27,191
New York	2,864	222,969	221	16,357	3,085	239,326
North Carolina	2,833	79,806	169	4,493	3,002	84,299
North Dakota	2,797	9,847	65	221	2,862	10,068
Pennsylvania	2,717	141,774	137	7,454	2,854	149,228
Rhode Island	2,696	11,995	133	582	2,829	12,577
South Carolina	2,863	49,988	185	3,340	3,048	53,328
Tennessee	1,998	57,433	112	3,725	2,110	61,157
Texas	2,454	238,075	288	29,116	2,742	267,191
Utah	2,733	31,893	138	1,712	2,871	33,605
Virginia	2,870	79,774	214	5,592	3,084	85,366
Washington	2,737	67,089	143	3,618	2,880	70,707
West Virginia	2,887	23,407	212	1,614	3,099	25,022
Wisconsin	2,719	68,292	181	4,370	2,900	72,662
Wyoming	2,699	7,398	116	330	2,815	7,728
Total	121,269	2,856,703	7,620	220,439	128,889	3,077,142

7.4.1 Weighted Distributions of Schools Before and After School Nonresponse

Table 7-2 shows the mean values of certain school characteristics for public schools, both before and after nonresponse. The means are weighted appropriately to reflect whether nonresponse adjustments have been applied (i.e., to respondents only) or not (to the full set of in-scope schools). The variables for which means are presented are the percentage of students in the school who are Black, the percentage who are Hispanic, the median income (1989) of the ZIP code area where the school is located, and the type of location. All variables were obtained from the sample frame, described in Chapter 3, with the exception of the type of location. This variable was derived for each sampled school using census data. The type of location variable has seven possible levels, which are defined in section 3.4.2. Although this variable is not interval-scaled, the mean value does give an indication of the degree of urbanization of the population represented by the school sample (lower values for type of location indicate a greater degree of urbanization).

Two sets of means are presented for these four variables. The first set shows the weighted mean derived from the full sample of in-scope schools selected for reading; that is, respondents and nonrespondents (for which there was no participating substitute). The weight for each sampled school is the product of the school base weight and the grade enrollment. This weight therefore represents the number of students in the state represented by the selected school. The second set of means is derived from responding schools only, after school substitution. In this case the weight for each school is the product of the nonresponse-adjusted school weight and the grade enrollment, and therefore indicates the number of students in the state represented by the responding school.

Table 7-3 shows some of these same statistics for all schools combined, for those states where both the public school participation rate prior to substitution, and the nonpublic school participation rate prior to substitution, exceeded 70 percent. These are the states for which assess.ment results have been published for both public and nonpublic schools combined. Data on minority enrollment were not available for nonpublic schools, and so are not included in Table 7.3.

The differences between these sets of means give an indication of the potential for nonresponse bias that has been introduced by nonresponding schools for which there was no participating substitute. For example, in Arkansas at grade 4 the mean percentage Black enrollment, estimated from the original sample of public schools, is 24.50 percent (Table 7-2). The estimate from the responding schools is 24.36 percent. Thus there may be a slight bias in the results for Arkansas because these two means differ. Note, however, that throughout these two tables the differences in the two sets of mean values are generally very slight, at least in absolute terms, suggesting that it is unlikely that substantial bias has been introduced by schools that did not participate and for which no substitute participated. Of course in a number of states (as indicated) there was no nonresponse at the school level, so that these sets of means are identical. Even in those jurisdictions where school nonresponse was relatively high (such as Tennessee, Nebraska, New Hampshire, and Michigan), the absolute differences in means are slight. Occasionally the relative difference is large (the "Percent Black" in Wisconsin, for example), but these are for small population subgroups, and thus are very unlikely to have a large impact on results for the jurisdiction as a whole.

Table 7-2
Weighted Mean Values Derived from Sampled Public Schools

Jurisdiction	Weighted Participation Rate After Substitution (%)	Weighted Mean Values Derived from Full Sample				Weighted Mean Values Derived from Responding Sample, with Substitutes and School Nonresponse Adjustment			
		Percent Black	Percent Hispanic	Median Income	Type of Location	Percent Black	Percent Hispanic	Median Income	Type of Location
Alabama	93.39	34.94	0.05	$23,860	4.37	35.16	0.05	$24,032	4.35
Arizona	99.04	3.49	23.22	$31,020	2.52	3.52	23.32	$31,058	2.51
Arkansas	94.09	24.50	0.23	$22,561	4.78	24.36	0.20	$22,648	4.81
California	90.52	6.79	36.85	$35,591	2.77	6.82	35.70	$35,766	2.75
Colorado	100.00	4.23	15.71	$32,485	3.38	4.19	15.92	$32,387	3.40
Connecticut	96.47	13.39	11.21	$44,520	3.34	12.87	10.81	$44,678	3.33
Delaware	100.00	28.53	1.99	$26,983	3.19	28.53	1.99	$26,983	3.19
DoDEA Overseas	99.25	—	—	—	—	—	—	—	—
Dist. of Columbia	100.00	88.78	5.06	$27,898	1.00	88.78	5.06	$27,898	1.00
Florida	100.00	25.10	11.32	$28,688	3.31	25.10	11.32	$28,688	3.31
Georgia	99.05	28.35	1.16	$30,537	4.57	28.39	1.16	$30,526	4.57
Guam	100.00	2.10	0.29	—	—	2.10	0.29	—	—
Hawaii	99.07	1.86	2.71	$35,436	4.34	1.88	2.72	$35,424	4.34
Idaho	91.45	0.18	5.39	$26,063	5.03	0.20	5.19	$26,091	4.97
Indiana	92.48	10.95	1.40	$27,947	3.87	10.96	1.40	$28,165	3.86
Iowa	99.05	2.70	1.05	$27,499	4.99	2.70	1.06	$27,404	5.01
Kentucky	96.16	9.64	0.10	$24,022	5.15	9.65	0.10	$23,782	5.16
Louisiana	100.00	43.87	1.32	$23,401	3.90	43.87	1.32	$23,401	3.90
Maine	96.99	0.09	0.50	$29,054	5.64	0.09	0.49	$29,191	5.65
Maryland	96.15	32.87	1.69	$40,496	2.67	32.94	1.76	$40,191	2.71
Massachusetts	97.02	8.65	7.91	$41,722	3.09	8.22	8.36	$41,567	3.11
Michigan	79.77	16.64	2.27	$33,009	3.72	17.37	2.26	$33,078	3.72
Minnesota	95.22	3.47	0.62	$33,478	4.11	3.57	0.64	$33,514	4.12
Mississippi	99.04	49.32	0.07	$21,249	5.37	49.46	0.07	$21,268	5.36
Missouri	98.40	10.06	1.29	$29,013	3.85	9.88	1.29	$29,107	3.86
Montana	88.58	0.29	1.39	$24,675	5.28	0.31	1.43	$24,682	5.28
Nebraska	77.35	4.02	2.59	$27,787	4.92	2.87	3.03	$26,479	5.09
New Hampshire	79.20	0.40	0.53	$39,829	4.52	0.38	0.51	$39,847	4.51
New Jersey	91.29	17.96	13.86	$42,647	3.22	18.83	13.95	$42,032	3.21
New Mexico	100.00	1.84	45.87	$24,273	4.27	1.84	45.87	$24,273	4.27
New York	90.57	20.72	15.60	$34,849	2.76	20.46	15.97	$34,351	2.75
North Carolina	99.05	29.69	0.70	$27,929	4.33	29.60	0.69	$28,036	4.33
North Dakota	91.19	0.62	0.51	$27,229	5.09	0.65	0.53	$27,203	5.04
Pennsylvania	83.69	14.91	2.20	$31,527	3.13	16.04	2.32	$31,238	3.13
Rhode Island	85.54	6.39	6.73	$31,585	2.94	6.80	6.67	$31,486	2.97
South Carolina	97.15	42.24	0.20	$26,573	4.55	42.17	0.20	$26,594	4.54
Tennessee	73.79	22.39	0.15	$25,243	3.63	24.23	0.16	$23,897	3.65
Texas	93.20	12.08	35.47	$27,869	2.74	11.69	35.49	$27,681	2.75
Utah	100.00	0.36	3.42	$32,643	3.95	0.36	3.42	$32,643	3.95
Virginia	99.05	21.27	1.77	$39,125	3.65	21.33	1.78	$39,124	3.65
Washington	100.00	4.12	5.13	$34,341	3.52	4.12	5.13	$34,341	3.52
West Virginia	100.00	2.93	0.06	$22,277	5.34	2.93	0.06	$22,277	5.34
Wisconsin	85.56	9.20	2.41	$32,677	3.86	5.75	2.62	$32,841	3.92
Wyoming	98.38	0.36	4.39	$31,446	5.19	0.37	4.45	$31,473	5.19

Table 7-3
Weighted Mean Values Derived from All Sampled Schools for Jurisdictions Achieving Minimal Required Public- and Nonpublic-school Participation, Before Substitution

Jurisdiction	Weighted Participation Rate After Substitution (%)	Weighted Mean Values Derived from Full Sample		Weighted Mean Values Derived from Responding Sample, with Substitutes and School Nonresponse Adjustment	
		Median Income	Type of Location	Median Income	Type of Location
Alabama	96.77	$24,022	4.32	$24,198	4.29
Arkansas	94.70	$22,837	4.69	$22,934	4.71
Colorado	91.54	$32,449	3.36	$32,386	3.38
Connecticut	84.14	$44,818	3.28	$45,008	3.28
Delaware	79.92	$29,135	3.01	$29,489	3.06
Georgia	93.69	$30,800	4.50	$30,876	4.50
Guam	95.61	—	—	—	—
Hawaii	90.07	$35,845	4.17	$35,709	4.14
Iowa	99.68	$27,450	4.99	$27,364	5.00
Indiana	76.54	$28,167	3.80	$28,403	3.78
Kentucky	87.42	$24,619	4.93	$24,129	4.96
Louisiana	91.83	$23,760	3.73	$23,662	3.74
Massachusetts	99.01	$41,268	3.01	$41,129	3.03
Maine	99.00	$29,046	5.59	$29,178	5.60
Minnesota	97.15	$33,409	4.11	$33,445	4.12
Missouri	94.50	$29,694	3.66	$29,675	3.68
North Dakota	88.65	$27,184	5.10	$27,224	5.04
New Jersey	71.12	$42,522	3.21	$41,661	3.21
New Mexico	100.00	$24,199	4.24	$24,199	4.24
Pennsylvania	64 42	$31,893	3.04	$31,453	3.13
Rhode Island	80.37	$31,702	2.98	$31,565	3.00
Virginia	80.92	$39,252	3.61	$39,428	3.60
West Virginia	91.84	$22,405	5.25	$22,430	5.26

7.4.2 Characteristics of Schools Related to Response

In an effort to evaluate the possibility that substantial bias remains as a results of school nonparticipation, following the use of nonresponse adjustments, a series of analyses were conducted on the response statuses for public schools. This analysis was restricted to those states with a participation rate below 90 percent (after substitution), since these are the states where the potential for nonresponse bias is likely to be the greatest. We did not include those states with an initial public school response rate was below 70 percent, since NAEP does not report results for these states because of concern about nonresponse bias. private schools were omitted from these analyses because of the small sample sizes involved, which mean that it is difficult to assess whether a potential for bias exists.

The seven states investigated were the following (with the public school participation rate shown in parentheses): Montana (89%), Nebraska (77%), New Hampshire (79%), Pennsylvania (84%), Rhode Island (86%), Tennessee (74%), and Wisconsin (86%). The approach used was to develop logistic regression models within each state, to predict the probability of participation as a function of the nonresponse adjustment classes, and other school characteristics. The aim was to determine whether the response rates are significantly related to school characteristics, after accounting for the effect of the nonresponse class. Thus dummy variables were created to indicate nonresponse class membership, and an initial model was created which predicted the probability of school participation as function of nonresponse class.

If there are *k* nonresponse classes within a state, let

X_{ij} = 1 if the school *j* is classified in nonresponse class *i*
= 0 otherwise, for $i = 1,...,(k-1)$

Let P_j denote the probability that school *j* is a participant, and let L_j denote the logit of P_j. That is,

$$L_j = ln(P_j/(1 - P_j)).$$

The initial model fitted for each state was

$$L_j = A + \Sigma B_i X_{ij.} \qquad (1)$$

The value of -2 log likelihood for this model, together with its degrees of freedom (k-1), are presented in Table 7-4, under the heading "Model with Only Nonresponse Classes". This constitutes a baseline, against which a second model was compared, as discussed below. Note that this model cannot be estimated if there are nonresponse classes in which all schools participated (so that no adjustments for nonresponse were made for schools in such a class). Even though this analysis was restricted to those states with relatively poor response, this occurred in a number of instances. When this happened, those (responding) schools in such classes were dropped from the analyses. Table 7-4 shows the proportion of the state public-school student population that is represented in the sample by schools from classes with less than 100 percent response. Thus in Nebraska, New Hampshire, and Tennessee, there was some nonresponse within every adjustment class, whereas for the other four states some portion of the population is not represented because schools were dropped from classes with no nonresponse.

Table 7-4

Results of Logistic Regression Analyses of School Nonresponse

Jurisdiction	School Participation Rate (%)	Percent of Population Covered by Models*	Model with Only Nonresponse Classes		Model with All Variables			Best Model - Significant Variables**
			-2 Log Likelihood	Degrees of Freedom	Change in -2 Log Likelihood	Change in Degrees of Freedom	Significance	
Montana	89	61.9	2.627	4	1.648	4	N.S.	None
Nebraska	77	100.0	0.618	3	25.701	5	$p < .005$	Y_5 - Median income ($p = .0001$) Y_1 - Percent Black ($p = .0157$)
New Hampshire	79	100.0	7.861	1	5.829	6	N.S.	None
Pennsylvania	84	78.8	0.291	3	7.813	5	N.S.	Z_1 - Type of locator ($p = .0323$)
Rhode Island	86	90.4	21.741	7	4.829	7	N.S.	None
Tennessee	74	100.0	2.917	5	11.700	5	$.025 < p < .05$	Y_3 - Median income ($p = .0132$)
Wisconsin	86	64.8	0.911	4	21.534	6	$p < .005$	Y_1 - Percent Black ($p = .0065$)

* For the remainder of the population (not covered by the models) there was 100 percent participation.

** Variables (in addition to the nonresponse classes) included in the best model obtained by a backwards stepwise procedure.

As an aside, these values for the log likelihood statistics show that in New Hampshire the response rates were significantly different between the two nonresponse classes, whereas the differences among the four classes in Nebraska were not statistically significant, nor among the six classes in Tennessee. This does not demonstrate that there was no benefit derived from the school nonresponse adjustments in Nebraska and Tennessee, as this analysis may be lacking in power, but it is suggestive of this possibility.

Within each state a second logistic model was fitted to the data on public school participation, In this model, the same indicator variables for nonresponse class were included, and also additional variables available for participating and nonparticipating schools alike. These variables were the percentage of Black students (Y_1), the percentage of Hispanic students (Y_2), the estimated grade 4 enrollment size of the school (Y_3), the median 1989 household income of the zip code area in which the school was located (Y_4), and a set of indicator variables indicating the type of location of the school. These type of location classes were the seven categories of the NCES type of location variable, described in Chapter 3. However, states did not each have six dummy variables for this classification for three reasons. First, most states are missing some of the categories. Second, it was necessary to collapse categories so that the collapsed classes did not have all schools as participants, and all as nonparticipants. Finally, since type of location classes were used in forming nonresponse adjustment classes, they are frequently confounded with the indicator variables for these classes. Thus the number of variables indicating type of location were 0 in Montana, 1 in Nebraska, 2 in New Hampshire, 1 in Pennsylvania, 3 in Rhode Island, 1 in Tennessee, and 2 in Wisconsin. These variables are denoted as Z_i, for i from 1 to the number given above. Thus in New Hampshire there are two variables, Z_1 and Z_2.

The model fitted in each state now was the following:

$$L_j = A + \Sigma\ B_i X_{ij} + \Sigma\_C_i Y_{ij} + \Sigma\_D_i Z_{ij}. \qquad (2)$$

The explanatory power of this model was compared with that of the initial one by examining the change in the value of -2 log likelihood, and assessing the statistical significance of this change. This evaluates whether, taken as a group, the Y, and Z variables are significantly related to the response probability, after accounting for nonresponse class. The results are shown in Table 7-4 under the heading "Model with All Variables".

The table shows that in Montana, New Hampshire, Pennsylvania, and Rhode Island, we are unable to detect any effect of the additional variables. In the other three states, however, these additional variables significantly explain variation in response rates, not accounted for by nonresponse class. This is in spite of the fact that functions of these variables were used in defining nonresponse adjustment classes, as described earlier in this chapter, and in Chapter 3 where the stratification for each state is described.

The final step in the analysis was to attempt to isolate which of the additional variables was able to contribute to the explanation of variation in response rate. This was done by fitting a logistic regression model, using a backwards stepwise elimination procedure to develop a parsimonious model. The starting point was the model (2) above, and nonsignificant variables *Y* and *Z* were removed until only the *X* variables, and significant *Y* and *Z* variables were retained.

The righthand column of Table 7-4 shows the ensuing variable selection for each state, along with the statistical significance of each retained variable.

This analysis shows that, for Nebraska, both the percent of Black students enrolled, and the median household income, were highly significant predictors, over and above nonresponse class. This occurs despite the fact that in Nebraska, minority enrollment was used in forming nonresponse adjustment classes within metropolitan areas (see Table 3-3). Median income classes were used to form nonresponse classes in nonmetropolitan areas, but evidently this did not capture the full explanatory power of this variable. The significance of these two variables is reflected in the results in Table 7-2. The full sample has a mean percent Black of 4.02 percent, whereas for the adjusted responding sample the mean percentage is 2.87 percent. The mean median household income for the full sample is $27,787, whereas for the respondents it is $26,247. Thus ιhere is indication that the final sample is somewhat under representative of schools with relatively high Black enrollment, and relatively high median household income.

For Pennsylvania, the single variable designating type of location is somewhat significant, even though this variable features prominently in the formation of nonresponse adjustment classes. This significance does not translate into the results of Table 7-2, since mean value for type of location for the full sample is 3.129, which is very close to the value of 3.133 for the respondents.

For Tennessee, the median income variable is somewhat significant. This variable was used in forming nonresponse adjustment classes in Tennessee only in rural areas, as minority enrollment was used in other areas (see Table 3-3). The median income for the full sample is $25,243, while for the respondents it is $23,897. This indicates that the final sample is somewhat underrepresentative of schools with relatively high median income.

For Wisconsin, the variable giving the percentage of Black enrollment is highly significant. Minority enrollment was used in forming nonresponse classes only with large central cities, and not elsewhere in the state (Table 3-3). This differential in nonresponse for schools with different levels of Black enrollment is reflected in Table 7-2. This shows that the mean percent Black for the full sample is 9.20 percent, but for the final sample it is only 5.75 percent. This indicates that the sample is likely under representative of schools with relatively high Black enrollment.

These results indicate that on occasicn there are differences between the original samples of schools, and those that participated, that are not fully removed by the process of creating nonresponse adjustments. Although these effects are not dramatic, they are statistically significant, and generally are reflected in noticeable differences in population characteristics estimated from the respondents, compared to those obtained for the full sample. However, the evidence presented here does not permit valid speculation about the likely size or even direction of the bias in the states where these sample differences are noticeable.

7.4.3 Weighted Distributions of Students Before and After Student Absenteeism

Table 7-5 shows, for the public schools in each state, the weighted sampled percentages of students by gender (male) and race/ethnicity (White, not Hispanic; Black, not Hispanic;

Table 7-5
Weighted Student Percentages Derived from Sampled Public Schools

Jurisdiction	Weighted Student Participation (%)	Weighted Estimates Derived from Full Sample							Weighted Estimates Derived from Assessed Sample, with Student Nonresponse Adjustment						
		Percent Male	Percent White	Percent Black	Percent Hispanic	Percent IEP	Percent LEP	Mean Age (Months)	Percent Male	Percent White	Percent Black	Percent Hispanic	Percent IEP	Percent LEP	Mean Age (Months)
Alabama	96.07	50.72	62.67	28.92	5.54	5.89	0.14	119.27	50.76	63.69	27.85	5.51	5.80	0.10	119.28
Arizona	94.27	49.76	58.32	3.63	28.58	6.44	9.02	119.23	49.52	58.32	3.63	28.58	6.40	9.15	119.23
Arkansas	95.96	50.06	70.66	20.58	6.08	6.14	0.30	119.19	49.90	71.28	19.56	6.14	6.05	6.40	119.19
California	93.86	50.86	44.03	7.01	33.38	5.81	16.09	116.42	50.90	45.80	6.75	32.18	5.14	16.33	116.40
Colorado	94.25	49.27	68.16	4.77	20.85	5.96	2.43	118.63	49.78	69.39	4.49	20.00	5.86	2.40	118.63
Connecticut	95.63	49.66	70.64	11.80	13.66	8.25	1.21	116.96	49.73	72.70	10.71	13.02	8.22	1.25	116.97
Delaware	95.58	49.46	63.21	23.81	9.18	9.16	0.60	116.42	49.22	67.09	20.71	8.62	9.13	0.63	116.40
DoDEA Overseas	94.58	49.84	47.87	19.10	16.90	3.79	1.55	117.00	50.06	47.87	19.10	16.90	3.57	1.57	117.00
Dist. of Columbia	94.52	50.16	5.33	79.73	12.15	1.57	1.75	117.95	49.81	8.36	76.74	11.90	1.51	1.85	117.91
Florida	93.96	48.87	57.08	21.18	18.78	9.87	3.56	119.55	48.84	59.20	19.40	18.02	9.79	3.55	119.58
Georgia	95.45	48.40	55.83	32.25	8.50	5.07	1.02	119.81	48.41	57.90	30.42	8.27	5.19	0.94	119.80
Guam	95.91	50.04	9.06	3.53	17.12	0.48	3.05	115.50	50.66	8.97	3.24	16.26	0.50	2.90	115.49
Hawaii	95.45	50.85	17.02	3.06	20.22	4.01	3.55	114.86	50.54	18.33	2.73	18.72	3.84	3.66	114.86
Idaho	96.12	50.62	81.09	0.57	13.30	6.05	1.79	118.18	50.16	81.45	0.55	13.01	5.92	1.74	118.17
Indiana	95.86	49.47	81.09	10.06	6.48	6.10	0.23	120.55	49.18	80.54	10.49	6.54	6.14	0.21	120.54
Iowa	95.54	50.95	88.36	2.73	5.84	6.74	0.20	119.78	51.07	88.56	2.60	6.04	6.50	0.21	119.79
Kentucky	96.68	51.21	83.84	9.73	4.44	4.02	0.24	118.84	51.19	84.10	9.14	4.70	3.95	0.25	118.84
Louisiana	96.08	50.04	50.86	38.83	7.47	5.16	0.56	120.52	49.51	54.58	35.27	7.02	5.03	0.58	120.45
Maine	94.32	50.13	92.04	0.74	4.33	7.30	0.17	119.39	50.12	92.08	0.75	4.25	7.23	0.18	119.39
Maryland	95.20	52.82	57.19	31.77	6.01	7.75	0.71	115.25	52.11	58.89	30.22	5.81	7.76	0.74	115.21
Massachusetts	95.43	50.35	76.94	7.44	10.69	10.36	1.43	117.65	49.71	76.89	7.55	10.56	10.11	1.43	117.65
Michigan	[illegible]1.87	49.22	72.78	15.15	7.80	3.54	0.51	118.24	48.92	72.78	15.15	7.80	3.60	0.46	118.21
Minnesota	95.49	51.09	83.93	2.93	7.98	7.09	0.83	119.32	50.89	84.97	2.83	7.47	7.29	0.84	119.32
Mississippi	96.68	48.84	45.98	45.35	6.70	3.46	0.15	120.92	48.60	48.90	42.69	6.44	3.57	0.15	120.90
Missouri	95.04	51.22	75.56	14.12	6.58	7.13	0.03	120.35	51.45	76.76	12.94	6.78	7.02	0.03	120.32
Montana	95.72	50.73	79.06	0.53	9.56	7.45	1.00	120.32	50.61	78.45	0.54	9.39	7.28	1.01	120.30
Nebraska	94.85	50.78	82.26	3.74	9.40	11.75	0.56	119.15	50.92	91.28	3.34	8.81	11.69	0.54	119.11
New Hampshire	95.58	49.69	91.28	0.98	4.56	9.71	0.17	119.52	49.63	60.74	0.98	4.56	9.77	0.17	119.51
New Jersey	95.30	48.80	60.35	16.08	17.08	5.37	1.52	117.47	48.62	39.03	15.41	17.59	5.28	1.58	117.44
New Mexico	94.68	47.88	41.14	2.90	43.88	8.64	2.57	118.93	47.86	56.69	3.01	42.01	8.59	2.65	118.91

Table 7-5 (continued)
Weighted Student Percentages Derived from Sampled Public Schools

Jurisdiction	Weighted Student Participation (%)	Weighted Estimates Derived from Full Sample							Weighted Estimates Derived from Assessed Sample, with Student Nonresponse Adjustment						
		Percent Male	Percent White	Percent Black	Percent Hispanic	Percent IEP	Percent LEP	Mean Age (Months)	Percent Male	Percent White	Percent Black	Percent Hispanic	Percent IEP	Percent LEP	Mean Age (Months)
New York	95.34	50.05	54.44	20.63	19.21	4.38	3.64	115.98	49.84	65.49	19.59	17.98	4.39	3.60	115.97
North Carolina	95.83	50.81	65.49	26.11	3.74	9.34	0.49	118.23	50.72	83.66	26.11	3.74	9.15	0.47	118.24
North Dakota	96.63	49.79	88.07	1.06	5.49	7.40	0.46	119.55	50.16	84.92	1.31	6.13	7.39	0.47	119.57
Pennsylvania	94.13	49.86	76.47	13.94	6.32	4.77	0.79	118.51	49.51	77.29	13.59	5.80	4.75	0.82	118.50
Rhode Island	94.70	49.15	80.22	5.83	9.12	8.28	2.26	117.14	48.96	80.90	5.63	9.01	8.14	2.36	117.12
South Carolina	96.39	50.78	53.10	36.72	7.48	6.69	0.28	118.22	50.65	54.75	35.14	7.30	6.62	0.29	118.23
Tennessee	95 63	49.48	74.26	19.73	4.00	7.37	0.07	119.38	49.12	74.26	19.73	4.00	6.78	0.08	119.37
Texas	96.45	49.86	50.39	12.06	33.91	6.89	8.64	119.78	49.91	50.39	12.06	33.91	6.90	8.74	119.79
Utah	94.82	50.74	82.05	0.65	11.60	6.60	0.88	118.00	50.54	82.05	0.65	11.60	6.37	0.90	117.98
Virginia	94.65	50.16	59.57	28.60	7.18	6.23	0.78	117.56	50.28	60.83	27.52	6.95	6.17	0.80	117.54
Washington	94.45	51.93	73.30	4.94	11.29	7.72	2.21	118.80	52.17	73.30	4.94	11.29	7.74	2.30	118.79
West Virginia	95.88	50.87	90.75	3.07	3.91	5.35	0.11	119.28	50.60	90.54	3.19	3.94	5.39	0.11	119.28
Wisconsin	96.34	49.19	83.45	4.74	7.30	4.61	1.67	119.25	48.97	84.43	4.52	6.91	4.54	1.74	119.24
Wyoming	95.92	51.01	82.01	0.83	12.21	7.31	0.31	119.66	50.85	82.01	0.83	12.21	7.31	0.29	119.68

Hispanic), Individualized Education Program (IEP) Status, and Limited English Proficient (LEP) Status for the full sample of students (after student exclusion) and for the assessed sample. The mean student age in months is also presented on each basis. Table 7-6 shows these results for all students, public and nonpublic, in those states having adequate school response rates to permit reporting of combined results for public and nonpublic students.

The weight used for the full sample is the adjusted student base weight, defined in section 7.3.5. The weight for the assessed students is the final student weight, also defined in section 7.3.5. The difference between the estimates of the population subgroups is an estimate of the bias in estimating the size of the subgroup, resulting from student absenteeism.

Care must be taken in interpreting these results, however. First, note that there is generally very little difference in the proportions estimated from the full sample and those estimated from the assessed students. While this is encouraging, it does not eliminate the possibility that bias exists, either within the state as a whole, or for results for gender and race/ethnicity subgroups, or for other subgroups. Second, on the other hand, where differences do exist they cannot be used to indicate the likely magnitude or direction of the bias with any reliability. For example, in Table 7-5, for New York the percentages of Black and Hispanic students in the full sample are respectively 20.62 and 19.21 percent. For assessed students, these percentages are 19.59 for Black students and 17.98 for Hispanic students. While these differences raise the possibility that some bias exists, it is not appropriate to speculate on the magnitude of this bias by considering the assessment results for Black and Hispanic students, in comparison to other students in the state. This is because the underrepresented Black and Hispanic students may not be typical of students that were included in the sample, and similarly those students within the same racial/ethnic groups who are disproportionately overrepresented may not be typical either. This is because not all students within the same race/ethnicity group receive the same student nonresponse adjustment.

One other feature to note is that, for assessed students, information as to the student's gender and race/ethnicity is provided by the student, while for absent students this information is provided by the school. Evidence from past NAEP assessments (see, for example, Rust & Johnson, 1992) indicates that there can be substantial discrepancies between those two sources, especially with regard to classifying grade 4 students as Hispanic.

7.5 VARIATION IN WEIGHTS

After computation of full-sample weights, an analysis was conducted on the distribution of the final student weights in each state. The analysis was intended to (1) check that the various weight components had been derived properly in each state, and (2) examine the impact of the variability of the sample weights on the precision of the sample estimates, both for the state as a whole and for major subgroups within the state.

The analysis was conducted by looking at the distribution of the final student weights for the assessed students in each state and for subgroups defined by age, sex, race, level of urbanization, and level of parents' education. Two key aspects of the distribution were considered in each case: the coefficient of variation (equivalently, the relative variance) of the

Table 7-6
Weighted Student Percentages Derived from All Schools Sampled

Jurisdiction	Weighted Student Participation (%)	Weighted Estimates Derived from Full Sample							Weighted Estimates Derived from Assessed Sample, with Student Nonresponse Adjustment						
		Percent Male	Percent White	Percent Black	Percent Hispanic	Percent IEP	Percent LEP	Mean Age (Months)	Percent Male	Percent White	Percent Black	Percent Hispanic	Percent IEP	Percent LEP	Mean Age (Months)
Alabama	95.99	50.56	62.35	28.89	5.76	5.54	0.13	119.20	50.59	63.32	27.86	5.74	5.46	0.09	119.20
Arkansas	95.90	50.08	70.16	20.69	6.35	5.97	0.34	119.11	49.98	70.75	19.67	6.42	5.86	0.36	119.12
Colorado	94.24	49.54	67.44	4.81	21.46	5.68	2.28	118.64	50.17	68.67	4.53	20.60	5.58	2.26	118.63
Connecticut	95.58	49.59	70.36	11.88	13.86	7.72	1.28	116.81	49.56	72.42	10.80	13.24	7.72	1.32	116.81
Delaware	95 84	49.18	63.12	23.46	9.47	7.84	0.49	116.42	48.93	66.98	20.43	8.88	7.78	0.52	116.40
Florida	94.28	49.33	56.55	21.31	19.10	9.10	3.27	119.46	49.25	58.68	19.52	18.33	8.95	3.27	119.49
Georgia	95.52	47.77	55.49	32.25	8.75	4.95	0.95	119.76	47.74	57.56	30.41	8.52	5.04	0.88	119.76
Guam	96.19	50.23	9.37	3.64	17.83	0.41	2.60	115.33	50.76	9.26	3.34	16.91	0.42	2.48	115.32
Hawaii	95.54	49.88	16.59	3.13	21.02	3.59	3.15	114.94	49.57	18.00	2.79	19.44	3.45	3.26	114.94
Indiana	96.12	49.71	80.71	10.15	6.68	5.72	0.31	120.53	49.48	80.18	10.59	6.74	5.76	0.25	120.51
Iowa	95.94	50.92	88.19	2.66	5.96	6.37	0.17	119.80	51.00	88.39	2.54	6.16	6.14	0.18	119.80
Kentucky	96.67	50.77	83.42	9.91	4.60	3.73	0.21	118.79	50.68	83.78	9.28	4.80	3.66	0.23	118.80
Louisiana	96.19	48.74	50.82	38.43	7.82	4.36	0.51	120.04	48.28	54.56	34.87	7.34	4.26	0.53	119.99
Maine	94.33	50.05	91.72	0.79	4.56	7.14	0.17	119.38	50.00	91.76	0.79	4.48	7.08	0.17	119.39
Massachusetts	95.50	50.46	76.71	7.38	10.87	9.62	1.29	117.64	49.86	76.63	7.52	10.72	9.41	1.29	117.64
Minnesota	95.56	50.45	83.69	2.94	8.21	6.46	0.73	119.34	50.25	84.72	2.85	7.68	6.64	0.74	119.35
Missouri	94.87	51.28	75.09	14.09	6.90	6.72	0.05	120.27	51.44	76.27	12.94	7.10	6.65	0.06	120.24
New Jersey	95.37	49.00	60.39	15.75	17.28	5.27	1.38	117.30	48.91	60.82	15.13	17.70	5.10	1.44	117.27
New Mexico	94 46	48.60	41.11	2.82	44.28	8.13	3.10	118.89	48.60	39.04	2.97	42.45	8.08	3.18	118.86
North Dakota	96.27	49.69	87.85	1.10	5.64	7.46	1.30	119.47	50.00	84.84	1.37	6.35	7.37	1.30	119.47
Pennsylvania	94.17	49.83	76.23	13.74	6.62	4.09	0.99	118.34	49.58	77.06	13.43	6.08	4.07	1.00	118.34
Rhode Island	94 89	49.45	79.85	5.91	9.38	7.58	2.21	116.97	49.32	80.50	5.72	9.27	7.47	2.31	116.95
Virginia	94.71	49.67	59.17	28.53	7.48	5.87	0.77	117.56	49.83	60.43	27.45	7.24	5.81	0.79	117.55
West Virginia	95.93	50.99	90.46	3.12	4.05	5.14	0.10	119.23	50.76	90.30	3.25	4.08	5.18	0.10	119.24

weight distribution; and the presence of outliers—that is, cases whose weights were several standard deviations away from the median weight.

It was important to examine the coefficient of variation of the weights because a large coefficient of variation reduces the effective size of the sample. Assuming that the variables of interest for individual students are uncorrelated with the weights of the students, the sampling variance of an estimated average or aggregate is approximately $(1+\left[\frac{C}{100}\right]^2)$ times as great as the corresponding sampling variance based on a self-weighting sample of the same size, where C is the coefficient of variation of the weights expressed as a percent. Outliers, or cases with extreme weights, were examined because the presence of such an outlier was an indication of the possibility that an error was made in the weighting procedure, and because it was likely that a few extreme cases would contribute substantially to the size of the coefficient of variation.

In most states, the coefficients of variation were 35 percent or less, both for the whole sample and for all major subgroups. This means that the quantity $(1+\left[\frac{C}{100}\right]^2)$ was generally below 1.1, and the variation in sampling weights had little impact on the precision of sample estimates.

A few relatively large student weights were observed in one state. These extreme weights were for students in a school for which the grade enrollment available at the time of sample selection proved to be several-fold short of the actual enrollment. An evaluation was made of the impact of trimming these largest weights back to a level consistent with the largest remaining weights found in the state. Such a procedure produced an appreciable reduction in the size of the coefficient of variation for the weights in this state, and hence this trimming was implemented in that state. We judged that this procedure had minimal potential to introduce bias, while the reduction in the coefficient of variation of the weights gives rise to an appreciable decrease in sampling error for the state.

7.6 CALCULATION OF REPLICATE WEIGHTS

A replication method known as *jackknife* was used to estimate the variance of statistics derived from the full sample. The process of replication involves repeatedly selecting portions of the sample (replicates) and calculating the desired statistic (replicate estimates). The variability among the calculated replicate estimates is then used to obtain the variance of the full-sample estimate.

In each state, replicates were formed in two steps. First, each school was assigned to one of a maximum of 62 replicate groups, each group containing at least one school. In the next step, a random subset of schools (or, in some cases, students within schools) in each replicate group was excluded. The remaining subset and all schools in the other replicate groups then constituted one of the 62 replicates. The process of forming these replicate groups, core to the process of variance estimation, is described below.

7.6.1 Defining Replicate Groups and Forming Replicates for Variance Estimation

Replicate groups were formed separately for public and nonpublic schools. Once replicate groups were formed for all schools, students were then assigned to their respective school replicate groups.

Public Schools. These schools were sorted according to the state, monitoring status, and, within monitoring status, the order in which they were selected from the sampling frame. The schools were then were grouped in pairs. Where there was an odd number of schools, the last replicate group contained three schools instead of two. The pairing was done such that no single pair contained schools with different monitoring status. In those states where the number of pairs exceeded 62 (Montana and Nebraska), the pair numbering proceeded up to 62, and then decreased back from 62 for the last few pairs.

Each of the certainty public schools (excluding those in Guam and the District of Columbia) was assigned to a single replicate group of its own. Here, schools were sorted by the estimated grade enrollment prior to group assignments. Again, depending on the state, a maximum of 62 certainty groups was formed. The group numbering resumed from the last group number used for the noncertainty schools if the total number of public school groups was less than 62. Otherwise, the numbering started from 62 down to the number needed for the last certainty public school. In the District of Columbia, which had only 117 certainty schools (no noncertainty schools), groups started at 1 and continued up to 62 and then back down to 8.

The purpose of this scheme was to assign as many replicates to a state's public schools as permitted by the design, to a maximum of 62. When more than 62 replicates were assigned, the procedure ensured that no subset of the replicate groups (pairs of noncertainty schools, individual certainty schools, or groups of these) was substantially larger than the other replicate groups. The aim was to maximize the degrees of freedom available for estimating variances for public-school data.

A single replicate was formed by dropping one member of a given pair. This process was repeated successively across pairs, giving up to 62 replicates.

Nonpublic Schools. Replicate groups for noncertainty nonpublic schools were formed in one of the two methods described below. If any of the following conditions was true for a given state, then the subsequent steps were taken to form replicate groups. Here, the numbering started at 62 down to the last needed number.

Conditions for Method 1:

- fewer than 11 nonpublic noncertainty schools;
- fewer than 2 Catholic noncertainty schools; or
- fewer than 2 nonCatholic noncertainty schools.

Steps for Method 1:

- all schools were grouped into a single replicate group;
- schools were randomly sorted; and
- starting with the second school, replicates were formed by consecutively leaving out one of the remaining *n - 1* schools; each replicate included the first school.

When a given state did not match conditions of the first method, i.e., when all of the following conditions were true, then the preceding steps were repeated separately for two replicate groups, one consisting of Catholic schools and on consisting of nonCatholic schools.

Conditions for Method 2:

- more than 10 nonpublic noncertainty schools;
- more than 1 Catholic noncertainty school; and
- more than 1 nonCatholic noncertainty school.

For states with certainty nonpublic schools (Delaware, District of Columbia, and Hawaii) each school was assigned to a single group. Prior to this assignment, schools were sorted in descending order of the estimated grade enrollment. The group numbering started at the last number where the noncertainty nonpublic schools ended. A replicate was formed by randomly deleting one half of the students in a certain school from the sample. This was repeated for each certainty school.

Again, the aim was to maximize the number of degrees of freedom for estimating sampling errors for nonpublic schools (and indeed for public and nonpublic schools combined) within the constraint of forming 62 replicate groups. Where a state had a significant contribution from both Catholic and nonCatholic schools, we ensured that the sampling error estimates reflected the stratification on this characteristic.

Guam. For Guam schools, the number of half-groups per school were obtained based on the number of students. For public schools, if the numbers of students were less than 60, between 60 and 119, and over 119, then the number of half-groups per school were set to 2, 4, and 6, respectively. For nonpublic schools, the limits were set to less than 70, between 70 and 119, and over 119.

7.6.2 School-level Replicate Weights

As mentioned above, each replicate sample had to be reweighted to compensate for the dropped unit(s) defining the replicate. This reweighting was done in two stages. At the first-stage, the ith school included in a particular replicate r was assigned a replicate-specific school base weight defined as

$$W_{ri}^{sch} = K_r \times W_i^{sch}$$

where W_i^{sch} is the full-sample base weight for school i, and, for public schools

$$K_r = \begin{cases} 1.5 & \text{if school } i \text{ was contained in a "pair" consisting of 3 units from which the complementary member was dropped to form replicate } r, \\ 2 & \text{if school } i \text{ was contained in a pair consisting of 2 units from which the complementary member was dropped to form replicate } r, \\ 0 & \text{if school } i \text{ was dropped to form replicate } r, \text{ and} \\ 1 & \text{otherwise.} \end{cases}$$

For private schools, Method 1:

$$K_r = \begin{cases} \frac{n}{n-1} & \text{if school } i \text{ was not dropped in forming replicate } r \\ 0 & \text{if school } i \text{ was dropped to form replicate } r \end{cases}$$

For private schools, Method 2 (with n_1 Catholic schools and n_2 nonCatholic schools):

$$K_r = \begin{cases} \frac{n_1}{n_1-1} & \text{if school } i \text{ was Catholic, not dropped from replicate } r, \text{ and replicate } r \text{ was formed by dropping a Catholic school} \\ 1 & \text{if school } i \text{ was Catholic and replicate } r \text{ was formed by dropping a nonCatholic school} \\ \frac{n_2}{n_2-1} & \text{if school } i \text{ was nonCatholic, not dropped from replicate } r, \text{ and replicate } r \text{ was formed by dropping a nonCatholic school} \\ 1 & \text{if school } i \text{ was nonCatholic and replicate } r \text{ was formed by dropping a Catholic school} \\ 0 & \text{if school } i \text{ was dropped to form replicate } r \end{cases}$$

Using the replicate-specific school base weights, W_{ri}^{sch}, the school-level nonresponse weighting adjustments were recalculated for each replicate r. That is, the school-level nonresponse adjustment factor for schools in replicate r and adjustment class k was computed as

$$F_{rk} = \frac{\sum_{i \in C_k} (W^{sch}_{rki} \times E_{ki})}{\sum_{i \in C_k} (W^{sch}_{rki} \times E_{ki} \times \delta_{rki})}$$

where

C_k = the subset of school records in adjustment class k;

W^{sch}_{rki} = the replicate-r base weight of the ith school in class k;

E_{ki} = the QED grade enrollment for the ith school in class k;

In the above formulation, the indicator variable δ_{rki} had a nonzero value only when the ith school in replicate r and adjustment class k participated in the assessment. The replicate-specific nonresponse-adjusted school weight for the ith school in replicate r in class k was then computed as

$$W^{adj}_{rki} = F_{rk} \times W^{sch}_{rki} \times \delta_{rki} \quad .$$

7.6.3 Student-level Replicate Weights

The replicate-specific adjusted student base weights were calculated by multiplying the replicate-specific adjusted school weights as described above by the corresponding within-school student weights. That is, the adjusted student base weight for the jth student in adjustment class k in replicate r was initially computed as

$$W_{rkij} = W^{adj}_{rki} \times W^{within}_{ij}$$

where

W^{adj}_{rki} = the nonresponse-adjusted school weight for school i in school adjustment class k and replicate r; and

W^{within}_{ij} = the within-school weight for the jth student in school i.

The final replicate-specific student weights were then obtained by applying the student nonresponse adjustment procedures to each set of replicate student weights. Let F_{rk} denote the student-level nonresponse adjustment factor for replicate r and adjustment class k. The final replicate-r student weight for student j in school i in adjustment class k was calculated as:

$$W^{final}_{rkij} = F_{rk} \times W^{adj}_{rki} \times W^{within}_{ij} \quad .$$

Finally, estimates of the variance of sample-based estimates were calculated as

$$Var_{JK}(\hat{x}) = \sum_{r=1}^{62} (\hat{x}_r - \hat{x})^2 ,$$

where

$$\hat{x} = \sum_{i=j}^{n} \times W_{kij}^{final} \times x_{kij}$$

denote an estimated total based on the full sample, and $\hat{x}_r$ denote the corresponding estimate based on replicate *r* with 62 replicates. The standard error of an estimate $\hat{x}$ is estimated by taking the square root of the estimated variance, $\mathrm{Var}_{JK}(\hat{x})$.

Chapter 8

THEORETICAL BACKGROUND AND PHILOSOPHY OF NAEP SCALING PROCEDURES

Eugene G. Johnson, Robert J. Mislevy, and Neal Thomas

Educational Testing Service

8.1 OVERVIEW

The primary method by which results from the Trial State Assessment are disseminated is scale-score reporting. With scaling methods, the performance of a sample of students in a subject area or subarea can be summarized on a single scale or a series of scales even when different studɛnts have been administered different items. This chapter presents an overview of the scaling methodologies employed in the analyses of the data from NAEP surveys in general and from the Trial State Assessment in reading in particular. Details of the scaling procedures specific to the Trial State Assessment are presented in Chapter 9.

8.2 BACKGROUND

The basic information from an assessment consists of the responses of students to the items presented in the assessment. For NAEP, these items are constructed to measure performance on sets of objectives developed by nationally representative panels of learning area specialists, educators, and concerned citizens. Satisfying the objectives of the assessment and ensuring that the tasks selected to measure each goal cover a range of difficulty levels typically requires many items. For example, the Trial State Assessment in reading required 84 items at grade 4. To reduce student burden, each assessed student was presented only a fraction of the full pool of items through multiple matrix sampling procedures.

The most direct manner of presenting the assessment results is to report separate statistics for each item. However, because of the vast amount of information, having separate results for each of the items in the assessment pool hinders the comparison of the general performance ɔf subgroups of the population. Item-by-item reporting masks similarities in trends and subgroup comparisons that are common across items.

An obvious summary of performance across a collection of items is the average of the separate item scores. The advantage of averaging is that it tends to cancel out the effects of peculiarities in items that can affect item difficulty in unpredictable ways. Furthermore, averaging makes it possible to compare ınore easily the general performances of subpopulations.

Despite their advantages, there are a number of significant problems with average item scores. First, the interpretation of these results depends on the selection of the items; the selection of easy or difficult items could make student performance appear to be overly high or low. Second, the average score is related to the particular items comprising the average, so that direct comparisons in performance between subpopulations require that those subpopulations have been administered the same set of items. Third, because this approach limits comparisons to average scores on specific sets of items, it provides no simple way to report trends over time when the item pool changes. Finally, direct estimates of parameters or quantities such as the proportion of students who would achieve a certain score across the items in the pool are not possible when every student is administered only a fraction of the item pool. While the mean average score across all items in the pool can be readily obtained (as the average of the individual item scores), statistics that provide distributional information, such as quantiles of the distribution of scores across the full set of items, cannot be readily obtained without additional assumptions.

These limitations can be overcome by the use of response scaling methods. If several items require similar skills, the regularities observed in response patterns can often be exploited to characterize both respondents and items in terms of a relatively small number of variables. These variables include a respondent-specific variable, called proficiency, which quantifies a respondent's tendency to answer items correctly (or, for multipoint items, to achieve a certain score) and item-specific variables that indicate characteristics of the item such as its difficulty, effectiveness in distinguishing between individuals with different levels of proficiency, and the chances of a very low proficiency respondent correctly answering a multiple-choice item. (These variables are discussed in more detail in the next section). When combined through appropriate mathematical formulas, these variables capture the dominant features of the data. Furthermore, all students can be placed on a common scale, even though none of the respondents takes all of the items within the pool. Using the common scale, it becomes possible to discuss distributions of proficiency in a population or subpopulation and to estimate the relationships between proficiency and background variables.

It is important to point out that any procedure of aggregation, from a simple average to a complex multidimensional scaling model, highlights certain patterns at the expense of other potentially interesting patterns that may reside within the data. Every item in a NAEP survey is of interest and can provide useful information about what young Americans know and can do. The choice of an aggregation procedure must be driven by a conception of just which patterns are salient for a particular purpose.

The scaling for the Trial State Assessment in reading was carried out separately within the two reading content areas specified in the framework for grade 4 reading. This scaling within subareas was done because it was anticipated that different patterns of performance might exist for these essential subdivisions of the subject area. The two content area scales correspond with two purposes of reading—Reading for Literary Experience and Reading to Gain Information. By creating a separate scale for each of these content areas, potential differences in subpopulation performance between the content areas are preserved.

The creation of a series of separate scales to describe reading performance does not preclude the reporting of a single index of overall reading performance—that is, an overall reading composite. A composite is computed as the weighted average of the two content area scales, where the weights correspond to the relative importance given to each content area as defined by the framework. The composite provides a global measure of performance within the subject area, while the constituent content area scales allow the measurement of important interactions within educationally relevant subdivisions of the subject area.

8.3 SCALING METHODOLOGY

This section reviews the scaling models employed in the analyses of data from the Trial State Assessment in reading and the 1994 national reading assessment, and the multiple imputation or "plausible values" methodology that allows such models to be used with NAEP's sparse item-sampling design. The reader is referred to Mislevy (1991) for an introduction to plausible values methods and a comparison with standard psychometric analyses, to Mislevy, Johnson and Muraki (1992) and Beaton and Johnson (1992) for additional information on how the models are used in NAEP, and to Rubin (1987) for the theoretical underpinnings of the approach. It should be noted that the imputation procedure used by NAEP is a mechanism for providing plausible values for proficiencies and not for filling in blank responses to background or cognitive variables.

While the NAEP procedures were developed explicitly to handle the characteristics of NAEP data, they build on other research, and are paralleled by other researchers. See, for example Dempster, Laird, and Rubin (1977); Little and Rubin (1983, 1987); Andersen (1980); Engelen (1987); Hoijtink (1991); Laird (1978); Lindsey, Clogg, and Grego (1991); Zwinderman (1991); Tanner and Wong (1987); and Rubin (1987, 1991).

The 84 reading items administered at grade 4 in the Trial State Assessment were also administered to fourth-grade students in the national reading assessment. However, because the administration procedures differed, the Trial State Assessment data were scaled independently from the national data. The national data also included results for students in grades 8 and 12. Details of the scaling of the Trial State Assessment and the subsequent linking to the results from the national reading assessment are provided in Chapter 9.

8.3.1 The Scaling Models

Three distinct scaling models, depending on item type and scoring procedure, were used in the analysis of the data from the Trial State Assessment. Each of the models is based on item response theory (IRT; e.g., Lord, 1980). Each is a "latent variable" model, defined separately for each of the scales, which express respondents' tendencies to achieve certain scores (such as correct/incorrect) on the items contributing to a scale as a function of a parameter that is not directly observed, called proficiency on the scale.

A three-parameter logistic (3PL) model was used for the multiple-choice items (which were scored correct/incorrect). The fundamental equation of the 3PL model is the probability that a person whose proficiency on scale k is characterized by the *unobservable* variable θ_k will respond correctly to item j:

$$P(X_j = 1|\theta_k,a_j,b_j,c_j) = c_j + \frac{(1 - c_j)}{1 + \exp[-1.7a_j\,(\theta_k - b_j)]} \equiv P_{j1}(\theta_k) \quad , \tag{8.1}$$

where

x_j is the response to item j, 1 if correct and 0 if not;

a_j where $a_j > 0$, is the slope parameter of item j, characterizing its sensitivity to proficiency;

b_j is the threshold parameter of item j, characterizing its difficulty; and

c_j where $0 \leq c_j < 1$, is the lower asymptote parameter of item j, reflecting the chances of students of very low proficiency selecting the correct option.

Further define the probability of an incorrect response to the item as

$$P_{j0} = P(x_j = 0|\theta_k,a_j,b_j,c_j) = 1 - P_{j1}(\theta_k) \tag{8.2}$$

A two-parameter logistic (2PL) model was used for short constructed-response items, which were scored correct or incorrect. The form of the 2PL model is the same as equations (8.1) and (8.2) with the c_j parameter fixed at zero.

Thirty-nine multiple-choice and 45 constructed-response items were presented in the Trial State and grade 4 national assessments. Of the latter, 37 were short constructed-response items, nine of which were scored on a three-point scale and 28 of which were dichotomously scored. The remaining eight constructed-response items were scored on a five-point scale with potential scores ranging from 0 to 4. Items that are scored on a multipoint scale are referred to as polytomous items, in contrast with the multiple-choice and short constructed-response items, which are scored correct/incorrect and referred to as dichotomous items.

The polytomous items were scaled using a generalized partial credit model (Muraki, 1992). The fundamental equation of this model is the probability that a person with proficiency θ_k on scale k will have, for the jth item, a response x_j that is scored in the ith of m_j ordered score categories:

$$P(X_j = i|\theta_k,a_j,b_j,d_{j,1},\ldots,d_{j,m_j-1}) = \frac{\exp(\sum_{v=0}^{i} 1.7a_j(\theta_k-b_j+d_{j,v})}{\sum_{g=0}^{m_j-1} \exp\ (\sum_{v=0}^{g} 1.7a_j(\theta_k-b_j+d_{j,v}))}$$

$$\equiv P_{ji}(\theta_k) \tag{8.3}$$

where

m_j is the number of categories in the response to item j

x_j is the response to item j, with possibilities $0,1,\ldots,m_j\text{-}1$

a_j is the slope parameter;

b_j is the item location parameter characterizing overall difficulty; and

$d_{j,i}$ is the category i threshold parameter (see below).

Indeterminacies in the parameters of the above model are resolved by setting $d_{j,o} = 0$ and setting

$$\sum_{i=1}^{m_j-1} d_{j,i} = 0.$$

Muraki (1992) points out that b_j - $d_{j,i}$ is the point on the θ_k scale at which the plots of $P_{j,i-1}(\theta_k)$ and $P_{ji}(\theta_k)$ intersect and so characterizes the point on the θ_k scale above which the category i response to item j has the highest probability of incurring a change from response category i-1 to i.

When $m_j = 2$, so that there are two score categories (0,1), it can be shown that $P_{ji}(\theta_k)$ of equation 8.3 for $i=0,1$ corresponds respectively to $P_{j0}(\theta_k)$ and $P_{j1}(\theta_k)$ of the 2PL model (equations 8.1 and 8.2 with $c_j=0$).

A typical assumption of item response theory is the conditional independence of the response by an individual to a set of items, given the individual's proficiency. That is, conditional on the individual's θ_k, the joint probability of a particular response pattern $\underset{\sim}{x} = (x_1,\ldots,x_n)$ across a set of n items is simply the product of terms based on (8.1), (8.2), and (8.3):

$$P(\underline{x}|\theta_k, item\ parameters) = \prod_{j=1}^{n} \prod_{i=0}^{m_j-1} P_{ji}(\theta_k)^{u_{ji}} \tag{8.4}$$

where $P_{ji}(\theta_k)$ is of the form appropriate to the type of item (dichotomous or polytomous), m_j is taken equal to 2 for the dichotomously scored items, and u_{ji} is an indicator variable defined by

$$u_{ji} = \begin{cases} 1 \text{ if response } x_j \text{ was in category } i \\ 0 \text{ otherwise.} \end{cases}$$

It is also typically assumed that response probabilities are conditionally independent of background variables ($\underline{y}$), given θ_k, or

$$P(\underline{x}|\theta_k, item\ parameters, \underline{y}) = p(\underline{x}|\theta_k, item\ parameters) \tag{8.5}$$

After $\underline{x}$ has been observed, equation 8.4 can be viewed as a likelihood function, and provides a basis for inference about θ_k or about item parameters. Estimates of item parameters were obtained by the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs, and which concurrently estimates parameters for all items (dichotomous and polytomous). The item parameters are then treated as known in subsequent calculations. The parameters of the items constituting each of the separate scales were estimated independently of the parameters of the other scales. Once items have been calibrated in this manner, a likelihood function for the scale proficiency θ_k is induced by a vector of responses to any subset of calibrated items, thus allowing θ_k-based inferences from matrix samples.

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

Although the IRT models are employed in NAEP only to summarize performance, a number of checks are made to detect serious violations of the assumptions underlying the models (such as conditional independence). When warranted, remedial efforts are made to mitigate the effects of such violations on inferences. These checks include comparisons of empirical and theoretical item response functions to identify items for which the IRT model may provide a poor fit to the data.

Scaling areas in NAEP are determined *a priori* by grouping items into content areas for which overall performance is deemed to be of interest, as defined by the frameworks developed by the National Assessment Governing Board. A proficiency scale θ_k is defined *a priori* by the collection of items representing that scale. What is important, therefore, is that the models capture salient information in the response data to effectively summarize the overall performance on the content area of the populations and subpopulations being assessed in the content area. NAEP routinely conducts differential item functioning (DIF) analyses to guard against potential biases in making subpopulation comparisons based on the proficiency distributions.

The local independence assumption embodied in equation 8.4 implies that item response probabilities depend only on θ and the specified item parameters, and not on the position of the item in the booklet, the content of items around an item of interest, or the test-administration and timing conditions. However, these effects are certainly present in any application. The practical question is whether inferences based on the IRT probabilities obtained via 8.4 are robust with respect to the ideal assumptions underlying the IRT model. Our experience with the 1986 NAEP reading anomaly (Beaton & Zwick, 1990) has shown that for measuring small changes over time, changes in item context and speededness conditions can lead to unacceptably large random error components. These can be avoided by presenting items used to measure change in identical test forms, with identical timings and administration conditions. Thus, we do *not* maintain that the item parameter estimates obtained in any particular booklet configuration are appropriate for other conceivable configurations. Rather, we assume that the parameter estimates are context-bound. (For this reason, we prefer common population equating to common item equating whenever equivalent random samples are available for linking.) This is the reason that the data from the Trial State Assessment were calibrated separately from the data from the national NAEP—since the administration procedures differed somewhat between the Trial State Assessment and the national NAEP, the values of the item parameters could be different. Chapter 9 provides details on the procedures used to link the results of the 1994 Trial State Assessment to those of the 1994 national assessment.

8.3.2 An Overview of Plausible Values Methodology

Item response theory was developed in the context of measuring individual examinees' abilities. In that setting, each individual is administered enough items (often 60 or more) to permit precise estimation of his or her θ, as a maximum likelihood estimate $\hat{\theta}$, for example. Because the uncertainty associated with each θ is negligible, the distribution of θ, or the joint distribution of θ with other variables, can then be approximated using individuals' $\hat{\theta}$ values as if they were θ values.

This approach breaks down in the assessment setting when, in order to provide broader content coverage in limited testing time, each respondent is administered relatively few items in a scaling area. The problem is that the uncertainty associated with individual θs is too large to ignore, and the features of the $\hat{\theta}$ distribution can be seriously biased as estimates of the θ distribution. (The failure of this approach was verified in early analyses of the 1984 NAEP reading survey; see Wingersky, Kaplan, & Beaton, 1987.) Plausible values were developed as a way to estimate key population features consistently, and appro: imate others no worse than standard IRT procedures would. A detailed development of plausible values methodology is given in Mislevy (1991). Along with theoretical justifications, that paper presents comparisons with standard procedures, discussions of biases that arise in some secondary analyses, and numerical examples.

The following provides a brief overview of the plausible values approach, focusing on its implementation in the Trial State Assessment analyses.

Let $\underline{y}$ represent the responses of all sampled examinees to background and attitude questions, along with design variables such as school membership, and let $\underline{\theta}$ represent the vector of scale proficiency values. If $\underline{\theta}$ were known for all sampled examinees, it would be possible to compute a statistic $t(\underline{\theta},\underline{y})$—such as a scale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient—to estimate a corresponding population quantity T. A function $U(\underline{\theta},\underline{y})$—e.g., a jackknife estimate—would be used to gauge sampling uncertainty, as the variance of t around T in repeated samples from the population.

Because the scaling models are latent variable models, however, $\underline{\theta}$ values are not observed even for sampled students. To overcome this problem, we follow Rubin (1987) by considering $\underline{\theta}$ as "missing data" and approximate $t(\underline{\theta},\underline{y})$ by its expectation given $(\underline{x},\underline{y})$, the data that actually were observed, as follows:

$$\begin{aligned} t^{*}(\underline{x},\underline{y}) &= E[t(\underline{\theta},\underline{y}) \mid \underline{x},\underline{y}] \\ &= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta} \mid \underline{x},\underline{y})\, d\underline{\theta} \ . \end{aligned} \tag{8.6}$$

It is possible to approximate t^{*} using random draws from the conditional distribution of the scale proficiencies given the item responses x_i, background variables y_i, and model parameters for sampled student i. These values are referred to as imputations in the sampling literature, and plausible values in NAEP. The value of $\underline{\theta}$ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from the respondent's conditional distribution. Rubin (1987) proposes that this process be carried out several times—multiple imputations—so that the uncertainty associated with imputation can be quantified. The average of the results of, for example, M estimates of t, each computed from a different set of plausible values, is a Monte Carlo approximation of (8.6); the variance among them, B, reflects uncertainty due to not observing θ, and must be added to the estimated expectation of $U(\underline{\theta},\underline{y})$, which reflects uncertainty due to testing only a sample of students from the population. Section 8.5 explains how plausible values are used in subsequent analyses.

It cannot be emphasized too strongly that **plausible values are *not* test scores for *individuals*** in the usual sense. Plausible values are offered only as intermediary computations for calculating integrals of the form of equation 8.6, in order to estimate *population* characteristics. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. The key idea lies in a contrast between plausible values and the more familiar θ estimates of educational measurement that are in some sense optimal for each examinee (e.g., maximum likelihood estimates, which are consistent estimates of an examinee's θ, and Bayes estimates, which provide minimum mean-squared errors with respect to a reference population): *Point estimates that are* *optimal for individual examinees have distributions that can produce decidedly nonoptimal (specifically, inconsistent) estimates of population characteristics* (Little & Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects. For further discussion see Mislevy, Beaton, Kaplan, and Sheehan (1992).

8.3.3 Computing Plausible Values in IRT-based Scales

Plausible values for each respondent *i* are drawn from the conditional distribution $p(\underline{\theta}_i|x_i,y_i,\Gamma,\Sigma)$, where Γ and Σ are regression model parameters defined in this subsection. This subsection describes how, in IRT-based scales, these conditional distributions are characterized, and how the draws are taken. An application of Bayes' theorem with the IRT assumption of conditional independence produces

$$p(\underline{\theta}_i|x_i,y_i,\Gamma,\Sigma) \propto P(x_i|\underline{\theta}_i,y_i,\Gamma,\Sigma)\, p(\underline{\theta}_i|y_i,\Gamma,\Sigma) = P(x_i|\underline{\theta}_i)\, p(\underline{\theta}_i|y_i,\Gamma,\Sigma) \quad , \tag{8.7}$$

where, for vector-valued $\underline{\theta}_i$, $P(x_i|\underline{\theta}_i)$ is the product over scales of the *independent likelihoods* induced by responses to items within each scale, and $p(\underline{\theta}_i|y_i,\Gamma,\Sigma)$ is the multivariate—and generally nonindependent—*joint density* of proficiencies for the scales, conditional on the observed value y_i of background responses, and the parameters Γ and Σ. The scales are determined by the item parameter estimates that constrain the population mean to zero and standard deviation to one. The item parameter estimates are fixed and regarded as population values in the computation described in this subsection.

In the analyses of the data from the Trial State Assessment and the data from the national reading assessment, a normal (Gaussian) form was assumed for $p(\underline{\theta}_i|y_i,\Gamma,\Sigma)$, with a common variance-covariance matrix, Σ, and with a mean given by a linear model with slope parameters, Γ, based on the first 134 to 200 principal components of 482 selected main effects and two-way interactions of the complete ector of background variables. The included principal components will be referred to as the *conditioning variables*, and will be denoted y^c. (The complete set of original background variables used in the Trial State Assessment reading analyses are listed in Appendix C.) The following model was fit to the data within each state:

$$\underline{\theta} = \Gamma' y^c + \underline{\varepsilon} \, , \tag{8.8}$$

where $\underline{\varepsilon}$ is multivariately normally distributed with mean zero and variance-covariance matrix Σ. The number of principal components of the conditioning variables used for each state was sufficient to account for 90 percent of the total variance of the full set of conditioning variables (after standardizing each variable). As in regression analysis, Γ is a matrix each of whose columns is the *effects* for one scale and Σ is the matrix *variance-covariance of residuals* between scales. By fitting the model (8.8) separately within each state, interactions between each state and the conditioning variables are automatically included in the conditional joint density of scale proficiencies.

Maximum likelihood estimates of Γ and Σ, denoted by $\hat{\Gamma}$ and $\hat{\Sigma}$, are obtained fror Sheehan's (1985) MGROUP computer program using the EM algorithm described in Mislevy

(1985). The EM algorithm requires the computation of the mean, $\bar{\underline{\theta}}_i$, and variance, Σ_i^p, of the posterior distribution in (8.7). These moments are computed using higher order asymptotic corrections (Thomas, 1992).

After completion of the EM algorithm, the plausible values are drawn in a three-step process from the joint distribution of the values of Γ for all sampled respondents. First, a value of Γ is drawn from a normal approximation to $P(\Gamma,\Sigma|x_i,y_i)$ that fixes Σ at the value $\hat{\Sigma}$, (Thomas, 1992). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean, $\bar{\underline{\theta}}_i$, and variance, Σ_i^p, of the posterior distribution in equation 8.7 (i.e., $p(\underline{\theta}_i|x_i,y_i,\Gamma,\Sigma)$) are computed using the same methods applied in the EM algorithm. In the third step, the $\underline{\theta}_i$ are drawn independently from a multivariate normal distribution with mean $\bar{\underline{\theta}}_i$ and variance Σ_i^p, approximating the distribution in (8.7). These three steps are repeated five times producing five imputations of $\bar{\underline{\theta}}_i$ for each sampled respondent.

8.4 NAGB ACHIEVEMENT LEVELS

Since its beginning, a goal of NAEP has been to inform the public about what students in American schools know and can do. While the NAEP scales provide information about the distributions of proficiency for the various subpopulations, they do not directly provide information about the meaning of various points on the scale. Traditionally, meaning has been attached to educational scales by norm-referencing—that is, by comparing students at a particular scale level to other students. Beginning in 1990, NAEP reports have also presented data using achievement levels. The reading achievement levels were developed and adopted by the National Assessment Governing Board (NAGB), as authorized by the NAEP legislation. The achievement levels describe selected points on the scale in terms of the types of skills that are or should be exhibited by students scoring at that level. The achievement level process was applied to the 1992 national NAEP reading composite and the 1994 national scales were linked to the 1992 national scales. Since the Trial State Assessment scales were linked to the national scales in both years, the interpretations of the selected levels also apply to the Trial State Assessment in 1994.

NAGB has determined that achievement levels shall be the first and primary way of reporting NAEP results. Setting achievement levels is a method for setting standards on the NAEP assessment that identify what students *should* know and be able to do at various points on the reading composite. For each grade in the national assessment and, here, for grade 4 in the Trial State Assessment, four levels were defined—*basic, proficient, advanced*, and the region *below basic*. Based on initial policy definitions of these levels, panelists were asked to determine operational descriptions of the levels appropriate with the content and skills assessed in the reading assessment. With these descriptions in mind, the panelists were then asked to rate the assessment items in terms of the expected performance of marginally acceptable examinees at each of these levels. These ratings were then mapped onto the NAEP scale to obtain the achievement level cutpoints for reporting. Further details of the achievement level-setting process appear in Appendix F.

8.5 ANALYSES

When survey variables are observed without error from every respondent, standard variance estimators quantify the uncertainty associated with sample statistics from the only source of uncertainty, namely the sampling of respondents. Item-level statistics for NAEP cognitive items meet this requirement, but scale-score proficiency values do not. The IRT models used in their construction posit an unobservable proficiency variable θ to summarize performance on the items in the subarea. The fact that θ values are not observed even for the respondents in the sample requires additional statistical analyses to draw inferences about θ distributions and to quantify the uncertainty associated with those inferences. As described above, Rubin's (1987) multiple imputations procedures were adapted to the context of latent variable models to produce the plausible values upon which many analyses of the data from the Trial State Assessment were based. This section describes how plausible values were employed in subsequent analyses to yield inferences about population and subpopulation distributions of proficiencies.

8.5.1 Computational Procedures

Even though one does not observe the θ value of respondent i, one does observe variables that are related to it: x_i, the respondent's answers to the cognitive items he or she was administered in the area of interest, and y_i, the respondent's answers to demographic and background variables. Suppose one wishes to draw inferences about a number $T(\underline{\theta},\underline{Y})$ that could be calculated explicitly if the θ and y values of each member of the population were known. Suppose further that if θ values were observable, we would be able to estimate T from a sample of N pairs of θ and y values by the statistic $t(\underline{\theta},\underline{y})$ [where $(\underline{\theta},\underline{y}) = (\theta_1,y_1,\ldots,\theta_N,y_N)$], and that we could estimate the variance in t around T due to sampling respondents by the function $U(\underline{\theta},\underline{y})$. Given that observations consist of (x,y_i) rather than $(\underline{\theta},y_i)$, we can approximate t by its expected value conditional on $(\underline{x},\underline{y})$, or

$$t^*(\underline{x},\underline{y}) = E[t(\underline{\theta},\underline{y})|\underline{x},\underline{y}] = \int t(\underline{\theta},\underline{y})\, p(\underline{\theta}|\underline{x},\underline{y})\, d\underline{\theta} .$$

It is possible to approximate t^* with random draws from the conditional distributions $p(\underline{\theta}_i|x_i,y_i)$, which are obtained for all respondents by the method described in section 8.3.3. Let $\underline{\hat{\theta}}_m$ be the mth such vector of plausible values, consisting of a multidimensional value for the latent variable of each respondent. This vector is a plausible representation of what the true $\underline{\theta}$ vector might have been, had we been able to observe it.

The following steps describe how an estimate of a scalar statistic $t(\underline{\theta},\underline{y})$ and its sampling variance can be obtained from M (> 1) such sets of plausible values. (Five sets of plausible values are used in NAEP analyses of the Trial State Assessment.)

1) Using each set of plausible values $\underline{\hat{\theta}}_m$ in turn, evaluate t as if the plausible values were true values of $\underline{\theta}$. Denote the results $\hat{t}_m$, for $m = 1,\ldots,M$.

2) Using the jackknife variance estimator defined in Chapter 7, compute the estimated sampling variance of $\hat{t}_m$, denoting the result U_m.

3) The final estimate of t is

$$t^* = \sum_{m=1}^{M} \frac{\hat{t}_m}{M}.$$

4) Compute the average sampling variance over the M sets of plausible values, to approximate uncertainty due to sampling respondents:

$$U^* = \sum_{m=1}^{M} \frac{U_m}{M}.$$

5) Compute the variance among the M estimates $\hat{t}_m$, to approximate uncertainty due to not observing θ values from respondents:

$$B = \sum_{m=1}^{M} \frac{(\hat{t}_m - t^*)^2}{(M - 1)}$$

6) The final estimate of the variance of t^* is the sum of two components:

$$V = U^* + (1 + M^{-1})\ B.$$

Note: Due to the excessive computation that would be required, NAEP analyses did not compute and average jackknife variances over all five sets of plausible values, but only on the first set. Thus, in NAEP reports, U^* is approximated by U_1.

8.5.2 Statistical Tests

Suppose that if θ values were observed for sampled students, the statistic $(t - T)/U^{1/2}$ would follow a t-distribution with d degrees of freedom. Then the incomplete-data statistic $(t^* - T)/V^{1/2}$ is approximately t-distributed, with degrees of freedom given by

$$\nu = \frac{1}{\frac{f^2}{M - 1} + \frac{(1 - f)^2}{d}}$$

where f is the proportion of total variance due to not observing θ values:

$$f_M = (1 + M^{-1})\ B_M / V_M.$$

When B is small relative to U^*, the reference distribution for incomplete-data statistics differs little from the reference distribution for the corresponding complete-data statistics. This is the case with main NAEP reporting variables. If, in addition, d is large, the normal approximation can be used to flag "significant" results.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U_m and U^* is a covariance matrix, and B is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T-t^*)\ V^{-1}\ (T-t^*)'$ is approximately F distributed, with degrees of freedom equal to k and ν, with ν defined as above but wi.h a matrix generalization of f:

$$f = (1+M^{-1})\ Trace\ (BV^{-1})/k\ .$$

By the same reasoning as used for the normal approximation for scalar t, a chi-square distribution on k degrees of freedom often suffices.

8.5.3 Biases in Secondary Analyses

Statistics t^* that involve proficiencies in a scaled content area and variables included in the conditioning variables y^c are consistent estimates of the corresponding population values T. Statistics involving background variables y that were *not* conditioned on, or relationships among proficiencies from *different* content areas, are subject to asymptotic biases whose magnitudes depend on the type of statistic and the strength of the relationships of the nonconditioned background variables to the variables that were conditioned on and to the proficiency of interest. That is, the large sample expectations of certain sample statistics need not equal the true population parameters.

The *direction* of the bias is typically to underestimate the effect of nonconditioned variables. For details and derivations see Beaton and Johnson (1990), Mislevy (1991), and Mislevy and Sheehan (1987, section 10.3.5). For a given statistic t^* involving one content area and one or more nonconditioned background variables, the *magnitude* of the bias is related to the extent to which observed responses x account for the latent variable θ, and the degree to which the nonconditioned background variables are explained by condit.oning background variables. The first factor—conceptually related to test reliability—acts consistently in that greater measurement precision reduces biases in *all* secondary analyses. The second factor acts to reduce biases in certain analyses but increase it in others. In particular,

- High shared variance between conditioned and nonconditioned background variables *mitigates* biases in analyses that involve only proficiency and nonconditioned variables, such as marginal means or regressions.

- High shared variance *exacerbates* biases in regression coefficients of conditional effects for nonconditioned variables, when nonconditioned and conditioned background variables are analyzed jointly as in multiple regression.

The large number of background variables that have been included in the conditioning vector for the Trial State Assessment allows a large number of secondary analyses to be carried out with little or no bias, and mitigates biases in analyses of the marginal distributions of θ in nonconditioned variables. Kaplan and Nelson's analysis of the 1988 NAEP reading data (some results of which are summarized in Mislevy, 1991), which had a similar design and fewer conditioning variables, indicates that the potential bias for nonconditioned variables in multiple regression analyses is below 10 percent, and biases in simple regression of such variables is below 5 percent. Additional research (summarized in Mislevy, 1990) indicates that most of the bias reduction obtainable from conditioning on a large number of variables can be captured by instead conditioning on the first several principal components of the matrix of all original conditioning variables. This procedure was adopted for the Trial State Assessment by replacing the conditioning effects by the first *K* principal components, where *K* was selected so that 90 percent of the total variance of the full set of conditioning variables (after standardization) was captured. Mislevy (1990) shows that this puts an upper bound of 10 percent on the average bias for all analyses involving the original conditioning variables.

Chapter 9

DATA ANALYSIS AND SCALING FOR THE 1994 TRIAL STATE ASSESSMENT IN READING[1]

Nancy L. Allen, John Mazzeo, Eddie H. S. Ip,
Spencer Swinton, Steven P. Isham, and Lois H. Worthington

Educational Testing Service

9.1 OVERVIEW

This chapter describes the analyses carried out in the development of the 1994 Trial State Assessment reading scales. The procedures used were similar to those employed in the analysis of the 1992 Trial State Assessment in reading (Allen, Mazzeo, Isham, Fong, & Bowker, 1994), and the 1990 and 1992 Trial State Assessments in mathematics (Mazzeo, 1991 and Mazzeo, Chang, Kulick, Fong, & Grima, 1993) and are based on the philosophical and theoretical underpinnings described in the previous chapter.

There were five major steps in the analysis of the Trial State Assessment reading data, each of which is described in a separate section:

- conventional item and test analyses (section 9.3);
- item response theory (IRT) scaling (section 9.4);
- estimation of state and subgroup proficiency distributions based on the "plausible values" methodology (section 9.5);
- linking of the 1994 Trial State Assessment scales to the corresponding scales from the 1994 national assessment (section 9.6); and
- creation of the Trial State Assessment reading composite scale (section 9.7).

To set the context within which to describe the methods and results of scaling procedures, a brief review of the assessment instruments and administration procedures is provided.

[1] Thanks to James Carlson, Huahua Chang, John Donoghue, David Freund, Frank Jenkins, Laura Jerry, Eugene Johnson, Ed Kulick, Jo-lin Liang, Eiji Muraki, Jennifer Nelson, and Neal Thomas for their help in completing the analysis. Thanks also to Angela Grima for her contributions to the original draft of this chapter

9.2 ASSESSMENT INSTRUMENTS AND SCORING

9.2.1 Items, Booklets, and Administration

The 1994 Trial State Assessment in reading was administered to fourth-grade public- and nonpublic-school students. The items in the instruments were based on the curriculum framework described in Chapter 2.

The fourth-grade item pool contained 84 items. They were categorized into one of two content areas: 43 were Reading for Literary Experience items and 41 were Reading to Gain Information items. These items, 39 of which were multiple-choice items, 37 of which were short constructed-response items, and 8 of which were extended constructed-response items, were divided into 8 mutually exclusive blocks. The composition of each block of items, in terms of content and format, is given in Table 9-1. Note that each block contained items from only one of the two content domains.

The 8 blocks were used to form 16 different booklets according to a partially balanced incomplete block (PBIB) design (see Chapter 2 for details). Each of these booklets contained two blocks of items, and each block of items appeared in exactly four booklets. To balance possible block position effect, each block appeared twice as the first block of reading items and twice as the second block. In addition, the design required that each block of items be paired in a booklet with every other block of items in the same content domain exactly once. Finally, each block of items was included in a booklet with a block of items from the other area.

Within each administration site, all booklets were "spiraled" together in a random sequence and distributed to students sequentially, in the order of the students' names on the Student Listing Form (see Chapter 4). As a result of the partial BIB design and the spiraling of booklets, a considerable degree of balance was achieved in the data collection process. Each block of items (and, therefore, each item) was administered to randomly equivalent samples of students of approximately equal size (i.e., about 4/16 or 1/4 of the total sample size) within each jurisdiction and across all jurisdictions. In addition, within and across jurisdictions, randomly equivalent samples of approximately equal size received each particular block of items as the first or second block within a booklet.

As described in Chapter 4, a randomly selected half of the administration sessions within each jurisdiction that had never participated in a Trial State assessment before were observed by Westat-trained quality control monitors. A randomly selected fourth of the administration sessions within each jurisdiction that had participated in previous Trial State assessments were observed by quality control monitors. Thus, within and across jurisdictions, randomly equivalent samples of students received each block of items in a particular position within a booklet under monitored and unmonitored administration conditions.

9.2.2 Scoring the Constructed-response Items

As indicated earlier, the reading assessment included constructed-response items (details of the professional scoring process are given in Chapter 5). Response to these items were included in the scaling process.

Table 9-1

1994 NAEP Reading Block Composition by Scale and Item Type
for Grade 4*

	Reading for Literary Experience				Reading to Gain Information				Total			
Block	Multiple Choice	Short Constructed Response	Extended Constructed Response	Total	Multiple Choice	Short Constructed Response	Extended Constructed Response	Total	Multiple Choice	Short Constructed Response	Extended Constructed Response	Total
R3	6	4	1	11	0	0	0	0	6	4	1	11
R4	5	6	1**	12	0	0	0	0	5	6	1	12
R5	7	3	1	11	0	0	0	0	7	3	1	11
R6	0	0	0	0	5	4	1	10	5	4	1	10
R7	0	0	0	0	4	5	1	10	4	5	1	10
R8	0	0	0	0	3	5	1**	9	3	5	1	9
R9	3	5	1	9	0	0	0	0	3	5	1	9
R10	0	0	0	0	6	5	1***	12	6	5	1	12
Total	21	18	4	43	18	19	4	41	39	37	8	84

* At grade 4, each block contained one reading passage.

** Two categories of response for this item were collapsed during the scaling process.

*** This item appears in the final position in the block.

Some of the constructed-response items were scored on a scale from 0 to 2 due to the short length of the responses expected. Other constructed-response items with short responses were scored on a scale from 0 to 3. One item per block was an extended constructed-response item. Each extended constructed-response item required about five minutes to complete and was scored by specially trained readers on a 0-to-4 scale. During the scaling process, the 0 (off-task) category was treated as "not administered" for each of the items so that the scaling model used for these items fit the data more closely. The remaining categories (1 to 4, 1 to 3, or 1 to 2) were transformed by subtracting 1; therefore, the categories used in the scaling model were 0 to 3 for the extended constructed-response items and either 0 to 2 or 0 to 1 for the short constructed-response items. (The categories of two of the extended constructed-response items were also collapsed.) The extended constructed-response items appeared in varying positions within each block. These items, including the recoding of the 0-to-4 scale, are described in more detail in section 9.4.1.

Table 5-5 in Chapter 5 provides the ranges for percent agreement between raters for the items as they were originally scored. Tables 9-2 and 9-3 present reliability data for items as they were used in scaling. The information in the tables includes, for each subject area and age/grade, the NAEP item numbers for each of the constructed-response items included in scaling, and the block that contains the item. The tables also indicate the codes from the NAEP database that denote the range of responses and the correct responses. A portion of the responses to the constructed-response items were scored twice for the purpose of examining rater reliability. For each item, the number of papers with responses that were scored a second time is listed, along with the percent agreement between raters and an index of reliability based on those responses. Cohen's Kappa (Cohen, 1968) is the reliability estimate used for the dichotomized short constructed-response items in Table 9-2. For the regular and extended constructed-response items, which were scored in either 3 or 4 categories, the intraclass correlation coefficient is used in Table 9-3 as the index of reliability.

9.2.3 Instrument Validity Evidence

Initial content validity evidence is provided by the consensus process used to formulate the framework and specifications for NAEP assessments. Broad-based committees are also involved in writing, selecting, and editing items for assessments. Further content validity evidence for the NAEP Trial State Assessment is provided by the National Academy of Education (1993b). Information about the validity of constructed-response items, as opposed to multiple-choice items, is available in Brennan (in press). Validity studies of NAEP are an ongoing interest of NCES.

9.3 ITEM ANALYSES

9.3.1 Conventional Item and Test Analyses

Tables 9-4 and 9-5 contain summary statistics for each block of items for public- and nonpublic-school sessions, respectively. Block-level statistics are provided both overall and by serial position of the block within booklet. To produce these tables, data from all 44 jurisdictions were aggregated and statistics were calculated using rescaled versions of the final

Table 9-2
Score Range, Percent Agreement, and Cohen's Kappa*
for the Short Constructed-response Reading Items Used in Scaling
Grade 4 Trial State Assessment

Item	Block	Range of Response Codes	Correct Response Codes	Sample Size	Percent Agreement	Cohen's Kappa
R012002	RC	1 - 2	2	7495	94.57	0.89
R012004	RC	1 - 2	2	7424	91.33	0.84
R012008	RC	1 - 2	2	6674	93.08	0.86
R012010	RC	1 - 2	2	6242	90.82	0.80
R012102	RD	1 - 2	2	7524	94.80	0.90
R012104	RD	1 - 2	2	7473	93.35	0.87
R012106	RD	1 - 2	2	7345	91.70	0.85
R012108	RD	1 - 2	2	7063	95.77	0.89
R012109	RD	1 - 2	2	6839	95.25	0.86
R012112	RD	1 - 2	2	5191	92.26	0.80
R012201	RF	1 - 2	2	7529	93.45	0.86
R012206	RF	1 - 2	2	6744	96.22	0.93
R012208	RF	1 - 2	2	6222	93.03	0.86
R012210	RF	1 - 2	2	5685	92.31	0.76
R012503	RJ	1 - 2	2	7507	90.17	0.81
R012504	RJ	1 - 2	2	7446	95.90	0.93
R012506	RJ	1 - 2	2	7281	92.39	0.86
R012508	RJ	1 - 2	2	6993	96.37	0.93
R012511	RJ	1 - 2	2	6389	94.79	0.88
R012601	RE	1 - 2	2	7458	90.48	0.79
R012604	RE	1 - 2	2	7291	95.23	0.89
R012611	RE	1 - 2	2	5874	94.26	0.89
R012702	RG	1 - 2	2	7469	94.43	0.85
R012703	RG	1 - 2	2	7404	91.82	0.85
R012705	RG	1 - 2	2	7051	94.43	0.88
R012706	RG	1 - 2	2	6828	90.95	0.80
R012710	RG	1 - 2	2	4954	93.74	0.88
R015802	RI	1 - 2	2	7380	91.37	0.80

* Cohen's Kappa is a measure of reliability that is appropriate for items that are dichotomized.

Table 9-3
Score Range, Percent Agreement, and Intraclass Correlation
for the Extended Constructed-response Reading Items Used in Scaling
Grade 4 Trial State Assessment

Item	Block	Range of Response Codes	Sample Size	Percent Agreement	Intraclass Correlation
R015702	RH	1 - 3	7837	85.76	0.83
R015703	RH	1 - 3	7540	88.36	0.84
R015704	RH	1 - 3	7453	84.17	0.87
R015705	RH	1 - 3	7271	90.07	0.94
R015709	RH	1 - 3	5760	90.63	0.91
R015803	RI	1 - 3	7391	84.10	0.79
R015806	RI	1 - 3	6583	81.45	0.81
R015807	RI	1 - 3	6222	78.54	0.83
R015809	RI	1 - 3	5466	78.01	0.73
R012006	RC	1 - 4	7059	83.61	0.92
R012111	RD	1 - 4	6174	89.75	0.94
R012204	RF	1 - 4	7422	78.70	0.89
R012512	RJ	1 - 4	6442	79.56	0.91
R012607	RE	1 - 4	6878	88.75	0.91
R012708	RG	1 - 4	6375	83.48	0.87
R015707	RH	1 - 4	2990	86.22	0.89
R015804	RI	1 - 4	7219	82.02	0.86

Table 9-4

Descriptive Statistics for Each Block of Items*
by Position Within Test Booklet and Overall
Public School

Statistic	Position	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted sample size	1 2 All	13928 13963 27891	13994 13992 27986	13809 13799 27608	13921 13926 27847	13933 13816 27749	14025 13958 27983	13945 13890 27835	13918 13923 27841
Average item score	1 2 All	.63 .61 .62	.66 .64 .65	.45 .43 .44	.58 .55 .57	.43 .41 .42	.62 .60 .61	.70 .67 .68	.66 .63 .65
Average r-polyserial	1 2 All	.71 .73 .72	.68 .71 .69	.61 .64 .63	.59 .62 .61	.68 .71 .69	.59 .62 .60	.55 .62 .58	.63 .66 .65
Proportion of students attempting last item	1 2 All	.71 .85 .79	.59 .74 .67	.72 .83 .78	.67 .83 .75	.54 .67 .61	.70 .82 .76	.63 .78 .71	.73 .81 .78

* The number and types of items contained in each block are shown in Table 9-1.

Table 9-5

Descriptive Statistics for Each Block of Items*
by Position Within Test Booklet and Overall
Nonpublic Schools

Statistic	Position	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted sample size	1	1124	1126	1141	1135	1136	1151	1136	1149
	2	1145	1155	1112	1150	1140	1142	1124	1122
	All	2269	2281	2253	2285	2276	2293	2260	2271
Average item score	1	.70	.72	.54	.66	.51	.68	.74	.72
	2	.69	.72	.51	.62	.51	.66	.73	.72
	All	.70	.72	.52	.64	.51	.67	.73	.72
Average r-polyserial	1	.70	.65	.62	.55	.66	.56	.54	.60
	2	.71	.69	.63	.61	.68	.59	.53	.64
	All	.70	.67	.62	.58	.67	.57	.53	.62
Proportion of students attempting last item	1	.80	.67	.78	.75	.62	.77	.71	.78
	2	.88	.81	.87	.88	.74	.87	.84	.88
	All	.85	.74	.83	.82	.68	.82	.78	.83

* The number and types of items contained in each block are shown in Table 9-1.

sampling weights provided by Westat. The rescaling, carried out within each jurisdiction, constrained the sum of the sampling weights within that jurisdiction to be equal to its sample size. The sample sizes for each jurisdiction were approximately equal. Use of the rescaled weights does nothing to alter the value of statistics calculated separately within each jurisdiction. However, for statistics obtained from samples that combine students from different jurisdictions, use of the rescaled weights results in a roughly equal contribution of each jurisdiction's data to the final value of the estimate. As discussed in Mazzeo (1991), equal contribution of each jurisdiction's data to the results of the IRT scaling was viewed as a desirable outcome and, as described in the scaling section below, these same rescaled weights were only adjusted slightly in carrying out that scaling. Hence, the item analysis statistics shown in Tables 9-4 and 9-5 are approximately consistent with the weighting used in scaling. The original final sampling weights provided by Westat were used in reporting.

Tables 9-4 and 9-5 show the number of students assigned each block of items, the average item score, the average polyserial correlation, and the proportion of students attempting the last item in the block. The average item score for the block is the average, over items, of the score means for each of the individual items in the block. For binary-scored multiple-choice and constructed-response items, these score means correspond to the proportion of students who correctly answered each item. For the other constructed-response items, the score means were calculated as item score mean divided by the maximum number of points possible.

In NAEP analyses (both conventional and IRT-based), a distinction is made between missing responses at the end of each block (i.e., missing responses subsequent to the last item the student answered) and missing responses prior to the last observed response. Missing responses before the last observed response were considered intentional omissions. Intentional omissions were treated as incorrect responses. When the last item in the block was a multiple-choice or short constructed-response item, missing responses at the end of the block were considered "not reached." When the last item in the block was an extended constructed-response item, missing responses at the end of the block were considered "not reached" if the responses to the next-to-last item were missing and were treated as if they had not been presented to the student. In calculating the average score for each item, only students classified as having been presented the item were included in the denominator of the statistic. The proportion of students attempting the last item of a block (or, equivalently, 1 minus the proportion of students not reaching the last item) is often used as an index of the degree of speededness associated with the administration of that block of items.

Standard practice at ETS is to treat all nonrespondents to the last item as if they had not reached the item. For multiple-choice and short constructed-response items, the use of such a convention most often produces a reasonable pattern of results in that the proportion reaching the last item is not dramatically smaller than the proportion reaching the next-to-last item. However, for the blocks that ended with extended constructed-response items, use of the standard ETS convention resulted in an extremely large drop in the proportion of students attempting the final item. A drop of such magnitude seemed somewhat implausible. Therefore, for blocks ending with an extended constructed-response items, students who answered the next-to-last item but did not respond to the extended constructed-response item were classified as having intentionally omitted the last item.

The average polyserial correlation is the average, over items, of the item-level polyserial correlations (r-polyserial). For each item-level r-polyserial, the total block number-correct score (including the item in question, and with students receiving zero points for all not-reached items) was used as the criterion variable for the correlation. For dichotomous items, the item-level r-polyserial correlations are standard r-biserial correlations. Data from students classified as not reaching the item were omitted from the calculation of the statistic.

As is evident from Tables 9-4 and 9-5, the difficulty and the internal consistency of the blocks varied somewhat. Such variability was expected since these blocks were not created to be parallel in either difficulty or content. Based on the proportion of students attempting the last item, all of the blocks seem to be somewhat speeded. Only 67 percent of the public-school students receiving block R4 and 61 percent of the public-school students receiving block R7 reached the last item in the block. The proportion of nonpublic-school students reaching the last item in blocks were generally higher. For example, 74 percent receiving block R4 and 68 percent receiving block R7 reached the last item in the block.

This table also indicates that there was little variability in average item scores or average polyserial correlations for each block by serial position within the assessment booklet. The differences in item statistics were small for items appearing in blocks in the first position and in the second position. However, differences were consistent in their direction. Average item scores were highest when each block was presented in the first position. Average polyserial correlations were highest when each block was presented in the second position. An aspect of block-level performance that did differ noticeably by serial position was the proportion of students attempting the last item in the block. As shown in Tables 9-4 and 9-5, the percentage of the students attempting the last item increased as the serial position of the block increased. Students may have learned to pace themselves through the later block after they had experienced the format of the first block they received. This was similar to what occurred in 1992. For the 1992 Trial State Assessment, a study was completed to examine the effect of the serial position differences on scaling. Due to the partial BIB design of the booklets, those effects were minimal.

As mentioned earlier, in an attempt to maintain rigorous standardized administration procedures across the jurisdictions, a randomly selected 50 percent of all sessions within each jurisdiction that had never participated in a Trial State Assessment was observed by a Westat-trained quality control monitor. A randomly selected 25 percent of the sessions within other jurisdictions were monitored. Observations from the monitored sessions provided information about the quality of administration procedures and the frequency of departures from standardized procedures in the monitored sessions (see Chapter 4, section 4.3.6, for a discussion of the substance of these observations.)

When public-school results were aggregated over all participating jurisdictions, there was little difference between the performance of students who attended monitored or unmonitored sessions. The average item score (over all 8 blocks and over all 44 participating jurisdictions) was .59 for both monitored and unmonitored public-school sessions. The average item score was .66 for monitored nonpublic-school sessions and .67 for unmonitored nonpublic-school sessions. Table 9-6 provides, for each block of items, the average item score, average r-polyserial, and the proportion of students attempting the last item for public-school students

Table 9-6
Block-level* Descriptive Statistics for Unmonitored and Monitored Public-school Sessions

Statistic	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted sample size								
Unmonitored	20377	20464	20216	20371	20309	20475	20369	20345
Monitored	7515	7523	7392	7476	7440	7508	7466	7496
Average item score								
Unmonitored	.62	.65	.44	.56	.42	.61	.68	.65
Monitored	.62	.65	.44	.57	.42	.61	.69	.65
Average r-polyserial								
Unmonitored	.72	.70	.62	.61	.69	.60	.58	.64
Monitored	.72	.69	.63	.61	.69	.61	.59	.65
Proportion of students attempting last item								
Unmonitored	.79	.66	.77	.75	.58	.76	.70	.77
Monitored	.79	.68	.78	.75	.58	.75	.70	.77

* The number and types of items contained in each block are shown in Table 9-1.

Table 9-7
Block-level* Descriptive Statistics for Monitored and Unmonitored Nonpublic-school Sessions

Statistic	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted sample size								
Unmonitored	1109	1123	1091	1111	1104	1111	1093	1089
Monitored	1160	1158	1162	1174	1172	1182	1167	1182
Average item score								
Unmonitored	.70	.73	.52	.65	.53	.67	.74	.73
Monitored	.69	.71	.52	.63	.49	.67	.73	.71
Average r-polyserial								
Unmonitored	.68	.66	.63	.58	.66	.56	.53	.61
Monitored	.72	.68	.62	.59	.68	.58	.54	.63
Proportion of students attempting last item								
Unmonitored	.85	.75	.82	.82	.67	.82	.76	.89
Monitored	.85	.73	.84	.82	.67	.82	.79	.89

* The number and types of items contained in each block are shown in Table 9-1

whose sessions were monitored and public-school students whose sessions were not monitored. A similar table for nonpublic-school students is provided in Table 9-7.

Figure 9-1 presents stem-and-leaf displays of the differences between unmonitored and monitored average item scores (over all eight blocks) on each of the two purpose-of-reading scales for each of the 44 jurisdictions participating in the public-school portion of the 1994 Trial State Assessment. Figure 9-2 presents similar displays of the differences between unmonitored and monitored item scores for each of the 34 jurisdictions participating in the nonpublic-school portion. Stem-and-leaf displays, developed by Tukey (1977), are similar to histograms. The combination of a stem with each of its leaves gives the actual value of one observation (i.e., the difference in average item scores for unmonitored and monitored sessions in a participating jurisdiction).

For public-school sessions, the median differences (unmonitored minus monitored) were -.0007 and .0011, respectively, for the Reading for Literary Experience scale and the Reading to Gain Information scale. In evaluating the magnitude of these differences, it should be noted that the standard error for a difference in proportions from independent simple random samples of size 1,250 (half the typical total state public-school sample size of 2,500) from a population with a true proportion of .5 is about .02. For samples with complex sampling designs like NAEP, the standard errors tend to be larger than those associated with simple random sampling. A design effect gives an indication of how much larger the standard errors are for a complex sample, rather than a random sample. A conservative estimate of the design effect for proportion-correct statistics based on past NAEP experience is about 2.0 (Johnson & Rust, 1992, Johnson, Rust & Wallace, 1994), which suggests that a typical estimate of the standard error of the difference between unmonitored and monitored sessions would be about .028 if 50 percent of the sessions were monitored for each jurisdiction. On the Reading for Literary Experience scale the absolute differences in item score means for 32 of the 44 participating jurisdictions were less than .02 in magnitude, and all but six were less than .028. The largest difference was positive, with a value of .057. The largest negative difference was -.051. On the Reading to Gain Information scale, the absolute differences in item score means for 33 of the 44 participants were less than .02 in magnitude. The differences with the largest magnitudes were -.061, .047, and .040. In summary, differences in results obtained from the two types of public-school sessions at the fourth grade were within the bounds expected due to sampling fluctuation.

For nonpublic-school sessions, the median differences (unmonitored minus monitored) were .0068 and .0161, respectively, for the Reading for Literary Experience scale and the Reading to Gain Information scale. The sample sizes in nonpublic-school sessions are much smaller than those in public-school sessions. With a typical sample size of 250 nonpublic schools per jurisdiction, the standard error and the standard error adjusted for design effects were respectively .045 and .063. On the Literary Experience scale, the absolute differences in item score means for 23 of the 34 participating jurisdictions that have nonpublic-school sessions were less than .045 and all but four were less than .063 in magnitude. The differences with the largest magnitude were positive, with values of .089 and .086. On the Gain Information scale, the absolute differences in item score means for 22 jurisdictions were less than .045 and for all but seven were less than .063. The differences with the largest magnitude were -.113 , .107 and .105. Although most of the differences seem to fall within the bounds expected due to sampling fluctuation, several outlying values raise concern.

Figure 9-1
Stem-and-leaf Display of State-by-state Differences
in Average Item Scores by Scale in Public Schools
(Unmonitored Minus Monitored)*

READING FOR LITERARY EXPERIENCE

N = 44, Median = -0.001, Quartiles = -0.010, 0.011
Decimal point is 2 places to the left of the colon

```
-11 :
-10 :
 -9 :
 -8 :
 -7 :
 -6 :
 -5 : 1
 -4 :
 -3 : 22
 -2 : 7
 -1 : 8553221
 -0 : 987644432211
  0 : 233345568
  1 : 1148
  2 : 023559
  3 : 1
  4 :
  5 : 7
  6 :
  7 :
  8 :
  9 :
 10 :
```

*Note: The *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading* uses (Monitored Minus Unmonitored) as the variable of interest.

Figure 9-1 (continued)
Stem-and-leaf Display of State-by-state Differences
in Average Item Scores by Scale in Public Schools
(Unmonitored Minus Monitored)*

READING TO GAIN INFORMATION

N = 44, Median = 0.001, Quartiles = -0.011, 0.013
Decimal point is 2 places to the left of the colon

```
-11  :
-10  :
 -9  :
 -8  :
 -7  :
 -6  :  1
 -5  :
 -4  :  7
 -3  :
 -2  :  50
 -1  :  95422210
 -0  :  998665410
  0  :  02334457899
  1  :  144
  2  :  0003569
  3  :
  4  :  07
  5  :
  6  :
  7  :
  8  :
  9  :
 10  :
```

*Note: The *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading* uses (Monitored Minus Unmonitored) as the variable of interest.

Figure 9-2
Stem-and-leaf Display of State-by-state Differences
in Average Item Scores by Scale in Nonpublic Schools
(Unmonitored Minus Monitored)*

READING FOR LITERARY EXPERIENCE

N = 34, Median = 0.007, Quartiles = -0.027, 0.040
Decimal point is 2 places to the left of the colon

```
-11  :
-10  :
 -9  :
 -8  :  0
 -7  :
 -6  :  2
 -5  :
 -4  :  7521
 -3  :  3
 -2  :  87
 -1  :  660
 -0  :  4
  0  :  03559
  1  :  05
  2  :  7
  3  :  0146
  4  :  0589
  5  :  01
  6  :
  7  :  5
  8  :  69
  9  :
 10  :
```

*Note: The *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading* uses (Monitored Minus Unmonitored) as the variable of interest.

Figure 9-2 (continued)
Stem-and-leaf Display of State-by-state Differences
in Average Item Scores by Scale in Nonpublic Schools
(Unmonitored Minus Monitored)*

READING TO GAIN INFORMATION

N = 34, Median = 0.016, Quartiles = -0.026, 0.032
Decimal point is 2 places to the left of the colon

```
 -11  :  3
 -10  :
  -9  :
  -8  :
  -7  :
  -6  :  5
  -5  :  73
  -4  :  96
  -3  :  92
  -2  :  6
  -1  :  9
  -0  :  851
   0  :  3
   1  :  12678
   2  :  127788
   3  :  22
   4  :  39
   5  :
   6  :  6
   7  :  8
   8  :
   9  :  6
  10  :  57
```

*Note: The *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading* uses (Monitored Minus Unmonitored) as the variable of interest.

A separate study was conducted at ETS to examine the monitoring effects across jurisdictions in nonpublic-school sessions. A description of the analyses conducted in the study on monitoring effect in nonpublic-school sessions is given in Appendix G.

9.3.2 Differential Item Functioning (DIF) Analyses

Prior to scaling, differential item functioning (DIF) analyses were carried out on 1994 NAEP reading data from the national cross-sectional samples at grades 4, 8, and 12 and the Trial State Assessment sample at grade 4. The purpose of these analyses was to identify items that were differentially difficult for various subgroups and to reexamine such items with respect to their fairness and their appropriateness for inclusion in the scaling process. The information in this section focuses mainly on the analyses conducted on the Trial State Assessment data. A description of the results based on the national assessment appears in the technical report for that assessment.

The DIF analyses of the dichotomous items were based on the Mantel-Haenszel chi-square procedure, as adapted by Holland and Thayer (1988). The procedure tests the statistical hypothesis that the odds of correctly answering an item are the same for two groups of examinees that have been matched on some measure of proficiency (usually referred to as the matching criterion). The DIF analyses of the polytomous items were based on the Mantel procedure (1963) and the Somes (1986) chi-square test. These procedures compare proportions of matched examinees from each group in each polytomous item response category. The groups being compared are often referred to as the focal group (usually a minority or other group of interest, such as Black examinees or female examinees) and the reference group (usually White examinees or male examinees).

For both types of analyses, the measure of proficiency used is typically the total item score on some collection of items. Since, by the nature of the BIB design, booklets comprise different combinations of blocks, there is no single set of items common to all examinees. Therefore, for each student, the measure of proficiency used was the total item score on the entire booklet. These scores were then pooled across booklets for each analysis. Note that all items were analyzed simultaneously. This procedure is described by Allen and Donoghue (1994, in press).

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

In the past, NAEP DIF analyses of polytomous items were completed by dichotomizing the responses to each item. This procedure was used because of a lack of validated techniques designed explicitly and demonstrated to be appropriate for such items. Polytomous items are being developed for NAEP because of the potential gain in assessment validity over instruments consisting solely of dichotomous items. ETS scales these polytomous items with a polytomous IRT model. Hence it is necessary to incorporate special DIF procedures for the analysis of these items.

ETS staff members have studied DIF procedures appropriate for polytomous items. These procedures include an extension of the MH procedure that can be used for ordinally scored polytomous items (Mantel, 1963), procedures based on the generalized Mantel-Haenszel statistic (Somes, 1986), and extensions of IRT-model based procedures (Bock, Muraki, and Pfeiffenberger, 1988; Thissen, Steinberg, and Wainer, 1988).

The ETS/NAEP DIF procedure for polytomous items incorporates both the MH ordinal procedure and the generalized MH statistic. The summary tables of identified polytomous items contain generalizations of the dichotomous A, B, and C categories: AA, BB, and CC, and an additional flagging category, EE, based on the Somes chi-square test. An item is coded EE when it is not significant in mean conditional difference by the Mantel procedure, but shows significantly different conditional category counts using the Somes chi-square test. This may occur when category curves for different groups cross one another.

For each block of items at grade 4 a single set of analyses was carried out based on equal-sized random samples of data from all participating jurisdictions. Each set of analyses involved four reference group/focal group comparisons: male/female, White/Asian American, White/Black, and White/Hispanic.

All analyses used rescaled sampling weights. A separate rescaled weight was defined for each comparison as:

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

where the total sample size is the total number of students for the two groups being analyzed (e.g., for the White/Hispanic comparison, the total number of Wh.te and Hispanic examinees in the sample at that grade), and the sum of the weights is the sum of the sampling weights of all the students in the sample for the two groups being analyzed. Four rescaled weights were computed for White examinees—one for the gender comparison and three for the race/ethnicity comparisons. Two rescaled overall weights were computed for the Asian American, Black, and Hispanic examinees—one for the gender comparison and another for the appropriate race/ethnicity comparison. The rescaled weights were used to ensure that the sum of the weights for each analysis equaled the number of students in that comparison, thus providing an accurate basis for significance testing.

In the calculation of total item scores for the matching criterion, both not-reached and omitted items were considered to be wrong responses. Polytomous items were weighted more heavily in the formation of the matching criterion, proportional to the number of score categories. For each item, calculation of the Mantel-Haenszel statistic did not include data from examinees who did not reach the item in question.

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

At grade 4, 84 items were analyzed. Table 9-8 provides a summary of the results of the DIF analyses for the collection of 67 dichotomous items grouped by content area. Although the items are grouped by content area, the criteria for the analyses included items from all of the content areas. The table provides two sets of five frequency distributions for the categorized Mantel-Haenszel statistics for the items in each of the scales. The leftmost frequency distribution gives the number (and percent) of items in each of five categories (C+, B+, A, B-, C-) based on the largest absolute DIF value obtained for the item across the four reference group/focal group comparisons that were carried out. The remaining four frequency distributions give the number of items with indices in each DIF category for each of the four reference group/focal group comparisons.

No dichotomous items were classified as C items for any of the analyses for the fourth-grade Trial State Assessment data. Four items were classified as B items in tl.e White/Asian American comparisons. Two were differentially more difficult for the Asian American examinees than for the White examinees, both items measuring Reading for Literary Experience. The other two items were in the Reading to Gain Information scale and were differentially more difficult for White examinees than for Asian American examinees. Four items were categorized as B items in the White/Black comparisons. One, on the Reading for Literary Experience scale, was relatively more difficult for Blacks than for Whites, as were two items on the Reading to Gain Information scale. One item on this latter scale was differentially more difficult for Whites than for Blacks.

Table 9-9 provides a summary of the results of the DIF analyses for the collection of 17 polytomous items grouped by content area. The table is in a format similar to that of Table 9-8, showing items in five categories (CC+, EE+, AA, EE-, CC-). A six categor , BB, could have occurred, but did not, and is not tabulated.

No polytomous items were classified as CC items for any of the analyses for the fourth-grade Trial State Assessment data. Except for the case of the White/Black comparison on the Reading to Gain Information scale, at least one item was classified as EE in each of the analyses. The only item categorized as EE in the White/Black comparisons was on the Reading for Literary Experience scale; this item was relatively easier for Black students than for White students. Five polytomous items (29%) were classified as EE in the Male/Female comparisons; three of them were on the Reading to Gain Information scale, with two of the three favoring

Table 9-8

Frequency Distributions of DIF Statistics for Grade 4 Dichotomous Items Grouped by Content Area

Category of Maximum Absolute DIF Value For All Comparisons			Number of Items in Category of DIF Value for Each Comparison (Reference Group/Focal Group)			
DIF Category*	Number	Percent	Male/Female	White/Black	White/Hispanic	White/Asian Amer.
Reading for Literary Experience						
C+	0	0.0	0	0	0	0
B+	0	0.0	0	0	0	0
A	33	94.3	35	34	35	33
B-	2	5.7	0	1	0	2
C-	0	0.0	0	0	0	0
Reading to Gain Information						
C+	0	0.0	0	0	0	0
B+	2	6.3	0	1	0	2
A	28	87.5	31	29	32	30
B-	2	6.3	1	2	0	0
C-	0	0.0	0	0	0	0

* Categories are A, B, and C. (+) indicates items in the category that are differentially easier for the focal group; (-) indicates items in the category that are differentially more difficult for the focal group. DIF categories are described on page 185.

Table 9-9

Frequency Distributions of DIF Statistics for Grade 4 Polytomous Items Grouped by Content Area

Category of Maximum Absolute DIF Value For All Comparisons			Number of Items in Category of DIF Value for Each Comparison (Reference Group/Focal Group)			
DIF Category*	Number	Percent	Male/Female	White/Black	White/Hispanic	White/Asian Amer.
Reading for Literary Experience						
CC+	0	0.0	0	0	0	0
EE+	3	37.5	1	1	1	1
AA	2	25.0	6	5	6	6
EE-	3	37.5	1	2	1	1
CC-	0	0.0	0	0	0	0
Reading to Gain Information						
CC+	0	0.0	0	0	0	0
EE+	3	33.3	2	0	1	1
AA	4	44.4	6	9	7	8
EE-	2	22.2	1	0	1	0
CC-	0	0.0	0	0	0	0

* Categories are AA, BB, CC, and EE. (+) indicates items in the category that are differentially easier for the focal group; (-) indicates items in the category that are differentially more difficult for the focal group. DIF categories are described on page 185.

females. In the White/Hispanic comparisons, two items on each scale were classified as EE. These were balanced in differential difficulty. The White/Asian American comparisons yielded three EE items. Two of these, one from each scale, were differentially easier for the Asian American examinees than for the White examinees.

Following standard practice at ETS for DIF analyses conducted on final test forms, all C, CC, and EE items were reviewed by a committee of trained test developers and subject-matter specialists. Such committees are charged with making judgments about whether or not the differential difficulty of an item is *unfairly* related to group membership. As pointed out by Zieky (1993):

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

The committee assembled to review NAEP items included both ETS staff and outside members with expertise in the field. It was the committee's judgment, based on a substantive review of the items identified by the statistical analyses, that none of the C, CC, or EE items for the national or Trial State Assessment data were functioning differentially due to factors irrelevant to test objectives. Hence, none of the items were removed from the scales due to differential item functioning.

9.4 ITEM RESPONSE THEORY (IRT) SCALING

Separate IRT-based scales were developed using the scaling models described in Chapter 8. Two scales were produced by separately calibrating the sets of items classified in each of the two content areas.

Figures 9-3 and 9-4 contain stem-and-leaf displays of the average scores for the items comprising each of the fourth-grade scales for public and nonpublic schools. The averages are based on the entire sample of students in the Trial State Assessment and use the same rescaled sampling weights described in section 9.3. As a whole, the fourth-grade students in the samples found the set of items in the Reading to Gain Information scale to be the most difficult.

For the reasons discussed in Mazzeo (1991), for each scale, a single set of item parameters for each item was estimated and used for all jurisdictions. Item parameter

Figure 9-3
Stem-and-leaf Display of Average Item Scores for Public-school Sessions

READING FOR LITERARY EXPERIENCE

N = 44, Median = 0.614, Quartiles = 0.589, 0.640
Decimal point is 2 places to the left of the colon

```
42 :
43 :
44 :
45 :
46 :
47 :
48 : 29
49 :
50 :
51 :
52 :
53 : 5
54 :
55 : 37
56 : 58
57 : 6
58 : 1899
59 : 02
60 : 0255778
61 : 3468
62 : 224
63 : 45599
64 : 0489
65 : 1468
66 : 12
67 : 3
68 :
69 :
70 :
71 :
72 :
```

Figure 9-3 (continued)
Stem-and-leaf Display of Average Item Scores for Public-school sessions

READING TO GAIN INFORMATION

N = 44, Median = 0.571, Quartiles = 0.547, 0.599
Decimal point is 2 places to the left of the colon

```
42  :  8
43  :  2
44  :
45  :
46  :
47  :
48  :
49  :
50  :  1
51  :  3
52  :  2778
53  :  7
54  :  11
55  :  2334
56  :  245779
57  :  012347
58  :
59  :  013589
60  :  055
61  :  0246
62  :  136
63  :  7
64  :
65  :
66  :
67  :
68  :
69  :
70  :
71  :
72  :
```

Figure 9-4
Stem-and-leaf Display of Average Item Scores for Nonpublic-school Sessions

READING FOR LITERARY EXPERIENCE

N = 34, Median = 0.691, Quartiles = 0.674, 0.710
Decimal point is 2 places to the left of the colon

```
42  :
43  :
44  :
45  :
46  :
47  :
48  :
49  :
50  :  7
51  :
52  :
53  :
54  :
55  :
56  :
57  :
58  :
59  :
60  :
61  :  4
62  :  68
63  :
64  :
65  :  67
66  :  35
67  :  45
68  :  35689
69  :  00124467
70  :  669
71  :  3577
72  :  0579
```

Figure 9-4 (continued)
Stem-and-leaf Display of Average Item Scores for Nonpublic-school Sessions

READING TO GAIN INFORMATION

N = 34, Median = 0.653, Quartiles = 0.630, 0.667
Decimal point is 2 places to the left of the colon

```
42  :
43  :
44  :
45  :
46  :
47  :
48  :
49  :  2
50  :
51  :
52  :
53  :
54  :
55  :
56  :
57  :  2
58  :
59  :  5
60  :  388
61  :  5
62  :  1
63  :  0158
64  :  227
65  :  234
66  :  233455677
67  :  24
68  :  888
69  :
70  :  3
71  :
72  :  4
```

estimation was carried out using a 25 percent systematic random sample of the students participating in the 1994 Trial State Assessment and included equal numbers of students from each participating jurisdiction, half from monitored sessions and half from unmonitored sessions. All students in the scaling sample were public-school students. The sample consisted of 28,072 students, with 638 students being sampled from each of the 44 participating jurisdictions. Of the 638 records sampled from each jurisdiction, 319 were drawn from the monitored sessions and 319 were drawn from the unmonitored sessions. The rescaled weights for the 25 percent sample of students used in item calibration were adjusted slightly to ensure that 1) each jurisdiction's data contributed equally to the estimation process, and 2) data from monitored and unmonitored sessions contributed equally. For each jurisdiction, the sum of the rescaled sampling weights for the set of monitored and unmonitored students selected for the sample was obtained (these sums are denoted as WM_s and WU_s, respectively). Then, for each jurisdiction, the rescaled weights for individuals in the sample (denoted as W_{si}) were adjusted so that the sum of the weights for the monitored and unmonitored sessions would each be equal to 319. Thus for the monitored students in the sample,

$$W_{si}^{*} = W_{si}(319/WM_s),$$

and for the unmonitored students

$$W_{si}^{*} = W_{si}(319/WU_s),$$

where

$$W_{si}^{*}$$

denotes the adjusted rescaled weight for individual *i* from jurisdiction *s*. These adjusted rescaled weights for the 25 percent sample of students were used only in item calibration.

As mentioned above, the sample used for item calibration was also constrained to contain an equal number of students from the monitored and unmonitored sessions from each of the participating jurisdictions. To the extent that items may have functioned differently in monitored and unmonitored sessions, the single set of item parameter estimates obtained define a sort of average item characteristic curve for the two types of sessions. Tables 9-6 and 9-7 (shown earlier) presented block-level item statistics that suggested little, if any, differences in item functioning by session type for the public-school samples.

9.4.1 Item Parameter Estimation

For each content area scale, item parameter estimates were obtained using the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs[2]. The program uses marginal maximum likelihood estimation procedures to estimate the parameters of the one-, two-, and three-parameter logistic models, and the generalized partial credit model described by Muraki (1992).

Multiple-choice items were dichotomously scored and were scaled using the three-parameter logistic model. Omitted responses to multiple-choice items were treated as fractionally correct, with the fraction being set to 1 over the number of response options. Short constructed-response items that were also in the 1992 assessment were dichotomously scored and scaled using the two-parameter logistic model. New short (regular) constructed-response items were scored on a three-point generalized partial credit scale. These items appear in blocks 8 and 9. Omitted responses to short constructed-response items were treated as incorrect.

There were a total of eight extended constructed-response items. Each of these items was also scaled using the generalized partial credit model. Four scoring levels were defined:

0 Unsatisfactory response or omitted;
1 Partial response;
2 Essential response; and
3 Extensive response.

Note that omitted responses were treated as the lowest possible score level. As stated earlier, not-reached and off-task responses were treated as if the item was not administered to the student. Table 9-10 provides a listing of the blocks, positions within the block, content area classifications, and NAEP identification numbers for all extended constructed-response items included in the 1994 assessment.

Table 9-10
Extended Constructed-response Items, 1994 Trial State Assessment in Reading

Block	Position In Block	Scale	NAEP ID
R3	6	Literary Experience	R012006
R4	11	Literary Experience	R012111
R5	7	Literary Experience	R012607
R6	4	Gain Information	R012204
R7	8	Gain Information	R012708
R8	4	Gain Information	R015804
R9	7	Literary Experience	R015707
R10	12	Gain Information	R012512

[2] Late in the analysis process, an error was discovered in the PARSCALE program documentation. This error affected the reading results, including those reported in the April 1995 version of the *First Look* report. The analyses and report were subsequently redone. Appendix H describes the error, its correction, and the revised results.

Bayes modal estimates of all item parameters were obtained from the BILOG/PARSCALE program. Prior distributions were imposed on item parameters with the following starting values: thresholds, normal [0,2]; slopes, log-normal [0,.5]; and asymptotes, two-parameter beta with parameter values determined as functions of the number of response options for an item and a weight factor of 50. The locations (but not the dispersions) were updated at each program estimation cycle in accordance with provisional estimates of the item parameters.

As was done for the 1990 and 1992 Trial State Assessments in mathematics and for the 1992 Trial State Assessment in reading, item parameter estimation proceeded in two phases. First, the subject ability distribution was assumed fixed (normal [0,1]) and a stable solution was obtained. Starting values for the item parameters were provided by item analysis routines. The parameter estimates from this initial solution were then used as starting values for a subsequent set of runs in which the subject ability distribution was freed and estimated concurrently with item parameter estimates. After each estimation cycle, the subject ability distribution was re-standardized to have a mean of zero and standard deviation of one. Item parameter estimates for that cycle were correspondingly linearly transformed.

During and subsequent to item parameter estimation, evaluations of the fit of the IRT models were carried out for each of the items in the grade 4 item pools. These evaluations were conducted to identify misfitting items, which would be excluded from the final item pool making up the scales. Evaluations of model fit were based primarily on a graphical analysis. For binary-scored items, model fit was evaluated by examining plots of nonmodel-based estimates of the expected conditional (on θ) proportion correct versus the proportion correct predicted by the estimated item characteristic curve (see Mislevy & Sheehan, 1987, p. 302). For the extended constructed-response items, similar plots were produced for each item category characteristic curve.

As with most procedures that involve evaluating plots of data versus model predictions, a certain degree of subjectivity is involved in determining the degree of fit necessary to justify use of the model. There are a number of reasons why evaluation of model fit relied primarily on analyses of plots rather than seemingly more objective procedures based on goodness-of-fit indices such as the "pseudo chi-squares" produced in BILOG (Mislevy & Bock, 1982). First, the exact sampling distributions of these indices when the model fits are not well understood, even for fairly long tests. Mislevy and Stocking (1987) point out that the usefulness of these indices appears particularly limited in situations like NAEP where examinees have been administered relatively short tests. Work reported by Stone, Ankenmann, Lane, and Liu (1993) using simulated data suggests that the correct reference chi-square distributions for these indices have considerably fewer degrees of freedom than the value indicated by the BILOG/PARSCALE program and require additional adjustments of scale. However, it is not yet clear how to estimate the correct number of degrees of freedom and necessary scale factor adjustment factors. Consequently, pseudo chi-square goodness-of-fit indices are used only as rough guides in interpreting the severity of model departures.

Second, as discussed in Chapter 8, it is almost certainly the case that, for most items, item-response models hold only to a certain degree of approximation. Given the large sample sizes used in NAEP and the Trial State Assessment, there will be sets of items for which one is almost certain to reject the hypothesis that the model fits the data even though departures are minimal in nature or involve kinds of misfit unlikely to impact on important model-based inferences. In practice, one is almost always forced to temper statistical decisions with judgments about the severity of model misfit and the potential impact of such misfit on final results.

In making decisions about excluding items from the final scales, a balance was sought between being too stringent, hence deleting too many items and possibly damaging the content representativeness of the pool of scaled items, and too lenient, hence including items with model fit poor enough to invalidate the types of model-based inferences made from NAEP results. Items that clearly did not fit the model were not included in the final scales; however, a certain degree of misfit was tolerated for a number of items included in the final scales.

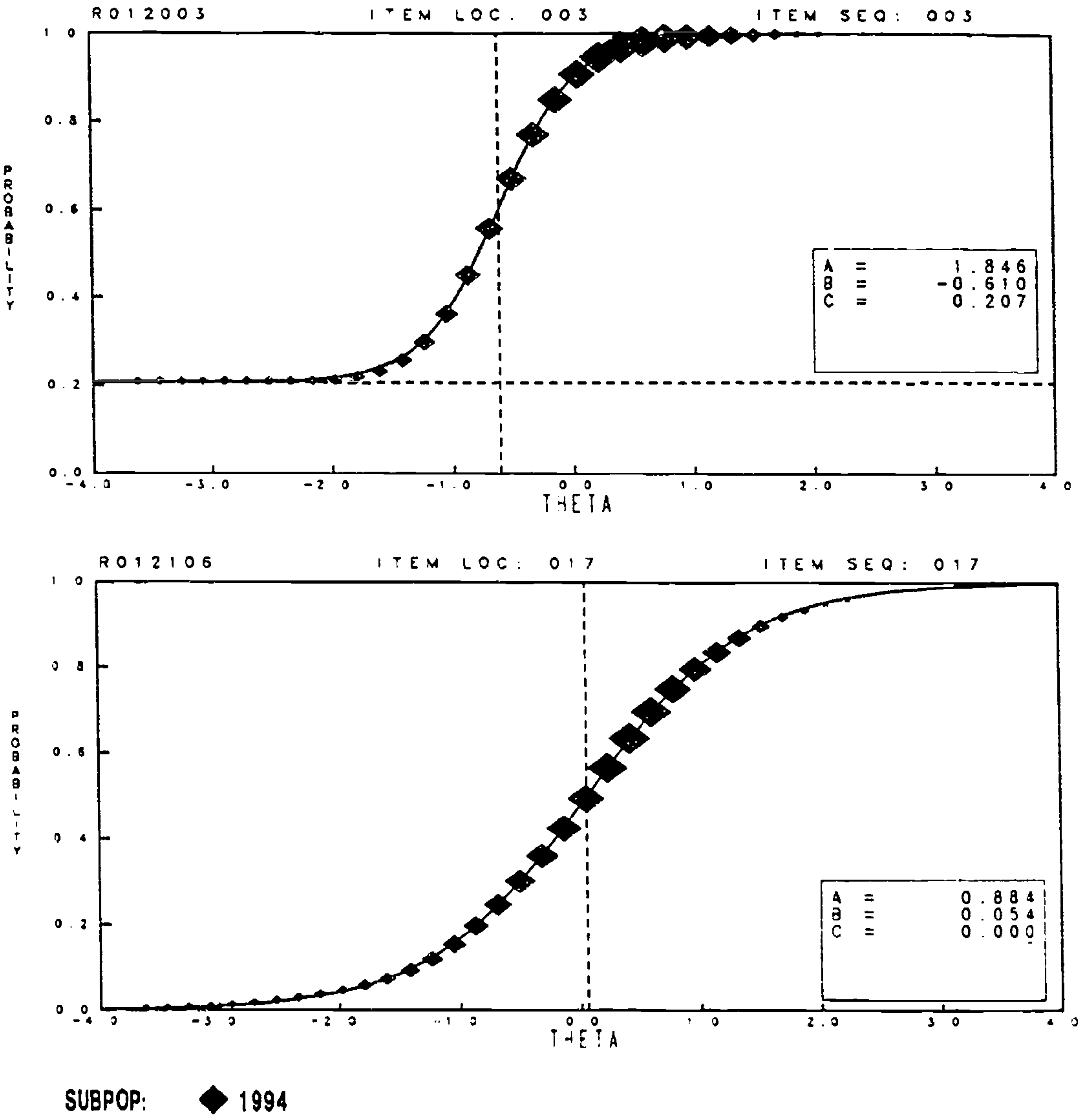
For the large majority of the grade 4 items, the fit of the model was extremely good. Figure 9-5 provides a typical example of what the plots look like for this class of items. The item at the top of the plot is a multiple-choice item; the item at the bottom of the plot is a binary-scored constructed-response item. In each plot, the y-axis indicates the probability of a correct response and the x-axis indicates proficiency level (theta). The diamonds show estimates of the conditional (on theta) probability of a correct response that do not assume a logistic form (referred to subsequently as nonlogistic-based estimates). The sizes of the diamonds are proportional to the number of students categorized as having thetas at or close to the indicated value. The solid curve shows the estimated item response function. The item response function provides estimates of the conditional probability of a correct response based on an assumed logistic form. The vertical dashed line indicates the estimated location parameter (b) for the item and the horizontal dashed line (top plot only) indicates the estimated lower asymptote (c). Also shown in the plot are the actual values of the item parameter estimates (lower right-hand corner) as well as the proportion of students that answered the item correctly (upper left-hand corner). As is evident from the plots, the nonlogistic-based estimates of conditional (diamonds) probabilities are in extremely close agreement with those given by the estimated item response function (the solid curves).

Figure 9-6 provides an example of a plot for a four-category extended constructed-response item exhibiting good model fit. Like the plots for the binary items, this plot shows two estimates of each item category characteristic curve, one set that does not assume the partial credit model (shown as diamonds) and one that does (the solid curves). The dashed vertical lines show the location of the estimated category thresholds for the item (d_1 to d_3; see Chapter 8, sections 8.3.1). The estimates for all parameters for the item in question are also indicated on the plot. As with Figure 9-5, the two sets of estimates agree quite well, although there are slight differences between the two. An aspect of Figure 9-6 worth noting is the large proportion of examinees that responded in the two lowest response categories for this item[3]. Although few student responses were categorized in the highest two categories, there were adequate data to estimate the model-based estimates for those categories (the solid curves). Such results were typical for the extended constructed-response items.

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

Figure 9-5

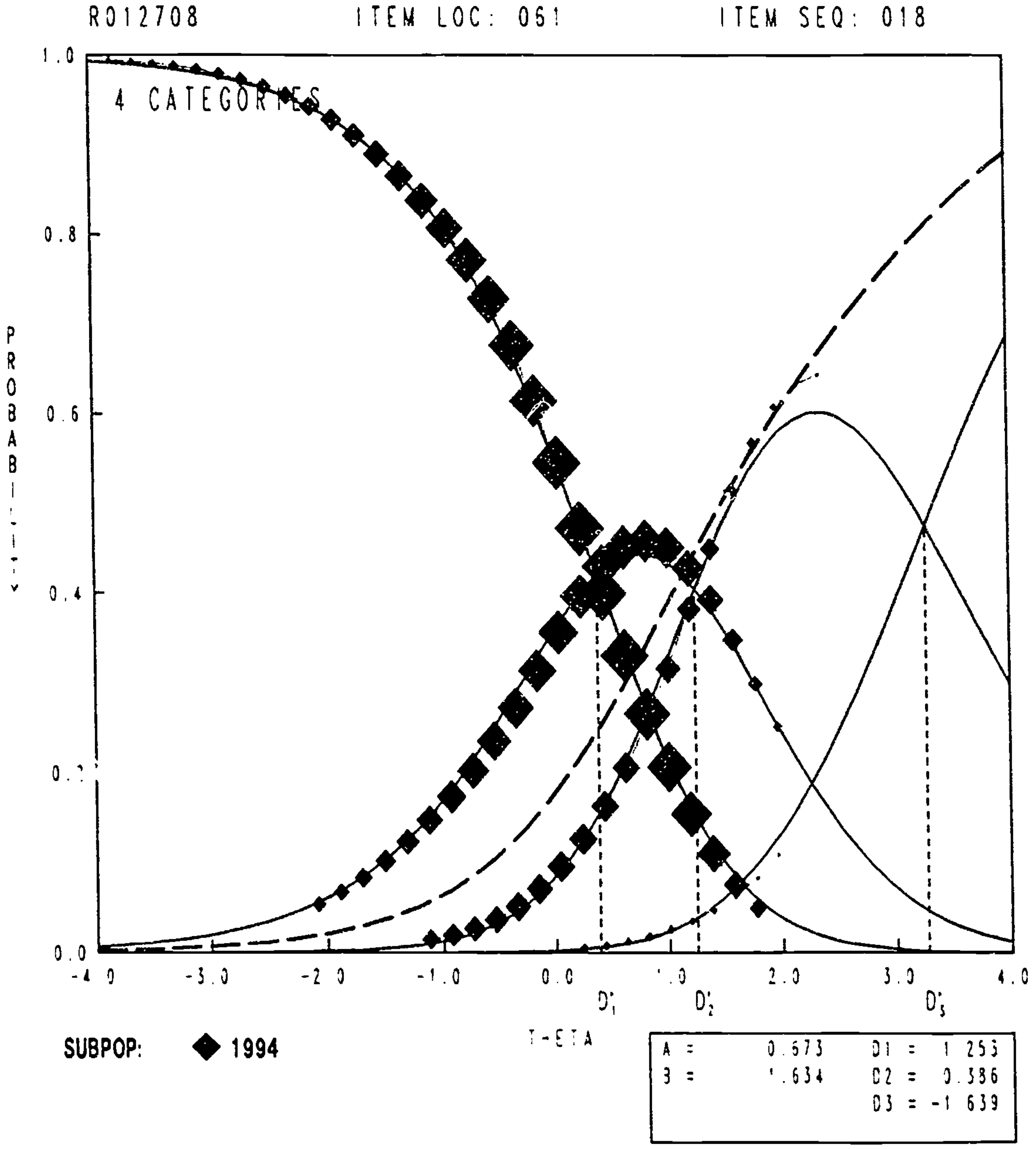
Plots* Comparing Empirical and Model-based Estimates of Item Response Functions for Binary-scored Items Exhibiting Good Model Fit



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

Figure 9-6

Plot* Comparing Empirical and Model-based Estimates of Item Category Characteristic Curves for a Polytomous Item Exhibiting Good Model Fit



*Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; the solid curves indicate estimated item response functions assuming a logistic form.

As discussed above, some of the items retained for the final scales display some degree of model misfit. Figures 9-7 (binary-scored items) and 9-8 (extended constructed-response item) provide typical examples of such items. In general, good agreement between nonlogistic and logistic estimates of conditional probabilities were found for the regions of the theta scale that includes most of the examinees. Misfit was confined to conditional probabilities associated with theta values in the tails of the subject ability distributions.

Only one item in the assessment received special treatment in the scaling process in both the 1992 and 1994 assessments. The generalized partial credit model did not fit the responses to the extended constructed-response item R012111 well. For this Reading for Literary Experience item, which appeared in the eleventh position in block R4, the categories 0 and 1 were combined and the other categories were relabeled. Therefore the codings for the three scoring levels were defined:

0 Unsatisfactory, partial response, or omitted;
- Essential response; and
2 Extensive response.

Plots for this item for the 1992 data are given ir Figures 9-9 and 9-10 before and after collapsing the unsatisfactory and partial response categories. The large differences between the estimates of the category characteristic curves when the partial credit model is assumed (shown as solid curves) and when the model is not assumed (shown as diamonds) indicate that the two lowest categories lack good model fit in Figure 9-9. In contrast, except for the tendency for the nonlogistic-based estimates to be somewhat different from the model-based estimates for theta values greater than 1, Figure 9-10 shows good model fit. Note that this item is functioning primarily as a dichotomous item due to the small frequencies in the top category. There were enough data, however, to calculate the model-based estimates of the category characteristic curve for this category (shown as the rightmost solid curve). Figure 9-11 is the plot for the 1994 data after collapsing the unsatisfactory and partial response categories.

In addition, one item that was administered only in 1994 received special treatment. As for item R012111, the general partial credit model did not fit the response to the extended constructed-response item R015707 well. This Reading for Information item was treated the same way as was item R012111. Plots for this item before and after collapsing the categories are displayed in Figure 9-12 and 9-13.

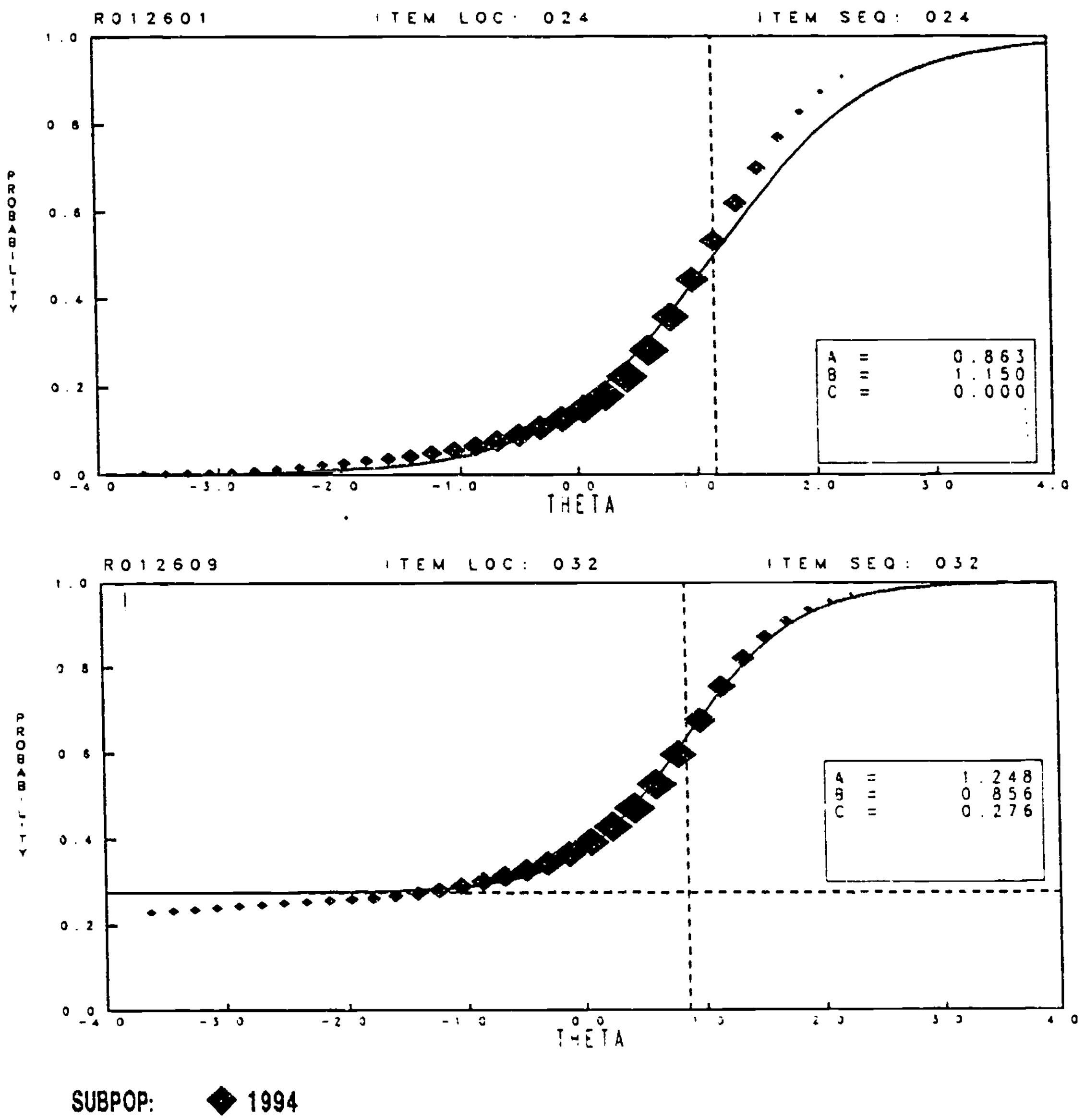
The IRT parameters for the items included in the Trial State Assessment are listed in Appendix D.

9.5 ESTIMATION OF STATE AND SUBGROUP PROFICIENCY DISTRIBUTIONS

The proficiency distributions in each jurisdiction (and for important subgroups within each jurisdiction) were estimated by using the multivariate plausible values methodology and the corresponding MGROUP computer program (described in Chapter 8; see also Mislevy, 1991). The MGROUP program (Sheehan, 1985; Rogers, 1991), which was originally based on the procedures described by Mislevy and Sheehan (1987), was used in the 1990 Trial State Assessment of mathematics. The 1992 and 1994 Trial State Assessments used an enhanced

Figure 9-7

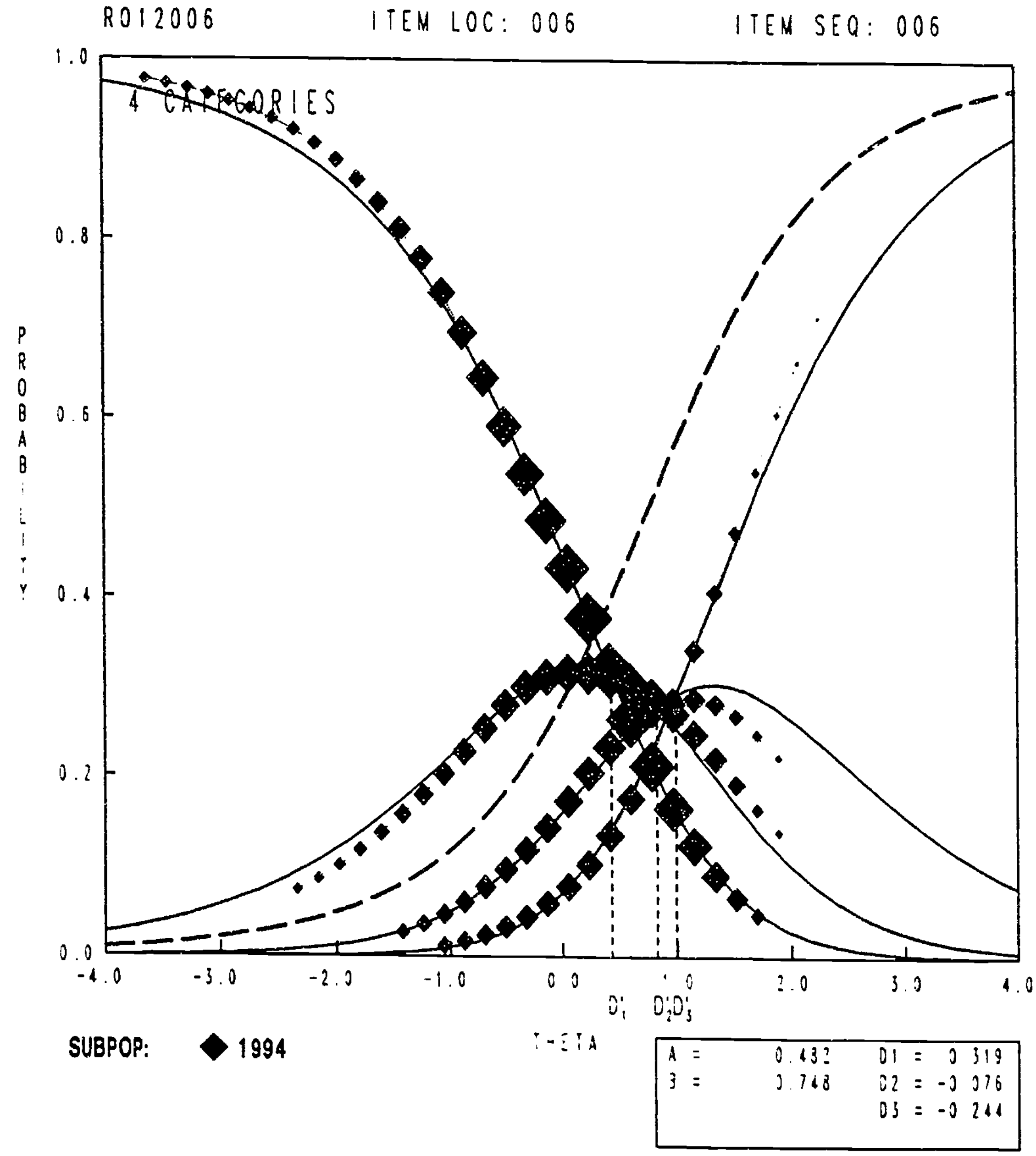
Plots* Comparing Empirical and Model-based Estimates of Item Response Functions for Binary-scored Items Exhibiting Some Model Misfit



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

Figure 9-8

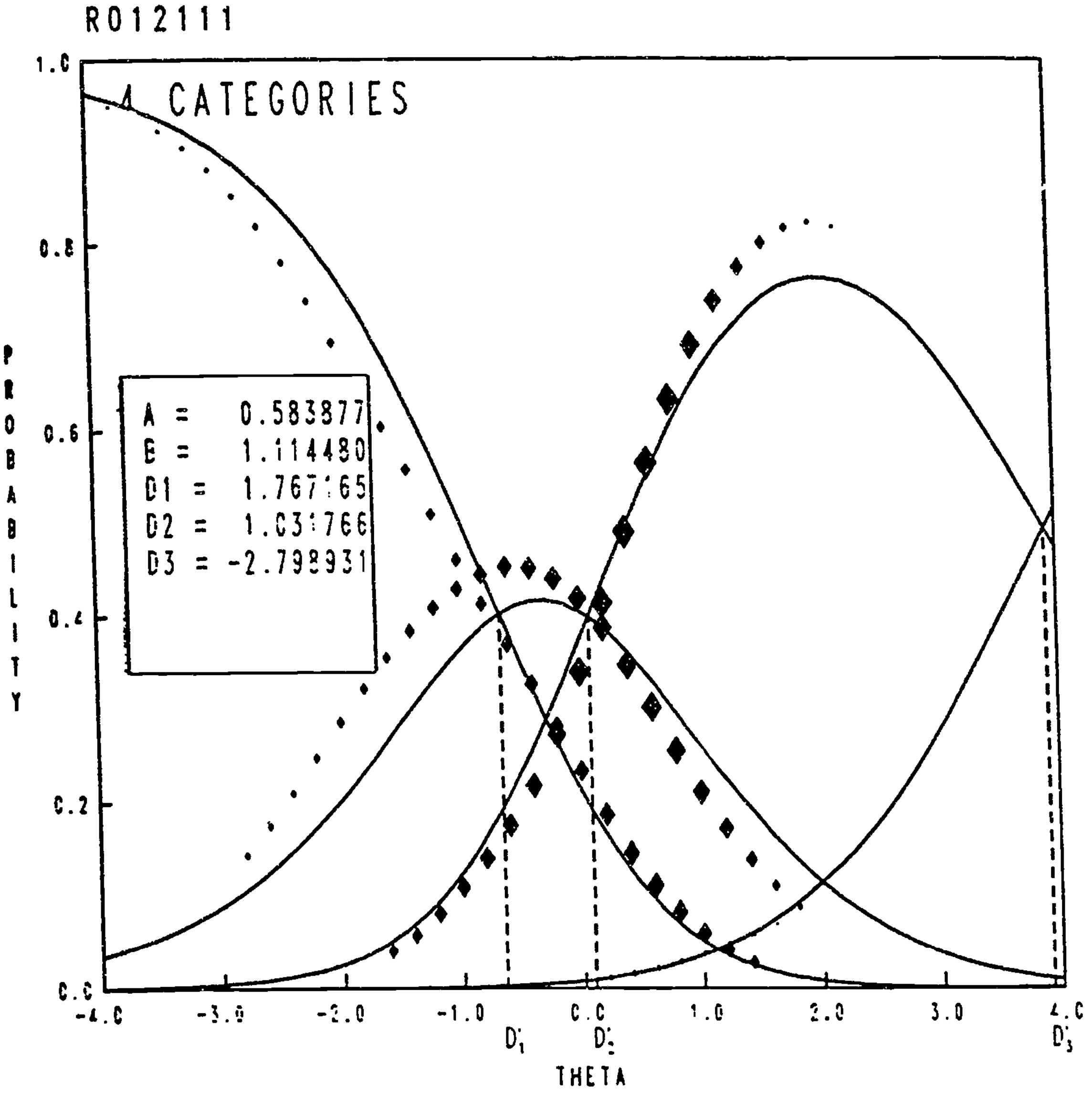
Plot* Comparing Empirical and Model-based Estimates of Item Category Characteristic Curves for a Polytomous Item Exhibiting Some Model Misfit



*Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; the solid curves indicate estimated item response functions assuming a logistic form.

Figure 9-9

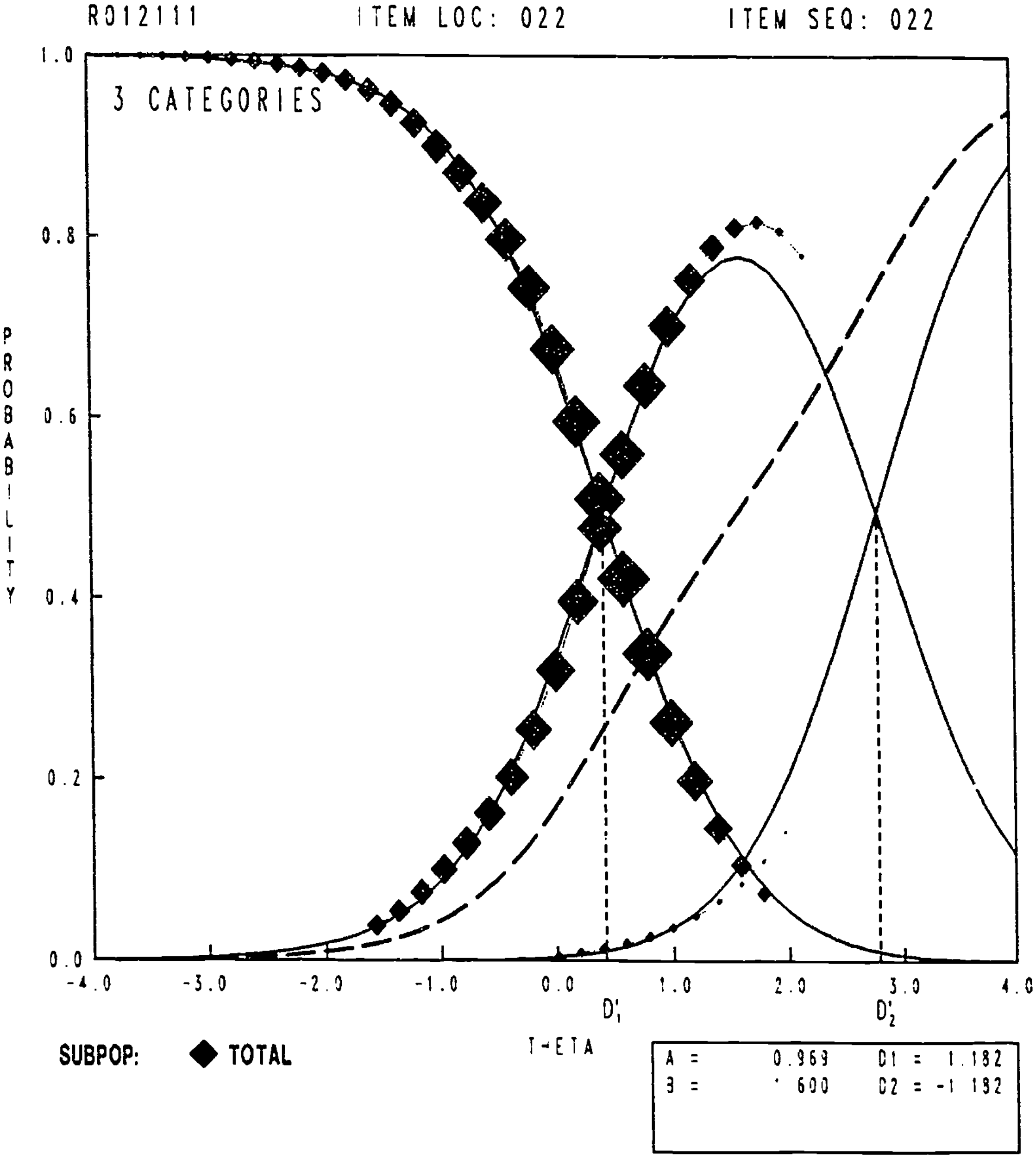
Plot* Comparing Empirical and Model-based Estimates of the Item Response Function for Item R012111 Using 1992 Assessment Data Before Collapsing Unsatisfactory and Partial Response Categories



* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; the solid curves indicate estimated item response functions assuming a logistic form.

Figure 9-10

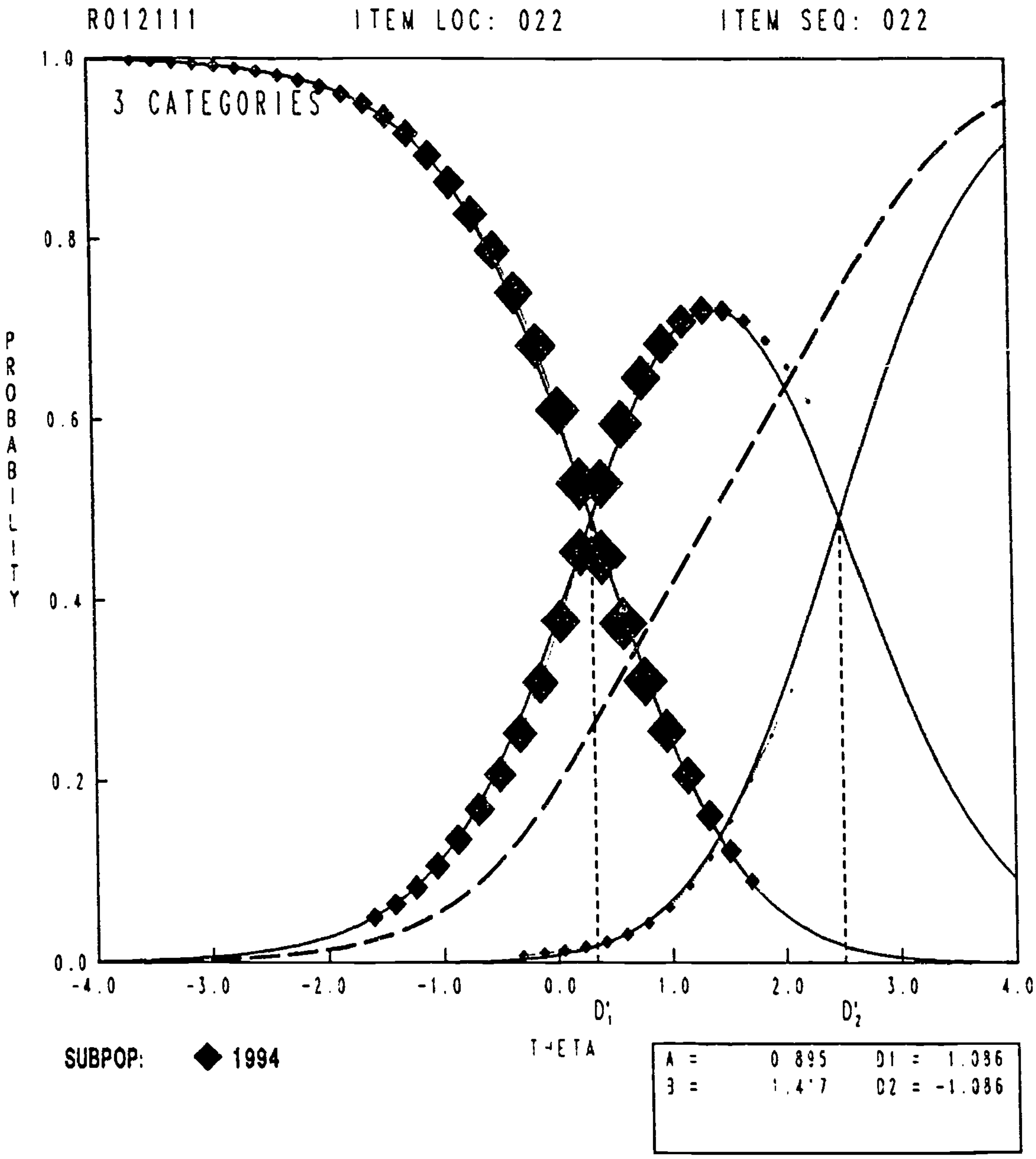
Plot* Comparing Empirical and Model-based Estimates of the Item Response Function for Item R012111 Using 1992 Assessment Data After Collapsing Unsatisfactory and Partial Response Categories



* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; the solid curves indicate estimated item response functions assuming a logistic form.

Figure 9-11

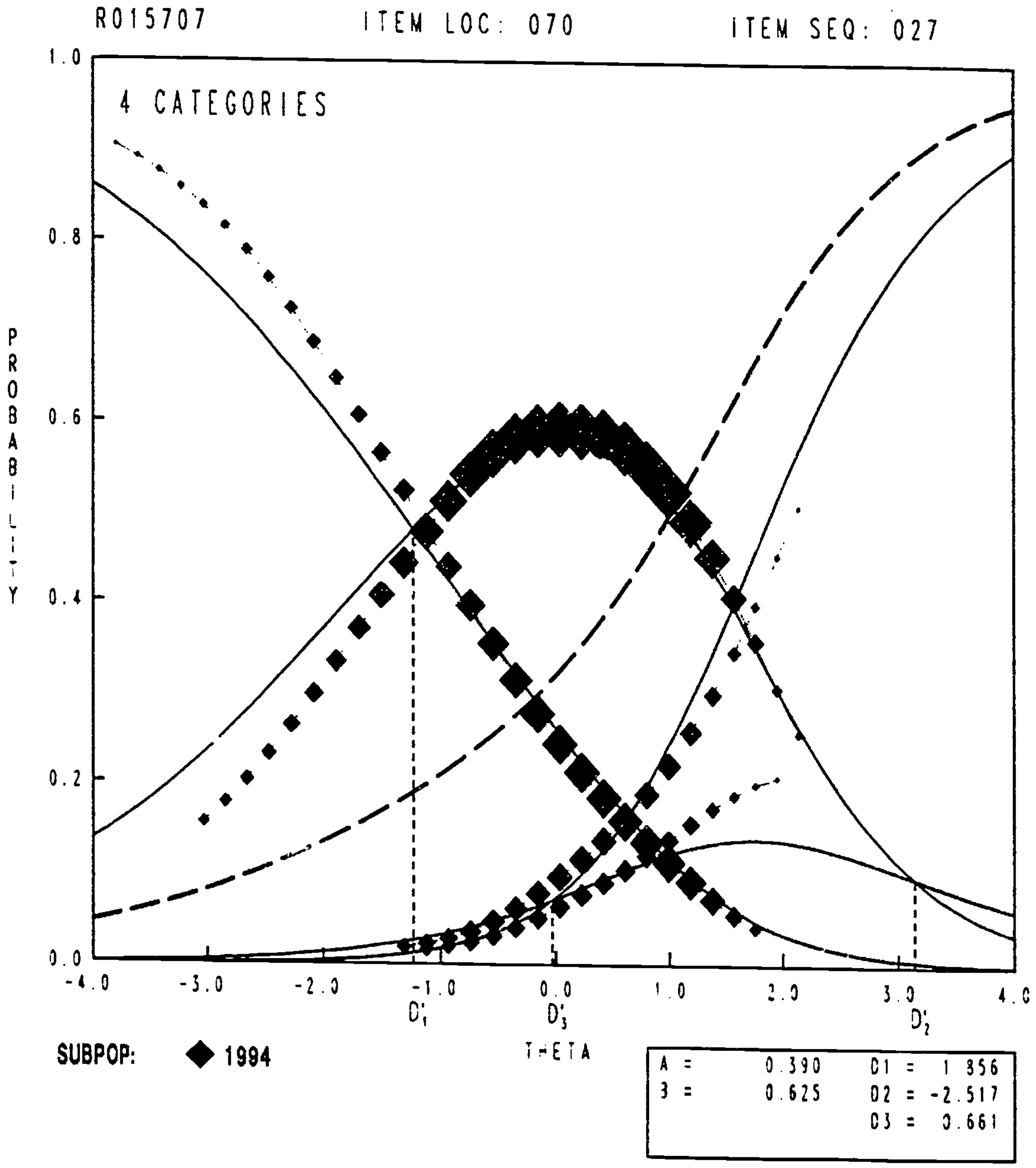
Plot* Comparing Empirical and Model-based Estimates of the Item Response Function for Item R012111 Using 1994 Assessment Data After Collapsing Unsatisfactory and Partial Response Categories



* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; the solid curves indicate estimated item response functions assuming a logistic form.

Figure 9-12

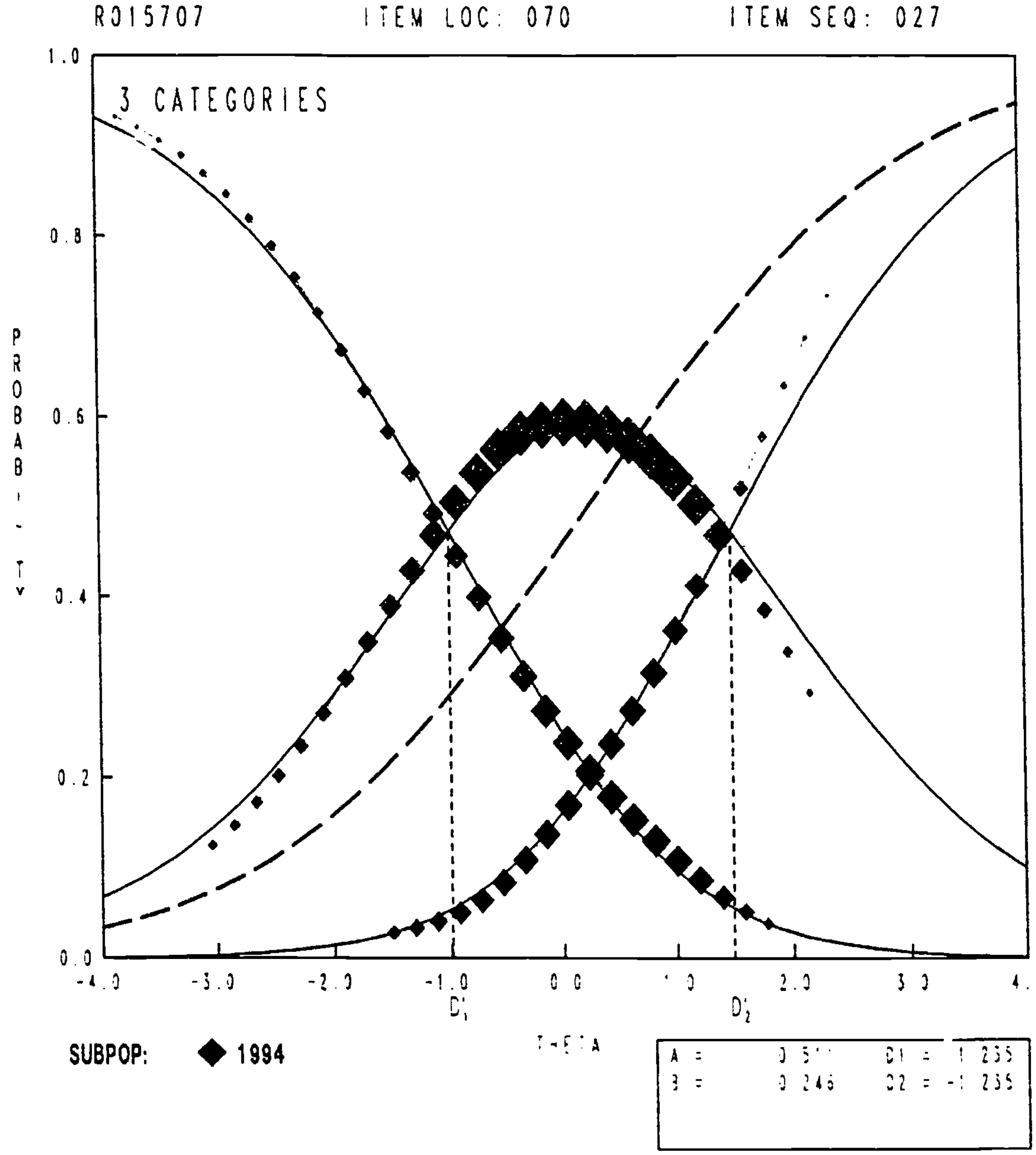
Plot* Comparing Empirical and Model-based Estimates of the Item Response Function for Item R015707 Using 1994 Assessment Data Before Collapsing Unsatisfactory and Partial Response Categories



* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; the solid curve indicate estimated item response functions assuming a logistic form.

Figure 9-13

Plot* Comparing Empirical and Model-based Estimates of the Item Response Function for Item R015707 Using 1994 Assessment Data After Collapsing Unsatisfactory and Partial Response Categories



* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; the solid curves indicate estimated item response functions assuming a logistic form.

version of MGROUP, based on modifications described by Thomas (1992), to estimate the fourth-grade proficiency distribution for each jurisdiction. As described in the previous chapter, MGROUP estimates proficiency distributions using information from students' item responses, students' background variables, and the item parameter estimates obtained from the BILOG/PARSCALE program.

As the result of research that indicated that the parameters estimated by the conditioning model differed across jurisdictions (Mazzeo, 1991), separate conditioning models were estimated for each jurisdiction. If a jurisdiction had a nonpublic-school sample, students from that sample were included in this part of the analysis, and a conditioning variable differentiating between public- and nonpublic-school students was included. This resulted in the estimation of 44 distinct conditioning models. The background variables included in each jurisdiction's model (denoted *y* in Chapter 8) were principal component scores derived from the within-jurisdiction correlation matrix of selected main-effects and two-way interactions associated with a wide range of student, teacher, school, and community variables. There were no interaction terms between independent variables in the 1992 Trial State Assessment in reading. However, in the 1994 assessment, interaction terms between certain independent variables that might be included in reports were added to the conditioning model. As was done for the 1992 Trial State Assessment, a set of five multivariate plausible values was drawn for each student who participated in the 1994 Trial State Assessment in reading.[4]

As was the case in previous assessments, plans for reporting each jurisdiction's results required analyses examining the relationships between proficiencies and a large number of background variables. The background variables included student demographic characteristics (e.g., the race/ethnicity of the student, highest level of education attained by parents), students' perceptions about reading, student behavior both in and out of school (e.g., amount of television watched daily, amount of homework done each day), and a variety of other aspects of the students' background and preparation, the background and preparation of their teachers, and the educational, social, and financial environment of the schools they attended. If a jurisdiction had a nonpublic-school sample, type of school was included as a background variable.

As described in the previous chapter, to avoid biases in reporting results and to minimize biases in secondary analyses, it is desirable to incorporate measures of a large number of independent variables in the conditioning model. When expressed in terms of contrast-coded main effects and interactions, the number of variables to be included totaled 482. Appendix C provides a listing of the full set of contrasts defined. These contrasts were the common starting point in the development of the conditioning models for each of the participating jurisdictions.

Because of the large number of these contrasts and the fact that, within each jurisdiction, some contrasts had zero variance, some involved relatively small numbers of individuals, and some were highly correlated with other contrasts or sets of contrasts, an effort was made to reduce the dimensionality of the predictor variables in each jurisdiction's MGROUP models. As

[4] There was one exception to this—in the 1994 public-school sample from Georgia. One student had an anomalous pattern of background characteristics that did not fit the conditioning model. After close scrutiny of the data for this student, it was determined that this outlying observation should be deleted from the principal component and conditioning portions of the analysis and from the results.

was done for the 1990 and 1992 Trial State Assessments in mathematics and the 1992 Trial State Assessment in reading, the original background variable contrasts were standardized and transformed into a set of linearly independent variables by extracting separate sets of principal components (one set for each of the 44 jurisdictions) from the within-jurisdiction correlation matrices of the original contrast variables. The principal components, rather than the original variables, were used as the independent variables in the conditioning model. As was done for the previous assessments, the number of principal components included for each jurisdiction was the number required to account for approximately 90 percent of the variance in the original contrast variables. Research based on data from the 1990 Trial State Assessment in mathematics suggests that results obtained using such a subset of the components will differ only slightly from those obtained using the full set (Mazzeo, Johnson, Bowker, & Fong, 1992).

Table 9-11 lists the number of principal components included in and the proportion of proficiency variance accounted for by the conditioning model for each participating jurisdictions. It is important to note that the proportion of variance accounted for by the conditioning model differs across scales within a jurisdiction, and across jurisdictions within a scale. Such variability is not unexpected for at least two reasons. First, there is no reason to expect the strength of the relationship between proficiency and demographics to be identical across all jurisdictions. In fact, one of the reasons for fitting separate conditioning models is that the strength and nature of this relationship may differ across jurisdictions. Second, the homogeneity of the demographic profile also differs across jurisdictions. As with any correlational analysis, the restriction of the range in the predictor variables will attenuate the relationship.

Table 9-11 also provides the estimated within-jurisdiction correlation between the two scales. The values, taken directly from the revised MGROUP program, are estimates of the within-jurisdiction correlations *conditional on the set of principal components included in the conditioning model*. The number and nature of the scales that were produced were consistent with the recommendations for reporting that were given by the National Assessment Planning Project (see Chapter 2). Reporting results on multiple scales is typically most informative when each of the scales provides unique information about the profile of knowledge and skills possessed by the students being assessed. In such cases, one would hope to see relatively low correlations among the scales. However, the correlations between the scales are high across all jurisdictions, always exceeding .7 and sometimes exceeding .9. This is particularly noteworthy when one considers that these are correlations *conditional* on a rather large set of background variables. The *marginal* correlations between content area scales would be higher, particularly for those correlations in the .7 to .8 range.

As discussed in Chapter 8, NAEP scales are viewed as summaries of consistencies and regularities that are present in item-level data. Such summaries should agree with other reasonable summaries of the item-level data. In order to evaluate the reasonableness of the scaling and estimation results, a variety of analyses were conducted to compare state-level and subgroup-level performance in terms of the content area scaled scores and in terms of the average proportion correct for the set of items in a content area. High agreement was found in all of these analyses. One set of such analyses is presented in Figure 9-14. The figure contains scatterplots of the state scaled score mean versus the state item score means, for each of the two reading content areas and the composite scale. As is evident from the figures, there is an extremely strong relationship between the estimates of state-level performance in the scale-score and item-score metrics for both content areas.

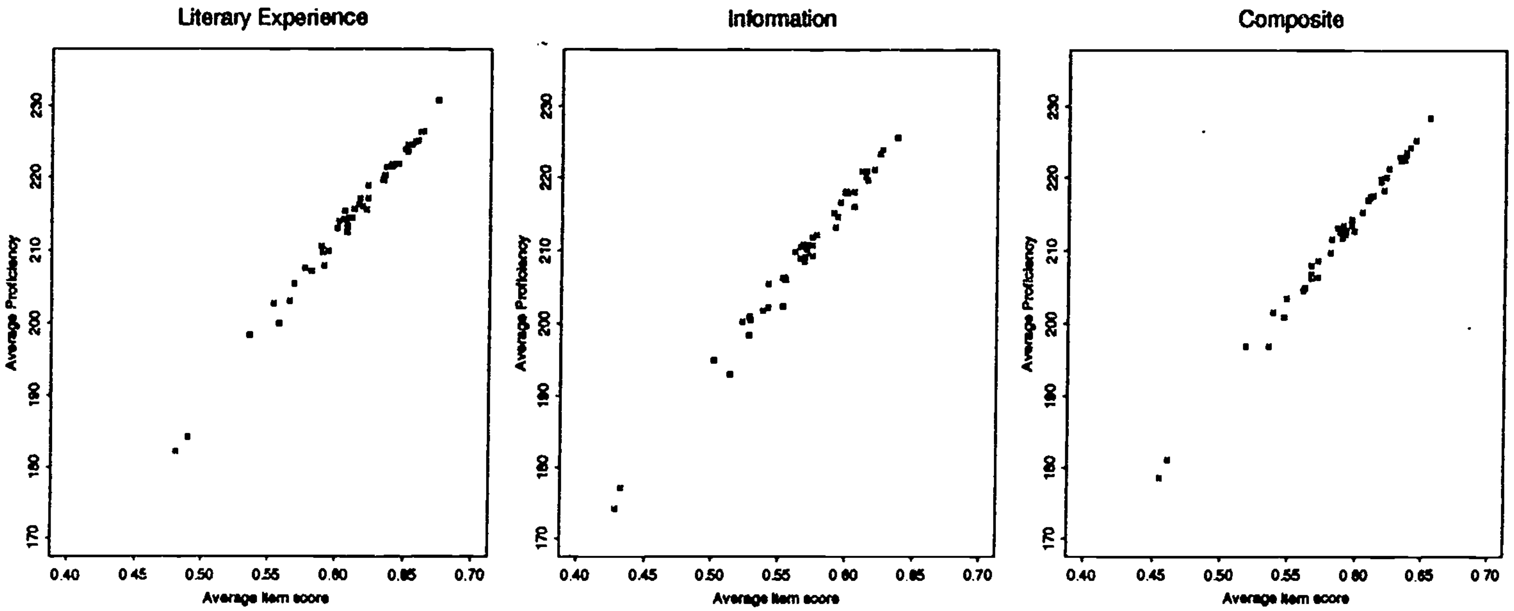
Table 9-11
Summary Statistics for Trial State Assessment Conditioning Models

Jurisdiction	Number of Principal Components	Proportion* of Proficiency Variance in the Reading for Literary Experience Scale Accounted for by the Conditioning Model	Proportion* of Proficiency Variance in the Reading to Gain Information Scale Accounted for by the Conditioning Model	Conditional Correlation Between Scales
Alabama	195	0.58	0.64	0.86
Arizona	195	0.59	0.64	0.93
Arkansas	191	0.65	0.69	0.84
California	186	0.67	0.72	0.90
Colorado	197	0.61	0.61	0.83
Connecticut	193	0.65	0.69	0.89
Delaware	188	0.68	0.67	0.85
District of Columbia	177	0.62	0.63	0.77
Florida	204	0.57	0.59	0.91
Georgia	197	0.63	0.64	0.88
Guam	133	0.53	0.54	0.77
Hawaii	206	0.61	0.63	0.91
Idaho	182	0.62	0.66	0.92
Indiana	190	0.61	0.64	0.83
Iowa	182	0.58	0.55	0.77
Kentucky	188	0.57	0.56	0.79
Louisiana	204	0.59	0.64	0.70
Maine	183	0.55	0.60	0.88
Maryland	186	0.66	0.68	0.86
Massachusetts	196	0.63	0.64	0.82
Michigan	175	0.65	0.65	0.75
Minnesota	192	0.61	0.63	0.80
Mississippi	192	0.66	0.62	0.78
Missouri	196	0.61	0.63	0.88
Montana	185	0.62	0.62	0.86
Nebraska	180	0.62	0.63	0.78
New Hampshire	176	0.59	0.64	0.82
New Jersey	186	0.61	0.73	0.84
New Mexico	193	0.61	0.64	0.89
New York	181	0.64	0.69	0.79
North Carolina	190	0.58	0.63	0.83
North Dakota	179	0.57	0.62	0.91
Pennsylvania	185	0.59	0.64	0.83
Rhode Island	184	0.61	0.64	0.93
South Carolina	193	0.60	0.59	0.81
Tennessee	181	0.65	0.68	0.90
Texas	195	0.67	0.61	0.92
Utah	191	0.60	0.63	0.87
Virginia	201	0.61	0.56	0.89
Washington	195	0.64	0.65	0.93
West Virginia	185	0.55	0.55	0.89
Wisconsin	186	0.55	0.59	0.86
Wyoming	185	0.56	0.58	0.88
DoDEA Overseas	178	0.54	0.61	0.87

* (Total Variance - Residual Variance)/Total Variance, where Total Variance consists of both sampling and measurement error variance.

Figure 9-14

Plot of Mean Proficiency Versus Mean Item Score by Jurisdiction



9.6 LINKING STATE AND NATIONAL SCALES

Data from the Trial State assessment and the national reading assessment were scaled separately for two major reasons: 1) because of a difference in administration procedures (Westat staff collected the data for the national assessment, while data collection for the Trial State assessment was the responsibility of individual jurisdictions) and 2) because of potential motivational differences between the samples of students participating in the national assessment and those participating in the Trial State assessment.

A major purpose of the Trial State Assessment Program was to allow each participating jurisdiction to compare its 1994 results with the nation as a whole and with the region of the country in which that jurisdiction is located. For meaningful comparisons to be made between each of the Trial State Assessment jurisdictions and the relevant national sample, results from these two assessments had to be expressed in terms of a similar system of scale units.

The purpose of this section is to describe the procedures used to align the 1994 Trial State scales with their 1994 national counterparts. The procedures that were used are similar to the common population equating procedures employed to link the 1990 national and state mathematics scales (Mazzeo, 1991; Yamamoto & Mazzeo, 1992) and the 1992 national and state mathematics and reading scales (Allen, Mazzeo, Isham, Fong & Bowker, 1994; Mazzeo, Chang, Kulick, Fong, & Grima, 1993).

Using the sampling weights provided by Westat, the combined sample of students from participating jurisdictions (a total sample size of 112,153) was used to estimate the distribution of proficiencies for the population of students enrolled in public schools in the participating states and the District of Columbia[5]. Data were also used from a subsample of 5,063 students in the national assessment at grade 4, consisting of grade-eligible public-school students from jurisdictions that contributed students to the combined sample from the Trial State Assessment. Appropriate weights were provided by Westat to obtain estimates of the distribution of proficiency for the same target population.

Thus, for each of the two scales, two sets of proficiency distributions were obtained. One set, based on the sample of combined data from the Trial State Assessment (referred to as the Trial State Assessment Aggregate Sample) and using item parameter estimates and conditioning results from that assessment, was in the metric of the 1994 Trial State Assessment. The other, based on the sample from the 1994 national assessment (referred to as the State Aggregate Comparison, or SAC, sample) and obtained using item parameters and conditioning results from that assessment, was in the metric of the 1994 national assessment. The latter metric had already been set using procedures described in the technical report of the 1994 national assessment. The two Trial State Assessment and national scales were made comparable by constraining the mean and standard deviation of the two sets of estimates to be equal.

[5] Students from Guam and DoDEA overseas schools were excluded from the definition of this target population; hence, data from students from these jurisdictions were not included in the combined Trial State Assessment samples

More specifically, the following steps were followed to linearly link the scales of the two assessments:

1) For each scale, estimates of the proficiency distribution for the Trial State Assessment Aggregate Sample were obtained using the full set of plausible values generated by the CGROUP program. The weights used were the final sampling weights provided by Westat, not the rescaled versions discussed in section 9.3. For each scale, the arithmetic mean of the five sets of plausible values was taken as the overall estimated mean and the standard deviations of the five sets of plausible values was taken as the overall estimated standard deviation.

2) For each scale, the estimated proficiency distribution of the State Aggregate Comparison sample was obtained, again using the full set of plausible values generated by the CGROUP program. The weights used were specially provided by Westat to allow for the estimation of proficiency for the same target population of students estimated by the state data. The means and standard deviations of the distributions for each scale were obtained for this sample in the same manner as described in step 1. These means and standard deviations were then linearly adjusted to reflect the reporting metric used for the national assessment (see the technical report for the NAEP 1994 national assessment.

3) For each scale, a set of linear transformation coefficients were obtained to link the state scale to the corresponding national scale. The linking was of the form

$$Y^* = k_1 + k_2Y$$

where

Y = a scale level in terms of the system of units of the provisional BILOG/PARSCALE scale of the Trial State Assessment scaling

Y^* = a scale level in terms of the system of units comparable to those used for reporting the 1994 national reading results

k_2 = [Standard Deviation$_{SAC}$]/[Standard Deviation$_{TSA}$]

k_1 = Mean$_{SAC}$ - k_2[Mean$_{TSA}$]

The final conversion parameters for transforming plausible values from the provisional BILOG/PARSCALE scales to the final Trial State Assessment reporting scales are given in Table 9-12. After the plausible values were linearly transformed to the new scale, any plausible value less than 1 was censored to 1 and any value greater than 500 was censored to 500. Fewer than .07 percent (unweighted) of the students had plausible values that were censored to one. So, the final Trial State Assessment reporting scale ranged from 1 to 500. All Trial State Assessment results, including those for nonpublic schools, are reported in terms of the Y^* metric using these transformations.

Table 9-12
Transformation Constants for the 1994 Trial State Assessment

Scale	k_1	k_2
Reading for Literary Experience	214.64	42.15
Reading to Gain Information	210.36	42.08

As evident from the discussion above, a linear method was used to link the scales from the Trial State and national assessments. While these linear methods ensure equality of means and standard deviations for the Trial State Assessment aggregate (after transformation) and the SAC samples, they do not guarantee that the shapes of the estimated proficiency distributions for the two samples will be the same. As these two samples are from a common target population, estimates of the proficiency distribution of that target population based on each of the samples should be quite similar in shape in order to justify strong claims of comparability for the Trial State and national scales. Substantial differences in the shapes of the two es imated distributions would result in differing estimates of the percentages of students above achievement levels or of percentile locations depending on whether Trial State or national scales were used—a clearly unacceptable result given claims about comparability of scales. In the face of such results, nonlinear linking methods would be required.

Analyses were carried out to verify the degree to which the linear linking process described above produced comparable scales for Trial State and national results. Comparisons were made between two estimated proficiency distributions, one based on the Trial State Assessment aggregate and one based on the SAC sample, for each of the two reading scales. The comparisons were carried out using slightly modified versions of what Wainer (1974) refers to as suspended rootograms. The final reporting scales for the Trial State and national assessments were each divided into 10-point intervals. Two sets of estimates of the percentage of students in each interval were obtained, one based on the Trial State Assessment aggregate sample and one based on the SAC sample. Following Tukey (1977), the square root of these estimated percentages were compared.[6]

The comparisons are shown in Figure 9-15 and 9-16 on the two content scales. The heights of each of the unshaded bars correspond to the square root of the percentage of students from the Trial State Assessment aggregate sample in each 10-point interval on the final reporting scale. The shaded bars show the differences in root percents between the Trial State Assessment and SAC estimates. Positive differences indicate intervals in which the estimated percentages from the State Aggregate Comparison sample are lower than those obtained from the Trial State Assessment aggregate. Conversely, negative differences indicate intervals in which the estimated percentages from the State Aggregate Comparison sample are higher. For both scales, differences in root percents are quite small, suggesting that the shapes of the two

[6] The square root transformation allows for more effective comparisons of counts (or equivalently, percentages) when the expected number in each interval is likely to vary greatly over the range of intervals, as is the case for the NAEP scales where the expected counts of individuals in intervals near the extremes of the scale (e.g., below 150 and above 350) are dramatically smaller than the counts obtained near the middle of the scale.

Figure 9-15

Rootogram Comparing Proficiency Distributions
for the Trial State Assessment Aggregate Sample
and the State Aggregate Comparison Sample from the National Assessment
for the Reading for Literary Experience Scale

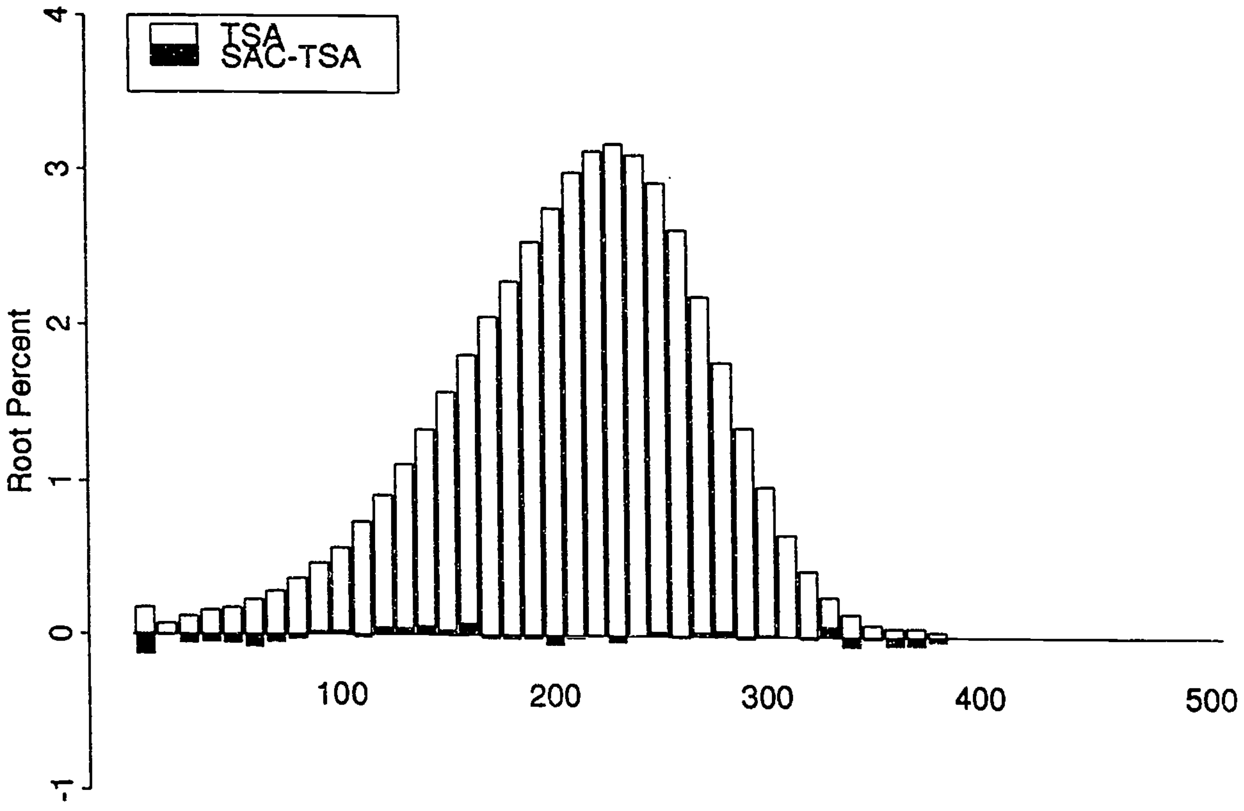
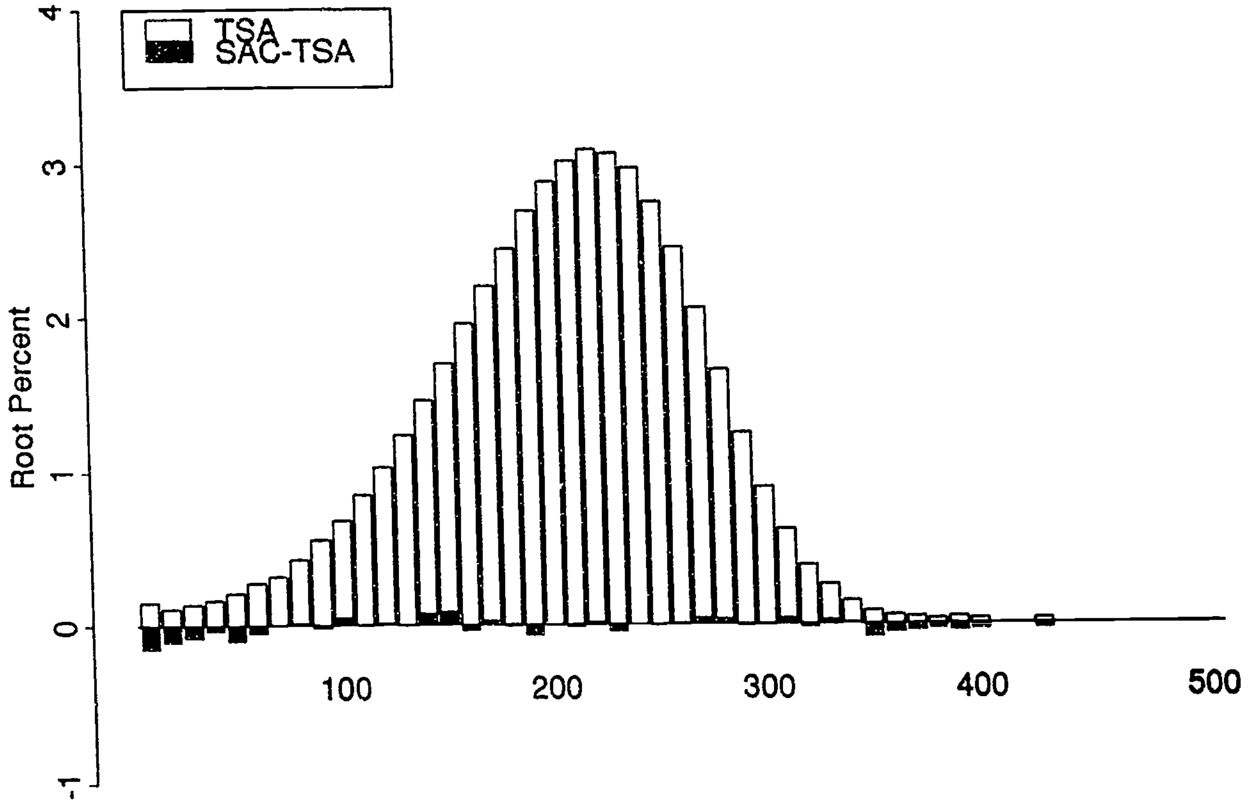


Figure 9-16

Rootogram Comparing Proficiency Distributions
for the Trial State Assessment Aggregate Sample
and the State Aggregate Comparison Sample from the National Assessment
for the Reading to Gain Information Scale



estimated distributions are quite similar (i.e., unimodal with slight negative skewness). There is some evidence that there are proportionately more proficiency for the Trial State Assessment data in the extreme lower and upper tails (below 100 and above 300). However, even these differences at the extremes are small in magnitude and have little impact on estimates of reported statistics such as percentages of students below the achievement levels. The results look similar to those in the 1992 Trial State Assessment.

9.7 PRODUCING A READING COMPOSITE SCALE

For the national assessment, a composite scale was created for the fourth grade as an overall measure of reading proficiency. The composite was a weighted average of plausible values on the two content area scales (Reading for Literary Experience and Reading to Gain Information). The weights for the national content area scales were proportional to the relative importance assigned to each content area for the fourth grade in the assessment specifications developed by the Reading Objectives Panel. Consequently, the weights for each of the content areas are similar to the actual proportion of items from that content area.

The Trial State Assessment composite scale was developed using weights identical to those used to produce the composites for the 1992 and 1994 national reading assessments. The weights are given in Table 9-13. In developing the Trial State Assessment composite the weights were applied to the plausible values for each content area scale as expressed in terms of the final Trial State Assessment scales (i.e., after transformation from the provisional BILOG/PARSCALE scales.)

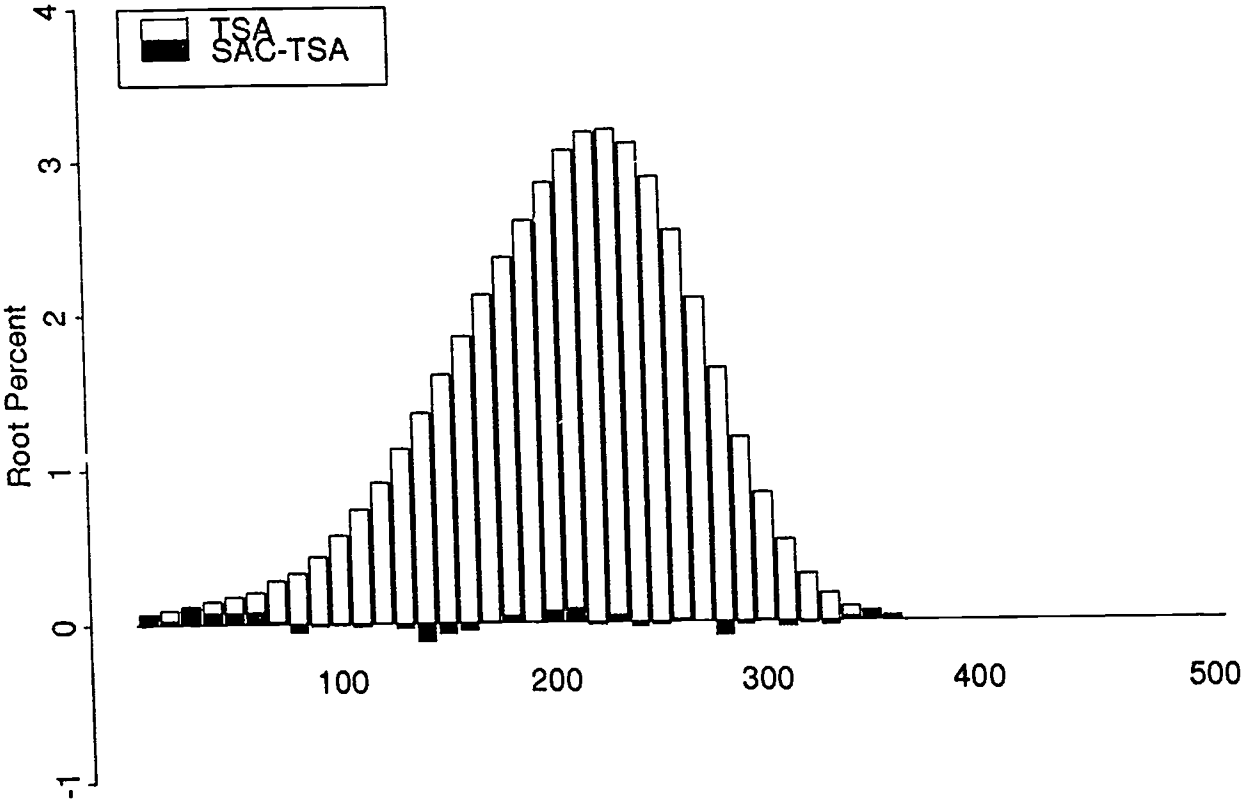
Table 9-13
Weights Used for Each Scale to Form the Reading Composite

Scale	Weights
Reading for Literary Experience	.55
Reading to Gain Information	.45

Figure 9-17 provides a rootogram comparing the estimated proficiency distributions based on the Trial State Assessment and SAC samples for the grade 4 composite. Consistent with the results presented separately by scale, the differences in root relative percents are small in magnitude.

Figure 9-17

Rootogram Comparing Proficiency Distributions
for the Trial State Assessment Aggregate Sample
and the State Aggregate Comparison Sample from the National Assessment
for the Reading Composite Scale



Chapter 10

CONVENTIONS USED IN REPORTING THE RESULTS OF THE 1994 TRIAL STATE ASSESSMENT IN READING

John Mazzeo arıd Clyde M. Reese

Educational Testing Service

10.1 OVERVIEW

Results for the 1994 Trial State Assessment in Reading were disseminated in several different reports: a *Reading State Report* for each jurisdiction, the brief report entitled *1994 NAEP Reading: A First Look, The 1994 Reading Report Card for the Nation and the States* , the *Cross-State Data Compendium for the NAEP 1994 Reading Assessment*, and a six-section almanac of data for each state.

The *Reading State Report* is a computer-generated report that provides, for each jurisdiction, reading results for its fourth-grade students. Although national and regional results[1] are included for comparison purposes, the major focus of each of these computer-generated reports is the results for a particular jurisdiction. Data about school and student participation rates are reported for each jurisdiction to provide information about the generalizability of the results. School participation rates are reported both in terms of the initially selected samples of schools and in terms of the finally achieved samples, including replacement schools. Several different student participation rates are reported, including the overall rate, the percentage of students excluded from the assessment, anc the exclusion rates for Limited English Proficiency (LEP) students and for students with Individualized Education Plans (IEP). In addition to 1994 results, the state reports contain comparisons of 1992 fourth-grade results to the 1992 fourth-grade results for the jurisdictions that participated in both assessments. Trend results are also provided for the nation and for the relevant region associated with each participating jurisdiction.

The state report text and tables were produced by a computerized report generation system developed by ETS report writers, statisticians, data analysts, graphic designers, and editors. Detailed technical documentation about the NAEP computer-generated reporting system can be found in the technical documentation of the 1994 NAEP computerized report generation system (Jerry, 1995). The reports contain state-level estimates of proficiency means,

[1] The national and regional results included in the state reports and in the portions of the *Cross-State Data Compendium for the NAEP 1994 Reading Assessment* are based on data from the 1992 and 1994 national reading assessment and include fourth-grade students enrolled in public and nonpublic schools.

proportions of students at or above achievement levels defined by the National Assessment Governing Board (NAGB) and selected percentiles for the state as a whole and for subgroups defined by four key reporting variables (referred to here as primary reporting variables)—gender, race/ethnicity, level of parents' education, and type of location. For jurisdictions that secured a sufficient level of participation (see Appendix B), means, achievement levels and percentile results were also reported for students in nonpublic schools (Catholic schools, other religious schools, and other private schools), and for the total in-school population (public-school students, nonpublic-school students, students from Domestic Department of Defense Schools and students attending Bureau of Indian Affairs schools). In addition, for public-school students, proficiency means were also reported for a variety of other subpopulations defined by responses to items from the student, teacher, and school questionnaires and by school and community demographic variables provided by Westat[2].

The second and third reports, *1994 NAEP Reading: A First Look* and the *NAEP 1994 Reading Report Card for the Nation and the States*, present key assessment results for the nation and summarizes results across jurisdictions participating in the assessment. The *First Look* report contains composite scale results (proficiency means, proportions at or above achievement levels, etc.) for the nation, each of the four regions of the country, and for public-school students within each jurisdiction participating in the Trial State Assessment, both overall and by the primary reporting variables. The *Report Card* expands upon the *First Look*, by including state results for nonpublic school students, results reported for each of the reading-purpose scales, and results pertaining to a variety of home, teacher and school contextual variables. Results in both reports include trend comparisons to 1992 for all grades in the national assessment and for grade 4 for those jurisdictions that participated in both the 1992 and 1994 Trial State Assessments. Both reports also contain a number of specially developed graphical displays that summarize and compare results for the full set of participating jurisdictions.

The fourth report is entitled *Cross-State Data Compendium from the NAEP 1994 Reading Assessment*. Like the *Report Card*, the *Compendium* reports results for the nation and for all of the jurisdictions participating in the Trial State Assessment. The *Compendium* is primarily tabular in nature and contains little in the way of interpretive text. The *Compendium* contains most of the tables presenting cross-state results for all the variables included in the *State Report* and the *Report Card*.

The fifth report is a six-section almanac that contains a detailed breakdown of the reading proficiency data according to the responses to the student, teacher, and school questionnaires for the public-school, nonpublic-school, and combined populations as a whole and for important subgroups of the public-school population. There are six sections to each almanac:

The Distribution Data Section provides information about the percentages of students at or above the three composite-scale achievement levels (and below basic). For the composite scale and each reading scale, this almanac also provides selected percentiles

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

for the public-school, nonpublic-school, and combined populations and for the standard demographic subgroups of the public-school population.

The Student Questionnaire Section provides a breakdown of the composite scale proficiency data according to the students' responses to questions in the three student questionnaires included in the assessment booklets.

The Teacher Questionnaire Section provides a breakdown of the composite-scale proficiency data according to the teachers' responses to questions in the reading teacher questionnaire.

The School Questionnaire Section provides a breakdown of the composite-scale proficiency data according to the principals' (or other administrators') responses to questions in the school characteristics and policies questionnaire.

The Scale Section provides a breakdown of the proficiency data for the two reading scales (Reading for Literary Experience, Reading to Gain Information) according to selected items from the questionnaires.

The Reading Item Section provides the response data for each reading item in the assessment.

The production of the state reports, the *First Look Report*, the *Report Card*, the *Cross-State Data Compendium*, and the almanacs required a large number of decisions about a variety of data analysis and statistical issues. For example, given the sample sizes obtained for each jurisdiction, certain categories of the reporting variables contained limited numbers of examinees. A decision was needed as to what constituted a sufficient sample size to permit the reliable reporting of subgroup results, and which, if any, estimates were sufficiently unreliable to need to be identified (or flagged) as a caution to readers. As a second example, the state report contained computer-generated text that described the results for a particular jurisdiction and compared total and subgroup performance within the jurisdiction to that of the region and nation. A number of inferential rules, based on logical and statistical considerations, had to be developed to ensure that the computer-generated reports were coherent from a substantive standpoint and were based on statistical principals of significance testing. As a third example, the *Report Card* contained tables that statistically compared performance between 1994 and 1992 for each of the participating jurisdictions. Practical multiple comparison procedures were required to control for Type I errors without paying too large a penalty with respect to the power to detect real and substantive differences.

The purpose of this chapter is to document the major conventions and statistical procedures used in generating the state reports, the *First Look Report*, the *Report Card*, the *Data Compendium*, and the almanacs. The principal focus of this chapter is on conventions used in the production of the computer-generated state reports. However, sections 10.2 to 10.4 contain material applicable to all four summary reports. Additional details about procedures relevant to the *Report Card* and *Data Compendium* can be found in the text and technical appendices of those reports.

10.2 MINIMUM SCHOOL AND STUDENT SAMPLE SIZES FOR REPORTING SUBGROUP RESULTS

In all of the reports, estimates of quantities such as composite and content area proficiency means, percentages of students at or above the achievement levels, and percentages of students indicating particular levels of background variables (as measured in the student, teacher, and school questionnaires) are reported for the total population of fourth-grade students in each jurisdiction, as well as for certain key subgroups of interest. The subgroups were defined by four primary NAEP reporting variables. NAEP reports results for six racial/ethnic subgroups (White, Black, Hispanic, Asian American,Pacific Islander, and American Indian/Alaskan Native), three types of locations (central cities, urban fringes/large towns, rural/small town areas), and four levels of parents' education (did not finish high school, high school graduate, some college, college graduate). However, in some jurisdictions, and for some regions of the country, school and/or student sample sizes were quite small for one or more of the categories of these variables. One would expect results for subgroups so defined to be imprecisely estimated.

It is common practice in reports generated by statistical agencies to suppress estimates for which the sampling error is so large that it is determined that no effective use can be made of the estimate, or that the potential for misinterpretation outweighs potential benefits of presenting results. A second, and equally important, consideration is whether the standard error estimate that accompanies a statistic is sufficiently accurate to adequately inform potential readers about the reliability of the statistic. The precision of a sample estimate (be it sample mean or standard error estimate) for a population subgroup from a two-stage sample design (such as was used to select the samples for the Trial State Assessment) is, in part, a function of the sample size of the subgroup and the distribution of that sample across first-stage sampling units (i.e., schools in the case of the Trial State Assessment). Hence, both of these factors were used in establishing minimum sample sizes for reporting.

For results to be reported for any subgroup, a minimum student sample size of 62 was required. This number was arrived at by determining the sample size necessary to detect an effect size of 0.5 with a probability of .8 or greater[3]. The effect size of 0.5 pertains to the "true" difference in mean proficiency between the subgroup in question and the total fourth-grade public-school population in the jurisdiction, divided by the standard deviation of proficiency in the total population. The same convention was used in reporting the 1990 and 1992 Trial State Assessment results. Further more, it was required that the students within a subgroup be adequately distributed across schools to allow for reasonably accurate estimation of standard errors. In consultation with Westat, a decision was reached to publish only those statistics that had standard errors estimates based on five or more degrees of freedom. Slightly different variance estimation procedures were used to obtain standard error estimates for public and nonpublic school statistics (see Chapter 7). These different procedures implied different minimum school sample sizes for public and nonpublic school results. For public school statistics, subgroup data was required to come from a minimum of 10 schools. For nonpublic school statistics, a six-school minimum was required.

[3] A design effect of 2 was assumed for this purpose, implying a sample design-based variance twice that of simple random sampling. This is consistent with previous NAEP experience (Johnson & Rust, 1992).

It should be noted that the full set of summary reports includes large numbers of tables that provide estimates of the proportion of the students responding to each category of a secondary reporting variable, as well as the mean proficiency of the students within each category. In several instances, the number of students in a particular category of these background variables was also less than 62 or was clustered within a small number of schools. The same minimum school and student sample sizes restrictions were applied to these subgroups as well.

10.3 ESTIMATES OF STANDARD ERRORS WITH LARGE MEAN SQUARED ERRORS

As noted above, standard errors of mean proficiencies, proportions, and percentiles play an important role in interpreting subgroup results and comparing the performances of two or more subgroups. The jackknife standard errors reported by NAEP are statistics whose quality depends on certain features of the sample from which the estimate is obtained. As discussed in the previous section, in certain cases, typically when the number of students upon which the standard error is based is small or when this group of students come from a small number of participating schools, the mean squared error[4] associated with the estimated standard errors may be quite large. Minimum school and student sizes were implemented which suppressed statistics in most instances where such problems existed. However, the possibility remained that some statistics based on sample sizes that exceed the minimum requirements might still be associated with standard errors that were not well estimated. Therefore, in the summary reports, estimated standard errors for published statistics that are subject to large mean squared errors are followed by the symbol "!".

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

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

where $\hat{N}$ is a point estimate of *N* and $SE(\hat{N})$ is the jackknife standard error of $\hat{N}$.

Experience with previous NAEP assessments suggests that when this coefficient exceeds 0.2, the mean squared error of the estimated standard errors of means and proportions based on samples of this size may be quite large. (Further discussion of this issue can be found in Johnson & Rust, 1992.) Therefore, the standard errors of means and proportions for all subgroups for which the coefficient of variation of the population size exceeds 0.2 are followed by "!" in the tables of all summary reports. These standard errors, and any confidence intervals or significance tests involving these standard errors, should be interpreted with caution. In the

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

First Look Report, *The Report Card*, the *Cross-State Data Compendium*, and the almanacs, statistical tests involving one or more quantities that have standard errors so flagged were not carried out.

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

Responses to the student, teacher, and school questionnaires played a prominent role in all reports. Although the return rate on all three types of questionnaire was high[5], there were missing data on each type. of questionnaire

The reported estimated percentages of students in the various categories of background variables, and the estimates of the mean proficiency of such groups, were based on only those students for whom data on the background variable were available. In the terminology of Little and Rubin (1987), the analyses pertaining to a particular background variable presented in the state reports and the *Data Compendium* assume the data are missing completely at random (i.e., the mechanism generating the missing data is independent of both the response to the particular background variables and to proficiency).

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

In order to analyze the relationships between teachers' questionnaire responses and their students' achievement, each teacher's questionnaire had to be matched to all of the students who were taught reading by that teacher. Table 10-1 provides the percentages of fourth-grade students that were matched to teacher questionnaires for each of the 44 jurisdictions. The first column presents match rates for public-school students, the second for nonpublic school students, the third for the combined within-jurisdiction sample. Note that these match rates do not reflect the additional missing data due to item-level nonresponse. The amount of additional item-level nonresponse in the returned teacher questionnaires can also be found in the almanacs produced for each jurisdiction.

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

Table 10-1
Weighted Percentage of Fourth-grade Students Matched to Teacher Questionnaires

Jurisdiction	Public	Nonpublic	Total
Alabama	99.0	99.7	99.0
Arizona	98.8	—	98.8
Arkansas	99.3	100.0	99.3
California	96.8	99.5	97.0
Colorado	96.7	100.0	96.9
Connecticut	98.0	99.7	98.2
Delaware	98.3	100.0	98.6
District of Columbia	93.9	91.6	93.5
Florida	94.0	92.7	93.8
Georgia	97.4	99.8	97.6
Guam	76.5	67.5	75.1
Hawaii	98.0	94.4	97.6
Idaho	97.7	100.0	97.8
Indiana	97.8	100.0	97.9
Iowa	96.2	94.3	95.9
Kentucky	97.1	99.6	97.1
ouisiana	98.9	100.0	99.0
aine	98.7	100.0	98.7
Maryland	97.6	100.0	97.8
Massachusetts	98.7	93.4	98.2
Michigan	96.0	—	96.0
Minnesota	95.5	100.0	96.1
Mississippi	98.4	100.0	98.5
Missouri	98.7	99.8	98.8
Montana	99.2	100.0	99.2
Nebraska	98.2	100.0	98.4
New Hampshire	99.3	—	99.3
New Jersey	97.9	95.3	97.6
New Mexico	96.4	100.0	95.1
New York	99.8	100.0	99.8
North Carolina	95.7	—	95.7
North Dakota	99.0	98.6	99.0
Pennsylvania	97.2	97.7	97.3
Rhode Island	98.1	87.4	96.8
South Carolina	96.7	100.0	96.8
Tennessee	99.2	—	99.2
Texas	98.0	—	98.0
Utah	99.4	—	99.4
Virginia	98.6	99.5	98.6
Washington	94.4	—	94.4
West Virginia	97.4	100.0	97.5
Wisconsin	97.2	100.0	97.6
Wyoming	96.2	—	96.2
DoDEA Overseas	96.8	—	96.8
TOTAL	97.3	98.3	97.4

10.5 STATISTICAL RULES USED FOR PRODUCING THE STATE REPORTS

As described earlier, the state reports contain state-level estimates of fourth-grade mean proficiencies, proportions of students at or above selected scale points, and percentiles for the jurisdiction as a whole and for the categories of a large number of reporting variables. Similar results are provided for the nation and, where sample sizes permitted, for the region to which each jurisdiction belongs[6]. The state reports were computer-generated. The tables and figures, as well as the text of the report, were automatically tailored for each jurisdiction based on the pattern of results obtained. The purpose of this section is to describe some of the procedures and rules used to produce these individually tailored reports. A complete and detailed presentation is available in the technical documentation of the 1994 NAEP computerized report generation system (Jerry, 1995).

In the 1994 state reports, the results are presented principally through a sequence of tables containing estimated means, proportions, and percentiles, along with their standard errors, for 1994 and, where appropriate, for 1992. In addition to the tables of results, computer-generated interpretive text is also provided. In some cases, the computer-generated interpretive text is primarily descriptive in nature and reports the total group and subgroup proficiency means and proportions of interest. However, some of the interpretive text focuses on interesting and potentially important group differences in reading proficiency or on the percentages of students responding in particular ways to the background questions. Additional interpretive text compares state-level results with those of the nation, and discusses changes in results from 1992 to 1994. For example, one question of considerable interest to each jurisdiction is whether, on average, its students performed higher than, lower than, or about the same as students in the nation. Additional interpretive text focuses on potentially interesting patterns of achievement across the reading-purpose scales or on the pattern of response to a particular background question in the jurisdiction. For example, do more students report spending 30 minutes or 15 minutes on homework each day?

Rules were developed to produce the computer-generated text for questions involving the comparison of results for subgroups and interpretations of patterns of results. These rules were based on a variety of considerations, including a desire for 1) statistical rigor in the identification of important group differences and patterns of results, and 2) solutions that were within the limitations imposed by the availability of computational resources and the time frame for the production of the report. The following sections describe some of these procedures and rules.

10.5.1 Comparing Means and Proportions for Mutually Exclusive Groups of Students

Many of the group comparisons explicitly commented on in the state reports involved mutually exclusive sets of students. One common example of such a comparison is the contrast between the mean composite proficiency in a particular jurisdiction and the mean composite proficiency in the nation. Other examples include comparisons within a jurisdiction of the

[6] Because United States territories are not classified into NAEP regions, no regional comparisons were provided for Guam Regional results are also not provided for Department of Defense Education Activity Overseas Schools.

average proficiency for male and female students; White and Hispanic students; students attending schools in central city and urban fringe/large town locations; and students who reported watching six or more hours of television each night and students who report watching less than one hour each night.

In the state reports, computer-generated text indicated that means or proportions from two groups were different only when the difference in the point estimates for the groups being compared was statistically significant at an approximate α level of .05. An approximate procedure was used for determining statistical significance that NAEP staff felt was reasonable from a statistical standpoint, as well as being computationally tractable. The procedure was as follows.

Let t_i be the statistic in question (i.e., a mean or proportion for group i) and let $SE(t_i)$ be the jackknife standard error of the statistic. The computer-generated text in the state report identified the means or proportions for groups i and j as being different if and only if:

$$\frac{|t_i - t_j|}{\sqrt{\hat{SE}^2(t_i) + \hat{SE}^2(t_j)}} \geq Z_{\frac{.05}{2c}}$$

where Z_α is the $(1 - \alpha)$ percentile of the standard normal distribution, and c is the number of contrasts being tested. In cases where group comparisons were treated as individual units (for example, comparing overall state results with overall national results or overall state results in 1994 with those of 1992, the value of c was taken as 1, and the test statistic was approximately equivalent to a standard two-tailed t-test for the difference between group means or proportions from large independent samples with the α level set at .05.

The procedures in this section assume that the data being compared are from independent samples. Because of the sampling design used for the Trial State Assessment, in which both schools and students within schools are randomly sampled, the data from mutually exclusive sets of students within a jurisdiction may not be strictly independent. Therefore, the significance tests employed are, in many cases, only approximate. As described in the next section, another procedure, one that does not assume independence, could have been conducted. However, that procedure is computationally burdensome and resources precluded its application for all the comparisons in the state reports. It was the judgment of NAEP staff that if the data were correlated across groups, in most cases the correlation was likely to be positive. Since, in such instances, significance tests based on assumptions of independent samples are conservative (because the estimated standard error of the difference based on independence assumptions is larger than the more complicated estimate based on correlated groups), the approximate procedure was used for most comparisons.

The procedures described above were used for testing differences of both means *and* proportions. The approximation for the test for proportions works best when sample sizes are large, and the proportions being tested have magnitude close to .5. Statements about group differences should be interpreted with caution if at least one of the groups being compared is small in size and/or if somewhat extreme proportions are being compared.

10.5.2 Multiple Comparison Procedures

Frequently, groups (or families) of comparisons were made and were presented as a single set. The appropriate text, usually a set of sentences or a paragraph, was selected for inclusion in the report based on the pattern of results for the entire set of comparisons. For example, Chapter 2 of the state report contained a section that compared average proficiencies for a majority group (in the case of race/ethnicity, for example, usually White students) to those obtained by each minority group meeting minimum school and student sample-size requirements. For families of contrasts like these, a Bonferroni procedure was used for determining the value of $Z_{c\alpha}$ where *c* was the number of contrasts in the set. In this example, *c* was taken to be the number of minority groups meeting minimum sample size requirements, and each statistical test was consequently carried out at an α level of .05/c.

10.5.3 Determining the Highest and Lowest Scoring Groups from a Set of Ranked Groups

Certain analyses in the state report consisted of determining which of a set of several groups had the highest or lowest proficiency among the set. For example, one analysis compared the average proficiency of students who reported reading various numbers of books outside of school during the past month. There were four levels of book reading–none, one or two, three or four, and five or more, Based on their answers to this question in the student background questionnaire, students were classified into one of the four levels of book reading, and the mean composite proficiency was obtained for students at each level. The analysis focused on which, if any, of the groups had the highest and lowest mean composite proficiency.

The analysis was carried out using the statistics described in the previous section. The groups were ranked from highest to lowest in terms of their estimated mean proficiency. Then, three separate significance tests were carried out: 1) the highest group was compared to the lowest group; 2) the highest group was compared to the second highest group; and 3) the lowest group was compared to the second lowest group. The following conclusions were drawn:

- If all three comparisons were statistically significant, the performance of the highest ranking group was described as *highest* and the performance of the lowest ranking group was described as *lowest*.
- If only the first and second tests were significant, the highest ranking group was described as *highest*, but no comment was made about the lowest ranking group.
- Similarly, if only the first and third tests were significant, the lowest ranking group was described as *lowest*, but no comment was made about the highest ranking group.
- If only the first test was significant, the highest group was described as performing better than the lowest group, but no *highest* and *lowest* group were designated.

The Bonferroni adjustment factor was taken as the number of possible pairwise comparisons because of the ranking of groups prior to the carrying out of significance tests.

10.5.4 Comparing 1994 and 1992 Results in State Report Tables

Since its inception, one of NAEP's central purposes has been the monitoring of trends in achievement. The 1994 Trial State Assessment provided the first opportunity to report on short-term trends (from 1992 to 1994) in fourth-grade reading achievement and instructional practices on a state-by-state basis, as well as for the nation and the relevant region of the country. As a result, one of the prominent features of the 1994 state report was the inclusion of a large number of trend comparisons in both the text and tables of the reports for those jurisdictions that participated in both the 1992 and 1994 Trial State Assessments.

The samples for the 1992 and 1994 Trial State Assessments were drawn independently and consisted of mutually exclusive groups of students. Therefore, the selections of text describing comparisons of 1992 and 1994 results were based on the types of significance testing procedures described in section 10.5.1. In sections of the report where trend comparisons were carried out for a number of subgroups (e.g., where 1994 results were compared to 1992 results for each race/ethnicity group within the jurisdiction, or for each of the purposes of reading scales), the significance testing procedures incorporated Bonferroni adjustments, like those described in section 10.5.2, which were based on the number of comparisons being made.

In addition, a large number of state report tables provided both 1992 and 1994 percentages of students and proficiency means for the subgroups of students defined by primary and secondary reporting variables. In most of these tables, three sets of trend results were reported, one set for the jurisdiction in question, one set for relevant region of the country, and one set for the nation. For each of these sets of results, symbols were included next to the 1994 results for each jurisdiction indicating which, if any, of the reported statistics represented a significant change from the 1992 results. A ">" sign was used to indicate 1994 results that were significantly higher than their corresponding 1992 levels. A "<" was used to indicate 1994 results that were significantly lower than their corresponding 1992 levels. No symbol appeared after results that did not differ significantly from their 1992 levels.

As was done for text selection, statistical tests were carried out using Bonferroni adjustments to significance levels when results for multiple groups were included in a table. For example, in a table containing 1992 and 1994 mean proficiencies for White, Black, and Hispanic students, statistical tests for differences were carried out at an α level of .05/3. It should be noted that national, regional, and jurisdiction comparisons were treated as separate families for the purposes of obtaining Bonferroni adjustments. Continuing with the race/ethnicity example, jurisdiction, national, and regional comparisons were treated as three separate families each consisting of three comparisons and each of the required statistical tests were carried out at an α level of .05/3.

10.5.5 Comparing Dependent Proportions

Certain analyses in the state report involved the comparison of dependent proportions. One example was the comparison of the proportion of students who reported that they spent 30 minutes a day on homework to the proportion of students who indicated that they spent 15 minutes a day on homework to determine which proportion was larger. For these types of analyses, NAEP staff determined that the dependencies in the data could not be ignored.

Unlike the case for analyses of the type described in section 10.5.1, the correlation between the proportion of students reporting 30 minutes of homework and the proportion reporting 15 minutes is likely to be negative. For a particular sample of students, it is likely that the higher the proportion of students reporting 30 minutes is, the lower the proportion of students reporting 15 minutes will be. A negative dependence will result in underestimates of the standard error if the estimation is based on independence assumptions (as is the case for the procedures described in the previous section). Such underestimation can result in too many "nonsignificant" differences being identified as significant.

The procedures of section 10.5.1 were modified for the state report analyses that involved comparisons of dependent proportions. The modification involved using a jackknife method for obtaining the standard error of the difference in dependent proportions. The standard error of the difference in proportions was obtained by first obtaining a separate estimate of the difference in question for each jackknife replicate, using the first plausible value only, then taking the standard deviation of the set of replicate estimates as the estimate. The procedures used for dependent proportions differed from the procedures of section 10.5.1 only with respect to estimating the standard error of the difference; all other aspects of the procedures were identical.

10.5.6 Statistical Significance and Estimated Effect Sizes

Whenever single comparisons were made between groups, an attempt was made to distinguish between group differences that were statistically significant but rather small in a practical sense and differences that were both statistically and practically significant. In order to make such distinctions, a procedure based on estimated effect sizes was used. The estimated effect size for comparing means from two groups was defined as:

$$\textit{estimated effect size} = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{\sqrt{\dfrac{S_i^2 + S_j^2}{2}}}$$

where $\hat{\mu}_i$ refers to the estimated mean for group *i*, and S_i refers to the estimated standard deviation within group *i*. The within-group estimated standard deviations were taken to be the standard deviation of the set of five plausible values for the students in subgroup *i* and were calculated using the Westat sampling weights.

The estimated effect size for comparing proportions was defined as

$|f_i - f_j|$, where $f_i = 2\ \text{arcsin}\sqrt{p_i}$, and p_i is the estimated proportion in group *i*.

For both means and proportions, no qualifying language was used in describing significant group differences when the estimated effect size exceeded .1. However, when a significant difference was found but the estimated effect size was less than .1, the qualifier *somewhat* was used. For example, if the mean proficiency for females was significantly higher than that for males but the estimated effect size of the difference was less than .1, females were described as performing *somewhat higher* than males.

10.5.7 Descriptions of the Magnitude of Percentage

Percentages reported in the text of the state reports are sometimes described using quantitative words or phrases. For example, the number of students being taught by teachers with master's degrees in English might be described as "relatively few" or "almost all," depending on the size of the percentage in question. Any convention for choosing descriptive terms for the magnitude of percentages is to some degree arbitrary. The rules used to select the descriptive phrases in the report are given in Figure 10-1.

Figure 10-1
Rules for Descriptive Terms in State Reports

Percentage	Description of Text in Report
$p = 0$	None
$0 < p \leq 8$	A small percentage
$8 < p \leq 12$	Relatively few
$12 < p \leq 18$	Less than one fifth
$18 < p \leq 22$	About one fifth
$22 < p \leq 27$	About one quarter
$27 < p \leq 30$	Less than a third
$30 < p \leq 36$	About one third
$36 < p \leq 47$	Less than half
$47 < p \leq 53$	About half
$53 < p \leq 64$	More than half
$64 < p \leq 70$	About two thirds
$70 < p \leq 79$	About three quarters
$79 < p \leq 89$	A large majority
$89 < p < 100$	Almost all
$p = 100$	All

10.6 COMPARISONS OF 1994 AND 1992 RESULTS IN THE *FIRST LOOK REPORT*, THE *READING REPORT CARD* AND THE *CROSS-STATE DATA COMPENDIUM*

The *First Look Report*, the *Reading Report Card* and the *Cross-State Data Compendium* contain many tables that compare fourth-grade public-school results for 1994 with those obtained in 1992 for the nation as a whole, for each of the four regions of the country, and for each of the 38 jurisdictions that participated in both the 1994 and 1992 Trial State Assessments. The national and regional results are based on the 1992 and 1994 national NAEP public-school samples. The results for the Trial State jurisdictions are based on the 1994 and 1992 Trial State

Assessment samples. Each jurisdiction's overall results are compared, as well as the results for both primary and secondary NAEP reporting subgroups. The following statistics are compared:

- the proportions of examinees in the various primary and secondary reporting subgroups;
- average proficiencies, overall and for the primary reporting subgroups, on the composite scale, and the purpose-of-reading scales;
- selected percentiles (10th, 25th, 50th, 75th, 90th) overall, for the NAEP composite scale and for the purpose-of-reading scales; and
- proportions of students at or above the achievement levels, overall and within the primary reporting subgroups, on the composite scale.

A number of different types of tables are included in these reports. For example, one type of table shows the average composite proficiency and the percentage of students at or above each of the achievement levels. A second type of table shows the percentage of students at or above achievement levels on the composite scales for each of the primary reporting subgroups. A third type of table shows average reading proficiency and five percentile locations for each of the reading-purpose scales. A fourth type of table shows average composite proficiencies for a particular set of primary or secondary reporting subgroups.

Because of the large volume of tables in the *First Look*, *Reading Report Card* and the *Cross-State Data Compendium*, most were computer-generated. To help readers focus on important outcomes, each of the tables containing results for both 1994 and 1992 are annotated with symbols indicating which 1994-to-1992 jurisdiction comparisons represent statistically significant changes[7]. The annotations to these tables were made automatically by the computer programs that produced them and were based on tests of statistical significance and Bonferroni adjustments like those described in sections 10.5.1 and 10.5.2. This section describes the rules and conventions used by the computer programs in annotating the tables. These rules and conventions were chosen based on feasibility considerations and a desire to balance statistical power with Type I error control within these feasibility constraints.

Two types of annotations were made. The first type of annotation ("<" or ">") was used to indicate a gain or loss that was statistically significant considering each jurisdiction as a separate entity and controlling for the number of tests conducted in a particular table within that jurisdiction. Since all tables were set up with jurisdictions as the row variable, the first type of annotation was used on significance tests that *separately controlled the Type I error rate within each row of the table*. The second type of annotation ("< <" or "> >") was used to indicate a gain or loss that was statistically significant after *simultaneously controlling the Type I error rate for the number of tests conducted across all jurisdictions within a table*. As a result of this

[7] Fourth-grade public-school results from the national assessment for the nation as a whole and for each region of the country are also shown in these tables. However, significance testing and table annotation was not carried out for these results. Statistical tests and annotations of differences for the national assessment were included in tables from the *Reading Report Card* that contain only national results.

simultaneous error-rate control, the latter tests were extremely conservative and annotations of the second type were infrequent.

Many of the tables contain two or more types of statistics. For example, a very common table in the *Cross-State Data Compendium* contains, for both 1994 and 1992, the proportion of examinees in each of a particular set of reporting subgroups (e.g., males and females, or each of the race/ethnicity groups) *and* the average composite proficiency for each subgroup. In a table of this nature, two distinct families of significance tests were distinguished. The first family consisted of the comparisons of 1994 and 1992 proportions within each of the subgroups; the second consisted of the comparisons of 1994 and 1992 subgroup means. For each of these families, Type I error rates were controlled separately within-row (for the determining the first type of annotation) and simultaneously across jurisdictions (for the second type of annotation).

As a second example, a different table contained the percentage of students in the top one-third of the schools, the average composite proficiency of these students, and the percentage of these students at or above each of the achievement levels. In this example, three families of significance tests were distinguished—tests comparing percentages in the top-third schools, tests comparing the average proficiencies of these students, and tests comparing the percentages exceeding the achievement levels. Again, Type I error rates were controlled separately within-row (for determining the first type of annotation) and simultaneously across jurisdictions (for the second type of annotation) for each of these three families.

To illustrate the rules and conventions that were used, two specific examples will be considered. Table 10-2 is taken from an early version of Chapter 1 of the 1994 *Cross-State Data Compendium*. It shows the 1994 average fourth-grade composite proficiencies and the percentages associated with each of the achievement levels for the nation, each of the four regions, and each jurisdiction that participated in the 1994 assessment. The same statistics are given for 1992, with "---" used for the jurisdictions that participated in 1994 but did not take part in the 1992 assessment. Two families of significance tests were distinguished for this table. The first family involved comparisons of 1994 average proficiencies with those obtained in 1992. The second family involved comparisons of 1994 and 1992 percentages at or above each of the achievement levels.

For the first family of tests (i.e., comparisons of average proficiencies), the annotations based on within-row control of Type I error required no Bonferroni adjustment. A t-test was carried out comparing each jurisdiction's 1994 average to the corresponding 1992 average at an α level of .05. For the annotation based on simultaneous control of Type I error, the family size was taken to be 38 (the 38 jurisdictions that participated in both the 1994 and 1992 assessments). In other words, the t-test for each jurisdiction was carried out at the $\alpha = .05/38$ level of significance.

For the second family of tests (i.e., comparisons of the percentage of students at or above each of the achievement levels), within-row control of Type I error *did* require adjustments to significance levels. Within each jurisdiction, three tests were carried out—one

Table 10-2

TABLE 3 1992 READING ASSESSMENT

POPULATION: 1992 Grade 4 Public School Students

REPORTED STATISTICS: Average Overall Reading Proficiency and Percentage of Students by Achievement Levels

1992 NAEP grade 4 public school student reading performance...	Average Overall Reading Proficiency	At or Above Advanced Achievement Level	At or Above Proficient Achievement Level	At or Above Basic Achievement Level	Below Basic Achievement Level
1992 JURISDICTIONS	PROF (SE)	PCT (SE)	PCT (SE)	PCT (SE)	PCT (SE)
Nation					
Nation	215 (1.0)	6 (0.6)	27 (1.3)	60 (1.1)	40 (1.1)
Northeast	220 (3.9)	9 (2.6)	32 (4.7)	65 (3.9)	35 (3.9)
Southeast	211 (2.5)	4 (0.8)	22 (2.6)	55 (3.5)	45 (3.5)
Central	218 (1.5)	6 (1.2)	29 (2.4)	65 (1.9)	35 (1.9)
West	212 (1.6)	5 (0.7)	24 (1.8)	56 (1.9)	44 (1.9)
States					
Alabama	207 (1.7)	3 (0.4)	20 (1.5)	51 (2.1)	49 (2.1)
Arizona	209 (1.2)	3 (0.4)	21 (1.2)	54 (1.8)	46 (1.8)
Arkansas	211 (1.2)	4 (0.6)	23 (1.2)	56 (1.5)	44 (1.5)
California	202 (2.0)	4 (0.7)	19 (1.7)	48 (2.2)	52 (2.2)
Colorado	217 (1.1)	4 (0.6)	25 (1.4)	64 (1.6)	36 (1.6)
Connecticut	222 (1.3)	6 (1.0)	34 (1.4)	69 (1.7)	31 (1.7)
Delaware†	213 (0.6)	5 (0.5)	24 (1.1)	57 (1.2)	43 (1.2)
Florida	208 (1.2)	3 (0.4)	21 (1.1)	53 (1.6)	47 (1.6)
Georgia	212 (1.5)	5 (0.8)	25 (1.5)	57 (1.7)	43 (1.7)
Hawaii	203 (1.7)	3 (0.5)	17 (1.5)	48 (1.9)	52 (1.9)
Indiana	221 (1.3)	6 (0.9)	30 (1.5)	68 (1.6)	32 (1.6)
Iowa	225 (1.1)	7 (0.7)	36 (1.6)	73 (1.4)	27 (1.4)
Kentucky	213 (1.3)	3 (0.5)	23 (1.6)	58 (1.7)	42 (1.7)
Louisiana	204 (1.2)	2 (0.4)	15 (1.1)	46 (1.6)	54 (1.6)
Maine†	227 (1.1)	6 (0.8)	36 (1.7)	75 (1.4)	25 (1.4)
Maryland	211 (1.6)	4 (0.6)	24 (1.2)	57 (1.8)	43 (1.8)
Massachusetts	226 (0.9)	7 (0.8)	36 (1.5)	74 (1.3)	26 (1.3)
Minnesota	221 (1.2)	6 (0.7)	31 (1.5)	68 (1.7)	32 (1.7)
Mississippi	199 (1.3)	2 (0.4)	14 (0.9)	41 (1.7)	59 (1.7)
Missouri	220 (1.2)	6 (0.7)	30 (1.5)	67 (1.5)	33 (1.5)
Montana	--- (---)	--- (---)	--- (---)	--- (---)	--- (---)
Nebraska†	221 (1.1)	6 (0.7)	31 (1.5)	68 (1.5)	32 (1.5)
New Hampshire†	228 (1.2)	8 (1.1)	38 (1.6)	76 (1.8)	24 (1.8)
New Jersey†	223 (1.4)	8 (1.0)	35 (1.8)	69 (1.8)	31 (1.8)
New Mexico	211 (1.5)	4 (0.7)	23 (1.7)	55 (1.7)	45 (1.7)
New York†	215 (1.4)	5 (0.6)	27 (1.3)	61 (1.4)	39 (1.4)
North Carolina	212 (1.1)	5 (0.7)	25 (1.3)	56 (1.4)	44 (1.4)
North Dakota	226 (1.1)	6 (0.8)	35 (1.5)	74 (1.8)	26 (1.8)
Pennsylvania	221 (1.3)	6 (0.8)	32 (1.7)	68 (1.7)	32 (1.7)
Rhode Island	217 (1.8)	5 (0.7)	28 (1.7)	63 (2.2)	37 (2.2)
South Carolina	210 (1.3)	4 (0.7)	22 (1.4)	53 (1.9)	47 (1.9)
Tennessee	212 (1.4)	4 (0.7)	23 (1.5)	57 (1.7)	43 (1.7)
Texas	213 (1.6)	4 (0.7)	24 (1.8)	57 (2.0)	43 (2.0)
Utah	220 (1.1)	5 (0.6)	30 (1.6)	67 (1.6)	33 (1.6)
Virginia	221 (1.4)	6 (1.0)	31 (1.6)	67 (1.8)	33 (1.8)
Washington	--- (---)	--- (---)	--- (---)	--- (---)	--- (---)
West Virginia	216 (1.3)	5 (0.7)	25 (1.4)	61 (1.4)	39 (1.4)
Wisconsin	224 (1.0)	6 (0.6)	33 (1.3)	71 (1.3)	29 (1.3)
Wyoming	223 (1.1)	5 (0.6)	33 (1.5)	71 (1.6)	29 (1.6)
Other Jurisdictions					
DoDEA Overseas	--- (---)	--- (---)	--- (---)	--- (---)	--- (---)
Guam	182 (1.4)	1 (0.3)	8 (0.8)	28 (1.2)	72 (1.2)

(continued on next page)

--- Montana, Washington, and the DoDEA Overseas jurisdiction did not participate in the 1992 Trial State Assessment.

† Did not satisfy one of the guidelines for school sample participation rates for the 1992 Trial State Assessment (see *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading*).

SOURCE: National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments.

Table 10-2 (continued)

TABLE 3 — 1994 READING ASSESSMENT

POPULATION: 1994 Grade 4 Publ:c School Students

REPORTED STATISTICS: Average Overall Reading Proficiency and Percentage of Students by Achievement Levels

1994 NAEP grade 4 public school student reading performance...	Average Overall Reading Proficiency	At or Above Advanced Achievement Level	At or Above Proficient Achievement Level	At or Above Basic Achievement Level	Below Basic Achievement Level
1994 JURISDICTIONS	PROF (SE)	PCT (SE)	PCT (SE)	PCT (SE)	PCT (SE)
Nation					
Nation	212 (1.1)	7 (0.7)	28 (1.2)	59 (1.1)	41 (1.1)
Northeast	212 (2.2)	7 (1.5)	28 (2.6)	58 (2.3)	42 (2.3)
Southeast	208 (2.0)	6 (0.6)	23 (2.1)	53 (2.4)	47 (2.4)
Central	218 (2.7)	7 (1.4)	33 (2.8)	65 (3.0)	35 (3.0)
West	212 (2.2)	7 (0.8)	28 (2.0)	59 (2.2)	41 (2.2)
States					
Alabama	208 (1.5)	5 (0.7)	23 (1.3)	52 (1.6)	48 (1.6)
Arizona	206 (1.9)	6 (0.8)>	24 (1.5)	52 (1.9)	48 (1.9)
Arkansas	209 (1.7)	5 (0.6)	24 (1.4)	54 (1.8)	46 (1.8)
California	197 (1.8)<	3 (0.5)	18 (1.3)	44 (2.0)	56 (2.0)
Colorado	213 (1.3)	6 (0.7)	28 (1.5)	59 (1.4)	41 (1.4)
Connecticut	222 (1.6)	11 (1.1)>	38 (1.6)	68 (1.7)	32 (1.7)
Delaware	206 (1.1)<<	5 (0.8)	23 (1.1)	52 (1.3)<	48 (1.3)>
Florida	205 (1.7)	5 (0.6)>	23 (1.5)	50 (1.8)	50 (1.8)
Georgia	207 (2.4)	7 (1.0)	26 (2.0)	52 (2.3)	48 (2.3)
Hawaii	201 (1.7)	4 (0.5)	19 (1.4)	46 (1.8)	54 (1.8)
Indiana	220 (1.3)	7 (0.8)	33 (1.5)	66 (1.6)	34 (1.6)
Iowa	223 (1.3)	8 (1.0)	35 (1.5)	69 (1.6)	31 (1.6)
Kentucky	212 (1.6)	6 (0.8)>	26 (1.9)	56 (1.6)	44 (1.6)
Louisiana	197 (1.3)<<	2 (0.5)	15 (1.2)	40 (1.5)<	60 (1.5)>
Maine	228 (1.3)	10 (1.0)>	41 (1.5)	75 (1.6)	25 (1.6)
Maryland	210 (1.5)	7 (0.7)>	26 (1.4)	55 (1.6)	45 (1.6)
Massachusetts	223 (1.3)	8 (1.0)	36 (1.7)	69 (1.5)<	31 (1.5)>
Minnesota	218 (1.4)	7 (0.7)	33 (1.4)	65 (1.5)	35 (1.5)
Mississippi	202 (1.6)	4 (0.6)>	18 (1.3)>	45 (1.7)	55 (1.7)
Missouri	217 (1.5)	7 (0.9)	31 (1.6)	62 (1.8)	38 (1.8)
Montana†	222 (1.4)	7 (0.7)	35 (1.5)	69 (1.7)	31 (1.7)
Nebraska†	220 (1.5)	8 (0.9)	34 (1.8)	66 (1.6)	34 (1.6)
New Hampshire†	223 (1.5)<	9 (1.0)	36 (1.6)	70 (1.9)	30 (1.9)
New Jersey	219 (1.2)	8 (0.8)	33 (1.6)	65 (1.5)	35 (1.5)
New Mexico	205 (1.7)<	4 (0.5)	21 (1.5)	49 (1.6)	51 (1.6)
New York	212 (1.4)	6 (0.8)	27 (1.5)	57 (1.7)	43 (1.7)
North Carolina	214 (1.5)	8 (0.8)	30 (1.7)	59 (1.5)	41 (1.5)
North Dakota	225 (1.2)	8 (0.8)	38 (1.5)	73 (1.4)	27 (1.4)
Pennsylvania†	215 (1.6)<	7 (0.8)	30 (1.3)	61 (1.6)<	39 (1.6)>
Rhode Island†	220 (1.3)	8 (1.0)	32 (1.4)	65 (1.6)	35 (1.6)
South Carolina	203 (1.4)<<	4 (0.6)	20 (1.3)	48 (1.5)	52 (1.5)
Tennessee†	2'3 (1.7)	6 (0.9)	27 (1.5)	58 (2.1)	42 (2.1)
Texas	212 (1.9)	6 (0.8)	26 (1.8)	58 (2.3)	42 (2.3)
Utah	217 (1.3)	6 (0.8)	30 (1.6)	64 (1.6)	36 (1.6)
Virginia	213 (1.5)<<	7 (0.7)	26 (1.7)	57 (1.8)<<	43 (1.8)>>
Washington	213 (1.5)	6 (0.7)	27 (1.2)	59 (1.6)	41 (1.6)
West Virginia	213 (1.1)	6 (0.6)	26 (1.4)	58 (1.4)	42 (1.4)
Wisconsin†	224 (1.1)	7 (0.7)	35 (1.6)	71 (1.6)	29 (1.6)
Wyoming	221 (1.2)	6 (0.6)	32 (1.4)	68 (1.7)	32 (1.7)
Other Jurisdictions					
DoDEA Overseas	218 (0.9)	6 (0.7)	28 (1.1)	63 (1.5)	37 (1.5)
Guam	181 (1.2)	1 (0.3)	8 (0.8)	27 (1.1)	73 (1.1)

<< The value for 1994 was significantly lower (>> higher) than the value for 1992 at or about the 95 percent confidence level. These notations indicate statistical significance from a multiple comparison procedure based on 38 jurisdictions participating in both 1992 and 1994. If looking at only one state, < indicates the value for 1994 was significantly lower (> higher) than the value for 1992 at or about the 95 percent confidence level.

† Did not satisfy one of the guidelines for school sample participation rates for the 1994 Trial State Assessment (see Appendix A).

SOURCE: National Assessment of Educational Progress (NAEP) 1992 and 1994 Reading Assessments.

test for each of three of the achievement levels: At or Above Advanced, At or Above Proficient, and At or Above Basic[8]. Hence, the first type of annotations was based on t-tests carried out at the $\alpha = .05/3$ level. Across all jurisdictions there were 114 total comparisons (38 jurisdictions times 3 achievement levels). Hence, the annotations based on simultaneous control of Type I error were based on t-tests conducted at the $\alpha = .05/114$ level.

Table 10-3 is taken from an early version of Chapter 2 of the *Cross-State Data Compendium*. For each jurisdiction it contains the 1994 and 1992 percentages of fourth-grade examinees in each race/ethnicity subgroup and, where minimum school and student sizes were obtained, the 1994 and 1992 average composite proficiencies. Again, two families of significance tests were distinguished—tests comparing subgroup percentages and tests comparing subgroup means. For the first family of tests (i.e, comparisons of percentages within each race/ethnicity subgroup), five tests were carried out (one for each race/ethnicity group). Hence, the first type of annotation was based on t-tests carried out at the $\alpha = .05/5$ level and the second type of annotation was based on t-tests carried out at the $\alpha = .05/190$ (i.e., 38 jurisdictions times 5 race/ethnicity groups). For the second family of tests (i.e., comparisons of the subgroup average proficiencies), within-row control of Type I error required adjustments to significance levels based on the number of race/ethnicity groups exceeding minimum sample sizes. The annotations based on simultaneous control of Type I error required adjustments based on the number of subgroups across all 38 jurisdictions with minimum sample sizes of 62.

[8] Testing the percentage Below Basic is isomorphic to testing the percentage At or Above Basic. Therefore, it need not be counted as a distinct significance test.

Table 10-3

TABLE 4 — 1992 READING ASSESSMENT

POPULATION: 1992 Grade 4 Public School Students

REPORTED STATISTICS: Percentage of Students and Average Overall Reading Proficiency

BREAKDOWNS BY: Race/Ethnicity*

Which best describes your race or your ethnic background?	White		Black		Hispanic		American Indian	
1992 JURISDICTIONS	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Nation								
Nation	69 (0.5)	223 (1.3)	17 (0.4)	192 (1.6)	10 (0.3)	199 (2.2)	2 (0.3)	205 (4.9)
Northeast	68 (3.4)	229 (3.9)	20 (3.2)	197 (3.8)	9 (1.3)	200 (4.9)	1 (0.4)	*** (***)
Southeast	63 (2.7)	220 (3.4)	29 (2.6)	194 (2.4)	5 (1.1)	194 (5.0)!	1 (0.4)	*** (***)
Central	79 (1.5)	224 (1.8)	11 (1.3)	187 (3.3)	7 (1.0)	209 (4.7)	2 (0.4)	*** (***)
West	65 (2.1)	220 (1.7)	11 (1.6)	185 (4.4)	16 (1.9)	196 (2.7)	2 (0.6)	*** (***)
States								
Alabama	61 (2.4)	218 (1.5)	31 (2.2)	188 (2.2)	5 (0.7)	190 (3.7)	2 (0.7)	*** (***)
Arizona	56 (1.9)	220 (1.1)	4 (0.6)	200 (4.3)	29 (1.6)	198 (2.0)	10 (1.8)	185 (3.1)
Arkansas	70 (1.8)	219 (1.1)	21 (1.5)	190 (1.7)	7 (0.7)	188 (3.8)	2 (0.3)	206 (4.8)
California	46 (1.9)	218 (2.0)	7 (0.8)	184 (3.2)	35 (1.6)	183 (2.7)	2 (0.3)	*** (***)
Colorado	70 (1.3)	222 (1.1)	4 (0.9)	202 (3.4)!	21 (0.9)	202 (1.9)	2 (0.3)	203 (4.7)
Connecticut	73 (1.7)	230 (1.0)	11 (1.3)	196 (3.1)	13 (1.1)	193 (3.4)	1 (0.3)	*** (***)
Delaware†	64 (1.1)	222 (0.8)	25 (1.0)	195 (1.6)	8 (0.5)	188 (3.2)	2 (0.4)	*** (***)
Florida	57 (1.9)	219 (1.1)	21 (2.0)	186 (2.7)	18 (1.4)	201 (2.7)	2 (0.3)	*** (***)
Georgia	57 (1.9)	224 (1.4)	34 (1.8)	196 (2.2)	5 (0.5)	192 (4.8)	1 (0.2)	*** (***)
Hawaii	20 (1.5)	215 (2.7)	5 (0.6)	192 (4.6)	11 (0.9)	193 (2.8)	2 (0.3)	*** (***)
Indiana	82 (1.4)	225 (1.2)	11 (1.4)	200 (2.3)	5 (0.6)	211 (3.7)	1 (0.3)	*** (***)
Iowa	88 (0.9)	227 (1.0)	3 (0.6)	209 (3.1)	6 (0.5)	211 (3.1)	1 (0.3)	*** (***)
Kentucky	86 (1.1)	215 (1.2)	9 (1.0)	197 (3.3)	3 (0.4)	195 (5.1)	1 (0.2)	*** (***)
Louisiana	51 (1.9)	216 (1.2)	41 (1.9)	191 (1.5)	5 (0.5)	188 (4.4)	1 (0.3)	*** (***)
Maine†	92 (0.6)	228 (1.1)	0 (0.1)	*** (***)	4 (0.7)	209 (3.2)	2 (0.3)	*** (***)
Maryland	60 (1.7)	221 (1.5)	29 (1.3)	193 (2.6)	6 (0.6)	197 (3.0)	1 (0.3)	*** (***)
Massachusetts	81 (1.2)	231 (0.9)	7 (0.6)	205 (2.7)	7 (0.6)	201 (2.2)	1 (0.2)	*** (***)
Minnesota	87 (1.2)	224 (1.1)	3 (0.5)	191 (5.9)	6 (0.6)	203 (3.5)	2 (0.2)	*** (***)
Mississippi	41 (2.0)	217 (1.4)	52 (2.2)	186 (1.6)	5 (1.0)	185 (3.7)	1 (0.3)	*** (***)
Missouri	7[illegible] (1.7)	226 (1.1)	14 (1.7)	196 (3.1)	5 (0.7)	202 (3.2)	2 (0.3)	*** (***)
Montana	— (—)	— (—)	— (—)	— (—)	— (—)	— (—)	— (—)	— (—)
Nebraska†	83 (1.2)	225 (1.2)	6 (0.6)	197 (3.2)	8 (1.1)	205 (2.9)	2 (0.3)	*** (***)
New Hampshire†	90 (1.0)	229 (1.2)	1 (0.2)	*** (***)	5 (0.6)	215 (3.1)	2 (0.3)	*** (***)
New Jersey†	[illegible] (2.2)	232 (1.4)	14 (1.6)	200 (2.7)	13 (1.4)	199 (2.8)	1 (0.2)	*** (***)
New Mexico	45 (2.0)	223 (1.8)	3 (0.4)	202 (5.6)	46 (1.7)	200 (1.5)	5 (1.2)	200 (3.8)!
New York†	61 (2.0)	226 (1.1)	14 (1.8)	202 (2.7)	20 (1.8)	187 (4.0)	2 (0.3)	*** (***)
North Carolina	63 (2.0)	221 (1.3)	28 (1.6)	194 (2.2)	5 (0.6)	192 (3.5)	3 (1.2)	204 (6.2)!
North Dakota	93 (1.1)	226 (1.1)	0 (0.1)	*** (***)	3 (0.5)	221 (4.8)	3 (0.8)	211 (4.7)!
Pennsylvania	79 (1.7)	227 (1.2)	11 (1.5)	190 (2.4)	8 (1.0)	200 (3.8)	1 (0.2)	*** (***)
Rhode Island	76 (2.2)	224 (1.3)	6 (1.0)	187 (3.7)	12 (1.3)	191 (4.3)	2 (0.3)	*** (***)
South Carolina	55 (1.9)	221 (1.4)	38 (2.0)	195 (1.6)	5 (0.7)	195 (2.4)	2 (0.3)	*** (***)
Tennessee	71 (1.8)	219 (1.3)	21 (1.6)	193 (2.2)	5 (0.7)	196 (4.4)	2 (0.3)	*** (***)
Texas	49 (2.1)	224 (2.1)	14 (1.7)	200 (2.5)	34 (2.3)	201 (1.8)	1 (0.2)	*** (***)
Utah	86 (1.1)	223 (1.0)	1 (0.1)	*** (***)	10 (0.9)	204 (2.3)	2 (0.5)	*** (***)
Virginia	67 (1 6)	228 (1.5)	24 (1.3)	203 (2.1)	5 (0.5)	202 (4.3)	2 (0.3)	*** (***)
Washington	— (—)	— (—)	— (—)	— (—)	— (—)	— (—)	— (—)	— (—)
West Virginia	91 (0 7)	217 (1.2)	2 (0.4)	204 (6.4)	4 (0.5)	196 (6.9)	2 (0.3)	*** (***)
Wisconsin	83 (1 4)	227 (1.0)	6 (0.8)	200 (2.4)	8 (0.9)	210 (3.3)	2 (0.8)	206 (5 0)!
Wyoming	83 (1.3)	226 (1.1)	1 (0.1)	*** (***)	12 (0.9)	209 (2.5)	4 (0.9)	211 (4.6)!
Other Jurisdictions								
DoDEA Overseas	— (—)	— (—)	— (—)	— (—)	— (—)	— (—)	— (—)	— (—)
Guam	12 (0 8)	195 (3.1)	4 (0.4)	166 (5.5)	18 (0.8)	165 (2.9)	1 (0.3)	*** (***)

(continued on next page)

* Due to significant changes in wording of the race/ethnicity question between the 1992 and 1994 assessments, the 1992 results for Asian and Pacific Islander students are not comparable to 1994 results. Therefore, 1992 results for these two subgroups are not presented. Also, the percentages for race/ethnicity may not add to 100% because a small percentage of students categorized themselves as "other."

-- Montana, Washington, and the DoDEA Overseas jurisdiction did not participate in the 1992 Trial State Assessment.

*** Sample size in the 1992 or 1994 assessment is insufficient to permit a reliable estimate.

! Interpret with caution -- the nature of the sample does not allow accurate determination of the variability of this statistic. For this reason, statistical comparisons between the 1992 and 1994 assessments were not conducted.

† Did not satisfy one of the guidelines for school sample participation rates for the 1992 Trial State Assessment (see *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading*).

SOURCE: National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments.

Table 10-3 (continued)

TABLE 4 — 1994 READING ASSESSMENT

POPULATION: 1994 Grade 4 Public School Students

REPORTED STATISTICS: Percentage of Students and Average Overall Reading Proficiency

BREAKDOWNS BY: Race/Ethnicity*

Which best describes your race or your ethnic background?	White		Black		Hispanic		American Indian	
[illegible]								
Nation								
Nation	68 (0.5)	223 (1.3)	16 (0.4)	186 (1.7)<	12 (0.3)>	188 (2.7)<	2 (0.1)	200 (3.6)
Northeast	62 (2.4)	224 (2.5)	22 (2.5)	184 (2.1)<	10 (1.4)	191 (4.2)	1 (0.3)	*** (***)
Southeast	63 (3.6)	219 (2.4)	26 (2.9)	188 (2.5)	8 (1.2)	184 (4.1)	1 (0.3)	*** (***)
Central	80 (2.2)	225 (2.8)	11 (1.6)	182 (6.4)	6 (0.8)	199 (6.7)	1 (0.3)	*** (***)
West	66 (2.0)	222 (2.0)	7 (1.4)	186 (4.8)!	20 (1.5)	186 (4.4)	2 (0.3)	*** (***)
States								
Alabama	62 (1.7)	220 (1.5)	29 (1.6)	188 (1.9)	6 (0.6)	178 (4.3)	2 (0.4)	*** (***)
Arizona	58 (1.9)	220 (1.6)	4 (0.4)	183 (5.7)	29 (1.6)	188 (2.6)<	8 (1.4)	181 (5.1)
Arkansas	70 (1.7)	218 (1.7)	21 (1.6)	183 (2.3)<	6 (0.7)	192 (4.2)	2 (0.3)	*** (***)
California	44 (2.3)	211 (2.0)	7 (1.0)	·182 (4.9)	33 (1.9)	174 (2.4)<	2 (0.4)	*** (***)
Colorado	67 (1.4)	222 (1.3)	5 (0.7)	191 (4.7)	21 (1.1)	193 (2.1)<	4 (0.4)	204 (5.2)
Connecticut	70 (1.4)	234 (1.3)	12 (1.1)	190 (4.8)	14 (1.1)	190 (3.9)	1 (0.2)	*** (***)
Delaware	63 (1.1)	215 (1.3)<<	23 (1.0)	188 (2.4)<	9 (0.6)	190 (3.1)	3 (0.4)	*** (***)
Florida	57 (1.8)	218 (1.6)	21 (1.8'	'83 (2.4)	19 (1.6)	189 (3.1)<	2 (0.2)	*** (***)
Georgia	56 (2.6)	222 (1.9)	32 (2..	.d5 (3.2)<	9 (0.8)>	184 (5.7)	1 (0.2)	*** (***)
Hawaii	17 (1.1)	219 (2.1)	3 (0.5)	189 (4.5)	11 (0.8)	185 (4.0)	2 (0.2)	*** (***)
Indiana	81 (1.1)	225 (1.4)	10 (0.8)	193 (2.5)	7 (0.7)	201 (3.5)	1 (0.3)	*** (***)
Iowa	88 (1.1)	225 (1.2)	3 (0.6)	186 (7.0)!	6 (0.7)	204 (4.1)	2 (0.3)	*** (***)
Kentucky	83 (1.2)	215 (1.6)	10 (1.0)	190 (3.4)	5 (0.6)	196 (4.1)	1 (0.2)	*** (***)
Louisiana	51 (1.8)	213 (1.4)	38 (1.9)	180 (1.6)<<	8 (0.9)>	175 (5.0)	2 (0.3)	*** (***)
Maine	92 (0.6)	229 (1.3)	1 (0.2)	*** (***)	5 (0.4)	218 (4.8)	2 (0.3)	*** (***)
Maryland	57 (1.8)	223 (1.5)	32 (1.8)	185 (2.3)	6 (0.7)	197 (3.5)	2 (0.3)	*** (***)
Massachusetts	77 (1.6)	231 (1.2)	7 (1.0)	199 (3.1)	11 (0.8)>>	194 (2.8)	2 (0.3)	*** (***)
Minnesota	84 (1.1)	222 (1.1)	3 (0.5)	173 (8.0)	8 (0.6)>	202 (4.4)	3 (0.5)	196 (6.7)
Mississippi	46 (1.7)	220 (2.0)	45 (1.8)<	187 (2.1)	7 (0.8)	181 (3.9)	1 (0.3)	*** (***)
Missouri	75 (2.1)	223 (1.3)	14 (1.7)	192 (4.1)	7 (0.7)	200 (3.9)	2 (0.3)	212 (4.9)
Montana†	79 (1.8)	226 (1.3)	1 (0.2)	*** (***)	10 (0.8)	208 (3.2)	9 (1.3)	203 (2.8)
Nebraska†	82 (1.8)	224 (1.4)	4 (1.1)	190 (5.5)!	10 (1.4)	205 (3.9)	3 (0.4)	202 (6.2)
New Hampshire†	91 (1.1)	224 (1 5)	1 (0.2)	*** (***)	5 (0.7)	213 (4.8)	2 (0.6)	*** (***)
New Jersey	80 (1.9)	231 (1.2)	16 (1.9)	193 (3.4)	17 (1.5)	200 (2.5)	1 (0.2)	*** (***)
New Mexico	41 (1.8)	219 (1.7)	3 (0.5)	196 (7.0)	44 (1.4)	196 (2.2)	10 (1.6)>	185 (5.3)
New York	54 (2.2)	226 (1.7)	21 (1.7)>	191 (1.9)<	19 (1.5)	193 (2.6)	2 (0.3)	*** (***)
North Carolina	65 (2.1)	225 (1.6)	26 (1.6)	193 (1.9)	4 (0.5)	189 (4.4)	3 (1.2)	201 (4.1)!
North Dakota	88 (1.4)<	228 (1.2)	1 (0.2)>	*** (***)	6 (0.6)>	212 (2.9)	4 (1.1)	197 (6.2)!
Pennsylvania†	76 (1.9)	224 (1.3)	14 (1.9)	180 (3.8)	7 (0.7)	187 (3.9)	1 (0.3)	*** (***)
Rhode Island†	80 (1.1)	226 (1.4)	6 (0.6)	197 (2.4)	9 (0.8)	195 (2.8)	1 (0.2)	*** (***)
South Carolina	53 (1.8)	219 (1 4)	37 (1.5)	184 (1.7)<<	8 (0.7)	182 (3.3)<	2 (0.3)	*** (***)
Tennessee†	74 (1.8)	220 (1.8)	19 (1.7)	188 (3.0)	4 (0.6)	196 (6.7)	1 (0.3)	*** (***)
Texas	50 (2.0)	227 (1.7)	12 (1.9)	191 (4.4)	34 (2.3)	198 (1.9)	1 (0.3)	*** (***)
Utah	82 (1.2)	221 (1.3)	1 (0.1)	*** (***)	12 (0.9)	199 (2.5)	3 (0.4)	195 (5.3)
Virginia	59 (2.0)<	224 (1.6)	29 (1.7)	192 (1.9)<<	7 (0.8)>	206 (3.4)	1 (0.2)	*** (***)
Washington	73 (1.7)	217 (1.5)	5 (0 8)	198 (3.1)	11 (1.1)	190 (3.6)	4 (0.4)	207 (4.2)
West Virginia	90 (0.8)	215 (1.0)	3 (0.5)	202 (4.2)	4 (0.5)	192 (4.8)	1 (0.2)	*** (***)
Wisconsin†	84 (1.4)	228 (1 1)	5 (0.9)	197 (3.5)	7 (0.8)	203 (4.3)	2 (0.4)	*** (***)
Wyoming	82 (1.6)	224 (1.2)	1 (0.2)	*** (***)	13 (1.0)	209 (3.1)	4 (1.0)	210 (3.3)!
Other Jurisdictions								
DoDEA Overseas	47 (1 1)	224 (1.2)	19 (0.7)	205 (1.9)	18 (0.9)	211 (1.7)	3 (0.4)	210 (4.2)
Guam	9 (0.6)<	192 (4.2)	4 (0.4)	171 (8.0)	18 (0.9)	171 (2.3)	1 (0 2)	*** (***)

(continued on next page)

* The percentages for race/ethnicity may not add to 100% because a small percentage of students categorized themselves as "other."

<< The value for 1994 was significantly lower (>> higher) than the value for 1992 at or about the 95 percent confidence level. These notations indicate statistical significance from a multiple comparison procedure based on 38 jurisdictions participating in both 1992 and 1994. If looking at only one state, < indicates the value for 1994 was significantly lower (> higher) than the value for 1992 at or about the 95 percent confidence level.

*** Sample size in the 1992 or 1994 assessment is insufficient to permit a reliable estimate.

! Interpret with caution -- the nature of the sample does not allow accurate determination of the variability of this statistic. For this reason, statistical comparisons between the 1992 and 1994 assessments were not conducted.

† Did not satisfy one of the guidelines for school sample participation rates for the 1994 Trial State Assessment (see Appendix A).

SOURCE: National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments.

Table 10-3 (continued)

TABLE 4 — 1994 READING ASSESSMENT

POPULATION: **1994 Grade 4 Public School Students**
REPORTED STATISTICS: **Percentage of Students and Average Overall Reading Proficiency**
BREAKDOWNS BY: **Race/Ethnicity***

Which best describes your race or your ethnic background?	Asian		Pacific Islander	
Nation				
Nation	2 (0.2)	231 (6.1)	1 (0.1)	216 (5.9)
Northeast	2 (0.6)	*** (***)	1 (0.3)	*** (***)
Southeast	1 (0.2)	*** (***)	0 (0.1)	*** (***)
Central	1 (0.3)	*** (***)	0 (0.2)	*** (***)
West	3 (0.6)	226 (7.0)!	1 (0.3)	*** (***)
States				
Alabama	1 (0.2)	*** (***)	0 (0.1)	*** (***)
Arizona	1 (0.2)	*** (***)	1 (0.2)	*** (***)
Arkansas	1 (0.2)	*** (***)	0 (0.1)	*** (***)
California	8 (1.1)	211 (6.0)	5 (1.0)	213 (4.5)!
Colorado	2 (0.3)	*** (***)	1 (0.2)	*** (***)
Connecticut	2 (0.3)	*** (***)	0 (0.1)	*** (***)
Delaware	1 (0.3)	*** (***)	0 (0.1)	*** (***)
Florida	1 (0.2)	*** (***)	1 (0.2)	*** (***)
Georgia	2 (0.3)	*** (***)	0 (0.1)	*** (***)
Hawaii	19 (1.3)	219 (2.6)	46 (1.6)	191 (2.0)
Indiana	1 (0.2)	*** (***)	0 (0.1)	*** (***)
Iowa	1 (0.2)	*** (***)	0 (0.1)	*** (***)
Kentucky	1 (0.1)	*** (***)	0 (0.1)	*** (***)
Louisiana	1 (0.7)	*** (***)	0 (0.0)	*** (***)
Maine	1 (0.1)	*** (***)	0 (0.1)	*** (***)
Maryland	3 (0.4)	232 (4.1)	1 (0.2)	*** (***)
Massachusetts	2 (0.6)	201 (9.2)!	0 (0.1)	*** (***)
Minnesota	2 (0.4)	*** (***)	0 (0.1)	*** (***)
Mississippi	0 (0.2)	*** (***)	0 (0.1	*** (***)
Missouri	1 (0.3)	*** (***)	0 (0.1)	*** (***)
Montana†	1 (0.1)	*** (***)	0 (0.1)	*** (***)
Nebraska†	1 (0.2)	*** (***)	1 (0.1)	*** (***)
New Hampshire†	1 (0.2)	*** (***)	0 (0.1)	*** (***)
New Jersey	4 (0.6)	237 (4.0)	1 (0.3)	*** (***)
New Mexico	1 (0.3)	*** (***)	0 (0.1)	*** (***)
New York	3 (0.5)	230 (6.3)	1 (0.2)	*** (***)
North Carolina	1 (0.3)	*** (***)	0 (0.1)	*** (***)
North Dakota	1 (0.2)	*** (***)	0 (0.1)	*** (***)
Pennsylvania†	1 (0.4)	*** (***)	1 (0.2)	*** (***)
Rhode Island†	3 (0.4)	203 (5.8)	0 (0.2)	*** (***)
South Carolina	0 (0.1)	*** (***)	1 (0.2)	*** (***)
Tennessee†	1 (0.2)	*** (***)	0 (0.1)	*** (***)
Texas	2 (0.4)	*** (***)	0 (0.2)	*** (***)
Utah	1 (0.2)	*** (***)	1 (0.3)	*** (***)
Virginia	2 (0.4)	*** (***)	1 (0.2)	*** (***)
Washington	4 (0.7)	220 (5.7)	2 (0.4)	208 (6.2)
West Virginia	1 (0.2)	*** (***)	0 (0.1)	*** (***)
Wisconsin†	2 (0.5)	*** (***)	0 (0.1)	*** (***)
Wyoming	1 (0.1)	*** (***)	0 (0.1)	*** (***)
Other Jurisdictions				
DoDEA Overseas	5 (0.5)	222 (3.6)	5 (0.6)	215 (3.8)
Guam	3 (0.4)	180 (6.0)	64 (0.9)	183 (1.3)

* Due to significant changes in wording of the race/ethnicity question between the 1992 and 1994 assessments, the 1992 results for Asian and Pacific Islander students are not comparable to 1994 results. Therefore, 1992 results for these two subgroups are not presented. The percentages for race/ethnicity may not add to 100% because a small percentage of students categorized themselves as "other."

*** Sample size in the 1992 or 1994 assessment is insufficient to permit a reliable estimate.

! Interpret with caution -- the nature of the sample does not allow accurate determination of the variability of this statistic.

† Did not satisfy one of the guidelines for school sample participation rates for the 1994 Trial State Assessment (see Appendix A).

SOURCE: National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments.

APPENDIX A

PARTICIPANTS IN THE OBJECTIVES AND ITEM DEVELOPMENT PROCESS

APPENDIX A

PARTICIPANTS IN THE OBJECTIVES AND ITEM DEVELOPMENT PROCESS

PROJECT STEERING COMMITTEE

American Association of
School Administrators
Gary Marx, Associate
Executive Director
Arlington, Virginia

American Educational
Research Association
Carole Perlman, Director
of Research and Evaluation
Chicago, Illinois

American Federation of Teachers
Marilyn Rauth, Director
Educational Issues
Washington, D.C.

Association of State Assessment
Programs
Edward Roeber, Co-Chairman
Lansing, Michigan

Association of Supervision
and Curriculum Development
Helene Hodges
Alexandria, Virginia

Council of Chief State
School Officers
H. Dean Evans, Superintendent
of Public Instruction
State Department of Education
Indianapolis, Indiana

National Alliance of Business
Esther Schaeffer
Washington, D.C.

National Association of
Elementary School Principals
Kathleen Holliday, Principal
Potomac, Maryland

National Educational Association
Ann Smith, NEA Board Member
Ormond Beach, Florida

National Governors' Association
Mike Cohen
Washington, D.C.

National Parent Teacher Association
Ann Kahn
Alexandria, Virginia

National Education of Secondary
School Principals
Scott Thompson, Executive Director
Reston, Virginia

National School Board Association
Harriet C. Jelnek, Director
Rhineland, Wisconsin

National Association of Test Directors
Paul Le Mahieu
Pittsburgh, Pennsylvania

National Catholic Educational Association
Brother Robert Kealey
Washington, D.C.

PROJECT PLANNING COMMITTEE

Marilyn Adams
BBN Laboratories
Cambridge, Massachusetts

Marsha Delain
South Carolina
Department of Education
Columbia, South Carolina

Lisa Delpit
Institute for Urban Research
Morgan State University
Baltimore, Maryland

William Feehan
Chase Manhattan Bank
New York, New York

Philip Gough
Department of Psychology
University of Texas at Austin
Austin, Texas

Edward Haertel
Stanford University
Stanford, California

Elfrieda Hiebert
School of Education
University of Colorado
Boulder, Colorado

Judith Langer
School of Education
State University of New York, Albany
Albany, New York

P. David Pearson
University of Illinois
College of Education
Champaign, Illinois

Charles Peters
Oakland Schools
Pontiac, Michigan

John P. Pikulski
College of Education
University of Delaware
Newark, Delaware

Keith Stanovich
Oakland University
Rochester, Michigan

Paul Randy Walker
Maine Department of Education
Augusta, Maine

Sheila Valencia
University of Washington
Seattle, Washington

Janet Jones
Charles County Public Schools
Waldorf, Maryland

1992 NAEP READING CCSSO PROJECT STAFF

Ramsay W. Selden, Director
State Education Assessment Center
Council of Chief State
School Officers

Barbara Kapinus
Project Coordinator

Diane Schilder
Project Associate

THE 1994 READING ITEM DEVELOPMENT COMMITTEE

Dr. Katherine H. Au
Honolulu, HI

Carmela Cocola
Yardley, PA

Dr. Janice Dole
University of Utah
Salt Lake City, UT

Dr. Alan Farstrup
Executive Director
International Reading Association
Newark, DE

Herberto Godina
Champaign, IL

Dr. Susan Hynds
Syracuse University
Syracuse, NY

Dr. Barbara Kapinus
Council of Chief State School Officers
Washington, DC

Dr. Judith Langer
State University of New York - Albany
Albany, NY

Dr. Susan Neuman
Temple University
Philadelphia, PA

Dr. David Pearson
University of Illinois
Champaign, IL

Dr. Jesse Perry
San Diego, CA

Dr. John Pikulski
University of Delaware
Newark, DE

Dr. Timothy Shanahan
University of Illinois at Chicago
Chicago, IL

Laura Tsosie
Pinon, AZ

APPENDIX B

SUMMARY OF PARTICIPATION RATES

Appendix B

SUMMARY OF PARTICIPATION RATES

Guidelines for Sample Participation and Explanation of the Derivation of Weighted Participation Rates for the 1994 Trial State Reading Assessment

Introduction

Since 1989, state representatives, the National Assessment Governing Board (NAGB), several committees of advisors external to the National Assessment of Educational Progress (NAEP), and the National Center for Education Statistics (NCES) have engaged in numerous discussions about the procedures for reporting the NAEP Trial State Assessment results. These discussions have continued and been extended in light of the addition of nonpublic school samples to the 1994 NAEP Trial State Assessment.

From the outset of these discussions, it was recognized that sample participation rates across jurisdictions have to be uniformly high to permit fair and valid comparisons. Therefore, NCES established four guidelines for school and student participation in the 1990 Trial State Assessment Program. Participation rate data were first presented in the appendix of the 1990 composite mathematics report (*The State of Mathematics Achievement*) and a notation was made in those appendix tables and in Table 2 of the appropriate state report for any jurisdiction with participation levels that did not meet the original NCES guidelines. Virtually every jurisdiction met or exceeded the four guidelines for the 1990 program.

For the 1992 Trial State Assessment Program, NCES continued to use four guidelines, the first two relating to school participation and the second two relating to student participation. Three of the guidelines for the 1992 program were identical to those used in 1990, while one guideline for school participation was slightly modified. After reviewing the policy of how participation rates should best be presented so that readers of reports could accurately assess the quality of the data being reported, NCES and NAGB decided that for reporting the results from the 1992 Trial State Assessment Program, tables again would have notations for the jurisdictions not meeting each guideline. They also decided that there would be a fuller discussion in the body of the 1992 composite reports and in the Trial State Assessment Program technical reports about the participation rates and nature of the samples for each of the participating jurisdictions.

The 1994 Trial State Assessment Program uses a set of guidelines that have been expanded in two ways. First, new guidelines were designed to preempt publication of results from jurisdictions for which participation rates suggest the possibility of appreciable nonresponse bias. The new guidelines are congruent both with NAGB policies as well as the resolutions of the Education Information Advisory Committee (EIAC). Second, existing guidelines have been extended to cover the presence of separate public and nonpublic school samples in the 1994 Trial State Assessment Program. Guidelines 4, 6, 8, and 10 are similar to the guidelines used for 1992 reading and mathematics state assessments.

This appendix provides:

- Participation rate information for the 1994 Trial State Assessment of reading at grade 4 for both public and nonpublic school samples. This information also appears in an appendix in the *NAEP 1994 Reading Report Card for the Nation and the States*.

- An explanation of the guidelines and notations used in 1994. In brief, the guidelines cover levels of school and student participation, both overall and for particular population classes, separately for both public and nonpublic school samples. Consistent with the NCES standards, weighted data are used to calculate all participation rates for sample surveys, and weighted rates are provided in the reports. The procedures used to derive the weighted school and student participation rates are provided immediately after the discussion of the guidelines and notations.

- A set of tables that provides the 1994 participation rate information for the 1994 Trial State Assessment of Reading. Separate information is provided for the public and nonpublic school samples. The nonpublic school classification includes schools not directed by traditional local or state government agencies. In this context the collection of nonpublic schools includes schools administered by Catholic dioceses and other religious and nonsectarian schools. Domestic schools administered by the Department of Defense[1], and schools administered by the Bureau of Indian Affairs were not classified in either public or nonpublic categories. Because the aggregation of either public or nonpublic school students across all participating jurisdictions is not representative of any meaningful sample, weighted participation rates across participating jurisdictions have not been analyzed. However, the national and regional counts from the national assessment have been included and do provide some context for interpreting the summary of activities in each individual state and territory and for each type of school. Results, for the BIA and domestic DoD schools were included in the overall national and regional results.

Notations for Use in Reporting School and Student Participation Rates

Unless the overall participation rate is sufficiently high for a jurisdiction, there is a risk that the assessment results for that jurisdiction are subject to appreciable nonresponse bias. Moreover, even if the overall participation rate is high, there may be significant nonresponse bias if the nonparticipation that does occur is heavily concentrated among certain types of schools or students. The following guidelines concerning school and student participation rates in the Trial State Assessment Program were established to address four significant ways in which

[1] In the 1994 Trial State Assessment Program, nondomestic (international) Department of Defense Educational Activity schools (DoDEA) were included for the first time. Together they comprise a separate jurisdiction. For all reports including 1994 state-level data, the nondomestic DoDEA schools will be treated as public schools.

nonresponse bias could be introduced into the jurisdiction sample estimates. The conditions that resulted in the publication of a jurisdiction's results are presented below. Also presented below are the conditions that resulted in a jurisdiction receiving a notation in the 1994 reports. Note that in order for a jurisdiction's results to be published with no notations, that jurisdiction musı satisfy all guidelines.

Guidelines on the Publication of NAEP Results

Guideline 1 - Publication of Public School Results:

A jurisdiction will have its public school results published in the 1994 NAEP reading reports if and only if its weighted participation rate for the initial sample of public schools is greater than or equal to 70 percent. Similarly, a jurisdiction will receive a separate NAEP State Report if and only if its weighted participation rate for the initial sample of public schools is greater than or equal to 70 percent.

Guideline 2 - Publication of Nonpublic School Results:

A jurisdiction will have its nonpublic school results published in the 1994 NAEP reading reports if and only if its weighted participation rate for the initial sample of nonpublic schools is greater than or equal to 70 percent AND meets minimum sample size requirements[1]. A jurisdiction eligible to receive a separate NAEP State Report under guideline 1 will have its nonpublic school results included in that report if and only if that jurisdiction's weighted participation rate for the initial sample of nonpublic schools is greater than or equal to 70 percent AND meets minimum sample size requirements. If a jurisdiction meets guideline 2 but fails to meet guideline 1, a separate state report will be produced containing only nonpublic school results.

Guideline 3 - Publication of Combined Public and Nonpublic School Results

A jurisdiction will have its combined results published in the 1994 NAEP reading reports if and only if both guidelines 1 and 2 are satisfied. Similarly, a jurisdiction eligible to receive a separate NAEP State Report under guideline 1 will have its combined results included in that report if and only if guideline 2 is also met.

Discussion: ıf a jurisdiction's public or nonpublic school participation rate for the initial sample of schools is below 70 percent there is a substantial possibility that bias will be introduced into the assessment results. This possibility remains even after making statistical adjustments to compensate for school nonparticipation. The likelihood remains that, in

[1] Minimum sample size requirements for reporting nonpublic school data consist of two components: (1) a school sample size of six or more participating schools and (2) an assessed student sample size of at least 62.

aggregate, the substitute schools are sufficiently dissimilar from the originals they are replacing and represent too great a proportion of the population to discount such a difference. Similarly, the assumptions underlying the use of statistical adjustments to compensate for nonparticipation are likely to be significantly violated if the initial response rate falls below the 70 percent level. Guidelines 1, 2, and 3 take this into consideration. These guidelines are congruent with current NAGB policy, which requires that data for jurisdictions that do not have a 70 percent before-substitution participation rate be reported "in a different format", and with the Education Information Advisory Committee (EIAC) resolution, which calls for data from such jurisdictions not to be published.

Guideline 4 - Notation for Overall Public School Participation Rate

A jurisdiction that meets guideline 1 will receive a notation if its weighted participation rate for the initial sample of public schools was below 85 percent <u>AND</u> the weighted public school participation rate after substitution was below 90 percent.

Guideline 5 - Notation for Overall Nonpublic School Participation Rate

A jurisdiction that meets guideline 2 will receive a notation if its weighted participation rate for the initial sample of nonpublic schools was below 85 percent <u>AND</u> the weighted nonpublic school participation rate after substitution was below 90 percent.

Discussion: For jurisdictions that did not use substitute schools, the participation rates are based on participating schools from the original sample. In these situations, the NCES standards specify weighted school participation rates of at least 85 percent to guard against potential bias due to school nonresponse. Thus, the first part of these guidelines, referring to the weighted school participation rate for the initial sample of schools, is in direct accordance with NCES standards.

To help ensure adequate sample representation for each jurisdiction participating in the 1994 Trial State Assessment Program, NAEP provided substitutes for nonparticipating public and nonpublic schools. When possible, a substitute school was provided for each initially selected school that declined participation before November 15, 1993. For jurisdictions that used substitute schools, the assessment results will be based on the student data from all schools participating from both the original sample and the list of substitutes (unless both an initial school and its substitute eventually participated, in which case only the data from the initial school will be used).

The NCES standards do not explicitly address the use of substitute schools to replace initially selected schools that decide not to participate in the assessment. However, considerable technical consideration was given to this issue. Even though the characteristics of the substitute schools were matched as closely as possible to the characteristics of the initially selected schools, substitution does not entirely eliminate bias due to the nonparticipation of initially selected schools. Thus, for the weighted school participation rates including substitute schools, the guidelines were set at 90 percent.

If a jurisdiction meets either standard (i.e., 85 percent or higher prior to substitution or 90 percent or higher after substitution) there will be no notation for the relevant overall school participation rate.

Guideline 6 - Notation for Strata-Specific Public School Participation Rate

A jurisdiction that is not already receiving a notation under guideline 4 will receive a notation if the nonparticipating public schools included a class of schools with similar characteristics, which together accounted for more than five percent of the jurisdiction's total fourth-grade weighted sample of public schools. The classes of schools from which a jurisdiction needed minimum school participation levels were determined by degree of urbanization, minority enrollment, and median household income of the area in which the school is located.

Guideline 7 - Notation for Strata-Specific Nonpublic School Participation Rate

A jurisdiction that is not already receiving a notation under guideline 5 will receive a notation if the nonparticipating nonpublic schools included a class of schools with similar characteristics, which together accounted for more than five percent of the jurisdiction's total fourth-grade weighted sample of nonpublic schools. The classes of schools from which a jurisdiction needed minimum school participation levels were determined by type of nonpublic school (Catholic versus non-Catholic) and location (metropolitan versus nonmetropolitan).

Discussion: The NCES standards specify that attention should be given to the representativeness of the sample coverage. Thus, if some important segment of the jurisdiction's population is not adequately represented, it is of concern, regardless of the overall participation rate.

These guidelines address the fact that, if nonparticipating schools are concentrated within a particular class of schools, the potential for substantial bias remains, even if the overall level of school participation appears to be satisfactory. Nonresponse adjustment cells for public schools have been formed within each jurisdiction, and the schools within each cell are similar with respect to minority enrollment, degree of urbanization, and/or median household income, as appropriate for each jurisdiction. For nonpublic schools, nonresponse adjustment cells are determined by type and location of school.

If more than five percent (weighted) of the sampled schools (after substitution) are nonparticipants from a single adjustment cell, the potential for nonresponse bias is too great. These guidelines are based on the NCES standard for stratum-specific school nonresponse rates.

Guideline 8 - Notation for Overall Student Participation Rate in Public Schools

A jurisdiction that meets guideline 1 will receive a notation if the weighted student response rate within participating public schools was below 85 percent.

Guideline 9 - Notation for Overall Student Participation Rate in Nonpublic Schools

A jurisdiction that meets guideline 2 will receive a notation if the weighted student response rate within participating nonpublic schools was below 85 percent.

Discussion: These guidelines follow the NCES standard of 85 percent for overall student participation rates. The weighted student participation rate is based on all eligible students from initially selected or substitute schools who participated in the assessment in either an initial session or a make-up session. If the rate falls below 85 percent, then the potential for bias due to students' nonresponse is too great.

Guideline 10 - Notation for Strata-Specific Student Participation Rate in Public Schools

A jurisdiction that is not already receiving a notation under guideline 8 will receive a notation if the nonresponding students within participating public schools included a class of students with similar characteristics, who together comprised more than five percent of the jurisdiction's weighted assessable public school student sample. Student groups from which a jurisdiction needed minimum levels of participation were determined by age of student and type of assessment session (unmonitored or monitored), as well as school level of urbanization, minority enrollment, and median household income of the area in which the school is located.

Guideline 11 - Notation for Strata-Specific Student Participation Rate in Nonpublic Schools

A jurisdiction that is not already receiving a notation under guideline 9 will receive a notation if the nonresponding students within participating nonpublic schools included a class of students with similar characteristics, who together comprised more than five percent of the jurisdiction's weighted assessable nonpublic school student sample. Student groups from which a jurisdiction needed minimum levels of participation were determined by age of student and type of assessment session (unmonitored or monitored), as well as type and location of school.

Discussion: These guidelines address the fact that if nonparticipating students are concentrated within a particular class of students, the potential for substantial bias remains, even if the overall student participation level appears to be satisfactory. Student nonresponse adjustment cells have been formed using the school-level nonresponse adjustment cells, together with the student's age and the nature of the assessment session (unmonitored or monitored). If more than five percent (weighted) of the invited students who do not participate in the assessment are from a single adjustment cell, then the potential for nonresponse bias is too great. These guidelines are based on the NCES standard for stratum-specific student nonresponse rates.

Derivation of Weighted Participation Rates

Weighted School Participation Rates. The weighted school participation rates within each jurisdiction provide the percentages of fourth-grade students in public (nonpublic) schools who are represented by the schools participating in the assessment, prior to statistical adjustments for school nonresponse.

Two sets of weighted school participation rates are computed for each jurisdiction, one for public schools and one for nonpublic schools. Each set consists of two weighted participation rates. The first is the weighted participation rate for the initial sample of schools. This rate is based only on those schools that were initially selected for the assessment. The numerator of this rate is the sum of the number of students represented by each initially selected school that participated in the assessment. The denominator is the sum of the number of students represented by each of the initially selected schools found to have eligible students enrolled. This includes both participating and nonparticipating schools.

The second participation rate is the weighted participation rate after substitution. The numerator of this rate is the sum of the number of students represented by each of the participating schools, whether originally selected or a substitute. The denominator is the same as that for the weighted participation rate for the initial sample. This means that, for a given jurisdiction and type of school, the weighted participation rate after substitution is always at least as great as the weighted participation rate for the initial sample of schools.

In general, different schools in the sample can represent different numbers of students in the jurisdiction's population. The number of students represented by an initially selected school (the school weight) is the fourth-grade enrollment of the school divided by the probability that the school was included in the sample. For instance, a selected school with a fourth-grade enrollment of 150 and a selection probability of 0.2 represents 750 students from that jurisdiction. The number of students represented by a substitute school is the number of students represented by the replaced nonparticipating school.

Because each selected school represents different numbers of students in the population, the weighted school participation rates may differ somewhat from the simple unweighted rates. (The unweighted rates are calculated from the counts of schools by dividing the number of participating schools by the number of schools in the sample with eligible students enrolled.) The difference between the weighted and the unweighted rates is potentially largest in smaller jurisdictions where all schools with fourth-grade students were included in the sample. In those jurisdictions, each school represents only its own students. Therefore, the nonparticipation of a large school reduces the weighted school participation rate by a greater amount than does the nonparticipation of a small school.

The nonparticipation of larger schools also has greater impact than that of smaller schools on reducing weighted school participation rates in larger jurisdictions where fewer than all of the schools were included in the sample. However, since the number of students represented by each school is more nearly constant in larger states, the difference between the impact of nonparticipation by either large or small schools is less marked than in jurisdictions where all schools were selected.

In general, the greater the population in the jurisdiction, the smaller the difference between the weighted and unweighted school participation rates. However, even in the less populous jurisdictions, the differences tend to be small.

Weighted Student Participation Rate. The weighted student participation rate provides the percentage of the eligible student population from participating schools within the jurisdiction that are represented by the students who participated in the assessment (in either an initial session or a make-up session). Separate weighted student participation rates were calculated for public and nonpublic school students. The eligible student population from participating schools (public or nonpublic) within a jurisdiction consists of all students who were in the fourth grade, who attended a school that, if selected, would have participated and who, if selected, would not have been excluded from the assessment. The numerator of this rate is the sum, across all assessed students, of the number of students represented by each assessed student (prior to adjustment for student nonparticipation). The denominator is the sum of the number of students represented by each selected student who was invited and eligible to participate (i.e., not excluded), including students who did not participate. Thus, the denominator is an estimate of the total number of assessable students in the group of schools within the jurisdiction that would have participated if selected.

The number of students represented by a single selected student (the student weight) is 1.0 divided by the overall probability that the student was selected for assessment. In general, the number of students from a jurisdiction's population represented by a sampled student is approximately constant across students. Consequently, there is little difference between the weighted student participation rate and the unweighted student participation rate.

Weighted Overall School and Student Participation Rate. An overall indicator of the effect of nonparticipation by both students and schools is given by the overall participation rate. Separate overall rates were calculated for public and nonpublic school samples. For each school type (public or nonpublic), these weights were calculated as the product of the weighted school participation rate (after substitution), and the weighted student participation rate. For jurisdictions having a high overall participation rate the potential is low for bias to be introduced through either school nonparticipation or student nonparticipation. This rate provides a summary measure that indicates the proportion of the jurisdiction's fourth-grade public or nonpublic student population that is directly represented by the final student sample. When the overall rate is high, the adjustments for nonresponse that are used in deriving the final survey weights are likely to be effective in maintaining nonresponse bias at a negligible level. Conversely, when the overall rate is relatively low there is a greater chance that a nonnegligible bias remains even after making such adjustments.

The overall rate is not used in establishing the guidelines/notations for school and student participation, since guidelines exist already covering school and student participation separately.

Derivation of Weighted Percentages for Excluded Students

Weighted Percentage of Excluded Students. The weighted percentage of excluded students estimates the percentage of the fourth-grade population in the jurisdiction's schools that is represented by the students who were excluded from the assessment, after accounting for school nonparticipation. The numerator is the sum, across all excluded students, of the number of students represented by each excluded student. The denominator is the sum of the number of students represented by each of the students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Students with an Individualized Education Plan (IEP). The weighted percentage of IEP students estimates the percentage of the fourth-grade population in the jurisdiction's schools represented by the students who were classified as IEP, after accounting for school nonparticipation. The numerator is the sum, across all students classified as IEP, of the number of students represented by each IEP student. The denominator is the sum of the number of students represented by each of the students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Excluded IEP Students. The weighted percentage of excluded IEP students estimates the percentage of students in the jurisdiction who are represented by those IEP students excluded from the assessment, after accounting for school nonparticipation. The numerator is the sum, across all students classified as IEP and excluded from the assessment, of the number of students represented by each excluded IEP student. The denominator is the sum of the number of students represented by each of the IEP students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Limited English Proficiency (LEP) Students. The weighted percentage of LEP students estimates the percentage of the fourth-grade population in the jurisdiction's schools represented by the students who were classified as LEP, after accounting for school nonparticipation. The numerator is the sum, across all students classified as LEP, of the number of students represented by each LEP student. The denominator is the sum of the number of students represented by each of the students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Excluded LEP Students. The weighted percentage of LEP students who were excluded estimates the percentage of students in the jurisdiction represented by those LEP students excluded from the assessment, after accounting for school nonparticipation. The numerator is the sum, across all students classified as LEP and excluded from the assessment, of the number of students represented by each excluded LEP student. The denominator is the sum of the number of students represented by each of the LEP students who was sampled (and had not withdrawn from the school at the time of the assessment).

Note: All percentages are based on student weights that have been adjusted for school-level nonresponse. All weighted percentages were calculated separately for public and nonpublic school samples.

TABLE B-1 Public School Participation Rates, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD naep 1992 1994 Reading Assessment	Weighted School Participation Rate Before Substitution	Weighted School Participation Rate After Substitution	Number of Schools in Original Sample	Number of Non-eligible Schools	Number of Participating Schools from Original Sample	Number of Substituted Schools Provided	Number of Participating Substituted Schools	Total Number of Participating Schools
Nation								
Nation	86	87	267	1	225	8	2	227
Northeast	93	93	53	0	49	0	0	49
Southeast	91	93	67	0	60	2	1	61
Central	85	87	59	0	51	1	1	52
West	77	77	88	1	65	5	0	65
States								
Alabama	87	93	107	1	92	14	7	99
Arizona	99	99	107	2	104	1	0	104
Arkansas	86	94	109	6	89	9	8	97
Califomia	80*	91*	106	0	85	21	12	97
Colorado	100	100	108	0	108	0	0	108
Connecticut	96*	96*	105	1	101	3	0	101
Delaware	100	100	54	3	51	0	0	51
Florida	100	100	107	0	107	0	0	107
Georgia	99*	99*	107	2	105	0	0	105
Hawaii	99	99	106	1	104	0	0	104
Idaho	69	91	109	1	74	27	24	98
Indiana[1]	83*	92*	107	0	89	18	11	100
Iowa	85	99	110	2	92	16	15	107
Kentucky	88	96	107	2	93	11	8	101
Louisiana	100	100	105	2	103	0	0	103
Maine	94	97	116	9	101	4	3	104
Maryland	94	96	106	2	98	6	2	100
Massachusetts	97	97	105	3	99	3	0	99
Michigan	63*	80*	106	3	66	37	17	83
Minnesota[1]	86	95	107	2	90	14	10	100
Mississippi	95	99	105	1	99	4	4	103
Missouri	96	98	109	2	103	4	2	105
Montana	85	89	13[illegible]	6	105	23	6	111
Nebraska[6]	71	77	144	2	101	38	8	109
New Hampshire[4]	71	79	109	0	77	25	9	86
New Jersey[4]	85	91	107	2	89	15	7	96
New Mexico	100	100	108	3	105	0	0	105
New York	75	91	106	0	79	26	17	96
North Carolina	99	99	108	2	105	1	0	105
North Dakota	80	91	129	1	101	22	16	117
Pennsylvania	80*	84*	107	2	85	19	4	89
Rhode Island[4]	80	86	109	2	86	14	6	92
South Carolina[4]	95	97	106	1	100	4	2	102
Tennessee	72	74	106	3	74	27	2	76
Texas[4]	91*	93*	108	4	96	8	2	98
Utah	100	100	106	1	105	0	0	105
Virginia	98*	99*	107	2	104	1	1	105
Washington	100	100	106	2	104	0	0	104
West Virginia	99	100	112	1	110	1	1	111
Wisconsin	79*	86*	108	3	84	20	7	91
Wyoming[4]	98	98	121	5	112	4	0	112
Other Jurisdictions								
DoDEA Overseas	99	99	83	1	81	1	0	81
Guam	100	100	21	0	21	0	0	21

The numeric footnote references appearing on juridisdictions above and their explanations given below do not appear sequentially as they correspond to the relevant participation rate guideline numbers. Refer to the preceding text section of this appendix for more detailed information on the notations and guidelines about sample representativeness, and for the derivation of weighted participation. For De.aware and Guam, the Trial State Assessment was based on all eligible public schools (i.e., there was no sampling of schools).

[1] The state's public school weighted participation rate for the initial sample was less than 70%.

[4] The state's public school weighted participation rate for the initial sample of schools was below 85% *and* the weighted school participation rate after substitution was below 90%.

[6] The nonparticipating public schools included a class of schools with sim!!ar characteristics, which together accounted for more than five percent of the state's total fourth-grade weighted sample of public schools.

* In California, Connecticut, Georgia, Indiana, Michigan, Pennsylvania, Texas, Virginia, and Wisconsin, the materials from one school that conducted an assessment were lost in shipping. The school is included in the counts of participating schools, both before and after substitution. However, in the weighted results, the school is treated in the same manner as a nonparticipating school because no student responses were available for analysis and reporting.

SOURCE National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments

TABLE B-2 Nonpublic School Participation Rates, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD naep 1992 1994 Reading Assessment	Weighted School Participation Rate Before Substitution	Weighted School Participation Rate After Substitution	Number of Schools in Original Sample	Number of Non-eligible Schools	Number of Participating Schools from Original Sample	Number of Substituted Schools Provided	Number of Participating Substituted Schools	Total Number of Participating Schools
Nation								
Nation	87	87	118	13	89	0	0	89
Northeast	82	82	37	5	26	0	0	26
Southeast	90	90	22	2	17	0	0	17
Central	97	97	30	2	27	0	0	27
West	80	80	29	4	19	0	0	19
States								
Alabama	92	96	11	0	8	3	1	9
Arizona[2]	35	35	11	0	3	8	0	3
Arkansas	81	94	9	0	6	3	1	7
Califomia[2]	42	51	15	4	5	6	1	6
Colorado[5]	71	85	11	1	7	3	1	8
Connecticut[5]	73	82	17	1	11	4	2	13
Delaware[5]	73	73	34	3	22	5	0	22
Florida[2]	52	73	16	1	8	7	3	11
Georgia[2]	74	84	12	1	8	3	1	9
Hawaii[5]	80	88	24	2	17	5	2	19
Idaho[5]	89	89	8	0	7	1	0	7
Indiana	85	85	18	4	10	4	0	10
Iowa	100	100	17	1	16	0	0	16
Kentucky	70	85	14	0	10	4	2	12
Louisiana[5]	82	91	21	0	17	4	2	19
Maine[7]	79	100	12	4	7	1	1	8
Maryland	63	70	19	2	10	7	1	11
Massachusotts[2]	95	100	17	2	14	1	1	15
Michigan	0	0	20	3	0	17	0	0
Minnesota[2]	91	99	21	0	18	3	2	20
Mississippi	64	64	12	1	7	4	0	7
Missouri[2]	90	90	21	0	19	2	0	19
Montana	65	65	14	2	7	5	0	7
Nebraska[2]	48	48	24	0	11	11	0	11
New Hampshire[2]	54	54	13	2	5	6	0	5
New Jersey[2]	76	76	23	1	17	5	0	17
New Mexico[5]	100	100	14	5	9	0	0	9
New York	40	62	25	0	10	15	5	15
North Carolina[2]	32	32	9	2	2	5	0	2
North Dakota[2]	77	91	17	2	12	2	2	14
Pennsylvania	72	72	31	5	17	9	0	17
Rhode Island[5]	93	93	20	1	17	2	0	17
South Carolina	69	86	12	3	5	4	2	7
Tennessee[2]	41	41	11	1	4	6	0	4
Texas[2]	24	39	8	1	2	5	1	3
Utah[2]	23	23	7	1	1	5	0	1
Virginia	81	81	11	1	8	1	0	8
Washington[5]	0	0	14	0	0	14	0	0
West Virginia	86	86	11	2	7	2	0	7
Wisconsin	66	66	36	4	20	12	0	20
Wyoming	0	0	8	0	0	8	0	0
Other Jurisdiction†								
Guam[2]	96	96	11	0	9	0	0	9

The numeric footnote references appearing on jurisdictions above and their explanations given below do not appear sequentially as they correspond to the relevant participation rate guideline numbers. Refer to the preceding text section of this appendix for more detailed information on the notations and guidelines about sample represantativeness, and for the derivation of weighted participation. For Guam, the Trial State Assessment was based on all eligible public schools (i.e., there was no sampling of schools)

[2] The state's nonpublic school weighted participation rate for the initial sample was less than 70%

[5] The state's nonpublic school weighted participation rate for the initial sample of schools was below 85% *and* the weighted school participation rate after substitution was below 90%

[7] The nonparticipating nonpublic schools included a class of schools with similar characteristics, which together accounted for more than five percent of the state's total fourth-grade weighted sample of nonpublic schools

† DoDEA nonpublic school data do not exist because all non-domestic schools are considered public schools.

SOURCE National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments

TABLE B-3	Public School Student Participation Rates, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD naep 1992 1994 Reading Assessment	Weighted Student Participation Rate After Make-ups	Number of Students in Original Sample	Number of Students in Sample Supplement	Number of Withdrawn Students	Number of Excluded Students	Number of Students to be Assessed	Number of Students Assessed in Initial Sessions	Number of Students Assessed in Make-up Sessions	Total Number of Assessed Students
Nation									
Nation	95	6,957	*	*	501	6,456	5,965	65	6,030
Northeast	94	1,583	*	*	98	1,485	1,357	10	1,367
Southeast	95	1,876	*	*	106	1,770	1,620	29	1,649
Central	95	1,325	*	*	75	1,250	1,172	12	1,184
West	95	2,173	*	*	222	1,951	1,816	14	1,830
States									
Alabama	96	2,942	66	97	162	2,749	2,640	6	2,646
Arizona	94	3,044	181	215	204	2,806	2,625	26	2,651
Arkansas	96	2,845	76	113	169	2,639	2,525	10	2,535
California	94	2,851	32	82	404	2,397	2,240	12	2,252
Colorado	94	3,126	134	140	221	2,899	2,720	10	2,730
Connecticut	96	2,988	60	104	248	2,696	2,510	67	2,577
Delaware	96	2,518	67	89	153	2,343	2,208	31	2,239
Florida	94	3,188	165	186	326	2,841	2,634	32	2,666
Georgia	95	3,111	114	153	171	2,901	2,749	17	2,766
Hawaii	95	3,060	88	128	154	2,866	2,689	43	2,732
Idaho	96	2,869	110	132	145	2,702	2,590	8	2,598
Indiana	96	2,925	84	90	153	2,766	2,645	10	2,655
Iowa	96	3,003	107	86	140	2,884	2,755	4	2,759
Kentucky	97	2,999	103	139	114	2,849	2,744	14	2,758
Louisiana	96	3,028	114	135	181	2,826	2,698	15	2,713
Maine	94	2,857	47	50	275	2,579	2,429	7	2,436
Maryland	95	2,979	86	168	216	2,681	2,529	26	2,555
Massachusetts	95	2,871	50	47	236	2,638	2,504	13	2,517
Michigan	95	2,431	52	82	139	2,262	2,089	53	2,142
Minnesota	95	2,919	58	62	133	2,782	2,638	17	2,655
Mississippi	97	3,039	96	113	169	2,853	2,748	14	2,762
Missouri	95	2,987	105	120	156	2,816	2,652	18	2,670
Montana	96	2,704	64	57	93	2,618	2,490	11	2,501
Nebraska	95	2,623	72	61	114	2,520	2,383	12	2,395
New Hampshire	96	2,448	44	51	145	2,296	2,165	32	2,197
New Jersey	95	2,823	42	66	162	2,637	2,495	14	2,509
New Mexico	95	3,026	121	125	241	2,781	2,618	17	2,635
New York	95	2,885	41	79	227	2,620	2,448	47	2,495
North Carolina	96	3,144	88	105	174	2,953	2,800	32	2,832
North Dakota	97	2,678	55	43	59	2,631	2,543	1	2,544
Pennsylvania	94	2,580	36	47	143	2,426	2,259	31	2,290
Rhode Island	95	2,720	77	183	140	2,474	2,312	29	2,341
South Carolina	96	3,028	78	107	190	2,809	2,677	30	2,707
Tennessee	96	2,256	44	83	127	2,090	1,979	19	1,998
Texas	96	2,878	142	158	317	2,545	2,449	5	2,454
Utah	95	3,045	107	118	153	2,881	2,702	31	2,733
Virginia	95	3,099	95	105	216	2,873	2,660	59	2,719
Washington	94	3,069	94	109	158	2,896	2,715	22	2,737
West Virginia	96	3,078	87	78	213	2,874	2,727	30	2,757
Wisconsin	96	2,618	37	46	189	2,420	2,322	9	2,331
Wyoming	96	2,902	145	104	130	2,813	2,660	39	2,699
Other Jurisdictions									
DoDEA Overseas	95	2,783	184	301	117	2,549	2,403	10	2,413
Guam	96	2,502	123	108	220	2,297	2,203	0	2,203

Refer to the preceding text section of this appendix for more detailed information on the notations and guidelines about sample representativeness, and for the derivation of weighted participation

* For the 1994 national NAEP there was no supplementary student sample taken

SOURCE: National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments

TABLE B-4 — Nonpublic School Student Participation Rates, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD NAEP 1992 1994 Reading Assessment	Weighted Student Participation Rate After Make-ups	Number of Students in Original Sample	Number of Students in Sample Supplement	Number of Withdrawn Students	Number of Excluded Students	Number of Students to be Assessed	Number of Students Assessed in Initial Sessions	Number of Students Assessed in Make-up Sessions	Total Number of Assessed Students
Nation									
Nation	97	1,410	*	*	6	1,404	1,346	6	1,352
Northeast	98	465	*	*	3	462	447	2	449
Southeast	94	255	*	*	3	252	238	1	239
Central	97	402	*	*	0	402	384	3	387
West	98	288	*	*	0	288	277	0	277
States									
Alabama	95	216	4	8	4	208	199	0	199
Arizona	---	107	0	7	25	75	69	0	69
Arkansas	95	165	1	2	1	163	154	0	154
Califomia	97	155	2	4	0	153	149	0	149
Colorado	94	135	7	3	0	139	130	0	130
Connecticut	95	312	2	4	5	305	290	0	290
Delaware	98	557	4	3	0	558	544	0	544
Florida	98	270	6	3	1	272	263	4	267
Georgia	97	227	1	3	0	225	217	0	217
Hawaii	96	427	6	3	2	428	409	6	415
Idaho	96	96	2	0	0	98	91	3	94
Indiana	95	234	0	3	2	229	219	0	219
Iowa	99	305	32	3	3	331	327	0	327
Kentucky	97	268	29	10	0	287	278	0	278
Louisiana	97	472	7	5	2	472	457	0	457
Maine	95	90	1	1	0	90	85	0	85
Maryland	97	288	1	3	3	283	275	0	275
Massachusetts	96	322	0	1	7	314	302	0	302
Michigan	---	0	0	0	0	0	0	0	0
Minnesota	96	417	2	4	8	407	390	0	390
Mississippi	96	167	5	3	8	161	156	0	156
Missouri	95	398	1	7	1	391	370	2	372
Montana	94	157	3	3	0	157	148	0	148
Nebraska	97	217	1	0	0	218	211	0	211
New Hampshire	---	118	3	1	0	120	116	0	116
New Jersey	96	397	4	2	3	396	370	9	379
New Mexico	92	232	2	5	22	207	187	4	191
New York	96	384	8	3	7	382	364	5	369
North Carolina	---	52	0	0	0	52	49	0	49
North Dakota	93	272	10	5	7	270	253	0	253
Pennsylvania	94	455	2	1	2	454	424	3	427
Rhode Island	96	372	1	4	1	368	354	0	354
South Carolina	98	162	1	3	0	160	156	0	156
Tennessee	---	89	0	0	0	89	82	1	83
Texas	---	80	0	1	0	79	79	0	79
Utah	---	32	0	0	0	32	32	0	32
Virginia	96	160	1	2	1	158	151	0	151
Washington	---	0	0	0	0	0	0	0	0
West Virginia	97	135	1	1	1	134	130	0	130
Wisconsin	95	407	4	4	1	406	385	3	388
Wyoming	---	0	0	0	0	0	0	0	0
Other Jurisdiction†									
Guam	98	389	1	10	0	380	372	0	372

Refer to the preceding text section of this appendix for more detailed information on the notations and guidelines about sample representativeness, and for the derivation of weighted participation

* For the 1994 national NAEP there was no supplementary student sample taken.

--- Due to the small number of schools comprising the state's nonpublic school sample, weighted student participation rates are not calculated.

† DoDEA nonpublic school data do not exist because all non-domestic schools are considered public schools.

SOURCE: National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments

TABLE B-5 Summary of Public School and Student Participation, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD naep 1992 1994 Reading Assessment	Weighted School Participation Rate Before Substitution	Notation Number 1	Weighted School Participation Rate After Substitution	Notation Number 4	Weighted Student Participation Rate After Make-ups	Notation Number 8	Weighted Overall Participation Rate
Nation							
Nation	86		87		95		82
Northeast	93		93		94		87
Southeast	91		93		95		88
Central	85		87		95		83
West	77		77		95		73
States							
Alabama	87		93		96		90
Arizona	99		99		94		93
Arkansas	86		94		96		90
California	80		91		94		85
Colorado	100		100		94		94
Connecticut	96		96		96		92
Delaware	100		100		96		96
Florida	100		100		94		94
Georgia	99		99		95		95
Hawaii	99		99		95		95
Idaho	69	•	91		96		88
Indiana	83		92		96		89
Iowa	85		99		96		95
Kentucky	88		96		97		93
Louisiana	100		100		96		96
Maine	94		97		94		91
Maryland	94		96		95		92
Massachusetts	97		97		95		93
Michigan	63	•	80		95		76
Minnesota	86		95		95		91
Mississippi	95		99		97		96
Missouri	96		98		95		94
Montana	85		89		96		85
Nebraska	71		77	•	95		73
New Hampshire	71		79	•	96		76
New Jersey	85		91		95		87
New Mexico	100		100		95		95
New York	75		91		95		86
North Carolina	99		99		96		95
North Dakota	80		91		97		88
Pennsylvania	80		84	•	94		79
Rhode Island	80		86	•	95		81
South Carolina	95		97		96		94
Tennessee	72		74	•	96		71
Texas	91		93		96		90
Utah	100		100		95		95
Virginia	98		99		95		94
Washington	100		100		94		94
West Virginia	99		100		96		96
Wisconsin	79		86	•	96		82
Wyoming	98		98		96		94
Other Jurisdictions							
DoDEA Overseas	99		99		95		94
Guam	100		100		96		96

Refer to the preceding text section of this appendix for more detailed information on the notations and guidelines about sample representativeness, and for the derivation of weighted participation.

Notation 1: The state's public school weighted participation rate for the initial sample of schools was less the 70%

Notation 4: The state's public school weighted participation rate for the initial sample of schools was less than 85% *and* the weighted school participation rate after substitution was below 90%

Notation 8: The weighted student resonse rate within participating public schools was below 85%.

SOURCE: National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments

TABLE B-6 Summary of Nonpublic School and Student Participation, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD naep 1992 1994 Reading Assessment	Weighted School Participation Rate Before Substitution	Notation Number 2	Weighted School Participation Rate After Substitution	Notation Number 5	Weighted Student Participation Rate After Make-ups	Notation Number 9	Weighted Overall Participation Rate
Nation							
Nation	87		87		97		84
Northeast	82		82		98		80
Southeast	90		90		94		85
Central	97		97		97		94
West	80		80		98		78
States							
Alabama	92		96		95		91
Arizona	35	*	35		*		*
Arkansas	81		94		95		89
California	42	*	51		97		50
Colorado	71		85	*	94		80
Connecticut	73		82	*	95		73
Delaware	73		73	*	98		71
Florida	52	*	73		98		72
Georgia	74		84	*	97		81
Hawaii	80		88	*	96		84
Idaho	89		89		96		86
Indiana	85		85		95		81
Iowa	100		100		99		99
Kentucky	70		85	*	97		82
Louisiana	82		91		97		88
Maine	79		100		95		95
Maryland	63	*	70		97		68
Massachusetts	95		100		96		96
Michigan	0	*	0		*		*
Minnesota	91		99		96		95
Mississippi	64	*	64		96		62
Missouri	90		90		95		86
Montana	65	*	65		94		61
Nebraska	48	*	48		97		47
New Hampshire	54	*	54		*		*
New Jersey	76		76	*	96		73
New Mexico	100		100		92		92
New York	40	*	62		96		59
North Carolina	32	*	32		*		*
North Dakota	77		91		93		85
Pennsylvania	72		72	*	94		68
Rhode Island	93		93		96		89
South Carolina	69	*	86		98		84
Tennessee	41	*	41		*		*
Texas	24	*	39		*		*
Utah	23	*	23		*		*
Virginia	81		81	*	96		77
Washington	0	*	0		*		*
West Virginia	86		86		97		84
Wisconsin	66	*	66		95		62
Wyoming	0	*	0		*		*
Other Jurisdiction†							
Guam	96		96		98		94

Refer to the preceding text section of this appendix for more detailed information on the notations and guidelines about sample representativeness, and for the derivation of weighted participation.

Notation 2: The state's nonpublic school weighted participation rate for the initial sample of schools was less the 70%.

Notation 5: The state's nonpublic school weighted participation rate for the initial sample of schools was less than 85% *and* the weighted school participation rate after substitution was below 90%.

Notation 9: The weighted student resonse rate within participating nonpublic schools was below 85%.

† DoDEA nonpublic school data do not exist because all non-domestic schools are considered public schools.

SOURCE: National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments.

TABLE B-7 Public School Weighted Percentages of Excluded Students (IEP or LEP) from Original Sample, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD NAEP 1992 1994 Reading Assessment	Total Percentage of Students Identified as IEP or LEP	Total Percentage of Excluded Students	Percentage of Students Identified as IEP	Percentage of Students Excluded and Identified as IEP	Percentage of Students Identified as LEP	Percentage of Students Excluded and Identified as LEP
Nation						
Nation	17	9	12	6	6	3
Northeast	14	8	13	7	1	1
Southeast	14	8	14	8	1	0
Central	14	8	12	7	2	1
West	25	11	10	5	15	7
States						
Alabama	11	5	11	5	0	0
Arizona	21	7	11	5	11	3
Arkansas	12	6	12	6	0	0
California	31	12	10	5	23	9
Colorado	15	7	12	6	4	2
Connecticut	17	8	13	6	4	3
Delaware	15	6	14	6	1	1
Florida	22	10	18	9	5	2
Georgia	11	5	10	5	2	1
Hawaii	12	5	8	4	5	1
Idaho	13	5	10	4	3	1
Indiana	11	5	11	5	0	0
Iowa	11	5	11	4	1	0
Kentucky	8	4	8	4	0	0
Louisiana	11	6	11	6	1	0
Maine	17	10	16	9	1	1
Maryland	15	7	14	7	1	1
Massachusetts	18	8	15	5	5	3
Michigan	10	6	9	6	1	0
Minnesota	12	4	10	4	2	1
Mississippi	9	6	9	6	0	0
Missouri	12	5	12	5	0	0
Montana	11	4	11	3	1	0
Nebraska	16	4	15	4	1	1
New Hampshire	15	6	15	6	0	0
New Jersey	12	6	9	4	3	2
New Mexico	18	8	14	6	4	2
New York	15	8	9	5	6	3
North Carolina	15	5	14	5	1	1
North Dakota	10	2	9	2	1	0
Pennsylvania	11	6	10	5	1	1
Rhode Island	15	5	12	4	3	1
South Carolina	13	7	13	7	0	0
Tennessee	13	6	13	6	0	0
Texas	24	11	13	7	13	5
Utah	12	5	11	5	2	1
Virginia	13	7	12	6	2	1
Washington	14	5	11	4	4	1
West Virginia	12	7	12	7	0	0
Wisconsin	13	7	11	7	2	1
Wyoming	11	4	11	4	1	0
Other Jurisdictions						
DoDEA Overseas	10	5	8	4	2	1
Guam	12	9	5	5	7	4

IEP denotes Individual Education Plan and LEP denotes Limited English Proficiency.
To be excluded, a student was supposed to be IEP or LEP *and* judged incapable of participating in the assessment. A student reported as both IEP and LEP is counted once in the overall rate (first column), once in the overall excluded rate (second column), and separately in the remaining columns.

Note: Weighted percentages for the nation and region are based on students sampled for all subject areas assessed in 1994 (reading, U.S. history, and world geography). However, based on the national sampling design, the rates shown also are the best estimates for the comparable rates for the reading assessment

SOURCE: **National Assessment of Educational Progress** (NAEP), 1992 and 1994 Reading Assessments.

TABLE B-8 Nonpublic School Weighted Percentages of Excluded Students (IEP or LEP) from Original Sample, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD 1992 1994 Reading Assessment

	Total Percentage of Students Identified as IEP or LEP	Total Percentage of Excluded Students	Percentage of Students Identified as IEP	Percentage of Students Excluded and Identified as IEP	Percentage of Students Identified as LEP	Percentage of Students Excluded and Identified as LEP
Nation						
Nation	2	1	2	1	0	0
Northeast	2	1	2	1	0	0
Southeast	6	1	5	1	0	0
Central	0	0	0	0	0	0
West	2	1	1	1	1	0
States						
Alabama	2	1	2	1	0	0
Arizona	---	---	---	---	---	---
Arkansas	5	1	3	1	1	0
California	0	0	0	0	0	0
Colorado	1	0	1	0	0	0
Connecticut	7	2	4	1	3	1
Delaware	2	0	2	0	0	0
Florida	4	1	3	1	1	0
Georgia	3	0	3	0	0	0
Hawaii	2	1	1	0	1	0
Idaho	14	0	14	0	0	0
Indiana	3	1	2	1	1	0
Iowa	5	1	5	1	0	0
Kentucky	1	0	1	0	0	0
Louisiana	1	0	1	0	0	0
Maine	2	0	2	0	0	0
Maryland	2	1	2	1	0	0
Massachusetts	5	2	5	2	0	0
Michigan	---	---	---	---	---	---
Minnesota	4	2	4	2	0	0
Mississippi	6	3	6	3	0	0
Missouri	4	0	4	0	0	0
Montana	1	0	1	0	0	0
Nebraska	2	0	2	0	0	0
New Hampshire	---	---	---	---	---	---
New Jersey	6	1	5	0	1	0
New Mexico	22	13	14	11	11	3
New York	2	2	1	1	1	1
North Carolina	---	---	---	---	---	---
North Dakota	17	4	11	3	9	1
Pennsylvania	3	0	1	0	2	0
Rhode Island	5	0	2	0	2	0
South Carolina	0	0	0	0	0	0
Tennessee	---	---	---	---	---	---
Texas	---	---	---	---	---	---
Utah	---	---	---	---	---	---
Virginia	1	1	1	1	1	0
Washington	---	---	---	---	---	---
West Virginia	2	1	2	1	1	1
Wisconsin	2	0	2	0	0	0
Wyoming	---	---	---	---	---	---
Other Jurisdiction†						
Guam	0	0	0	0	0	0

IEP denotes Individual Education Plan and LEP denotes Limited English Proficiency.

To be excluded, a student was supposed to be IEP or LEP *and* judged incapable of participating in the assessment. A student reported as both IEP and LEP is counted once in the overall rate (first column), once in the overall excluded rate (second column), and separately in the remaining columns.

Note: Weighted percentages for the nation and region are based on students sampled for all subject areas assessed in 1994 (reading, U.S. history, and world geography). However, based on the national sampling design, the rates shown also are the best estimates for the comparable rates for the reading assessment.

--- Due to the small number of schools comprising the state's nonpublic school sample, weighted student participation rates are not calculated.

† DoDEA nonpublic school data do not exist because all non-domestic schools are considered public schools.

SOURCE: **National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments.**

TABLE B-9 Public School Weighted Percentages of Absent, IEP, or LEP Students Based on Those Invited to Participate in the Assessment, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD 1992 1994 Reading Assessment	Weighted Percentage of Student Participation After Make-up	Weighted Percentage of Absent Students	Weighted Percentage of Assessed IEP Students	Weighted Percentage of Absent IEP Students	Weighted Percentage of Assessed LEP Students	Weighted Percentage of Absent LEP Students
Nation						
Nation	95	5	92	8	92	8
Northeast	94	6	91	9	87	13
Southeast	95	5	93	7	100	0
Central	95	5	96	4	100	0
West	95	5	91	9	92	8
States						
Alabama	96	4	94	6	67	33
Arizona	94	6	94	6	96	4
Arkansas	96	4	94	6	100	0
California	94	6	82	18	95	5
Colorado	94	6	93	7	95	5
Connecticut	96	4	95	5	97	3
Delaware	96	4	95	5	100	0
Florida	94	6	93	7	93	7
Georgia	95	5	98	2	88	12
Hawaii	95	5	91	9	99	1
Idaho	96	4	94	6	93	7
Indiana	96	4	96	4	86	14
Iowa	96	4	92	8	100	0
Kentucky	97	3	95	5	100	0
Louisiana	96	4	94	6	100	0
Maine	94	6	94	6	100	0
Maryland	95	5	96	4	100	0
Massachusetts	95	5	93	7	95	5
Michigan	95	5	96	4	84	16
Minnesota	95	5	98	2	97	3
Mississippi	97	3	99	1	100	0
Missouri	95	5	93	7	100	0
Montana	96	4	93	7	97	3
Nebraska	95	5	95	5	92	8
New Hampshire	96	4	95	5	100	0
New Jersey	95	5	93	7	98	2
New Mexico	95	5	93	7	97	3
New York	95	5	96	4	93	7
North Carolina	96	4	93	7	93	7
North Dakota	97	3	96	4	100	0
Pennsylvania	94	6	94	6	97	3
Rhode Island	95	5	93	7	97	3
South Carolina	96	4	95	5	100	0
Tennessee	96	4	88	12	100	0
Texas	96	4	97	3	98	2
Utah	95	5	92	8	97	3
Virginia	95	5	93	7	97	3
Washington	94	6	94	6	97	3
West Virginia	96	4	96	4	100	0
Wisconsin	96	4	94	6	100	0
Wyoming	96	4	96	4	88	12
Other Jurisdictions						
DoDEA Overseas	95	5	88	12	95	5
Guam	96	4	100	0	91	9

IEP denotes Individual Education Plan and LEP denotes Limited English Proficiency.

Note: Weighted percentages for the nation and region are based on students sampled for all subject areas assessed in 1994 (reading, U.S. history, and world geography). However, based on the national sampling design, the rates shown also are the best estimates for the comparable rates for the reading assessment.

SOURCE **National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments.**

TABLE B-10 Nonpublic School Weighted Percentages of Absent, IEP, or LEP Students Based on Those Invited to Participate in the Assessment, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD naep 1992 1994 Reading Assessment	Weighted Percentage of Student Participation After Make-up	Weighted Percentage of Absent Students	Weighted Percentage of Assessed IEP Students	Weighted Percentage of Absent IEP Students	Weighted Percentage of Assessed LEP Students	Weighted Percentage of Absent LEP Students
Nation						
Nation	97	3	69	31	100	0
Northeast	98	2	82	18	100	0
Southeast	94	6	63	37	***	***
Central	97	3	100	0	***	***
West	98	2	50	50	100	0
States						
Alabama	95	5	100	0	***	***
Arizona	---	---	---	---	---	---
Arkansas	95	5	80	20	100	0
California	97	3	***	***	***	***
Colorado	94	6	100	0	***	***
Connecticut	95	5	100	0	100	0
Delaware	98	2	87	13	***	***
Florida	98	2	67	33	100	0
Georgia	97	3	86	14	***	***
Hawaii	96	4	100	0	100	0
Idaho	96	4	89	11	***	***
Indiana	95	5	100	0	63	37
Iowa	99	1	94	6	***	***
Kentucky	97	3	100	0	***	***
Louisiana	97	3	100	0	100	0
Maine	95	5	100	0	***	***
Maryland	97	3	100	0	***	***
Massachusetts	96	4	100	0	***	***
Michigan	---	---	---	---	---	---
Minnesota	96	4	100	0	***	***
Mississippi	96	4	63	37	***	***
Missouri	96	4	100	0	100	0
Montana	94	6	100	0	***	***
Nebraska	97	3	100	0	***	***
New Hampshire	---	---	---	---	---	---
New Jersey	96	4	83	17	100	0
New Mexico	92	8	89	11	92	8
New York	96	4	100	0	***	***
North Carolina	---	---	---	---	---	---
North Dakota	93	7	86	14	94	6
Pennsylvania	94	6	100	0	90	10
Rhode Island	96	4	100	0	100	0
South Carolina	98	2	***	***	***	***
Tennessee	---	---	---	---	---	---
Texas	---	---	---	---	---	---
Utah	---	---	---	---	---	---
Virginia	96	4	***	***	100	0
Washington	---	---	---	---	---	---
West Virginia	97	3	100	0	***	***
Wisconsin	95	5	75	25	***	***
Wyoming	---	---	---	---	---	---
Other Jurisdiction†						
Guam	98	2	***	***	***	***

IEP denotes Individual Education Plan and **LEP** denotes Limited English Proficiency.

Note: Weighted percentages for the nation and region are based on students sampled for all subject areas assessed in 1994 (reading, U.S. history, and world geography). However, based on the nationa' sampling design, the rates shown also are the best estimates for the comparable rates for the reading assessment.

--- Due to the small number of schools comprising the state's nonpublic school sample, weighted student participation rates are not calculated.

*** There were no students in the state's sample comprising the denominator of the percentage denoted by the column heading.

† DoDEA nonpublic school data do not exist because all non-domestic schools are considered public schools.

SOURCE. **National Assessment of Educational Progress (NAEP)**, 1992 and 1994 Reading Assessments.

TABLE B-11	Public School Questionnaire Response Rates, 1994 Grade 4 Reading Assessment

THE NATION'S REPORT CARD 1992 1994 Reading Assessment	Weighted Percentage of Students Matched to Reading Teacher Questionnaires	Percentage of Reading Teacher Questionnaires Returned	Weighted Percentage of Students Matched to School Characteristics and Policies Questionnaire	Weighted Percentage of School Characteristics and Policies Questionnaires Returned	Percentage of Excluded Student Questionnaires Returned
Nation					
Nation	94.7	95.0	94.5	95.1	93.3
Northeast	94.5	93.5	91.2	91.7	92.8
Southeast	96.2	96.3	97.7	96.7	95.4
Central	94.7	96.6	100.0	100.0	95.1
West	93.5	93.8	90.1	92.4	91.6
States					
Alabama	99.0	99.0	100.0	100.0	98.4
Arizona	98.8	98.7	99.2	99.0	99.7
Arkansas	99.3	99.4	99.0	99.0	99.1
California	96.8	96.2	99.0	99.0	97.3
Colorado	96.7	98.1	97.8	98.0	97.6
Connecticut	98.0	98.7	98.1	97.9	99.4
Delaware	98.3	98.2	100.0	100.0	98.4
Florida	94.0	98.4	99.4	99.4	99.5
Georgia	98.0	99.8	100.0	100.0	98.2
Hawaii	98.2	99.2	97.9	97.9	99.4
Idaho	97.7	98.5	100.0	100.0	99.4
Indiana	97.8	99.7	100.0	100.0	98.8
Iowa	96.5	99.7	100.0	100.0	100.0
Kentucky	97.1	98.6	100.0	100.0	99.2
Louisiana	98.9	99.5	100.0	100.0	99.7
Maine	98.7	99.5	100.0	100.0	99.6
Maryland	97.6	99.0	100.0	100.0	100.0
Massachusetts	98.7	98.4	100.0	100.0	99.8
Michigan	96.0	98.1	97.9	97.7	100.0
Minnesota	95.5	97.3	100.0	100.0	95.0
Mississippi	98.4	99.4	100.0	100.0	98.8
Missouri	98.7	99.8	99.2	99.3	100.0
Montana	99.2	99.1	100.0	100.0	95.3
Nebraska	97.7	100.0	99.2	99.4	99.8
New Hampshire	99.3	98.5	100.0	100.0	99.7
New Jersey	97.9	98.7	100.0	100.0	98.8
New Mexico	96.4	98.3	98.8	99.0	100.0
New York	99.8	99.8	100.0	100.0	100.0
North Carolina	95.7	99.5	98.9	98.8	99.8
North Dakota	99.0	99.6	100.0	100.0	98.9
Pennsylvania	97.2	99.4	100.0	100.0	99.6
Rhode Island	98.1	99.4	97.9	97.6	99.7
South Carolina	96.7	99.4	100.0	100.0	99.7
Tennessee	99.2	99.4	98.9	98.5	100.0
Texas	98.0	99.0	100.0	100.0	98.3
Utah	99.4	98.7	98.2	98.5	99.4
Virginia	98.6	99.1	100.0	100.0	97.4
Washington	94.4	99.3	100.0	100.0	98.8
West Virginia	95.2	99.0	100.0	100.0	100.0
Wisconsin	97.1	96.9	100.0	100.0	99.7
Wyoming	96.2	99.7	97.9	96.2	99.4
Other Jurisdictions					
DoDEA Overseas	96.8	99.1	96.5	96.1	99.2
Guam	76.5	100.0	100.0	100.0	98.2

Note: For the nation and regions, the percentage of excluded student questionnaires returned is based on students sampled for all subjects assessed in 1994 (reading, U.S. history, and world geography). However, based on the national sampling design, the rates shown also are the best estimates of the comparable rates for the reading assessment.

SOURCE: **National Assessment of Educational Progress (NAEP)**, 1992 and 1994 Reading Assessments.

TABLE B-12 Nonpublic School Questionnaire Response Rates, 1994 Grade 4 Reading Assessment

	Weighted Percentage of Students Matched to Reading Teacher Questionnaires	Percentage of Reading Teacher Questionnaires Returned	Weighted Percentage of Students Matched to School Characteristics and Policies Questionnaire	Weighted Percentage of School Characteristics and Policies Questionnaires Returned	Percentage of Excluded Student Questionnaires Returned
Nation					
Nation	95.9	97.7	100.0	100.0	95.1
Northeast	100.0	97.2	100.0	100.0	91.9
Southeast	92.9	100.0	100.0	100.0	85.7
Central	96.8	97.2	100.0	100.0	99.1
West	91.5	96.5	100.0	100.0	100.0
States					
Alabama	99.7	100.0	100.0	100.0	100.0
Arizona	---	100.0	---	100.0	100.0
Arkansas	100.0	100.0	100.0	100.0	100.0
California	99.5	100.0	100.0	100.0	***
Colorado	100.0	100.0	100.0	100.0	***
Connecticut	99.7	100.0	100.0	100.0	100.0
Delaware	100.0	100.0	100.0	100.0	100.0
Florida	92.7	100.0	100.0	100.0	100.0
Georgia	99.8	100.0	100.0	100.0	100.0
Hawaii	94.4	97.1	94.3	92.1	100.0
Idaho	100.0	100.0	100.0	100.0	100.0
Indiana	100.0	100.0	100.0	100.0	100.0
Iowa	94.3	100.0	100.0	100.0	100.0
Kentucky	97.6	100.0	100.0	100.0	100.0
Louisiana	100.0	97.1	100.0	100.0	100.0
Maine	100.0	100.0	100.0	100.0	100.0
Maryland	100.0	100.0	100.0	100.0	100.0
Massachusetts	93.4	96.0	100.0	100.0	100.0
Michigan	---	***	---	***	***
Minnesota	100.0	100.0	100.0	100.0	100.0
Mississippi	100.0	100.0	100.0	100.0	100.0
Missouri	99.8	100.0	100.0	100.0	100.0
Montana	100.0	100.0	100.0	100.0	100.0
Nebraska	100.0	100.0	100.0	100.0	100.0
New Hampshire	---	100.0	---	100.0	100.0
New Jersey	95.3	95.7	100.0	100.0	100.0
New Mexico	83.1	100.0	100.0	100.0	100.0
New York	100.0	100.0	100.0	100.0	100.0
North Carolina	---	100.0	---	100.0	***
North Dakota	99.2	100.0	100.0	100.0	100.0
Pennsylvania	97.7	100.0	100.0	100.0	100.0
Rhode Island	87.4	100.0	100.0	100.0	100.0
South Carolina	100.0	100.0	100.0	100.0	***
Tennessee	---	100.0	---	100.0	***
Texas	---	100.0	---	100.0	***
Utah	---	100.0	---	100.0	***
Virginia	99.6	100.0	100.0	100.0	100.0
Washington	---	***	---	***	***
West Virginia	100.0	100.0	100.0	100.0	100.0
Wisconsin	100.0	87.5	100.0	100.0	100.0
Wyoming	---	***	---	***	***
Other Jurisdiction†					
Guam	67.5	100.0	100.0	100.0	***

Note: For the nation and regions, the percentage of excluded student questionnaires returned is based on students sampled for all subjects assessed in 1994 (reading, U.S. history, and world geography). However, based on the national sampling design, the rates shown also are the best estimates of the comparable rates for the reading assessment.

--- Due to the small number of schools comprising the state's nonpublic school sample, weighted student participation rates are not calculated.

*** There were no students in the state's sample comprising the denominator of the percentage denoted by the column heading.

† DoDEA nonpublic school data do not exist because all non-domestic schools are considered public schools.

SOURCE: **National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments.**

APPENDIX C

CONDITIONING VARIABLES AND CONTRAST CODINGS

APPENDIX C

Conditioning Variables and Contrast Codings

This appendix contains information about the conditioning variables used in scaling/plausible value estimation for the 1994 NAEP Trial State Assessment Program in reading. The initial step in construction of conditioning variables involves forming primary student-based vectors of response data from answers to student, teacher, and school questionnaires, demographic and background data such as supplied by Westat, and other student information known prior to scaling. The initial conditioning vectors concatenate this student background information into a series of identifying "contrasts" comprising:

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

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

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

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

As described in Chapter 9, the linear conditioning model employed for the estimation of plausible values in each jurisdiction does not directly use the conditioning variable vectors derived from the procedures described above (see actual specifications listed in this appendix). A second step is employed to eliminate the inherent instabilities in estimation encountered when using a large number of correlated variables by performing a principal component transformation of the correlation matrix of the primary conditioning variable contrasts. The principal components scores based on this transformation are then used as the predictor variables in estimating the linear conditioning/scaling model.

The remainder of this appendix gives a complete list of the original questionnaire, demographic, and background variables used in constructing the primary 1994 NAEP student-based reading conditioning vectors, defining the specifications employed for transforming the raw responses to the primary conditioning contrast values. Table C-1 preceding the list defines the labeling used in presenting the specifications listed for each conditioning variable.

Table C-1
Description of Specifications Provided for Each Conditioning Variable

Title	Description
CONDITIONING ID	An unique eight-character ID assigned to identify each conditioning variable corresponding to a particular background or subject area question within the entire pool of conditioning variables. The first four characters identify the origin of the variable: BACK (background questionnaire), READ (student reading questionnaire), SCHL (school questionnaire), TCHR (background part of teacher questionnaire), and TSUB (subject classroom part of teacher questionnaire). The second four digits represent the sequential position within each origin group.
DESCRIPTION	A short description of the conditioning variable.
GRADES/ASSESSMENTS	Three characters identifying assessment ("S" for state, "N" for national) and grade (04, 08, and 12) in which the conditioning variable was used.
CONDITIONING VAR LABEL	A descriptive eight-character label identifying the conditioning variable.
NAEP ID	The seven-character NAEP database identification for the conditioning variable.
TYPE OF CONTRAST	The type of conditioning variable. "CLASS" identifies a categorical conditioning variable and "SCALE" identifies continuous or quasi-continuous conditioning variables. "INTERACTION" identifies a set of orthogonal contrasts formed from two or more "CLASS" variables. "OTHER" conditioning variables do not fall into any of the above types.
TOTAL NUMBER OF SPECIFIED CONTRASTS	Each conditioning variable forms a set of one or more contrasts. For each valid response value of conditioning variable a contrast must be defined. One or more response values may be collapsed together to form one contrast. The number of response value "sets" of a conditioning variable forming a unique contrast is the value given in this field.
NUMBER OF INDEPENDENT CONTRASTS	The number of degree of freedom in a set of contrasts formed from a conditioning variable. For a categorical conditioning variable this number would be the number of response options minus one if each response option formed its own unique contrast.

```
CONDITIONING VARIABLE ID:  BACK0001
DESCRIPTION:               GRAND MEAN
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    OVERALL
NAEP ID:                   BKSER                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     1
TYPE OF CONTRAST:          OTHER                    NUMBER OF INDEPENDENT CONTRASTS:         1

001 OVERALL  (@          ) 1                        GRAND MEAN

CONDITIONING VARIABLE ID:  BACK0002
DESCRIPTION:               DERIVED SEX
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    GENDER
NAEP ID:                   DSEX                     TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                    NUMBER OF INDEPENDENT CONTRASTS:         1

001 MALE     (1          ) 0                        MALE
002 FEMALE   (2          ) 1                        FEMALE

CONDITIONING VARIABLE ID:  BACK0003
DESCRIPTION:               DERIVED RACE/ETHNICITY
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    RACE/ETH
NAEP ID:                   DRACE7                   TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                    NUMBER OF INDEPENDENT CONTRASTS:         4

001 WHI/AI/O (1,6,7      ) 0000                     RACE/ETHNICITY:  WHITE, AMERICAN INDIAN/ALASKAN NATIVE, OTHER, MISSING, UNCLASSIFIED
002 BLACK    (2          ) 1000                     RACE/ETHNICITY:  BLACK
003 HISPANIC (3          ) 0100                     RACE/ETHNICITY:  HISPANIC
004 ASIAN    (4          ) 0010                     RACE/ETHNICITY:  ASIAN
005 PAC ISLD (5          ) 0001                     RACE/ETHNICITY:  PACIFIC ISLANDER

CONDITIONING VARIABLE ID:  BACK0004
DESCRIPTION:               IF HISPANIC, WHAT IS YOUR HISPANIC BACKGROUND?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    HISPANIC
NAEP ID:                   B003101                  TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                    NUMBER OF INDEPENDENT CONTRASTS:         4

001 NOT HISP (1          ) 0000                     HISPANIC:  NOT HISPANIC
002 MEXICAN  (2          ) 1000                     HISPANIC:  MEXICAN, MEXICAN AMERICAN, CHICANO
003 PUER RIC (3          ) 0100                     HISPANIC:  PUERTO RICAN
004 CUBN,OTH (4,5        ) 0010                     HISPANIC:  CUBAN, OTHER
005 HISP-?   (M          ) 0001                     HISPANIC:  MISSING

CONDITIONING VARIABLE ID:  BACK0006
DESCRIPTION:               TYPE OF LOCALE (5 CATEGORIES)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
```

```
CONDITIONING VAR LABEL:     TOL5
NAEP ID:                    TOL5                          TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:           CLASS                         NUMBER OF INDEPENDENT CONTRASTS:         4

001 BIG CTY5 (1           ) 0000                          TOL5:  LARGE CITY
002 MID CTY5 (2,M         ) 1000                          TOL5:  MID-SIZE CITY
003 FR/BTWN5 (3           ) 0100                          TOL5:  URBAN FRINGE OF LARGE CITY, URBAN FRINGE OF MID-SIZE CITY
004 SML TWN5 (4           ) 0010                          TOL5:  SMALL TOWN
005 RURAL5   (5           ) 0001                          TOL5:  RURAL (MSA AND NON-MSA)

CONDITIONING VARIABLE ID:   BACK0007
DESCRIPTION:                DESCRIPTION OF COMMUNITY
GRADES/ASSESSMENTS:         N04, S04, N08, N12
CONDITIONING VAR LABEL:     DOC
NAEP ID:                    DOC                           TOTAL NUMBER OF SPECIFIED CONTRASTS:     4
TYPE OF CONTRAST:           CLASS                         NUMBER OF INDEPENDENT CONTRASTS:         3

001 BIG CITY (1           ) 000                           DOC:  BIG CITY
002 URBAN FR (2           ) 100                           DOC:  URBAN FRINGE
003 MED CITY (3,M         ) 010                           DOC:  MEDIUM CITY
004 SM PLACE (4           ) 001                           DOC:  SMALL PLACE

CONDITIONING VARIABLE ID:   BACK0008
DESCRIPTION:                PARENTS' HIGHEST LEVEL OF EDUCATION
GRADES/ASSESSMENTS:         N04, S04, N08, N12
CONDITIONING VAR LABEL:     PARED
NAEP ID:                    PARED                         TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:           CLASS                         NUMBER OF INDEPENDENT CONTRASTS:         4

001 < HS     (1           ) 0000                          PARED:  LESS THAN HIGH SCHOOL
002 HS GRAD  (2           ) 1000                          PARED:  HIGH SCHOOL GRADUATE
003 POST HS  (3           ) 0100                          PARED:  POST HIGH SCHOOL
004 COL GRAD (4           ) 0010                          PARED:  COLLEGE GRADUATE
005 PARED-?  (5,M         ) 0001                          PARED:  MISSING, I DON'T KNOW

CONDITIONING VARIABLE ID:   BACK0010
DESCRIPTION:                SCHOOL TYPE (PQ)
GRADES/ASSESSMENTS:         N04, S04, N08, N12
CONDITIONING VAR LABEL:     SCHTYPE
NAEP ID:                    SCHTYPE                       TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:           CLASS                         NUMBER OF INDEPENDENT CONTRASTS:         2

001 PUBLIC   (1           ) 0U                            SCHOOL TYPE:  PUBLIC
002 PRIVATE  (2,4,5,M     ) 10                            SCHOOL TYPE:  PRIVATE, BUREAU OF INDIAN AFFAIRS, DEPARTMENT OF DEFENSE, MISSING
003 CATHOLIC (3           ) 01                            SCHOOL TYPE:  CATHOLIC

CONDITIONING VARIABLE ID:   BACK0011
DESCRIPTION:                INDIVIDUALIZED EDUCATION PLAN
GRADES/ASSESSMENTS:         N04, S04, N08, N12
```

```
CONDITIONING VAR LABEL:    IEP
NAEP ID:                   IEP                          TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         1

001 IEP-YES  (1          ) 0                            IEP:  YES
002 IEP-NO   (2          ) 1                            IEP:  NO

CONDITIONING VARIABLE ID:  BACK0012
DESCRIPTION:               LIMITED ENGLISH PROFICIENCY
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    LEP
NAEP ID:                   LEP                          TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         1

001 LEP-YES  (1          ) 0                            LEP:  YES
002 LEP-NO   (2          ) 1                            LEP:  NO

CONDITIONING VARIABLE ID:  BACK0013
DESCRIPTION:               CHAPTER 1 (BOOK COVER)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    CHAPTER1
NAEP ID:                   CHAP1                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         1

001 CHAP1-Y  (1          ) 0                            CHAPTER 1:  YES
002 CHAP1-N  (2          ) 1                            CHAPTER 1:  NO

CONDITIONING VARIABLE ID:  BACK0014
DESCRIPTION:               PERCENT WHITE STUDENTS IN SCHOOL (FROM QED)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PCTWHITE
NAEP ID:                   PCTWHTQ                      TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         2

001 PREDOM/? (80-110,M   ) 00                           PREDOMINANTLY WHITE, MISSING
002 INTEGRAT (50-79      ) 10                           INTEGRATED
003 MINORITY (0-49       ) 01                           WHITE MINORITY

CONDITIONING VARIABLE ID:  BACK0015
DESCRIPTION:               DO YOU RECEIVE A FREE OR REDUCED-PRICE LUNCH?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    LUNCH
NAEP ID:                   B008101                      TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         2

001 FR LUNCH (1          ) 00                           LUNCH PROGRAM:  FREE/REDUCED
002 NO LUNCH (2          ) 10                           LUNCH PROGRAM:  NOT FREE/REDUCED
003 LUNCH-?  (3,M        ) 01                           LUNCH PROGRAM:  I DON'T KNOW, MISSING
```

```
CONDITIONING VARIABLE ID:  BACK0017
DESCRIPTION:               HOW OFTEN DO THE PEOPLE IN YOUR HOME SPEAK A LANGUAGE OTHER THAN ENGLISH?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    HOMELANG
NAEP ID:                   B003201                          TOTAL NUMBER OF SPECIFIED CONTRASTS:    4
TYPE OF CONTRAST:          CLASS                            NUMBER OF INDEPENDENT CONTRASTS:        3

001 HL-NEVER (1          ) 000                              HOMELANG:  NEVER
002 HL-SOME  (2          ) 100                              HOMELANG:  SOMETIMES
003 HL-ALWAY (3          ) 010                              HOMELANG:  ALWAYS
004 HL-?     (M          ) 001                              HOMELANG:  MISSING

CONDITIONING VARIABLE ID:  BACK0018
DESCRIPTION:               HOW MUCH TELEVISION DO YOU USUALLY WATCH EACH DAY? (LINEAR)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    TVWATCHL
NAEP ID:                   B001801                          TOTAL NUMBER OF SPECIFIED CONTRASTS:    7
TYPE OF CONTRAST:          LINEAR                           NUMBER OF INDEPENDENT CONTRASTS:        1

001 TVLIN-0  (1          ) 0                                TV WATCHING (LINEAR) (0 TO 6+ HOURS PER DAY)
002 TVLIN-1  (2          ) 1                                TV WATCHING (LINEAR)
003 TVLIN-2  (3          ) 2                                TV WATCHING (LINEAR)
004 TVL:N-3  (4,M        ) 3                                TV WATCHING (LINEAR)
005 TVLIN-4  (5          ) 4                                TV WATCHING (LINEAR)
006 TVLIN-5  (6          ) 5                                TV WATCHING (LINEAR)
007 TVLIN-6  (7          ) 6                                TV WATCHING (LINEAR)

CONDITIONING VARIABLE ID:  BACK0019
DESCRIPTION:               HOW MUCH TELEVISION DO YOU USUALLY WATCH EACH DAY? (QUADRATIC)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    TVWATCHQ
NAEP ID:                   B001801                          TOTAL NUMBER OF SPECIFIED CONTRASTS:    1
TYPE OF CONTRAST:          QUADRATIC                        NUMBER OF INDEPENDENT CONTRASTS:        1

001 TV-QUAD  (1-7,M=4    )  1.0 + -2.0*X +  1.0*X**2        TV WATCHING (QUADRATIC)

CONDITIONING VARIABLE ID:  BACK0020
DESCRIPTION:               HOMEWORK ASSIGNED?:  BASED ON TIME SPENT ON HOMEWORK EACH DAY.
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    HWASSIGN
NAEP ID:                   B006601                          TOTAL NUMBER OF SPECIFIED CONTRASTS:    3
TYPE OF CONTRAST:          CLASS                            NUMBER OF INDEPENDENT CONTRASTS:        2

001 HW-MISS  (M          ) 00                               HOMEWORK ASSIGNED?:  MISSING
002 HW-NO    (1          ) 10                               HOMEWORK ASSIGNED?:  NO
003 HW-YES   (2-5        ) 01                               HOMEWORK ASSIGNED?:  YES

CONDITIONING VARIABLE ID:  BACK0021
DESCRIPTION:               HOW MUCH TIME DO YOU USUALLY SPEND ON HOMEWORK EACH DAY? (LINEAR)
```

```
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    HOMEWRKL
NAEP ID:                   B006601                         TOTAL NUMBER OF SPECIFIED CONTRASTS:     4
TYPE OF CONTRAST:          LINEAR                          NUMBER OF INDEPENDENT CONTRASTS:         1

001 HWLIN-0  (1,2,M      ) 0                               HOMEWORK (LINEAR):  DON'T HAVE ANY, DON'T DO ANY, MISSING
002 HWLIN-1  (3          ) 1                               HOMEWORK (LINEAR):  1/2 HOUR OR LESS
003 HWLIN-2  (4          ) 2                               HOMEWORK (LINEAR):  1 HOUR
004 HWLIN-3  (5          ) 3                               HOMEWORK (LINEAR):  MORE THAN 1 HOUR

CONDITIONING VARIABLE ID:  BACK0022
DESCRIPTION:               HOW MUCH TIME DO YOU USUALLY SPEND ON HOMEWORK EACH DAY (QUADRATIC)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    HOMEWRKQ
NAEP ID:                   B006601                         TOTAL NUMBER OF SPECIFIED CONTRASTS:     4
TYPE OF CONTRAST:          SCALE                           NUMBER OF INDEPENDENT CONTRASTS:         1

001 HWQUAD-0 (1,2,M      ) 0                               HOMEWORK (QUADRATIC):  DON'T HAVE ANY, DON'T DO ANY, MISSING
002 HWQUAD-1 (3          ) 1                               HOMEWORK (QUADRATIC):  1/2 HOUR OR LESS
003 HWQUAD-2 (4          ) 4                               HOMEWORK (QUADRATIC):  1 HOUR
004 HWQUAD-3 (5          ) 9                               HOMEWORK (QUADRATIC):  MORE THAN 1 HOUR

CONDITIONING VARIABLE ID:  BACK0023
DESCRIPTION:               NUMBER OF ITEMS IN THE HOME (NEWSPAPER, > 25 BOOKS, ENCYCLOPEDIA, MAGAZINES) (DERIVED)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    HOMEITMS
NAEP ID:                   HOMEEN2                         TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                           NUMBER OF INDEPENDENT CONTRASTS:         2

001 HITEM<=2 (1,M        ) 00                              ITEMS IN HOME:  ZERO TO TWO ITEMS, MISSING
002 HITEM=3  (2          ) 10                              ITEMS IN HOME:  THREE ITEMS
003 HITEM=4  (3          ) 01                              ITEMS IN HOME:  FOUR ITEMS

CONDITIONING VARIABLE ID:  BACK0025
DESCRIPTION:               DOES MOTHER OR STEPMOTHER LIVE AT HOME WITH YOU?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    MOM@HOME
NAEP ID:                   B005601                         TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                           NUMBER OF INDEPENDENT CONTRASTS:         2

001 MOMHOM-Y (1          ) 00                              MOTHER AT HOME:  YES
002 MOMHOM-N (2          ) 10                              MOTHER AT HOME:  NO
003 MOMHOM-? (M          ) 01                              MOTHER AT HOME:  MISSING

CONDITIONING VARIABLE ID:  BACK0026
DESCRIPTION:               DOES FATHER OR STEPFATHER LIVE AT HOME WITH YOU?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    DAD@HOME
NAEP ID:                   B005701                         TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
```

```
TYPE OF CONTRAST:         CLASS                            NUMBER OF INDEPENDENT CONTRASTS:        2

001 DADHOM-Y (1          ) 00                              FATHER AT HOME:  YES
002 DADHOM-N (2          ) 10                              FATHER AT HOME:  NO
003 DADHOM-? (M          ) 01                              FATHER AT HOME:  MISSING

CONDITIONING VARIABLE ID:  BACK0027
DESCRIPTION:               HOW MANY DAYS OF SCHOOL MISSED LAST MONTH?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    SCH MISS
NAEP ID:                   S004001                         TOTAL NUMBER OF SPECIFIED CONTRASTS:    2
TYPE OF CONTRAST:          CLASS                           NUMBER OF INDEPENDENT CONTRASTS:        1

001 MISS->2  (3,4,5,M    ) 0                               DAYS OF SCHOOL MISSED:  3-4, 5-10, 10 OR MORE DAYS, MISSING
002 MISS-2<  (1,2        ) 1                               DAYS OF SCHOOL MISSED:  0-1, 2 DAYS

CONDITIONING VARIABLE ID:  BACK0028
DESCRIPTION:               HOW LONG LIVED IN THE UNITED STATES?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    YRSINUSA
NAEP ID:                   B008001                         TOTAL NUMBER OF SPECIFIED CONTRASTS:    4
TYPE OF CONTRAST:          CLASS                           NUMBER OF INDEPENDENT CONTRASTS:        3

001 USA >5   (1          ) 000                             LIVED IN US MORE THAN 5 YEARS
002 USA 3-5  (2          ) 100                             LIVED IN US 3-5 YEARS
003 USA <3   (3          ) 010                             LIVED IN US LESS THAN 3 YEARS
004 USA-?    (M          ) 001                             LIVED IN US MISSING

CONDITIONING VARIABLE ID:  BACK0029
DESCRIPTION:               HOW MANY GRADES IN THIS STATE? (4TH GRADE)
GRADES/ASSESSMENTS:        N04, S04
CONDITIONING VAR LABEL:    STGRADE4
NAEP ID:                   B007601                         TOTAL NUMBER OF SPECIFIED CONTRASTS:    3
TYPE OF CONTRAST:          CLASS                           NUMBER OF INDEPENDENT CONTRASTS:        2

001 STGRD<1  (1,M        ) 00                              GRADES IN STATE:  LESS THAN 1 GRADE, MISSING
002 STGRD1-2 (2          ) 10                              GRADES IN STATE:  1-2 GRADES
003 STGRD3>  (3          ) 01                              GRADES IN STATE:  3 OR MORE GRADES

CONDITIONING VARIABLE ID:  BACK0032
DESCRIPTION:               DID YOU GO TO PRESCHOOL, NURSERY OR DAYCARE?
GRADES/ASSESSMENTS:        N04, S04
CONDITIONING VAR LABEL:    PRESCH
NAEP ID:                   B004201                         TOTAL NUMBER OF SPECIFIED CONTRASTS:    2
TYPE OF CONTRAST:          CLASS                           NUMBER OF INDEPENDENT CONTRASTS:        1

001 PRESCH-Y (2,3,M      ) 0                               PRESCHOOL:  NO, I DON'T KNOW, MISSING
002 PRESCH-N (1          ) 1                               PRESCHOOL:  YES
```

```
CONDITIONING VARIABLE ID:  BACK0033
DESCRIPTION:               HOW MANY TIMES HAVE YOU CHANGED SCHOOLS IN PAST TWO YEARS BECAUSE YOU MOVED?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    SCH CHGS
NAEP ID:                   B007301                     TOTAL NUMBER OF SPECIFIED CONTRASTS:    4
TYPE OF CONTRAST:          CLASS                       NUMBER OF INDEPENDENT CONTRASTS:        3

001 SCHCHG-0 (1          ) 000                         SCHOOL CHANGES:  NONE
002 SCHCHG-1 (2          ) 100                         SCHOOL CHANGES:  ONE
003 SCHCHG-2 (3          ) 010                         SCHOOL CHANGES:  TWO
004 SCHCHG-3 (4,M        ) 001                         SCHOOL CHANGES:  THREE OR MORE, MISSING

CONDITIONING VARIABLE ID:  BACK0034
DESCRIPTION:               HOW OFTEN DO YOU DISCUSS THINGS STUDIED IN SCHOOL WITH SOMEONE AT HOME?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    DISC@HOM
NAEP ID:                   B007401                     TOTAL NUMBER OF SPECIFIED CONTRASTS:    4
TYPE OF CONTRAST:          CLASS                       NUMBER OF INDEPENDENT CONTRASTS:        3

001 DIS@HOM1 (1          ) 000                         DISCUSS STUDIES AT HOME:  ALMOST EVERY DAY
002 DIS@HOM2 (2          ) 100                         DISCUSS STUDIES AT HOME:  ONCE OR TWICE A WEEK
003 DIS@HOM3 (3          ) 010                         DISCUSS STUDIES AT HOME:  ONCE OR TWICE A MONTH
004 DIS@HOM4 (4,M        ) 001                         DISCUSS STUDIES AT HOME:  NEVER OR HARDLY EVER, MISSING

CONDITIONING VARIABLE ID:  BACK0035
DESCRIPTION:               ABOUT HOW MANY PAGES A DAY DO YOU HAVE TO READ FOR SCHOOL AND HOMEWORK?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PGSREAD1
NAEP ID:                   B001101                     TOTAL NUMBER OF SPECIFIED CONTRASTS:    2
TYPE OF CONTRAST:          CLASS                       NUMBER OF INDEPENDENT CONTRASTS:        1

001 PGS<6,?  (5,M        ) 0                           PAGES READ:  5 OR FEWER A DAY, MISSING
002 PGS>5    (1,2,3,4    ) 1                           PAGES READ:  6-10, 11-15, 16-20, 20 OR MORE

CONDITIONING VARIABLE ID:  BACK0036
DESCRIPTION:               ABOUT HOW MANY PAGES A DAY DO YOU HAVE TO READ FOR SCHOOL AND HOMEWORK?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PGSREAD2
NAEP ID:                   B001101                     TOTAL NUMBER OF SPECIFIED CONTRASTS:    2
TYPE OF CONTRAST:          CLASS                       NUMBER OF INDEPENDENT CONTRASTS:        1

001 PGS<11,? (4,5,M      ) 0                           PAGES READ:  6-10, 5 OR FEWER A DAY, MISSING
002 PGS>10   (1,2,3      ) 1                           PAGES READ:  11-15, 16-20, 20 OR MORE

CONDITIONING VARIABLE ID:  BACK0037
DESCRIPTION:               HOW OFTEN DO YOU USE A COMPUTER FOR SCHOOLWORK?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    COMP@SCH
NAEP ID                    B007501                     TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
```

```
TYPE OF CONTRAST:         CLASS                    NUMBER OF INDEPENDENT CONTRASTS:        4

001 COMP-DAY (1          ) 0000                    USE COMPUTER AT SCHOOL:  ALMOST EVERY DAY
002 COMP-WK  (2          ) 1000                    USE COMPUTER AT SCHOOL:  ONCE OR TWICE A WEEK
003 COMP-MO  (3          ) 0100                    USE COMPUTER AT SCHOOL:  ONCE OR TWICE A MONTH
004 COMP-NEV (4          ) 0010                    USE COMPUTER AT SCHOOL:  NEVER OR HARDLY EVER
005 COMP-?   (M          ) 0001                    USE COMPUTER AT SCHOOL:  MISSING

CONDITIONING VARIABLE ID:  BACK0054
DESCRIPTION:               INTERACTION:  GENDER BY RACE/ETHNICITY
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    GEND/RAC
NAEP ID:                   N/A                     TOTAL NUMBER OF SPECIFIED CONTRASTS:    10
TYPE OF CONTRAST:          INTERACTION             NUMBER OF INDEPENDENT CONTRASTS:         4

001 G/R 11   (11         ) 01010101                GEND/RAC INTACT: 1. MALE     1. WHI/AI/O
002 G/R 12   (12         ) -1000000                GEND/RAC INTACT: 1. MALE     2. BLACK
003 G/R 13   (13         ) 00-10000                GEND/RAC INTACT: 1. MALE     3. HISPANIC
004 G/R 14   (14         ) 0000-100                GEND/RAC INTACT: 1. MALE     4. ASIAN
005 G/R 15   (15         ) 000000-1                GEND/RAC INTACT: 1. MALE     5. PAC ISLD
006 G/R 21   (21         ) -1-1-1-1                GEND/RAC INTACT: 2. FEMALE   1. WHI/AI/O
007 G/R 22   (22         ) 01000000                GEND/RAC INTACT: 2. FEMALE   2. BLACK
008 G/R 23   (23         ) 00010000                GEND/RAC INTACT: 2. FEMALE   3. HISPANIC
009 G/R 24   (24         ) 00000100                GEND/RAC INTACT: 2. FEMALE   4. ASIAN
010 G/R 25   (25         ) 00000001                GEND/RAC INTACT: 2. FEMALE   5. PAC ISLD

CONDITIONING VARIABLE ID:  BACK0055
DESCRIPTION:               INTERACTION:  GENDER BY TYPE OF LOCALE (5 CATEGORIES)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    GEND/TOL
NAEP ID:                   N/A                     TOTAL NUMBER OF SPECIFIED CONTRASTS:    10
TYPE OF CONTRAST:          INTERACTION             NUMBER OF INDEPENDENT CONTRASTS:         4

001 G/T 11   (11         ) 01010101                GEND/TOL INTACT: 1. MALE     1. BIG CTY5
002 G/T 12   (12         ) -1000000                GEND/TOL INTACT: 1. MALE     2. MID CTY5
003 G/T 13   (13         ) 00-10000                GEND/TOL INTACT: 1. MALE     3. FR/BTWN5
004 G/T 14   (14         ) 0000-100                GEND/TOL INTACT: 1. MALE     4. SML TWN5
005 G/T 15   (15         ) 000000-1                GEND/TOL INTACT: 1. MALE     5. RURAL5
006 G/T 21   (21         ) -1-1-1-1                GEND/TOL INTACT: 2. FEMALE   1. BIG CTY5
007 G/T 22   (22         ) 01000000                GEND/TOL INTACT: 2. FEMALE   2. MID CTY5
008 G/T 23   (23         ) 00010000                GEND/TOL INTACT: 2. FEMALE   3. FR/BTWN5
009 G/T 24   (24         ) 00000100                GEND/TOL INTACT: 2. FEMALE   4. SML TWN5
010 G/T 25   (25         ) 00000001                GEND/TOL INTACT: 2. FEMALE   5. RURAL5

CONDITIONING VARIABLE ID:  BACK0056
DESCRIPTION:               INTERACTION:  GENDER BY PARENTS' EDUCATION
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    GEND/PAR
NAEP ID:                   N/A                     TOTAL NUMBER OF SPECIFIED CONTRASTS:    10
```

```
TYPE OF CONTRAST:          INTERACTION                     NUMBER OF INDEPENDENT CONTRASTS:        4

001 G/P 11    (11          ) 01010101                      GEND/PAR INTACT: 1. MALE     1. < HS
002 G/P 12    (12          ) -1000000                      GEND/PAR INTACT: 1. MALE     2. HS GRAD
003 G/P 13    (13          ) 00-10000                      GEND/PAR INTACT: 1. MALE     3. POST HS
004 G/P 14    (14          ) 0000-100                      GEND/PAR INTACT: 1. MALE     4. COL GRAD
005 G/P 15    (15          ) 000000-1                      GEND/PAR INTACT: 1. MALE     5. PARED-?
006 G/P 21    (21          ) -1-1-1-1                      GEND/PAR INTACT: 2. FEMALE   1. < HS
007 G/P 22    (22          ) 01000000                      GEND/PAR INTACT: 2. FEMALE   2. HS GRAD
008 G/P 23    (23          ) 00010000                      GEND/PAR INTACT: 2. FEMALE   3. POST HS
009 G/P 24    (24          ) 00000100                      GEND/PAR INTACT: 2. FEMALE   4. COL GRAD
010 G/P 25    (25          ) 00000001                      GEND/PAR INTACT: 2. FEMALE   5. PARED-?

CONDITIONING VARIABLE ID:  BACK0057
DESCRIPTION:               INTERACTION:  GENDER BY SCHOOL TYPE
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    GEND/SCH
NAEP ID:                   N/A                             TOTAL NUMBER OF SPECIFIED CONTRASTS:    6
TYPE OF CONTRAST:          INTERACTION                     NUMBER OF INDEPENDENT CONTRASTS:        2

001 G/S 11    (11          ) 0101                          GEND/SCH INTACT: 1. MALE     1. PUBLIC
002 G/S 12    (12          ) -100                          GEND/SCH INTACT: 1. MALE     2. PRIVATE
003 G/S 13    (13          ) 00-1                          GEND/SCH INTACT: 1. MALE     3. CATHOLIC
004 G/S 21    (21          ) -1-1                          GEND/SCH INTACT: 2. FEMALE   1. PUBLIC
005 G/S 22    (22          ) 0100                          GEND/SCH INTACT: 2. FEMALE   2. PRIVATE
006 G/S 23    (23          ) 0001                          GEND/SCH INTACT: 2. FEMALE   3. CATHOLIC

CONDITIONING VARIABLE ID:  BACK0058
DESCRIPTION:               INTERACTION:  RACE/ETHNICITY BY TYPE OF LOCALE (5 CATEGORIES)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    RACE/TOL
NAEP ID:                   N/A                             TOTAL NUMBER OF SPECIFIED CONTRASTS:   25
TYPE OF CONTRAST:          INTERACTION                     NUMBER OF INDEPENDENT CONTRASTS:       16

001 R/T 11    (11          ) 01010101010101010101010101010101 RACE/TOL INTACT: 1. WHI/AI/O 1. BIG CTY5
002 R/T 12    (12          ) -1000000-1000000-1000000-1000000 RACE/TOL INTACT: 1. WHI/AI/O 2. MID CTY5
003 R/T 13    (13          ) 00-1000000-1000000-1000000-10000 RACE/TOL INTACT: 1. WHI/AI/O 3. FR/BTWN5
004 R/T 14    (14          ) 0000-1000000-1000000-1000000-100 RACE/TOL INTACT: 1. WHI/AI/O 4. SML TWN5
005 R/T 15    (15          ) 000000-1000000-1000000-1000000-1 RACE/TOL INTACT: 1. WHI/AI/O 5. RURAL5
006 R/T 21    (21          ) -1-1-1-1000000000000000000000000 RACE/TOL INTACT: 2. BLACK    1. BIG CTY5
007 R/T 22    (22          ) 01000000000000000000000000000000 RACE/TOL INTACT: 2. BLACK    2. MID CTY5
008 R/T 23    (23          ) 00010000000000000000000000000000 RACE/TOL INTACT: 2. BLACK    3. FR/BTWN5
009 R/T 24    (24          ) 00000100000000000000000000000000 RACE/TOL INTACT: 2. BLACK    4. SML TWN5
010 R/T 25    (25          ) 00000001000000000000000000000000 RACE/TOL INTACT: 2. BLACK    5. RURAL5
011 R/T 31    (31          ) 00000000-1-1-1-10000000000000000 RACE/TOL INTACT: 3. HISPANIC 1. BIG CTY5
012 R/T 32    (32          ) 00000000010000000000000000000000 RACE/TOL INTACT: 3. HISPANIC 2. MID CTY5
013 R/T 33    (33          ) 00000000000100000000000000000000 RACE/TOL INTACT: 3. HISPANIC 3. FR/BTWN5
014 R/T 34    (34          ) 00000000000001000000000000000000 RACE/TOL INTACT: 3. HISPANIC 4. SML TWN5
015 R/T 35    (35          ) 00000000000000010000000000000000 RACE/TOL INTACT: 3. HISPANIC 5. RURAL5
```

```
016 R/T 41   (41          ) 0000000000000000-1-1-1-100000000 RACE/TOL INTACT: 4. ASIAN    1. BIG CTY5
017 R/T 42   (42          ) 00000000000000000100000000000000 RACE/TOL INTACT: 4. ASIAN    2. MID CTY5
018 R/T 43   (43          ) 00000000000000000001000000000000 RACE/TOL INTACT: 4. ASIAN    3. FR/BTWN5
019 R/T 44   (44          ) 00000000000000000000010000000000 RACE/TOL INTACT: 4. ASIAN    4. SML TWN5
020 R/T 45   (45          ) 00000000000000000000000100000000 RACE/TOL INTACT: 4. ASIAN    5. RURAL5
021 R/T 51   (51          ) 000000000000000000000000-1-1-1-1 RACE/TOL INTACT: 5. PAC ISLD 1. BIG CTY5
022 R/T 52   (52          ) 00000000000000000000000001000000 RACE/TOL INTACT: 5. PAC ISLD 2. MID CTY5
023 R/T 53   (53          ) 00000000000000000000000000010000 RACE/TOL INTACT: 5. PAC ISLD 3. FR/BTWN5
024 R/T 54   (54          ) 00000000000000000000000000000100 RACE/TOL INTACT: 5. PAC ISLD 4. SML TWN5
025 R/T 55   (55          ) 00000000000000000000000000000001 RACE/TOL INTACT: 5. PAC ISLD 5. RURAL5

CONDITIONING VARIABLE ID:  BACK0059
DESCRIPTION:               INTERACTION:  RACE/ETHNICITY BY PARENTS' EDUCATION
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    RACE/PAR
NAEP ID:                   N/A                          TOTAL NUMBER OF SPECIFIED CONTRASTS:   25
TYPE OF CONTRAST:          INTERACTION                  NUMBER OF INDEPENDENT CONTRASTS:       16

001 R/P 11   (11          ) 01010101010101010101010101010101 RACE/PAR INTACT: 1. WHI/AI/O 1. < HS
002 R/P 12   (12          ) -1000000-1000000-1000000-1000000 RACE/PAR INTACT: 1. WHI/AI/O 2. HS GRAD
003 R/P 13   (13          ) 00-1000000-1000000-1000000-10000 RACE/PAR INTACT: 1. WHI/AI/O 3. POST HS
004 R/P 14   (14          ) 0000-1000000-1000000-1000000-100 RACE/PAR INTACT: 1. WHI/AI/O 4. COL GRAD
005 R/P 15   (15          ) 000000-1000000-1000000-1000000-1 RACE/PAR INTACT: 1. WHI/AI/O 5. PARED-?
006 R/P 21   (21          ) -1-1-1-1000000000000000000000000 RACE/PAR INTACT: 2. BLACK    1. < HS
007 R/P 22   (22          ) 01000000000000000000000000000000 RACE/PAR INTACT: 2. BLACK    2. HS GRAD
008 R/P 23   (23          ) 00010000000000000000000000000000 RACE/PAR INTACT: 2. BLACK    3. POST HS
009 R/P 24   (24          ) 00000100000000000000000000000000 RACE/PAR INTACT: 2. BLACK    4. COL GRAD
010 R/P 25   (25          ) 00000001000000000000000000000000 RACE/PAR INTACT: 2. BLACK    5. PARED-?
011 R/P 31   (31          ) 00000000-1-1-1-10000000000000000 RACE/PAR INTACT: 3. HISPANIC 1. < HS
012 R/P 32   (32          ) 00000000010000000000000000000000 RACE/PAR INTACT: 3. HISPANIC 2. HS GRAD
013 R/P 33   (33          ) 00000000000100000000000000000000 RACE/PAR INTACT: 3. HISPANIC 3. POST HS
014 R/P 34   (34          ) 00000000000001000000000000000000 RACE/PAR INTACT: 3. HISPANIC 4. COL GRAD
015 R/P 35   (35          ) 00000000000000010000000000000000 RACE/PAR INTACT: 3. HISPANIC 5. PARED-?
016 R/P 41   (41          ) 0000000000000000-1-1-1-100000000 RACE/PAR INTACT: 4. ASIAN    1. < HS
017 R/P 42   (42          ) 00000000000000000100000000000000 RACE/PAR INTACT: 4. ASIAN    2. HS GRAD
018 R/P 43   (43          ) 00000000000000000001000000000000 RACE/PAR INTACT: 4. ASIAN    3. POST HS
019 R/P 44   (44          ) 00000000000000000000010000000000 RACE/PAR INTACT: 4. ASIAN    4. COL GRAD
020 R/P 45   (45          ) 00000000000000000000000100000000 RACE/PAR INTACT: 4. ASIAN    5. PARED-?
021 R/P 51   (51          ) 000000000000000000000000-1-1-1-1 RACE/PAR INTACT: 5. PAC ISLD 1. < HS
022 R/P 52   (52          ) 00000000000000000000000001000000 RACE/PAR INTACT: 5. PAC ISLD 2. HS GRAD
023 R/P 53   (53          ) 00000000000000000000000000010000 RACE/PAR INTACT: 5. PAC ISLD 3. POST HS
024 R/P 54   (54          ) 00000000000000000000000000000100 RACE/PAR INTACT: 5. PAC ISLD 4. COL GRAD
025 R/P 55   (55          ) 00000000000000000000000000000001 RACE/PAR INTACT: 5. PAC ISLD 5. PARED-?

CONDITIONING VARIABLE ID:  BACK0060
DESCRIPTION:               INTERACTION:  RACE/ETHNICITY BY SCHOOL TYPE
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    RACE/SCH
NAEP ID:                   N/A                          TOTAL NUMBER OF SPECIFIED CONTRASTS:   15
```

```
TYPE OF CONTRAST:          INTERACTION                  NUMBER OF INDEPENDENT CONTRASTS:       8

001 R/S 11   (11          ) 0101010101010101            RACE/SCH INTACT: 1. WHI/AI/O 1. PUBLIC
002 R/S 12   (12          ) -100-100-100-100            RACE/SCH INTACT: 1. WHI/AI/O 2. PRIVATE
003 R/S 13   (13          ) 00-100-100-100-1            RACE/SCH INTACT: 1. WHI/AI/O 3. CATHOLIC
004 R/S 21   (21          ) -1-1000000000000            RACE/SCH INTACT: 2. BLACK    1. PUBLIC
005 R/S 22   (22          ) 0100000000000000            RACE/SCH INTACT: 2. BLACK    2. PRIVATE
006 R/S 23   (23          ) 0001000000000000            RACE/SCH INTACT: 2. BLACK    3. CATHOLIC
007 R/S 31   (31          ) 0000-1-100000000            RACE/SCH INTACT: 3. HISPANIC 1. PUBLIC
008 R/S 32   (32          ) 0000010000000000            RACE/SCH INTACT: 3. HISPANIC 2. PRIVATE
009 R/S 33   (33          ) 0000000100000000            RACE/SCH INTACT: 3. HISPANIC 3. CATHOLIC
010 R/S 41   (41          ) 00000000-1-10000            RACE/SCH INTACT: 4. ASIAN    1. PUBLIC
011 R/S 42   (42          ) 0000000001000000            RACE/SCH INTACT: 4. ASIAN    2. PRIVATE
012 R/S 43   (43          ) 0000000000010000            RACE/SCH INTACT: 4. ASIAN    3. CATHOLIC
013 R/S 51   (51          ) 000000000000-1-1            RACE/SCH INTACT: 5. PAC ISLD 1. PUBLIC
014 R/S 52   (52          ) 0000000000000100            RACE/SCH INTACT: 5. PAC ISLD 2. PRIVATE
015 R/S 53   (53          ) 0000000000000001            RACE/SCH INTACT: 5. PAC ISLD 3. CATHOLIC

CONDITIONING VARIABLE ID:  BACK0061
DESCRIPTION:               INTERACTION:  TYPE OF LOCALE (5 CATEGORIES) BY PARENT'S EDUCATION
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    TOL5/PAR
NAEP ID:                   N/A                          TOTAL NUMBER OF SPECIFIED CONTRASTS:   25
TYPE OF CONTRAST:          INTERACTION                  NUMBER OF INDEPENDENT CONTRASTS:       16

001 T/P 11   (11          ) 01010101010101010101010101010101 TOL5/PAR INTACT: 1. BIG CTY5 1. < HS
002 T/P 12   (12          ) -1000000-1000000-1000000-1000000 TOL5/PAR INTACT: 1. BIG CTY5 2. HS GRAD
003 T/P 13   (13          ) 00-1000000-1000000-1000000-10000 TOL5/PAR INTACT: 1. BIG CTY5 3. POST HS
004 T/P 14   (14          ) 0000-1000000-1000000-1000000-100 TOL5/PAR INTACT: 1. BIG CTY5 4. COL GRAD
005 T/P 15   (15          ) 000000-1000000-1000000-1000000-1 TOL5/PAR INTACT: 1. BIG CTY5 5. PARED-?
006 T/P 21   (21          ) -1-1-1-1000000000000000000000000 TOL5/PAR INTACT: 2. MID CTY5 1. < HS
007 T/P 22   (22          ) 01000000000000000000000000000000 TOL5/PAR INTACT: 2. MID CTY5 2. HS GRAD
008 T/P 23   (23          ) 00010000000000000000000000000000 TOL5/PAR INTACT: 2. MID CTY5 3. POST HS
009 T/P 24   (24          ) 00000100000000000000000000000000 TOL5/PAR INTACT: 2. MID CTY5 4. COL GRAD
010 T/P 25   (25          ) 00000001000000000000000000000000 TOL5/PAR INTACT: 2. MID CTY5 5. PARED-?
011 T/P 31   (31          ) 00000000-1-1-1-10000000000000000 TOL5/PAR INTACT: 3. FR/BTWN5 1. < HS
012 T/P 32   (32          ) 00000000010000000000000000000000 TOL5/PAR INTACT: 3. FR/BTWN5 2. HS GRAD
013 T/P 33   (33          ) 00000000000100000000000000000000 TOL5/PAR INTACT: 3. FR/BTWN5 3. POST HS
014 T/P 34   (34          ) 00000000000001000000000000000000 TOL5/PAR INTACT: 3. FR/BTWN5 4. COL GRAD
015 T/P 35   (35          ) 00000000000000010000000000000000 TOL5/PAR INTACT: 3. FR/BTWN5 5. PARED-?
016 T/P 41   (41          ) 0000000000000000-1-1-1-100000000 TOL5/PAR INTACT: 4. SML TWN5 1. < HS
017 T/P 42   (42          ) 00000000000000000100000000000000 TOL5/PAR INTACT: 4. SML TWN5 2. HS GRAD
018 T/P 43   (43          ) 00000000000000000001000000000000 TOL5/PAR INTACT: 4. SML TWN5 3. POST HS
019 T/P 44   (44          ) 00000000000000000000010000000000 TOL5/PAR INTACT: 4. SML TWN5 4. COL GRAD
020 T/P 45   (45          ) 00000000000000000000000100000000 TOL5/PAR INTACT: 4. SML TWN5 5. PARED-?
021 T/P 51   (51          ) 000000000000000000000000-1-1-1-1 TOL5/PAR INTACT: 5. RURAL5   1. < HS
022 T/P 52   (52          ) 00000000000000000000000001000000 TOL5/PAR INTACT: 5. RURAL5   2. HS GRAD
023 T/P 53   (53          ) 00000000000000000000000000010000 TOL5/PAR INTACT: 5. RURAL5   3. POST HS
024 T/P 54   (54          ) 00000000000000000000000000000100 TOL5/PAR INTACT: 5. RURAL5   4. COL GRAD
```

```
025 T/P 55   (55        ) 00000000000000000000000000000001 TOL5/PAR INTACT: 5. RURAL5   5. PARED-?

CONDITIONING VARIABLE ID:  BACK0062
DESCRIPTION:               INTERACTION:  TYPE OF LOCALE (5 CATEGORIES) BY SCHOOL TYPE
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    TOL5/SCH
NAEP ID:                   N/A                          TOTAL NUMBER OF SPECIFIED CONTRASTS:   15
TYPE OF CONTRAST:          INTERACTION                  NUMBER OF INDEPENDENT CONTRASTS:        8

001 T/S 11   (11        ) 0101010101010101              TOL5/SCH INTACT: 1. BIG CTY5 1. PUBLIC
002 T/S 12   (12        ) -100-100-100-100              TOL5/SCH INTACT: 1. BIG CTY5 2. PRIVATE
003 T/S 13   (13        ) 00-100-100-100-1              TOL5/SCH INTACT: 1. BIG CTY5 3. CATHOLIC
004 T/S 21   (21        ) -1-1000000000000              TOL5/SCH INTACT: 2. MID CTY5 1. PUBLIC
005 T/S 22   (22        ) 0100000000000000              TOL5/SCH INTACT: 2. MID CTY5 2. PRIVATE
006 T/S 23   (23        ) 0001000000000000              TOL5/SCH INTACT: 2. MID CTY5 3. CATHOLIC
007 T/S 31   (31        ) 0000-1-100000000              TOL5/SCH INTACT: 3. FR/BTWN5 1. PUBLIC
008 T/S 32   (32        ) 0000010000000000              TOL5/SCH INTACT: 3. FR/BTWN5 2. PRIVATE
009 T/S 33   (33        ) 0000000100000000              TOL5/SCH INTACT: 3. FR/BTWN5 3. CATHOLIC
010 T/S 41   (41        ) 00000000-1-10000              TOL5/SCH INTACT: 4. SML TWN5 1. PUBLIC
011 T/S 42   (42        ) 0000000001000000              TOL5/SCH INTACT: 4. SML TWN5 2. PRIVATE
012 T/S 43   (43        ) 0000000000010000              TOL5/SCH INTACT: 4. SML TWN5 3. CATHOLIC
013 T/S 51   (51        ) 000000000000-1-1              TOL5/SCH INTACT: 5. RURAL5   1. PUBLIC
014 T/S 52   (52        ) 0000000000000100              TOL5/SCH INTACT: 5. RURAL5   2. PRIVATE
015 T/S 53   (53        ) 0000000000000001              TOL5/SCH INTACT: 5. RURAL5   3. CATHOLIC

CONDITIONING VARIABLE ID:  BACK0063
DESCRIPTION:               INTERACTION:  PARENTS' EDUCATION BY SCHOOL TYPE
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PARE/SCH
NAEP ID:                   N/A                          TOTAL NUMBER OF SPECIFIED CONTRASTS:   15
TYPE OF CONTRAST:          INTERACTION                  NUMBER OF INDEPENDENT CONTRASTS:        8

001 P/S 11   (11        ) 0101010101010101              PARE/SCH INTACT: 1. < HS     1. PUBLIC
002 P/S 12   (12        ) -100-100-100-100              PARE/SCH INTACT: 1. < HS     2. PRIVATE
003 P/S 13   (13        ) 00-100-100-100-1              PARE/SCH INTACT: 1. < HS     3. CATHOLIC
004 P/S 21   (21        ) -1-1000000000000              PARE/SCH INTACT: 2. HS GRAD  1. PUBLIC
005 P/S 22   (22        ) 0100000000000000              PARE/SCH INTACT: 2. HS GRAD  2. PRIVATE
006 P/S 23   (23        ) 0001000000000000              PARE/SCH INTACT: 2. HS GRAD  3. CATHOLIC
007 P/S 31   (31        ) 0000-1-100000000              PARE/SCH INTACT: 3. POST HS  1. PUBLIC
008 P/S 32   (32        ) 0000010000000000              PARE/SCH INTACT: 3. POST HS  2. PRIVATE
009 P/S 33   (33        ) 0000000100000000              PARE/SCH INTACT: 3. POST HS  3. CATHOLIC
010 P/S 41   (41        ) 00000000-1-10000              PARE/SCH INTACT: 4. COL GRAD 1. PUBLIC
011 P/S 42   (42        ) 0000000001000000              PARE/SCH INTACT: 4. COL GRAD 2. PRIVATE
012 P/S 43   (43        ) 0000000000010000              PARE/SCH INTACT: 4. COL GRAD 3. CATHOLIC
013 P/S 51   (51        ) 000000000000-1-1              PARE/SCH INTACT: 5. PARED-?  1. PUBLIC
014 P/S 52   (52        ) 0000000000000100              PARE/SCH INTACT: 5. PARED-?  2. PRIVATE
015 P/S 53   (53        ) 0000000000000001              PARE/SCH INTACT: 5. PARED-?  3. CATHOLIC

CONDITIONING VARIABLE ID:  BACK0065
```

```
DESCRIPTION:                MSA/NON-MSA
GRADES/ASSESSMENTS:         S04
CONDITIONING VAR LABEL:     MSATSA
NAEP ID:                    MA92FLG                       TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:           CLASS                         NUMBER OF INDEPENDENT CONTRASTS:         2

001 MSA      (0           ) 00                            MSA
002 NON MSA  (1           ) 10                            NON-MSA
003 MSA-MISS (M           ) 01                            MSA MISSING

CONDITIONING VARIABLE ID:   BACK0066
DESCRIPTION:                STATE ADMINISTRATION MONITORED/UNMONITORED SESSION
GRADES/ASSESSMENTS:         S04
CONDITIONING VAR LABEL:     MONITOR
NAEP ID:                    MONSTUD                       TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:           CLASS                         NUMBER OF INDEPENDENT CONTRASTS:         1

001 UNMONIT  (0           ) 0                             UNMONITORED SESSION
002 MONITOR  (1           ) 1                             MONITORED SESSION

CONDITIONING VARIABLE ID:   BACK0067
DESCRIPTION                 INTERACTION:  SCHOOL TYPE BY MONITORED/UNMONITORED SESSION
GRADES/ASSESSMENTS:         S04
CONDITIONING VAR LABEL:     SCHT/MON
NAEP ID:                    N/A                           TOTAL NUMBER OF SPECIFIED CONTRASTS:     6
TYPE OF CONTRAST:           INTERACTION                   NUMBER OF INDEPENDENT CONTRASTS:         2

001 S/M 11   (11          ) 0101                          SCHT/MON INTACT: 1. PUBLIC   1. UNMONIT
002 S/M 12   (12          ) -100                          SCHT/MON INTACT: 1. PUBLIC   2. MONITOR
003 S/M 21   (21          ) 00-1                          SCHT/MON INTACT: 2. PRIVATE  1. UNMONIT
004 S/M 22   (22          ) -1-1                          SCHT/MON INTACT: 2. PRIVATE  2. MONITOR
005 S/M 31   (31          ) 0100                          SCHT/MON INTACT: 3. CATHOLIC 1. UNMONIT
006 S/M 32   (32          ) 0001                          SCHT/MON INTACT: 3. CATHOLIC 2. MONITOR

CONDITIONING VARIABLE ID:   SUBJ0001
DESCRIPTION:                DURING THE PAST MONTH, HOW MANY BOOKS HAVE YOU READ ON YOUR OWN OUTSIDE OF SCHOOL?
GRADES/ASSESSMENTS:         N04, S04, N08, N12
CONDITIONING VAR LABEL:     NBOOKSRD
NAEP ID:                    R810801                       TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:           CLASS                         NUMBER OF INDEPENDENT CONTRASTS:         4

001 NBOOK-0  (1           ) 0000                          NUMBER OF BOOKS READ:  NONE
002 NBOOK-12 (2           ) 1000                          NUMBER OF BOOKS READ:  ONE OR TWO
003 NBOOK-34 (3           ) 0100                          NUMBER OF BOOKS READ:  THREE OR FOUR
004 NBOOK-5+ (4           ) 0010                          NUMBER OF BOOKS READ:  FIVE OR MORE
005 NBOOK-?  (M           ) 0001                          NUMBER OF BOOKS READ:  MISSING

CONDITIONING VARIABLE ID:   SUBJ0002
DESCRIPTION:                WHAT KIND OF READER DO YOU THINK YOU ARE?
```

```
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    KIND RDR
NAEP ID:                   R810201                   TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                     NUMBER OF INDEPENDENT CONTRASTS:        4

001 VGD RDR  (1          ) 0000                      A VERY GOOD READER
002 GOOD RDR (2          ) 1000                      A GOOD READER
003 AVG RDR  (3          ) 0100                      AN AVERAGE READER
004 POOR RDR (4          ) 0010                      A POOR READER
005 RDR-MISS (M          ) 0001                      KIND OF READER:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0003
DESCRIPTION:               HOW OFTEN DO YOU READ FOR FUN ON YOUR OWN TIME?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    READ4FUN
NAEP ID:                   R810901                   TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                     NUMBER OF INDEPENDENT CONTRASTS:        4

001 RD4FUN-1 (1          ) 0000                      READ FOR FUN:  ALMOST EVERY DAY
002 RD4FUN-2 (2          ) 1000                      READ FOR FUN:  ONCE OR TWICE A WEEK
003 RD4FUN-3 (3          ) 0100                      READ FOR FUN:  ONCE OR TWICE A MONTH
004 RD4FUN-4 (4          ) 0010                      READ FOR FUN:  NEVER OR HARDLY EVER
005 RD4FUN-? (M          ) 0001                      READ FOR FUN:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0004
DESCRIPTION:               HOW OFTEN DO YOU TALK TO FRIENDS/FAMILY ABOUT WHAT YOU READ?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    FAMLYRED
NAEP ID:                   R810902                   TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                     NUMBER OF INDEPENDENT CONTRASTS:        4

001 FAMRED-1 (1          ) 0000                      TALK ABOUT READING:  ALMOST EVERY DAY
002 FAMRED-2 (2          ) 1000                      TALK ABOUT READING:  ONCE OR TWICE A WEEK
003 FAMRED-3 (3          ) 0100                      TALK ABOUT READING:  ONCE OR TWICE A MONTH
004 FAMRED-4 (4          ) 0010                      TALK ABOUT READING:  NEVER OR HARDLY EVER
005 FAMRED-? (M          ) 0001                      TALK ABOUT READING:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0006
DESCRIPTION:               HOW OFTEN DO YOU TAKE BOOKS OUT OF THE LIBRARY FOR YOUR OWN ENJOYMENT?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    USELIBRY
NAEP ID:                   R810903                   TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                     NUMBER OF INDEPENDENT CONTRASTS:        4

001 USELIB-1 (1          ) 0000                      USE THE LIBRARY:  ALMOST EVERY DAY
002 USELIB-2 (2          ) 1000                      USE THE LIBRARY:  ONCE OR TWICE A WEEK
003 USELIB-3 (3          ) 0100                      USE THE LIBRARY:  ONCE OR TWICE A MONTH
004 USELIB-4 (4          ) 0010                      USE THE LIBRARY:  NEVER OR HARDLY EVER
005 USELIB-? (M          ) 0001                      USE THE LIBRARY:  MISSINGR HAROLY EVER
```

```
CONDITIONING VARIABLE ID:  SUBJ0009
DESCRIPTION:               HOW OFTEN DOES YOUR TEACHER DISCUSS NEW OR DIFFICULT VOCABULARY?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    VOCAB
NAEP ID:                   R811001                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         4

001 VOCAB-1  (1          ) 0000                           DISCUSS VOCABULARY:  ALMOST EVERY DAY
002 VOCAB-2  (2          ) 1000                           DISCUSS VOCABULARY:  ONCE OR TWICE A WEEK
003 VOCAB-3  (3          ) 0100                           DISCUSS VOCABULARY:  ONCE OR TWICE A MONTH
004 VOCAB-4  (4          ) 0010                           DISCUSS VOCABULARY:  NEVER OR HARDLY EVER
005 VOCAB-?  (M          ) 0001                           DISCUSS VOCABULARY:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0010
DESCRIPTION:               HOW OFTEN DOES YOUR TEACHER ASK STUDENTS TO TALK TO EACH OTHER ABOUT WHAT THEY HAVE READ?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    TALKREAD
NAEP ID:                   R811002                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         4

001 TALKRD-1 (1          ) 0000                           TEACHER ASK TO TALK ABOUT READING:  ALMOST EVERY DAY
002 TALKRD-2 (2          ) 1000                           TEACHER ASK TO TALK ABOUT READING:  ONCE OR TWICE A WEEK
003 TALKRD-3 (3          ) 0100                           TEACHER ASK TO TALK ABOUT READING:  ONCE OR TWICE A MONTH
004 TALKRD-4 (4          ) 0010                           TEACHER ASK TO TALK ABOUT READING:  NEVER OR HARDLY EVER
005 TALKRD-? (M          ) 0001                           TEACHER ASK TO TALK ABOUT READING:  MISSING

CONDITIONING VAR'ABLE ID:  SUBJ0011
DESCRIPTION:               HOW OFTEN DOES YOUR TEACHER ASK YOU TO WORK IN A READING WORKBOOK OR ON A WORKSHEET?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    WBK/WSHT
NAEP ID:                   R811003                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         4

001 WBK/WS-1 (1          ) 0000                           READING WORKBOOK/WORKSHEET:  ALMOST EVERY DAY
002 WBK/WS-2 (2          ) 1000                           READING WORKBOOK/WORKSHEET:  ONCE OR TWICE A WEEK
003 WBK/WS-3 (3          ) 0100                           READING WORKBOOK/WORKSHEET:  ONCE OR TWICE A MONTH
004 WBK/WS-4 (4          ) 0010                           READING WORKBOOK/WORKSHEET:  NEVER OR HARDLY EVER
005 WBK/WS-? (M          ) 0001                           READING WORKBOOK/WORKSHEET:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0012
DESCRIPTION:               HOW OFTEN DOES YOUR TEACHER ASK YOU TO WRITE SOMETHING ABOUT WHAT YOU HAVE READ?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    WRITREAD
NAEP ID:                   R811004                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         4

001 WRTRED-1 (1          ) 0000                           WRITE ABOUT READING:  ALMOST EVERY DAY
002 WRTRED-2 (2          ) 1000                           WRITE ABOUT READING:  ONCE OR TWICE A WEEK
```

```
003 WRTRED-3 (3          ) 0100                           WRITE ABOUT READING:  ONCE OR TWICE A MONTH
004 WRTRED-4 (4          ) 0010                           WRITE ABOUT READING:  NEVER OR HARDLY EVER
005 WRTRED-? (M          ) 0001                           WRITE ABOUT READING:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0013
DESCRIPTION:               HOW OFTEN DOES YOUR TEACHER ASK STUDENTS TO DO A GROUP ACTIVITY/PROJECT ABOUT WHAT THEY HAVE READ?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    READPROJ
NAEP ID:                   R811005                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         4

001 RDPRJ-S1 (1          ) 0000                           PROJECT ABOUT READING: ALMOST EVERY DAY
002 RDPRJ-S2 (2          ) 1000                           PROJECT ABOUT READING: ONCE OR TWICE A WEEK
003 RDPRJ-S3 (3          ) 0100                           PROJECT ABOUT READING: ONCE OR TWICE A MONTH
004 RDPRJ-S4 (4          ) 0010                           PROJECT ABOUT READING: NEVER OR HARDLY EVER
005 RDPRJ-S? (M          ) 0001                           PROJECT ABOUT READING: MISSING

CONDITIONING VARIABLE ID:  SUBJ0014
DESCRIPTION:               HOW OFTEN DOES YOUR TEACHER ASK STUDENTS TO READ ALOUD?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    RD ALOUD
NAEP ID:                   R811006                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         4

001 ALOUD-S1 (1          ) 0000                           READ ALOUD: ALMOST EVERY DAY
002 ALOUD-S2 (2          ) 1000                           READ ALOUD: ONCE OR TWICE A WEEK
003 ALOUD-S3 (3          ) 0100                           READ ALOUD: ONCE OR TWICE A MONTH
004 ALOUD-S4 (4          ) 0010                           READ ALOUD: NEVER OR HARDLY EVER
005 ALOUD-S? (M          ) 0001                           READ ALOUD: MISSING

CONDITIONING VARIABLE ID:  SUBJ0015
DESCRIPTION:               HOW OFTEN DOES YOUR TEACHER ASK YOU TO READ SILENTLY?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    RD SILNT
NAEP ID:                   R811007                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         4

001 SILNT-S1 (1          ) 0000                           READ SILENTLY: ALMOST EVERY DAY
002 SILNT-S2 (2          ) 1000                           READ SILENTLY: ONCE OR TWICE A WEEK
003 SILNT-S3 (3          ) 0100                           READ SILENTLY: ONCE OR TWICE A MONTH
004 SILNT-S4 (4          ) 0010                           READ SILENTLY: NEVER OR HARDLY EVER
005 SILNT-S? (M          ) 0001                           READ SILENTLY: MISSING

CONDITIONING VARIABLE ID:  SUBJ0016
DESCRIPTION:               HOW OFTEN DOES YOUR TEACHER GIVE YOU TIME TO READ BOOKS YOU HAVE CHOSEN YOURSELF?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    RDOWNBKS
NAEP ID:                   R811009                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         4
```

```
001 OWNBK-S1 (1          ) 0000                         BOOKS CHOSEN YOURSELF:  ALMOST EVERY DAY
002 OWNBK-S2 (2          ) 1000                         BOOKS CHOSEN YOURSELF:  OR TWICE A WEEK
003 OWNBK-S3 (3          ) 0100                         BOOKS CHOSEN YOURSELF:  ONCE OR TWICE A MONTH
004 OWNBK-S4 (4          ) 0010                         BOOKS CHOSEN YOURSELF:  NEVER OR HARDLY EVER
005 OWNBK-S? (M          ) 0001                         BOOKS CHOSEN YOURSELF:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0025
DESCRIPTION:               HOW OFTEN DOES YOUR TEACHER TAKE YOU TO THE SCHOOL LIBRARY?
GRADES/ASSESSMENTS:        N04, S04
CONDITIONING VAR LABEL:    TAKE2LIB
NAEP ID:                   R811013                      TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:        4

001 TK2LIB-1 (1          ) 0000                         TEACHER TAKE TO LIBRARY:  ALMOST EVERY DAY
002 TK2LIB-2 (2          ) 1000                         TEACHER TAKE TO LIBRARY:  ONCE OR TWICE A WEEK
003 TK2LIB-3 (3          ) 0100                         TEACHER TAKE TO LIBRARY:  ONCE OR TWICE A MONTH
004 TK2LIB-4 (4          ) 0010                         TEACHER TAKE TO LIBRARY:  NEVER OR HARDLY EVER
005 TK2LIB-? (M          ) 0001                         TEACHER TAKE TO LIBRARY:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0026
DESCRIPTION:               DO YOU OR YOUR TEACHER SAVE YOUR READING WORK IN A FOLDER OR PORTFOLIO?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PRTFOLIO
NAEP ID:                   R820001                      TOTAL NUMBER OF SPECIFIED CONTRASTS:    3
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:        2

001 PFOLIO-Y (1          ) 00                           READING PORTFOLIO:  YES
002 PFOLIO-N (2          ) 10                           READING PORTFOLIO:  NO
003 PFOLIO-? (M          ) 01                           READING PORTFOLIO:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0028
DESCRIPTION:               ABOUT HOW MANY QUESTIONS DID YOU GET RIGHT ON THE READING TEST?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    #QUESTN+
NAEP ID:                   RM00101                      TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:        4

001 #QUEST+1 (1          ) 0000                         NUMBER QUESTIONS RIGHT:  ALMOST ALL
002 #QUEST+2 (2          ) 1000                         NUMBER QUESTIONS RIGHT:  MORE THAN HALF
003 #QUEST+3 (3          ) 0100                         NUMBER QUESTIONS RIGHT:  ABOUT HALF
004 #QUEST+4 (4          ) 0010                         NUMBER QUESTIONS RIGHT:  LESS THAN HALF
005 #QUEST+? (M          ) 0001                         NUMBER QUESTIONS RIGHT:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0029
DESCRIPTION:               HOW HARD WAS THIS READING TEST COMPARED TO OTHERS?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    TEST DIF
NAEP ID:                   RM00201                      TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:        4
```

```
001 TESTOIF1 (1          ) 0000                      TEST DIFFICULTY:  MUCH HARDER THAN OTHERS
002 TESTDIF2 (2          ) 1000                      TEST DIFFICULTY:  HARDER THAN OTHERS
003 TESTDIF3 (3          ) 0100                      TEST DIFFICULTY:  ABOUT AS HARD AS OTHERS
004 TESTDIF4 (4          ) 0010                      TEST DIFFICULTY:  EASIER THAN OTHERS
005 TESTDIF? (M          ) 0001                      TEST DIFFICULTY:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0030
DESCRIPTION:               HOW HARD DID YOU TRY ON THIS TEST COMPARED TO OTHER READING TESTS?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    TEST EFF
NAEP ID:                   RM00301                   TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                     NUMBER OF INDEPENDENT CONTRASTS:         4

001 TESTEFF1 (1          ) 0000                      TEST EFFORT:  MUCH HARDER THAN OTHERS
002 TESTEFF2 (2          ) 1000                      TEST EFFORT:  HARDER THAN OTHERS
003 TESTEFF3 (3          ) 0100                      TEST EFFORT:  ABOUT AS HARD AS OTHERS
004 TESTEFF4 (4          ) 0010                      TEST EFFORT:  NOT AS HARD AS OTHERS
005 TESTEFF? (M          ) 0001                      TEST EFFORT:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0031
DESCRIPTION:               HOW IMPORTANT WAS IT TO YOU TO DO WELL ON THE READING TEST?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    TEST IMP
NAEP ID:                   RM00401                   TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                     NUMBER OF INDEPENDENT CONTRASTS:         4

001 TESTIMP1 (1          ) 0000                      TEST IMPORTANCE:  VERY IMPORTANT
002 TESTIMP2 (2          ) 1000                      TEST IMPORTANCE:  IMPORTANT
003 TESTIMP3 (3          ) 0100                      TEST IMPORTANCE:  SOMEWHAT IMPORTANT
004 TESTIMP4 (4          ) 0010                      TEST IMPORTANCE:  NOT VERY IMPORTANT
005 TESTIMP? (M          ) 0001                      TEST IMPORTANCE:  MISSING

CONDITIONING VARIABLE ID:  SUBJ0032
DESCRIPTION:               HOW OFTEN WERE YOU ASKED TO WRITE LONG ANSWERS ON READING TESTS?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    LONG ANS
NAEP ID:                   RM00501                   TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                     NUMBER OF INDEPENDENT CONTRASTS:         4

001 LNGANSW1 (1          ) 0000                      LONG ANSWERS:  AT LEAST ONCE A WEEK
002 LNGANSW2 (2          ) 1000                      LONG ANSWERS:  ONCE OR TWICE A MONTH
003 LNGANSW3 (3          ) 0100                      LONG ANSWERS:  ONCE OR TWICE A YEAR
004 LNGANSW4 (4          ) 0010                      LONG ANSWERS:  NEVER
005 LNGANSW? (M          ) 0001                      LONG ANSWERS:  MISSING

CONDITIONING VARIABLE ID:  SCHL0001
DESCRIPTION:               SCHOOL LEVEL AVERAGE READING NORMIT (MISSING VS NON-MISSING)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    SCH NORM
```

```
NAEP ID:                  SCHNORM                         TOTAL NUMBER OF SPECIFIED CONTRASTS:    2
TYPE OF CONTRAST:         CLASS                           NUMBER OF INDEPENDENT CONTRASTS:        1

001 SCHNRM-? (M         ) 0                               SCHOOL LEVEL AVERAGE READING NORMIT MISSING
002 SCHNRM-Y (@         ) 1                               SCHOOL LEVEL AVERAGE READING NORMIT NOT-MISSING

CONDITIONING VARIABLE ID: SCHL0002
DESCRIPTION:              SCHOOL LEVEL AVERAGE READING NORMIT
GRADES/ASSESSMENTS:       N04, S04, N08, N12
CONDITIONING VAR LABEL:   SNRM-LIN
NAEP ID:                  SCHNORM                         TOTAL NUMBER OF SPECIFIED CONTRASTS:    2
TYPE OF CONTRAST:         SCALE                           NUMBER OF INDEPENDENT CONTRASTS:        1

001 SNRM-LIN (#         ) (F8.4)                          SCHOOL LEVEL AVERAGE READING NORMIT MEAN
002 SNRM-LIN (M         ) 0                               SCHOOL LEVEL AVERAGE READING NORMIT MISSING

CONDITIONING VARIABLE ID: SCHL0003
DESCRIPTION:              HOW IS 4TH GRADE ORGANIZED AT YOUR SCHOOL?
GRADES/ASSESSMENTS:       N04, S04
CONDITIONING VAR LABEL:   SCH ORG4
NAEP ID:                  C030900                         TOTAL NUMBER OF SPECIFIED CONTRASTS:    4
TYPE OF CONTRAST:         CLASS                           NUMBER OF INDEPENDENT CONTRASTS:        3

001 SELFCONT (1         ) 000                             4TH GRADE ORGANIZATION:  SELF CONTAINED
002 DEPTLIZD (2         ) 100                             4TH GRADE ORGANIZATION:  DEPARTMENTALIZED
003 REGROUPD (3         ) 010                             4TH GRADE ORGANIZATION:  REGROUPED
004 SCH4ORG? (M         ) 001                             4TH GRADE ORGANIZATION:  MISSING

CONDITIONING VARIABLE ID: SCHL0005
DESCRIPTION:              ARE 4TH GRADERS ASSIGNED TO CLASSES BY ABILITY?
GRADES/ASSESSMENTS:       N04, S04
CONDITIONING VAR LABEL:   CLA ABL4
NAEP ID:                  C031100                         TOTAL NUMBER OF SPECIFIED CONTRASTS:    3
TYPE OF CONTRAST:         CLASS                           NUMBER OF INDEPENDENT CONTRASTS:        2

001 CLAABL-Y (1         ) 00                              4TH GRADERS ASSIGNED BY ABILITY:  YES
002 CLAABL-N (2         ) 10                              4TH GRADERS ASSIGNED BY ABILITY:  NO
003 CLAABL-? (M         ) 01                              4TH GRADERS ASSIGNED BY ABILITY:  MISSING

CONDITIONING VARIABLE ID: SCHL0008
DESCRIPTION:              HAS READING BEEN IDENTIFIED AS A PRIORITY? (GRADES 4 AND 8)
GRADES/ASSESSMENTS:       N04, S04, N08
CONDITIONING VAR LABEL:   RD PRIOR
NAEP ID:                  C031601                         TOTAL NUMBER OF SPECIFIED CONTRASTS:    3
TYPE OF CONTRAST:         CLASS                           NUMBER OF INDEPENDENT CONTRASTS:        2

001 RD PRI-Y (1         ) 00                              READING PRIORITY:  YES
002 RD PRI-N (2         ) 10                              READING PRIORITY:  NO
003 RD PRI-? (M         ) 01                              READING PRIORITY:  MISSING
```

```
CONDITIONING VARIABLE ID:  SCHL0009
DESCRIPTION:               WHAT PERCENT OF STUDENTS RECEIVE SUBSIDIZED LUNCH?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    %SUB LUN
NAEP ID:                   C032001                       TOTAL NUMBER OF SPECIFIED CONTRASTS:    6
TYPE OF CONTRAST:          CLASS                         NUMBER OF INDEPENDENT CONTRASTS:        5

001 %SUBLUN1 (1,2,3      ) 00000                         PERCENT SUBSIDIZED LUNCH:  NONE, 1-5%, 6-10%
002 %SUBLUN2 (4          ) 10000                         PERCENT SUBSIDIZED LUNCH:  11-25%
003 %SUBLUN3 (5          ) 01000                         PERCENT SUBSIDIZED LUNCH:  26-50%
004 %SUBLUN4 (6          ) 00100                         PERCENT SUBSIDIZED LUNCH:  51-75%
005 %SUBLUN5 (7,8        ) 00010                         PERCENT SUBSIDIZED LUNCH:  76-90%, OVER 90%
006 %SUBLUN? (M          ) 00001                         PERCENT SUBSIDIZED LUNCH:  MISSING

CONDITIONING VARIABLE ID:  SCHL0010
DESCRIPTION:               WHAT PERCENT OF STUDENTS RECEIVE REMEDIAL READING INSTRUCTION?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    %REMREAD
NAEP ID:                   C032002                       TOTAL NUMBER OF SPECIFIED CONTRASTS:    6
TYPE OF CONTRAST:          CLASS                         NUMBER OF INDEPENDENT CONTRASTS:        5

001 %REMRED1 (1,2        ) 00000                         PERCENT REMEDIAL READING:  NONE, 1-5%
002 %REMRED2 (3          ) 10000                         PERCENT REMEDIAL READING:  6-10%
003 %REMRED3 (4          ) 01000                         PERCENT REMEDIAL READING:  11-25%
004 %REMRED4 (5,6        ) 00100                         PERCENT REMEDIAL READING:  26-50%, 51-75%
005 %REMRED5 (7,8        ) 00010                         PERCENT REMEDIAL READING:  76-90%, OVER 90%
006 %REMRED? (M          ) 00001                         PERCENT REMEDIAL READING:  MISSING

CONDITIONING VARIABLE ID:  SCHL0015
DESCRIPTION:               WHAT PERCENTAGE OF STUDENTS ARE ENROLLED AT BEGINNING AND END OF SCHOOL YEAR?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    %ENR@EOY
NAEP ID:                   C033700                       TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                         NUMBER OF INDEPENDENT CONTRASTS:        4

001 %ENREOY1 (1          ) 0000                          PERCENT ENROLLED AT END OR YEAR: 98-100%
002 %ENREOY2 (2          ) 1000                          PERCENT ENROLLED AT END OF YEAR: 95-97%
003 %ENREOY3 (3          ) 0100                          PERCENT ENROLLED AT END OF YEAR: 90-94%
004 %ENREOY4 (4          ) 0010                          PERCENT ENROLLED AT END OF YEAR: LESS THAN 90%
005 %ENREOY? (M          ) 0001                          PERCENT ENROLLED AT END OF YEAR: MISSING

CONDITIONING VARIABLE ID:  SCHL0016
DESCRIPTION:               DOES SCHOOL INVOLVE PARENTS AS AIDES IN CLASS?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PAR AIDE
NAEP ID:                   C032207                       TOTAL NUMBER OF SPECIFIED CONTRASTS:    4
TYPE OF CONTRAST:          CLASS                         NUMBER OF INDEPENDENT CONTRASTS:        3
```

337

338

```
001 PARAID-R (1          ) 000                        PARENTS AS AIDES IN CLASS:  ROUTINELY
002 PARAID-O (2          ) 100                        PARENTS AS AIDES IN CLASS:  OCCASIONALLY
003 PARAID-N (3          ) 010                        PARENTS AS AIDES IN CLASS:  NO
004 PARAID-? (M          ) 001                        PARENTS AS AIDES IN CLASS:  MISSING

CONDITIONING VARIABLE ID:  SCHL0017
DESCRIPTION:               DOES YOUR SCHOOL HAVE PARENTS REVIEW OR SIGN STUDENTS' HOMEWORK?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PARREVHW
NAEP ID:                   C032209                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         2

001 PARHW-RO (1          ) 00                         PARENTS REVIEW HOMEWORK:  YES, ROUTINELY
002 PARHW-OC (2          ) 10                         PARENTS REVIEW HOMEWORK:  YES, OCCASIONALLY
003 PARHW-N? (3,M        ) 01                         PARENTS REVIEW HOMEWORK:  NO, MISSING

CONDITIONING VARIABLE ID:  SCHL0018
DESCRIPTION:               DOES YOUR SCHOOL ASSIGN HOMEWORK FOR STUDENTS TO DO WITH PARENTS?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    STUPARHW
NAEP ID:                   C032210                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         2

001 S/PHW-RO (1          ) 00                         STUDENT/PARENT HOMEWORK:  YES, ROUTINELY
002 S/PHW-OC (2          ) 10                         STUDENT/PARENT HOMEWORK:  YES, OCCASIONALLY
003 S/PHW-N? (3,M        ) 01                         STUDENT/PARENT HOMEWORK:  NO, MISSING

CONDITIONING VARIABLE ID:  SCHL0019
DESCRIPTION:               DOES YOUR SCHOOL HAVE A PARENT VOLUNTEER PROGRAM?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PARVOLPG
NAEP ID:                   C032211                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         2

001 VOLPG-RO (1          ) 00                         PARENT VOLUNTEER PROGRAM:  YES, ROUTINELY
002 VOLPG-OC (2          ) 10                         PARENT VOLUNTEER PROGRAM:  YES, OCCASIONALLY
003 VOLPG-N? (3,M        ) 01                         PARENT VOLUNTEER PROGRAM:  NO, MISSING

CONDITIONING VARIABLE ID:  SCHL0020
DESCRIPTION:               DOES YOUR SCHOOL RECEIVE CHAPTER 1 FUNDING?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    SCHCHAP1
NAEP ID:                   C036701                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         2

001 SCHCH1-Y (1          ) 00                         SCHOOL CHAPTER 1 FUNDING:  YES
002 SCHCH1-N (2          ) 10                         SCHOOL CHAPTER 1 FUNDING:  NO
003 SCHCH1-? (M          ) 01                         SCHOOL CHAPTER 1 FUNDING:  MISSING
```

```
CONDITIONING VARIABLE ID:  SCHL0021
DESCRIPTION:               WHAT PERCENTAGE OF STUDENTS IN YOUR SCHOOL ARE CHAPTER 1 ELIGIBLE?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    SCH%CHP1
NAEP ID:                   C036801                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     6
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         5

001 %CHAP1-1 (1          ) 00000                          PERCENT CHAPTER 1 ELIGIBLE STUDENTS:  10% OR BELOW
002 %CHAP1-2 (2          ) 10000                          PERCENT CHAPTER 1 ELIGIBLE STUDENTS:  11-25%
003 %CHAP1-3 (3          ) 01000                          PERCENT CHAPTER 1 ELIGIBLE STUDENTS:  26-75%
004 %CHAP1-4 (4          ) 00100                          PERCENT CHAPTER 1 ELIGIBLE STUDENTS:  76-99%
005 %CHAP1-5 (5          ) 00010                          PERCENT CHAPTER 1 ELIGIBLE STUDENTS:  100%
006 %CHAP1-? (M          ) 00001                          PERCENT CHAPTER 1 ELIGIBLE STUDENTS:  MISSING

CONDITIONING VARIABLE ID:  SCHL0022
DESCRIPTION:               PERCENT OF STUDENTS FROM A RURAL AREA (MISSING VS NON-MISSING)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    SCH%RURL
NAEP ID:                   C036201                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         1

001 SCH%RURL (M          ) 0                              PERCENT OF STUDENTS - RURAL:  MISSING
002 SCH%RURL (@          ) 1                              PERCENT OF STUDENTS - RURAL:  NOT MISSING

CONDITIONING VARIABLE ID:  SCHL0023
DESCRIPTION:               PERCENT OF STUDENTS FROM A RURAL AREA OF LESS THAN 2,500 (LINEAR)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    RUR%-LIN
NAEP ID:                   C036201                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     1
TYPE OF CONTRAST:          LINEAR                         NUMBER OF INDEPENDENT CONTRASTS:         1

001 RUR%-LIN (0-100,M=0  )  0.0 +  1.0*X                  PERCENT OF STUDENTS FROM A RURAL AREA OF LESS THAN 2,500 (LINEAR)

CONDITIONING VARIABLE ID:  SCHL0026
DESCRIPTION:               PERCENT OF STUDENTS FROM A CITY (MISSING VS NON-MISSING)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    SCH%CITY
NAEP ID:                   C036203                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:         1

001 SCH%CITY (M          ) 0                              PERCENT OF STUDENTS - CITY:  MISSING
002 SCH%CITY (@          ) 1                              PERCENT OF STUDENTS - CITY:  NOT MISSING

CONDITIONING VARIABLE ID:  SCHL0027
DESCRIPTION:               PERCENT OF STUDENTS FROM A TOWN OF 10,000 OR MORE (LINEAR)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    CITY%-LN
NAEP ID:                   C036203                        TOTAL NUMBER OF SPECIFIED CONTRASTS:     1
TYPE OF CONTRAST:          LINEAR                         NUMBER OF INDEPENDENT CONTRASTS:         1
```

```
001 CITY%-LN (0-100,M=0  )  0.0 +  1.0*X              PERCENT OF STUDENTS FROM A TOWN OF 10,000 OR MORE (LINEAR)

CONDITIONING VARIABLE ID:  SCHL0028
DESCRIPTION:               PERCENT OF STUDENTS WITH PROFESSIONAL PARENTS (MISSING VS NON-MISSING)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PAR%PROF
NAEP ID:                   C036301                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         1

001 PAR%PROF (M          ) 0                          PERCENT OF STUDENTS - PARENTS IN PROFESSIONS:  MISSING
002 PAR%PROF (@          ) 1                          PERCENT OF STUDENTS - PARENTS IN PROFESSIONS:  NOT MISSING

CONDITIONING VARIABLE ID:  SCHL0029
DESCRIPTION:               PERCENT OF STUDENTS WITH PARENTS IN PROFESSIONAL/MANAGEMENT POSITIONS (LINEAR)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PROF%-LN
NAEP ID:                   C036301                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     1
TYPE OF CONTRAST:          LINEAR                     NUMBER OF INDEPENDENT CONTRASTS:         1

001 PROF%-LN (0-100,M=0  )  0.0 +  1.0*X              PERCENT OF STUDENTS WITH PARENTS IN PROFESSIONAL/MANAGEMENT POSITIONS (LINEAR)

CONDITIONING VARIABLE ID:  SCHL0032
DESCRIPTION:               PERCENT OF STUDENTS WITH PARENTS IN BLUE-COLLAR POSITIONS (MISSING VS NON-MISSING)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PAR%BCOL
NAEP ID:                   C036303                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         1

001 PAR%BCOL (M          ) 0                          PERCENT OF STUDENTS - PARENTS IN BLUE-COLLAR POSITIONS:  MISSING
002 PAR%BCOL (@          ) 1                          PERCENT OF STUDENTS - PARENTS IN BLUE-COLLAR POSITIONS:  NOT MISSING

CONDITIONING VARIABLE ID:  SCHL0033
DESCRIPTION:               PERCENT OF STUDENTS WITH PARENTS IN BLUE-COLLAR POSITIONS (LINEAR)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    BCOL%-LN
NAEP ID:                   C036303                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     1
TYPE OF CONTRAST:          LINEAR                     NUMBER OF INDEPENDENT CONTRASTS:         1

001 BCOL%-LN (0-100,M=0  )  0.0 +  1.0*X              PERCENT OF STUDENTS WITH PARENTS IN BLUE-COLLAR POSITIONS (LINEAR)

CONDITIONING VARIABLE ID:  SCHL0034
DESCRIPTION:               PERCENT OF STUDENTS WITH PARENTS WHO ARE FARM WORKERS (MISSING VS NON-MISSING)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PAR%FARM
NAEP ID:                   C036304                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         1

001 PAR%FARM (M          ) 0                          PERCENT OF STUDENTS - PARENTS WHO ARE FARM WORKERS:  MISSING
002 PAR%FARM (@          ) 1                          PERCENT OF STUDENTS - PARENTS WHO ARE FARM WORKERS:  NOT MISSING
```

```
CONDITIONING VARIABLE ID:  SCHL0035
DESCRIPTION:               PERCENT OF STUDENTS WITH PARENTS WHO ARE FARM WORKERS (LINEAR)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    FARM%-LN
NAEP ID:                   C036304                   TOTAL NUMBER OF SPECIFIED CONTRASTS:     1
TYPE OF CONTRAST:          LINEAR                    NUMBER OF INDEPENDENT CONTRASTS:         1

001 FARM%-LN (0-100,M=0  )  0.0 +  1.0*X             PERCENT OF STUDENTS WITH PARENTS WHO ARE FARM WORKERS (LINEAR)

CONDITIONING VARIABLE ID:  SCHL0036
DESCRIPTION:               PERCENT OF STUDENTS WITH PARENTS WHO ARE IRREGULARLY EMPLOYED (MISSING VS NON-MISSING)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PAR%IRRE
NAEP ID:                   C036305                   TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                     NUMBER OF INDEPENDENT CONTRASTS:         1

001 PAR%IRRE (M          ) 0                         PERCENT OF STUDENTS - PARENTS WHO ARE IRREGULARLY EMPLOYED:  MISSING
002 PAR%IRRE (@          ) 1                         PERCENT OF STUDENTS - PARENTS WHO ARE IRREGULARLY EMPLOYED:  NOT MISSING

CONDITIONING VARIABLE ID:  SCHL0037
DESCRIPTION:               PERCENT OF STUDENTS WITH PARENTS WHO ARE IRREGULARLY EMPLOYED BUT NOT ON WELFARE (LINEAR)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    IRRE%-LN
NAEP ID:                   C036305                   TOTAL NUMBER OF SPECIFIED CONTRASTS:     1
TYPE OF CONTRAST:          LINEAR                    NUMBER OF INDEPENDENT CONTRASTS:         1

001 IRRE%-LN (0-100,M=0  )  0.0 +  1.0*X             PERCENT OF STUDENTS WITH PARENTS WHO ARE IRREGULARLY EMPLOYED BUT NOT ON WELFARE (LINEAR)

CONDITIONING VARIABLE ID:  SCHL0038
DESCRIPTION:               PERCENT OF STUDENTS WITH PARENTS WHO ARE WELFARE RECIPIENTS (MISSING VS NON-MISSING)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    PAR%WELF
NAEP ID:                   C036306                   TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                     NUMBER OF INDEPENDENT CONTRASTS:         1

001 PAR%WELF (M          ) 0                         PERCENT OF STUDENTS - PARENTS WHO ARE WELFARE RECIPIENTS:  MISSING
002 PAR%WELF (@          ) 1                         PERCENT OF STUDENTS - PARENTS WHO ARE WELFARE RECIPIENTS:  NOT MISSING

CONDITIONING VARIABLE ID:  SCHL0039
DESCRIPTION:               PERCENT OF STUDENTS WITH PARENTS WHO ARE WELFARE RECIPIENTS (LINEAR)
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    WELF%-LN
NAEP ID:                   C036306                   TOTAL NUMBER OF SPECIFIED CONTRASTS:     1
TYPE OF CONTRAST:          LINEAR                    NUMBER OF INDEPENDENT CONTRASTS:         1

001 WELF%-LN (0-100,M=0  )  0.0 +  1.0*X             PERCENT OF STUDENTS WITH PARENTS WHO ARE WELFARE RECIPIENTS (LINEAR)

CONDITIONING VARIABLE ID:  SCHL0042
```

```
DESCRIPTION:                 WHAT IS THE PRIMARY WAY YOUR LIBRARY IS STAFFED?
GRADES/ASSESSMENTS:          N04, S04, N08, N12
CONDITIONING VAR LABEL:      LIBSTAFF
NAEP ID:                     C036601                   TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:            CLASS                     NUMBER OF INDEPENDENT CONTRASTS:         4

001 NO LIBRY (1          ) 0000                        LIBRARY STAFFING:  NO LIBRARY IN SCHOOL
002 LSTAF-NO (2          ) 1000                        LIBRARY STAFFING:  LIBRARY IN SCHOOL BUT NO/VOLUNTEER STAFF
003 LSTAF-PT (3          ) 0100                        LIBRARY STAFFING:  PART-TIME STAFF
004 LSTAF-FT (4          ) 0010                        LIBRARY STAFFING:  FULL-TIME STAFF
005 LSTAF-?  (M          ) 0001                        LIBRARY STAFFING:  MISSING

CONDITIONING VARIABLE ID:  SCHL0047
DESCRIPTION:               ARE COMPUTERS AVAILIABE AT ALL TIMES IN CLASSROOMS?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    COMP CLS
NAEP ID:                   C035701                     TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                       NUMBER OF INDEPENDENT CONTRASTS:         1

001 CMPCLS-Y (1          ) 0                           COMPUTERS AVAILABLE IN CLASS:  YES
002 CMPCL-N? (2,M        ) 1                           COMPUTERS AVIALABLE IN CLASS:  NO, MISSING

CONDITIONING VARIABLE ID:  SCHL0048
DESCRIPTION:               ARE COMPUTERS GROUPED IN A SEPARATE COMPUTER LABORATORY AVAILABLE TO CLASSES?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    COMP LAB
NAEP ID:                   C035702                     TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                       NUMBER OF INDEPENDENT CONTRASTS:         1

001 CMPLAB-Y (1          ) 0                           COMPUTERS IN A LAB:  YES
002 CMPLB-N? (2,M        ) 1                           COMPUTERS IN A LAB:  NO, MISSING

CONDITIONING VARIABLE ID:  SCHL0049
DESCRIPTION:               ARE COMPUTERS AVAILABLE TO BRING TO CLASSROOMS WHEN NEEDED?
GRADES/ASSESSMENTS:        N04, S04, N08, N12
CONDITIONING VAR LABEL:    COMP BRG
NAEP ID:                   C035703                     TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                       NUMBER OF INDEPENDENT CONTRASTS:         1

001 CMPBRG-Y (1          ) 0                           BRING COMPUTERS TO CLASS:  YES
002 CMPBR-N? (2,M        ) 1                           BRING COMPUTERS TO CLASS:  NO, MISSING

CONDITIONING VARIABLE ID:  TCHR0001
DESCRIPTION:               TEACHER MATCH STATUS WITH STUDENT
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_MATCH
NAEP ID:                   TCHMTCH                     TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                       NUMBER OF INDEPENDENT CONTRASTS:         2
```

```
001 TMCH-NO  (1,M        ) 00                              TEACHER MATCH:  NO MATCH
002 TMCH-PAR (2          ) 10                              TEACHER MATCH:  PARTIAL MATCH
003 TMCH-COM (3          ) 01                              TEACHER MATCH:  COMPLETE MATCH

CONDITIONING VARIABLE ID:  TCHR0002
DESCRIPTION:               TEACHER GENDER
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_GENDER
NAEP ID:                   T040001                         TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                           NUMBER OF INDEPENDENT CONTRASTS:         2

001 T_MALE   (1          ) 00                              TEACHER GENDER:  MALE
002 T_FEMALE (2          ) 10                              TEACHER GENDER:  FEMALE
003 T_SEX-?  (M          ) 01                              TEACHER GENDER:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0003
DESCRIPTION:               YEARS TEACHING ELEMENTARY/SECONDARY SCHOOL
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_YRSEXP
NAEP ID:                   T040301                         TOTAL NUMBER OF SPECIFIED CONTRASTS:     6
TYPE OF CONTRAST:          CLASS                           NUMBER OF INDEPENDENT CONTRASTS:         5

001 T_YREXP1 (1          ) 00000                           YEARS TEACHING:  2 OR LESS YEARS
002 T_YREXP2 (2          ) 10000                           YEARS TEACHING:  3-5 YEARS
003 T_YREXP3 (3          ) 01000                           YEARS TEACHING:  6-10 YEARS
004 T_YREXP4 (4          ) 00100                           YEARS TEACHING:  11-24 YEARS
005 T_YREXP5 (5          ) 00010                           YEARS TEACHING:  25 OR MORE YEARS
006 T_YREXP? (M          ) 00001                           YEARS TEACHING:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0004
DESCRIPTION:               TEACHER RACE/ETHNICITY
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_RACE
NAEP ID:                   T050801                         TOTAL NUMBER OF SPECIFIED CONTRASTS:     7
TYPE OF CONTRAST:          CLASS                           NUMBER OF INDEPENDENT CONTRASTS:         6

001 T_WHITE  (1          ) 000000                          TEACHER RACE/ETHNICITY:  WHITE
002 T_BLACK  (2          ) 100000                          TEACHER RACE/ETHNICITY:  BLACK
003 T_HISP   (3          ) 010000                          TEACHER RACE/ETHNICITY:  HISPANIC
004 T_ASIAN  (4          ) 001000                          TEACHER RACE/ETHNICITY:  ASIAN
005 T_PAC IS (5          ) 000100                          TEACHER RACE/ETHNICITY:  PACIFIC ISLANDER
006 T_AM IND (6          ) 000010                          TEACHER RACE/ETHNICITY:  AMERICAN INDIAN/ALASKAN NATIV
007 T_RACE-? (M          ) 000001                          TEACHER RACE/ETHNICITY:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0005
DESCRIPTION:               TEACHER GENERAL CERTIFICATION (ELEMENTARY, MIDDLE/JUNIOR, HIGH SCHOOL EDUCATION)
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    CERT GEN
NAEP ID:                   T040501                         TOTAL NUMBER OF SPECIFIED CONTRASTS:     4
```

```
TYPE OF CONTRAST:         CLASS                          NUMBER OF INDEPENDENT CONTRASTS:        3

001 CERTG-Y  (1          ) 000                           TEACHER GENERAL CERTIFICATION:  YES
002 CERTG-N  (2          ) 100                           TEACHER GENERAL CERTIFICATION:  NO
003 CERTG-NS (3          ) 010                           TEACHER GENERAL CERTIFICATION:  NOT OFFERED IN STATE
004 CERTG-?  (M          ) 001                           TEACHER GENERAL CERTIFICATION:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0006
DESCRIPTION:               TEACHER CERTIFICATION IN READING
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    CERT RED
NAEP ID:                   T040502                       TOTAL NUMBER OF SPECIFIED CONTRASTS:    4
TYPE OF CONTRAST:          CLASS                         NUMBER OF INDEPENDENT CONTRASTS:        3

001 CERTR-Y  (1          ) 000                           TEACHER READING CERTIFICATION:  YES
002 CERTR-N  (2          ) 100                           TEACHER READING CERTIFICATION:  NO
003 CERTR-NS (3          ) 010                           TEACHER READING CERTIFICATION:  NOT OFFERED IN STATE
004 CERTR-?  (M          ) 001                           TEACHER READING CERTIFICATION:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0007
DESCRIPTION:               TEACHER CERTIFICATION MIDDLE/JUNIOR HIGH SCHOOL/SECONDARY ENGLISH/LANGUAGE ARTS
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    CERT LAN
NAEP ID:                   T040508                       TOTAL NUMBER OF SPECIFIED CONTRASTS:    4
TYPE OF CONTRAST:          CLASS                         NUMBER OF INDEPENDENT CONTRASTS:        3

001 CERTL-Y  (1          ) 000                           TEACHER ENGLISH/LANGUAGE ARTS CERTIFICATION: YES
002 CERTL-N  (2          ) 100                           TEACHER ENGLISH/LANGUAGE ARTS CERTIFICATION: NO
003 CERTL-NS (3          ) 010                           TEACHER ENGLISH/LANGUAGE ARTS CERTIFICATION: NOT OFFERED IN STATE
004 CERTL-?  (M          ) 001                           TEACHER ENGLISH/LANGUAGE ARTS CERTIFICATION: MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0009
DESCRIPTION:               HOW WELL DOES YOUR SCHOOL PROVIDE YOU WITH INSTRUCTIONAL MATERIAL/RESOURCES YOU NEED?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    RESOURCE
NAEP ID:                   T041201                       TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                         NUMBER OF INDEPENDENT CONTRASTS:        4

001 RESOURC1 (1          ) 0000                          RESOURCES:  GET ALL
002 RESOURC2 (2          ) 1000                          RESOURCES:  GET MOST
003 RESOURC3 (3          ) 0100                          RESOURCES:  GET SOME
004 RESOURC4 (4          ) 0010                          RESOURCES:  DON'T GET
005 RESOURC? (M          ) 0001                          RESOURCES:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0010
DESCRIPTION:               TEACHER UNDERGRADUATE MAJOR IN EDUCATION
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    UGRAD ED
NAEP ID:                   T040701                       TOTAL NUMBER OF SPECIFIED CONTRASTS:    2
```

```
TYPE OF CONTRAST:          CLASS                               NUMBER OF INDEPENDENT CONTRASTS:        1

001 UGR ED-? (0,M        ) 0                                   TEACHER UNDERGRAD EDUCATION MAJOR:  MISSING, DOES NOT APPLY
002 UGR ED-Y (1          ) 1                                   TEACHER UNDERGRAD EDUCATION MAJOR:  YES

CONDITIONING VARIABLE ID:  TCHR0011
DESCRIPTION:               TEACHER GRADUATE MAJOR IN EDUCATION
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    GRAD ED
NAEP ID:                   T040801                             TOTAL NUMBER OF SPECIFIED CONTRASTS:    2
TYPE OF CONTRAST:          CLASS                               NUMBER OF INDEPENDENT CONTRASTS:        1

001 GRA ED-? (0,M        ) 0                                   TEACHER GRADUATE EDUCATION MAJOR:  MISSING, DOES NOT APPLY
002 GRA ED-Y (1          ) 1                                   TEACHER GRADUATE EDUCATION MAJOR:  YES

CONDITIONING VARIABLE ID:  TCHR0012
DESCRIPTION:               NO TEACHER GRADUATE-LEVEL STUDY
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    NO GRAD
NAEP ID:                   T040806                             TOTAL NUMBER OF SPECIFIED CONTRASTS:    2
TYPE OF CONTRAST:          CLASS                               NUMBER OF INDEPENDENT CONTRASTS:        1

001 NOGRAD-? (0,M        ) 0                                   NO TEACHER GRADUATE STUDY:  MISSING, DOES NOT APPLY
002 NOGRAD-Y (1          ) 1                                   NO TEACHER GRADUATE STUDY:  YES

CONDITIONING VARIABLE ID:  TCHR0013
DESCRIPTION:               HOW MANY YEARS IN TOTAL HAVE YOU TAUGHT READING? (4TH GRADE)
GRADES/ASSESSMENTS:        N04, S04
CONDITIONING VAR LABEL:    T4REDYRS
NAEP ID:                   T049901                             TOTAL NUMBER OF SPECIFIED CONTRASTS:    6
TYPE OF CONTRAST:          CLASS                               NUMBER OF INDEPENDENT CONTRASTS:        5

001 T4REDYR1 (1          ) 00000                               YEARS TEACHING READING:  2 YEARS OR LESS
002 T4REDYR2 (2          ) 10000                               YEARS TEACHING READING:  3-5 YEARS
003 T4REDYR3 (3          ) 01000                               YEARS TEACHING READING:  6-10 YEARS
004 T4REDYR4 (4          ) 00100                               YEARS TEACHING READING:  11-24 YEARS
005 T4REDYR5 (5          ) 00010                               YEARS TEACHING READING:  25 YEARS OR MORE
006 T4REDYR? (M          ) 00001                               YEARS TEACHING READING:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0015
DESCRIPTION:               WHAT TYPE OF TEACHING CERTIFICATION DO YOU HAVE THAT IS RECOGNIZED BY THE STATE IN WHICH YOU TEACH?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    TYP CERT
NAEP ID:                   T050001                             TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                               NUMBER OF INDEPENDENT CONTRASTS:        4

001 CERT-NO  (1          ) 0000                                TYPE OF TEACHING CERTIFICATION:  NONE
002 CERT-TMP (2          ) 1000                                TYPE OF TEACHING CERTIFICATION:  TEMPORARY, PROBATIONAL, PROVISIONAL, EMERGENCY
003 CERT-REG (3          ) 0100                                TYPE OF TEACHING CERTIFICATION:  REGULAR, BUT NOT HIGHEST
```

```
004 CERT-HGH (4         ) 0010                  TYPE OF TEACHING CERTIFICATION:  HIGHEST AVAILABLE
005 CERT-?   (M         ) 0001                  TYPE OF TEACHING CERTIFICATION:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0016
DESCRIPTION:               WHAT IS THE HIGHEST ACADEMIC DEGREE YOU HOLD?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_OEGREE
NAEP ID:                   T050101              TOTAL NUMBER OF SPECIFIED CONTRASTS:     8
TYPE OF CONTRAST:          CLASS                NUMBER OF INDEPENDENT CONTRASTS:         7

001 TDEG-HSD (1         ) 0000000               TEACHER HIGHEST DEGREE:  HIGH SCHOOL DIPLOMA
002 TDEG-ASC (2         ) 1000000               TEACHER HIGHEST DEGREE:  ASSOCIATES/VOCATIONAL
003 TDEG-BAC (3         ) 0100000               TEACHER HIGHEST DEGREE:  BACHELOR'S HIGHEST DEGREE
004 TDEG-MAS (4         ) 0010000               TEACHER HIGHEST DEGREE:  MASTER'S HIGHEST DEGREE
005 TDEG-EDS (5         ) 0001000               TEACHER HIGHEST DEGREE:  EDUCATION SPECIALIST
006 TDEG-DOC (6         ) 0000100               TEACHER HIGHEST DEGREE:  DOCTORATE
007 TDEG-PRO (7         ) 0000010               TEACHER HIGHEST DEGREE:  PROFESSIONAL HIGHEST DEGREE
008 TDEG-?   (M         ) 0000001               TEACHER HIGHEST DEGREE:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0017
DESCRIPTION:               TEACHER ENGLISH UNDERGRADUATE MAJOR
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    UGRD ENG
NAEP ID:                   T040706              TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                NUMBER OF INDEPENDENT CONTRASTS:         1

001 UGRENG-Y (1         ) 0                     TEACHER ENGLISH UNDERGRADUATE MAJOR:  YES
002 UGRENG-? (0,M       ) 1                     TEACHER ENGLISH UNDERGRADUATE MAJOR:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0018
DESCRIPTION:               TEACHER READING/LANGUAGE ARTS UNDERGRADUATE MAJOR
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    UGRD RED
NAEP ID:                   T040707              TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                NUMBER OF INDEPENDENT CONTRASTS:         1

001 UGRRED-Y (1         ) 0                     TEACHER READING UNDERGRADUATE MAJOR:  YES
002 UGRRED-? (0,M       ) 1                     TEACHER READING UNDERGRADUATE MAJOR:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0019
DESCRIPTION:               TEACHER ENGLISH GRADUATE MAJOR
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    GRAD ENG
NAEP ID:                   T040807              TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                NUMBER OF INDEPENDENT CONTRASTS:         1

001 GRDENG-Y (1         ) 0                     TEACHER ENGLISH GRADUATE MAJOR:       YES
002 GRDENG-? (0,M       ) 1                     TEACHER ENGLISH GRAOUATE MAJOR:       MISSING, OOES NOT APPLY
```

```
CONDITIONING VARIABLE ID:  TCHR0020
DESCRIPTION:               TEACHER READING/LANGUAGE ARTS GRADUATE MAJOR
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    GRAD RED
NAEP ID:                   T040808                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     2
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         1

001 GRDRED-Y (1        ) 0                            TEACHER READING GRADUATE MAJOR:      YES
002 GRDRED-? (0,M      ) 1                            TEACHER READING GRADUATE MAJOR:      MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0021
DESCRIPTION:               HOW MUCH TIME HAVE YOU SPENT LAST YEAR IN READING DEVELOPMENT WORKSHOPS/SEMINARS?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    READ DEV
NAEP ID:                   T050201                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     6
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         5

001 READDEV1 (1        ) 00000                        TIME IN READING DEVELOPMENT WORKSHOPS:  NONE
002 READDEV2 (2        ) 10000                        TIME IN READING DEVELOPMENT WORKSHOPS:  LESS THAN 6 HOURS
003 READDEV3 (3        ) 01000                        TIME IN READING DEVELOPMENT WORKSHOPS:  6-15 HOURS
004 READDEV4 (4        ) 00100                        TIME IN READING DEVELOPMENT WORKSHOPS:  16-35 HOURS
005 READDEV5 (5        ) 00010                        TIME IN READING DEVELOPMENT WORKSHOPS:  MORE THAN 35 HOURS
006 READDEV5 (M        ) 00001                        TIME IN READING DEVELOPMENT WORKSHOPS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0022
DESCRIPTION:               HOW MANY HOURS DO YOU HAVE DESIGNATED AS PREPERATION PERIODS PER WEEK?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    PREP PER
NAEP ID:                   T051101                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         4

001 PRPER-0  (1        ) 0000                         TEACHER WEEKLY PREPARTION PERIODS:  NONE
002 PRPER-<1 (2        ) 1000                         TEACHER WEEKLY PREPARTION PERIODS:  LESS THAN 1
003 PRPER-12 (3        ) 0100                         TEACHER WEEKLY PREPARTION PERIODS:  1 TO 2
004 PRPER->2 (4        ) 0010                         TEACHER WEEKLY PREPARTION PERIODS:  MORE THAN 2
005 PRPER-?  (M        ) 0001                         TEACHER WEEKLY PREPARTION PERIODS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TCHR0023
DESCRIPTION:               ARE CURRICULUM SPECIALISTS AVAILABLE FOR READING?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    CURSPECS
NAEP ID:                   T041301                    TOTAL NUMBER OF SPECIFIED CONTRASTS:     3
TYPE OF CONTRAST:          CLASS                      NUMBER OF INDEPENDENT CONTRASTS:         2

001 CSPECS-Y (1        ) 00                           READING CURRICULUM SPECIALISTS:  YES
002 CSPECS-N (2        ) 10                           READING CURRICULUM SPECIALISTS:  NO
003 CSPECS-? (M        ) 01                           READING CURRICULUM SPECIALISTS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0001
```

```
DESCRIPTION:                WHAT IS YOUR AVERAGE READING CLASS SIZE? (4TH GRADE)
GRADES/ASSESSMENTS:         N04, S04
CONDITIONING VAR LABEL:     CLASSIZ4
NAEP ID:                    T050701                  TOTAL NUMBER OF SPECIFIED CONTRASTS:    6
TYPE OF CONTRAST:           CLASS                    NUMBER OF INDEPENDENT CONTRASTS:        5

001 CLASIZ-1 (1          ) 00000                     AVERAGE READING CLASS SIZE:  1-20 STUDENTS
002 CLASIZ-2 (2          ) 10000                     AVERAGE READING CLASS SIZE:  21-25 STUDENTS
003 CLASIZ-3 (3          ) 01000                     AVERAGE READING CLASS SIZE:  26-30 STUDENTS
004 CLASIZ-4 (4          ) 00100                     AVERAGE READING CLASS SIZE:  31-35 STUDENTS
005 CLASIZ-5 (5          ) 00010                     AVERAGE READING CLASS SIZE:  36 OR MORE STUDENTS
006 CLASIZ-? (M          ) 00001                     AVERAGE READING CLASS SIZE:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0003
DESCRIPTION:                ARE STUDENTS ASSIGNED TO THIS READING CLASS BY ABILITY? (TEACHER)
GRADES/ASSESSMENTS:         N04, S04, N08
CONDITIONING VAR LABEL:     T_ABILTY
NAEP ID:                    T046101                  TOTAL NUMBER OF SPECIFIED CONTRASTS:    3
TYPE OF CONTRAST:           CLASS                    NUMBER OF INDEPENDENT CONTRASTS:        2

001 T_ABIL-Y (1          ) 00                        STUDENTS ASSIGNED TO READING CLASS BY ABILITY:  YES
002 T_ABIL-N (2          ) 10                        STUDENTS ASSIGNED TO READING CLASS BY ABILITY:  NO
003 T_ABIL-? (M          ) 01                        STUDENTS ASSIGNED TO READING CLASS BY ABILITY:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0004
DESCRIPTION:                WHAT IS THE READING ABILITY LEVEL OF STUDENTS IN THIS CLASS?
GRADES/ASSESSMENTS:         N04, S04, N08
CONDITIONING VAR LABEL:     ABILRED
NAEP ID:                    T046201                  TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:           CLASS                    NUMBER OF INDEPENDENT CONTRASTS:        4

001 ABILRED1 (1          ) 0000                      READING ABILITY:  PRIMARILY HIGH
002 ABILRED2 (2          ) 1000                      READING ABILITY:  PRIMARILY AVERAGE
003 ABILRED3 (3          ) 0100                      READING ABILITY:  PRIMARILY LOW
004 ABILRED4 (4          ) 0010                      READING ABILITY:  WIDELY MIXED
005 ABILRED? (M          ) 0001                      READING ABILITY:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0005
DESCRIPTION:                HOW MUCH TIME DO YOU SPEND WITH THIS CLASS FOR READING INSTRUCTION EACH DAY?
GRADES/ASSESSMENTS:         N04, S04, N08
CONDITIONING VAR LABEL:     INSTTIME
NAEP ID:                    T046301                  TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:           CLASS                    NUMBER OF INDEPENDENT CONTRASTS:        4

001 INSTIME1 (1          ) 0000                      READING INSTRUCTION TIME:  30 MINUTES A DAY
002 INSTIME2 (2          ) 1000                      READING INSTRUCTION TIME:  45 MINUTES A DAY
003 INSTIME3 (3          ) 0100                      READING INSTRUCTION TIME:  60 MINUTES A DAY
004 INSTIME4 (4          ) 0010                      READING INSTRUCTION TIME:  90 MINUTES A DAY
005 INSTIME? (M          ) 0001                      READING INSTRUCTION TIME:  MISSING, DOES NOT APPLY
```

```
CONDITIONING VARIABLE ID:  TSUB0006
DESCRIPTION:               HOW MANY INSTRUCTIONAL GROUPS DO YOU DIVIDE YOUR READING CLASS?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    INSTGRPS
NAEP ID:                   T046401                      TOTAL NUMBER OF SPECIFIED CONTRASTS:     8
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         7

001 1 GROUP  (1          ) 0000000                      NUMBER INSTRUCTIONAL GROUPS:  WHOLE-CLASS ACTIVITY
002 FLEX GRP (2          ) 1000000                      NUMBER INSTRUCTIONAL GROUPS:  FLEXIBLE GROUPING
003 2 GROUPS (3          ) 0100000                      NUMBER INSTRUCTIONAL GROUPS:  2 GROUPS
004 3 GROUPS (4          ) 0010000                      NUMBER INSTRUCTIONAL GROUPS:  3 GROUPS
005 4 GROUPS (5          ) 0001000                      NUMBER INSTRUCTIONAL GROUPS:  4 GROUPS
006 5+GROUPS (6          ) 0000100                      NUMBER INSTRUCTIONAL GROUPS:  5 OR MORE GROUPS
007 INDIVLZD (7          ) 0000010                      NUMBER INSTRUCTIONAL GROUPS:  INDIVIDUALIZED INSTR
008 GROUPS-? (M          ) 0000001                      NUMBER INSTRUCTIONAL GROUPS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0007
DESCRIPTION:               WHAT TYPE OF MATERIALS FORM THE CORE OF YOUR READING PROGRAM?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    READMATS
NAEP ID:                   T046501                      TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         4

001 BASAL    (1          ) 0000                         TYPE OF READING MATERIALS:  BASAL
002 TRADE    (2          ) 1000                         TYPE OF READING MATERIALS:  TRADE
003 BAS/TRAD (3          ) 0100                         TYPE OF READING MATERIALS:  BASAL AND TRADE
004 OTHER RM (4          ) 0010                         TYPE OF READING MATERIALS:  OTHER
005 RDMATS-? (M          ) 0001                         TYPE OF READING MATERIALS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0008
DESCRIPTION:               HOW OFTEN ARE CHILDREN'S NEWSPAPERS/MAGAZINES USED IN READING CLASS?
GRADES/ASSESSMENTS:        N04, S04
CONDITIONING VAR LABEL:    CHILDMAG
NAEP ID:                   T046601                      TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         4

001 CHIMAG-1 (1          ) 0000                         NEWSPAPERS/MAGAZINES (TEACHER):  ALMOST EVERY DAY
002 CHIMAG-2 (2          ) 1000                         NEWSPAPERS/MAGAZINES (TEACHER):  ONCE OR TWICE A WEEK
003 CHIMAG-3 (3          ) 0100                         NEWSPAPERS/MAGAZINES (TEACHER):  ONCE OR TWICE A MONTH
004 CHIMAG-4 (4          ) 0010                         NEWSPAPERS/MAGAZINES (TEACHER):  NEVER OF HARDLEY EVER
005 CHIMAG-? (M          ) 0001                         NEWSPAPERS/MAGAZINES (TEACHER):  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0009
DESCRIPTION:               HOW OFTEN ARE READING KITS USED IN READING CLASS?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    READKITS
NAEP ID:                   T046602                      TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         4
```

```
001 RDKITS-1 (1          ) 0000                        READING KITS (TEACHER):  ALMOST EVERY OAY
002 RDKITS-1 (2          ) 1000                        READING KITS (TEACHER):  ONCE OR TWICE A WEEK
003 RDKITS-2 (3          ) 0100                        READING KITS (TEACHER):  ONCE OR TWICE A MONTH
004 RDKITS-3 (4          ) 0010                        READING KITS (TEACHER):  NEVER OF HAROLEY EVER
005 RDKITS-? (M          ) 0001                        READING KITS (TEACHER):  MISSING, OOES NOT APPLY

CONOITIONING VARIABLE IO:  TSUB0010
DESCRIPTION:               HOW OFTEN IS READING COMPUTER SOFTWARE IN READING CLASS?
GRADES/ASSESSMENTS:        N04, S04, N08
CONOITIONING VAR LABEL:    SOFTWARE
NAEP IO:                   T046603                     TOTAL NUMBER OF SPECIFIEO CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                       NUMBER OF INOEPENOENT CONTRASTS:         4

001 SOFTWR-1 (1          ) 0000                        READING COMPUTER SOFTWARE (TEACHER):  ALMOST EVERY OAY
002 SOFTWR-2 (2          ) 1000                        READING COMPUTER SOFTWARE (TEACHER):  ONCE OR TWICE A WEEK
003 SOFTWR-3 (3          ) 0100                        READING COMPUTER SOFTWARE (TEACHER):  ONCE OR TWICE A MONTH
004 SOFTWR-4 (4          ) 0010                        READING COMPUTER SOFTWARE (TEACHER):  NEVER OF HAROLEY EVER
005 SOFTWR-? (M          ) 0001                        READING COMPUTER SOFTWARE (TEACHER):  MISSING, OOES NOT APPLY

CONOITIONING VARIABLE IO:  TSUB0011
OESCRIPTION:               HOW OFTEN ARE VARIETY OF BOOKS USEO IN READING CLASS?
GRADES/ASSESSMENTS:        N04, S04, N08
CONOITIONING VAR LABEL:    VARTYBKS
NAEP IO:                   T046604                     TOTAL NUMBER OF SPECIFIEO CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                       NUMBER OF INOEPENOENT CONTRASTS:         4

001 VARBKS-1 (1          ) 0000                        VARIETY OF BOOKS (TEACHER):  ALMOST EVERY OAY
002 VARBKS-2 (2          ) 1000                        VARIETY OF BOOKS (TEACHER):  ONCE OR TWICE A WEEK
003 VARBKS-3 (3          ) 0100                        VARIETY OF BOOKS (TEACHER):  ONCE OR TWICE A MONTH
004 VARBKS-4 (4          ) 0010                        VARIETY OF BOOKS (TEACHER):  NEVER OF HAROLEY EVER
005 VARBKS-? (M          ) 0001                        VARIETY OF BOOKS (TEACHER):  MISSING, OOES NOT APPLY

CONOITIONING VARIABLE IO:  TSUB0012
OESCRIPTION:               HOW OFTEN ARE MATERIALS FROM OTHER SUBJECTS USEO IN READING CLASS?
GRADES/ASSESSMENTS:        N04, S04, N08
CONOITIONING VAR LABEL:    OTHRMATS
NAEP IO:                   T046605                     TOTAL NUMBER OF SPECIFIEO CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                       NUMBER OF INOEPENOENT CONTRASTS:         4

001 OTHMAT-1 (1          ) 0000                        OTHER SUBJECT MATERIALS (TEACHER):  ALMOST EVERY OAY
002 OTHMAT-2 (2          ) 1000                        OTHER SUBJECT MATERIALS (TEACHER):  ONCE OR TWICE A WEEK
003 OTHMAT-3 (3          ) 0100                        OTHER SUBJECT MATERIALS (TEACHER):  ONCE OR TWICE A MONTH
004 OTHMAT-4 (4          ) 0010                        OTHER SUBJECT MATERIALS (TEACHER):  NEVER OR HAROLY EVER
005 OTHMAT-? (M          ) 0001                        OTHER SUBJECT MATERIALS (TEACHER):  MISSING, OOES NOT APPLY

CONOITIONING VARIABLE IO:  TSUB0013
DESCRIPTION:               HOW OFTEN DO YOU DISCUSS NEW OR OIFFICULT VOCABULARY?
GRADES/ASSESSMENTS:        N04, S04, N08
CONOITIONING VAR LABEL:    T_VOCAB
```

```
NAEP ID:                     T046701                             TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:            CLASS                               NUMBER OF INDEPENDENT CONTRASTS:         4

001 VOCAB-T1 (1          ) 0000                                  DISCUSS VOCABULARY:  ALMOST EVERY DAY
002 VOCAB-T2 (2          ) 1000                                  DISCUSS VOCABULARY:  ONCE OR TWICE A WEEK
003 VOCAB-T3 (3          ) 0100                                  DISCUSS VOCABULARY:  ONCE OR TWICE A MONTH
004 VOCAB-T4 (4          ) 0010                                  DISCUSS VOCABULARY:  NEVER OR HARDLY EVER
005 VOCAB-T? (M          ) 0001                                  DISCUSS VOCABULARY:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0014
DESCRIPTION:               HOW OFTEN DO YOU ASK STUDENTS TO READ ALOUD?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_ALOUD
NAEP ID:                   T046702                               TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                                 NUMBER OF INDEPENDENT CONTRASTS:         4

001 ALOUD-T1 (1          ) 0000                                  READ ALOUD: ALMOST EVERY DAY
002 ALOUD-T2 (2          ) 1000                                  READ ALOUD: ONCE OR TWICE A WEEK
003 ALOUD-T3 (3          ) 0100                                  READ ALOUD: ONCE OR TWICE A MONTH
004 ALOUD-T4 (4          ) 0010                                  READ ALOUD: NEVER OR HARDLY EVER
005 ALOUD-T? (M          ) 0001                                  READ ALOUD: MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0015
DESCRIPTION:               HOW OFTEN DO YOU ASK STUDENTS TO TALK TO EACH OTHER ABOUT WHAT THEY HAVE READ?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_TALKRD
NAEP ID:                   T046703                               TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                                 NUMBER OF INDEPENDENT CONTRASTS:         4

001 TLKRD-T1 (1          ) 0000                                  TALK ABOUT READING:  ALMOST EVERY DAY
002 TLKRD-T2 (2          ) 1000                                  TALK ABOUT READING:  ONCE OR TWICE A WEEK
003 TLKRD-T3 (3          ) 0100                                  TALK ABOUT READING:  ONCE OR TWICE A MONTH
004 TLKRD-T4 (4          ) 0010                                  TALK ABOUT READING:  NEVER OR HARDLY EVER
005 TLKRD-T? (M          ) 0001                                  TALK ABOUT READING:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0016
DESCRIPTION:               HOW OFTEN DO YOU ASK STUDENTS TO WRITE SOMETHING ABOUT WHAT THEY HAVE READ?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_WRITRD
NAEP ID:                   T046704                               TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                                 NUMBER OF INDEPENDENT CONTRASTS:         4

001 WRTRD-T1 (1          ) 0000                                  WRITE ABOUT READING:  ALMOST EVERY DAY
002 WRTRD-T2 (2          ) 1000                                  WRITE ABOUT READING:  ONCE OR TWICE A WEEK
003 WRTRD-T3 (3          ) 0100                                  WRITE ABOUT READING:  ONCE OR TWICE A MONTH
004 WRTRD-T4 (4          ) 0010                                  WRITE ABOUT READING:  NEVER OR HARDLY EVER
005 WRTRD-T? (M          ) 0001                                  WRITE ABOUT READING:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0017
```

```
DESCRIPTION:                HOW OFTEN DO YOU ASK STUDENTS TO WORK IN A READING WORKBOOK OR ON A WORKSHEET?
GRADES/ASSESSMENTS:         N04, S04, N08
CONDITIONING VAR LABEL:     T_WBKWSH
NAEP ID:                    T046705                     TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:           CLASS                       NUMBER OF INDEPENDENT CONTRASTS:         4

001 WB/WS-T1 (1           ) 0000                        READING WORKBOOK/WORKSHEET:  ALMOST EVERY DAY
002 WB/WS-T2 (2           ) 1000                        READING WORKBOOK/WORKSHEET:  ONCE OR TWICE A WEEK
003 WB/WS-T3 (3           ) 0100                        READING WORKBOOK/WORKSHEET:  ONCE OR TWICE A MONTH
004 WB/WS-T4 (4           ) 0010                        READING WORKBOOK/WORKSHEET:  NEVER OR HARDLY EVER
005 WB/WS-T? (M           ) 0001                        READING WORKBOOK/WORKSHEET:  MISSING, MISSING NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0018
DESCRIPTION:                HOW OFTEN DO YOU ASK STUDENTS TO READ SILENTLY?
GRADES/ASSESSMENTS:         N04, S04, N08
CONDITIONING VAR LABEL:     T_SILENT
NAEP ID:                    T046706                     TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:           CLASS                       NUMBER OF INDEPENDENT CONTRASTS:         4

001 SILNT-T1 (1           ) 0000                        READ SILENTLY: ALMOST EVERY DAY
002 SILNT-T2 (2           ) 1000                        READ SILENTLY: ONCE OR TWICE A WEEK
003 SILNT-T3 (3           ) 0100                        READ SILENTLY: ONCE OR TWICE A MONTH
004 SILNT-T4 (4           ) 0010                        READ SILENTLY: NEVER OR HARDLY EVER
005 SILNT-T? (M           ) 0001                        READ SILENTLY: MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0019
DESCRIPTION:                HOW OFTEN DO YOU GIVE STUDENTS TIME TO READ BOOKS OF THEIR OWN CHOOSING?
GRADES/ASSESSMENTS:         N04, S04, N08
CONDITIONING VAR LABEL:     T_OWNBKS
NAEP ID:                    T046707                     TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:           CLASS                       NUMBER OF INDEPENDENT CONTRASTS:         4

001 OWNBK-T1 (1           ) 0000                        BOOKS CHOSEN YOURSELF:  ALMOST EVERY DAY
002 OWNBK-T2 (2           ) 1000                        BOOKS CHOSEN YOURSELF:  OR TWICE A WEEK
003 OWNBK-T3 (3           ) 0100                        BOOKS CHOSEN YOURSELF:  ONCE OR TWICE A MONTH
004 OWNBK-T4 (4           ) 0010                        BOOKS CHOSEN YOURSELF:  NEVER OR HARDLY EVER
005 OWNBK-T? (M           ) 0001                        BOOKS CHOSEN YOURSELF:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0020
DESCRIPTION:                HOW OFTEN DO YOU ASK STUDENTS TO DO A GROUP ACTIVITY/PROJECT ABOUT WHAT THEY HAVE READ?
GRADES/ASSESSMENTS:         N04, S04, N08
CONDITIONING VAR LABEL:     T_PROJCT
NAEP ID:                    T046709                     TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:           CLASS                       NUMBER OF INDEPENDENT CONTRASTS:         4

001 PRJCT-T1 (1           ) 0000                        PROJECT ABOUT READING:  ALMOST EVERY DAY
002 PRJCT-T2 (2           ) 1000                        PROJECT ABOUT READING:  ONCE OR TWICE A WEEK
003 PRJCT-T3 (3           ) 0100                        PROJECT ABOUT READING:  ONCE OR TWICE A MONTH
004 PRJCT-T4 (4           ) 0010                        PROJECT ABOUT READING:  NEVER OR HARDLY EVER
```

```
005 PRJCT-T? (M          ) 0001                    PROJECT ABOUT READING:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0021
DESCRIPTION:               HOW OFTEN DO YOU ASK STUDENTS TO DISCUSS DIFFERENT INTERPRETATIONS OF WHAT THEY HAVE READ?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_INTERP
NAEP ID:                   T046710                 TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                   NUMBER OF INDEPENDENT CONTRASTS:         4

001 INTRP-T1 (1          ) 0000                    DISCUSS READING INTERPRETATIONS:  ALMOST EVERY DAY
002 INTRP-T2 (2          ) 1000                    CISCUSS READING INTERPRETATIONS:  ONCE OR TWICE A WEEK
003 INTRP-T3 (3          ) 0100                    DISCUSS READING INTERPRETATIONS:  ONCE OR TWICE A MONTH
004 INTRP-T4 (4          ) 0010                    DISCUSS READING INTERPRETATIONS:  NEVER OR HARDLY EVER
005 INTRP-T? (M          ) 0001                    DISCUSS READING INTERPRETATIONS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0022
DESCRIPTION:               HOW OFTEN DO YOU ASK STUDENTS TO EXPLAIN OR SUPPORT THEIR UNDERSTANDING OF WHAT THEY HAVE READ?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_EXPLAN
NAEP ID:                   T046711                 TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                   NUMBER OF INDEPENDENT CONTRASTS:         4

001 EXPLA-T1 (1          ) 0000                    EXPLAIN/SUPPORT READING:  ALMOST EVERY DAY
002 EXPLA-T2 (2          ) 1000                    EXPLAIN/SUPPORT READING:  ONCE OR TWICE A WEEK
003 EXPLA-T3 (3          ) 0100                    EXPLAIN/SUPPORT READING:  ONCE OR TWICE A MONTH
004 EXPLA-T4 (4          ) 0010                    EXPLAIN/SUPPORT READING:  NEVER OR HARDLY EVER
005 EXPLA-T? (M          ) 0001                    EXPLAIN/SUPPORT READING:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0023
DESCRIPTION:               HOW OFTEN DO YOU GIVE READING QUIZZES OR TESTS?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_QUIZES
NAEP ID:                   T046712                 TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                   NUMBER OF INDEPENDENT CONTRASTS:         4

001 QUIZS-T1 (1          ) 0000                    READING QUIZZES OR TESTS:  ALMOST EVERY DAY
002 QUIZS-T2 (2          ) 1000                    READING QUIZZES OR TESTS:  ONCE OR TWICE A WEEK
003 QUIZS-T3 (3          ) 0100                    READING QUIZZES OR TESTS:  ONCE OR TWICE A MONTH
004 QUIZS-T4 (4          ) 0010                    READING QUIZZES OR TESTS:  NEVER OR HARDLY EVER
005 QUIZA-T? (M          ) 0001                    READING QUIZZES OR TESTS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0024
DESCRIPTION:               HOW OFTEN DO YOU USE MOVIES, VIDEOS, FILMSTRIPS, TV, TAPES, CDS, OR RECORDS?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    T_MOVIES
NAEP ID:                   T046713                 TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTR^ST:          CLASS                   NUMBER OF INDEPENDENT CONTRASTS:         4

001 MOVIE-T1 (1          ) 0000                    MOVIES, VIDEOS, TV, CDS:  ALMOST EVERY DAY
```

```
002 MOVIE-T2 (2          ) 1000                         MOVIES, VIDEOS, TV, CDS:  ONCE OR TWICE A WEEK
003 MOVIE-T3 (3          ) 0100                         MOVIES, VIDEOS, TV, CDS:  ONCE OR TWICE A MONTH
004 MOVIE-T4 (4          ) 0010                         MOVIES, VIDEOS, TV, CDS:  NEVER OR HARDLY EVER
005 MOVIE-T? (M          ) 0001                         MOVIES, VIDEOS, TV, CDS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0026
DESCRIPTION:               HOW OFTEN DO YOU USE MULTIPLE-CHOICE TESTS TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    MC TESTS
NAEP ID:                   T047001                      TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         4

001 MCTEST-1 (1          ) 0000                         MULTIPLE-CHOICE TESTS:  ONCE OR TWICE A WEEK
002 MCTEST-2 (2          ) 1000                         MULTIPLE-CHOICE TESTS:  ONCE OR TWICE A MONTH
003 MCTEST-3 (3          ) 0100                         MULTIPLE-CHOICE TESTS:  ONCE OR TWICE A YEAR
004 MCTEST-4 (4          ) 0010                         MULTIPLE-CHOICE TESTS:  NEVER OR HARDLY EVER
005 MCTEST-? (M          ) 0001                         MULTIPLE-CHOICE TESTS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0027
DESCRIPTION:               HOW OFTEN DO YOU USE SHORT-ANSWER TESTS TO ASSESS STUDFNTS IN READING?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    SA TESTS
NAEP ID:                   T047002                      TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         4

001 SATEST-1 (1          ) 0000                         SHORT-ANSWER TESTS:  ONCE OR TWICE A WEEK
002 SATEST-2 (2          ) 1000                         SHORT-ANSWER TESTS:  ONCE OR TWICE A MONTH
003 SATEST-3 (3          ) 0100                         SHORT-ANSWER TESTS:  ONCE OR TWICE A YEAR
004 SATEST-4 (4          ) 0010                         SHORT-ANSWER TESTS:  NEVER OR HARDLY EVER
005 SATEST-? (M          ) 0001                         SHORT-ANSWER TESTS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0028
DESCRIPTION:               HOW OFTEN DO YOU USE WRITING PARAGRAPHS TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    WRI TEST
NAEP ID:                   T047003                      TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:         4

001 WRITST-1 (1          ) 0000                         ASSESS BY WRITING PARAGRAPHS:  ONCE OR TWICE A WEEK
002 WRITST-2 (2          ) 1000                         ASSESS BY WRITING PARAGRAPHS:  ONCE OR TWICE A MONTH
003 WRITST-3 (3          ) 0100                         ASSESS BY WRITING PARAGRAPHS:  ONCE OR TWICE A YEAR
004 WRITST-4 (4          ) 0010                         ASSESS BY WRITING PARAGRAPHS:  NEVER OR HARDLY EVER
005 WRITST-? (M          ) 0001                         ASSESS BY WRITING PARAGRAPHS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0029
DESCRIPTION:               HOW OFTEN DO YOU USE INDIVIDUAL OR GROUP PROJECTS/PRESENTATIONS TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    PRJ TEST
NAEP ID:                   T047006                      TOTAL NUMBER OF SPECIFIED CONTRASTS:     5
```

372

371

```
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:        4

001 PRJTST-1 (1          ) 0000                         ASSESS BY PROJECTS/PRESENTATIONS:  ONCE OR TWICE A WEEK
002 PRJTST-2 (2          ) 1000                         ASSESS BY PROJECTS/PRESENTATIONS:  ONCE OR TWICE A MONTH
003 PRJTST-3 (3          ) 0100                         ASSESS BY PROJECTS/PRESENTATIONS:  ONCE OR TWICE A YEAR
004 PRJTST-4 (4          ) 0010                         ASSESS BY PROJECTS/PRESENTATIONS:  NEVER OR HARDLY EVER
005 PRJTST-? (M          ) 0001                         ASSESS BY PROJECTS/PRESENTATIONS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0030
DESCRIPTION:               HOW OFTEN DO YOU USE READING PORTFOLIOS TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    PRT TEST
NAEP ID:                   T047007                      TOTAL NUMBER OF SPECIFIED CONTRASTS:    5
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:        4

001 PRTTST-1 (1          ) 0000                         ASSESS BY READING PORTFOLIOS:  ONCE OR TWICE A WEEK
002 PRTTST-2 (2          ) 1000                         ASSESS BY READING PORTFOLIOS:  ONCE OR TWICE A MONTH
003 PRTTST-3 (3          ) 0100                         ASSESS BY READING PORTFOLIOS:  ONCE OR TWICE A YEAR
004 PRTTST-4 (4          ) 0010                         ASSESS BY READING PORTFOLIOS:  NEVER OR HARDLY EVER
005 PRTTST-? (M          ) 0001                         ASSESS BY READING PORTFOLIOS:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0031
DESCRIPTION:               HOW OFTEN DO YOU SEND OR TAKE THE CLASS TO THE LIBRARY?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    CLAS2LIB
NAEP ID:                   T047101                      TOTAL NUMBER OF SPECIFIED CONTRASTS:    6
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:        5

001 CLSLIB-1 (1          ) 00000                        TAKE CLASS TO LIBRARY:  ALMOST EVERY DAY
002 CLSLIB-2 (2          ) 10000                        TAKE CLASS TO LIBRARY:  ONCE OR TWICE A WEEK
003 CLSLIB-3 (3          ) 01000                        TAKE CLASS TO LIBRARY:  ONCE OR TWICE A MONTH
004 CLSLIB-4 (4          ) 00100                        TAKE CLASS TO LIBRARY:  NEVER OR HARDLY EVER
005 CLSLIB-5 (5          ) 00010                        TAKE CLASS TO LIBRARY:  THERE IS NO LIBRARY
006 CLSLIB-? (M          ) 00001                        TAKE CLASS TO LIBRARY:  MISSING, DOES NOT APPLY

CONDITIONING VARIABLE ID:  TSUB0032
DESCRIPTION:               HOW OFTEN DO YOU ASSIGN STUDENTS TO READ A BOOK FROM THE LIBRARY?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    LIB BOOK
NAEP ID:                   T047102                      TOTAL NUMBER OF SPECIFIED CONTRASTS:    6
TYPE OF CONTRAST:          CLASS                        NUMBER OF INDEPENDENT CONTRASTS:        5

001 CLALIB-1 (1          ) 00000                        ASSIGN STUDENTS TO READ LIBRARY BOOK:  ALMOST EVERY DAY
002 CLALIB-2 (2          ) 10000                        ASSIGN STUDENTS TO READ LIBRARY BOOK:  ONCE OR TWICE A WEEK
003 CLALIB-3 (3          ) 01000                        ASSIGN STUDENTS TO READ LIBRARY BOOK:  ONCE OR TWICE A MONTH
004 CLALIB-4 (4          ) 00100                        ASSIGN STUDENTS TO READ LIBRARY BOOK:  NEVER OR HARDLY EVER
005 CLALIB-5 (5          ) 00010                        ASSIGN STUDENTS TO READ LIBRARY BOOK:  THERE IS NO LIBRARY
006 CLALIB-? (M          ) 00001                        ASSIGN STUDENTS TO READ LIBRARY BOOK:  MISSING, DOES NOT APPLY
```

```
CONDITIONING VARIABLE ID:  TSUB0033
DESCRIPTION:               ARE COMPUTERS AVAILABLE FOR USE BY STUDENTS IN READING CLASS?
GRADES/ASSESSMENTS:        N04, S04, N08
CONDITIONING VAR LABEL:    COMP4RED
NAEP ID:                   T047201                        TOTAL NUMBER OF SPECIFIED CONTRASTS:    4
TYPE OF CONTRAST:          CLASS                          NUMBER OF INDEPENDENT CONTRASTS:        3

001 COMP-NA  (1          ) 000                            COMPUTERS IN READING CLASS:  NOT AVAILABLE
002 COMP-DIF (2          ) 100                            COMPUTERS IN READING CLASS:  AVAILABLE BUT DIFFICULT TO ACCESS
003 COMP-AVL (3          ) 010                            COMPUTERS IN READING CLASS:  AVAILABLE IN THE CLASSROOM
004 COMP-?   (M          ) 001                            COMPUTERS IN READING CLASS:  MISSING, DOES NOT APPLY
```

APPENDIX D

IKT PARAMETERS FOR READING ITEMS

APPENDIX D

IRT Parameters for Reading Items

This appendix contains two tables of IRT (item response theory) parameters for the items that were used in each reading scale for the 1994 fourth-grade Trial State Assessment.

For each of the binary scored items used in scaling (i.e., multiple-choice items and short constructed-response items), the tables provide estimates of the IRT parameters (which correspond to a_j, b_j, and c_j in equation 8.1 in Chapter 8) and their associated standard errors (s.e.) of the estimates. For each of the polytomously scored items (i.e., the extended constructed-response items), the tables also show the estimates of the d_{jv} parameters (see equation 8.1) and their associated standard errors.

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

Note that the item parameters in this appendix are in the metrics used for the original calibration of the scales. The transformations needed to represent these parameters in terms of the metrics of the final reporting scales are given in Chapter 9.

Table D-1
IRT Parameters for Grade 4 Reading Items
Reading for Literary Experience

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*
R012001	RC	1	1.537 (0.066)	0.567 (0.019)	0.129 (0.009)			
R012002	RC	2	1.367 (0.034)	-0.186 (0.015)	0.000 (0.000)			
R012003	RC	3	1.846 (0.072)	-0.610 (0.027)	0.207 (0.017)			
R012004	RC	4	0.752 (0.024)	0.337 (0.022)	0.000 (0.000)			
R012005	RC	5	1.124 (0.049)	0.061 (0.033)	0.184 (0.016)			
R012006	RC	6	0.482 (0.013)	0.748 (0.019)	0.000 (0.000)	0.319 (0.039)	-0.076 (0.047)	-0.244 (0.055)
R012007	RC	7	0.620 (0.033)	-0.935 (0.101)	0.176 (0.033)			
R012008	RC	8	0.569 (0.021)	-0.588 (0.036)	0.000 (0.000)			
R012009	RC	9	1.258 (0.065)	-0.781 (0.060)	0.354 (0.027)			
R012010	RC	10	1.066 (0.031)	-0.438 (0.022)	0.000 (0.000)			
R012011	RC	11	1.718 (0.076)	-0.358 (0.032)	0.238 (0.018)			
R012101	RD	1	1.712 (0.079)	-1.060 (0.040)	0.299 (0.023)			
R012102	RD	2	0.662 (0.021)	-0.102 (0.024)	0.000 (0.000)			
R012103	RD	3	1.246 (0.048)	-0.639 (0.038)	0.188 (0.019)			
R012104	RD	4	0.659 (0.021)	-0.354 (0.027)	0.000 (0.000)			
R012105	RD	5	0.810 (0.041)	-0.084 (0.056)	0.185 (0.022)			
R012106	RD	6	0.884 (0.025)	0.054 (0.019)	0.000 (0.000)			
R012107	RD	7	1.346 (0.064)	0.228 (0.030)	0.241 (0.014)			
R012108	RD	8	0.661 (0.022)	-1.208 (0.043)	0.000 (0.000)			
R012109	RD	9	0.509 (0.020)	-1.594 (0.064)	0.000 (0.000)			
R012110	RD	10	0.826 (0.045)	-1.175 (0.101)	0.288 (0.038)			
R012111	RD	11	0.895 (0.023)	1.417 (0.020)	0.000 (0.000)	1.086 (0.020)	-1.086 (0.052)	
R012112	RD	12	0.734 (0.027)	-0.823 (0.039)	0.000 (0.000)			
R012601	RE	1	0.863 (0.031)	1.150 (0.032)	0.000 (0.000)			
R012602	RE	2	1.573 (0.083)	1.285 (0.029)	0.175 (0.007)			
R012603	RE	3	1.524 (0.066)	0.147 (0.025)	0.215 (0.013)			
R012604	RE	4	1.180 (0.038)	1.053 (0.023)	0.000 (0.000)			
R012605	RE	5	0.959 (0.086)	1.095 (0.046)	0.302 (0.015)			
R012606	RE	6	1.580 (0.086)	0.424 (0.028)	0.321 (0.013)			
R012607	RE	7	0.900 (0.024)	1.274 (0.017)	0.000 (0.000)	0.686 (0.021)	-0.045 (0.031)	-0.641 (0.057)
R012608	RE	8	0.556 (0.042)	-0.713 (0.164)	0.304 (0.045)			
R012609	RE	9	1.248 (0.091)	0.856 (0.034)	0.276 (0.014)			
R012610	RE	10	1.671 (0.109)	0.621 (0.029)	0.363 (0.013)			
R012611	RE	11	0.796 (0.027)	0.214 (0.023)	0.000 (0.000)			
R015801	RI	1	0.968 (0.045)	-1.534 (0.081)	0.241 (0.034)			
R015802	RI	2	0.405 (0.017)	-1.275 (0.062)	0.000 (0.000)			
R015803	RI	3	0.594 (0.011)	-0.137 (0.023)	0.000 (0.000)	1.679 (0.038)	-1.679 (0.032)	
R015804	RI	4	0.571 (0.011)	0.705 (0.020)	0.000 (0.000)	2.414 (0.040)	-0.392 (0.032)	-2.022 (0.072)
R015805	RI	5	0.965 (0.058)	0.242 (0.050)	0.288 (0.020)			
R015806	RI	6	0.617 (0.013)	0.358 (0.022)	0.000 (0.000)	1.377 (0.033)	-1.377 (0.036)	
R015807	RI	7	0.585 (0.015)	-0.162 (0.023)	0.000 (0.000)	1.170 (0.039)	-1.170 (0.033)	
R015808	RI	8	0.610 (0.035)	-1.662 (0.145)	0.219 (0.046)			
R015809	RI	9	0.580 (0.014)	0.017 (0.026)	0.000 (0.000)	1.435 (0.042)	-1.435 (0.038)	

Table D-2
IRT Parameters for Grade 4 Reading Items
Reading for Information

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*
R012201	RF	1	0.206 (0.014)	-1.355 (0.115)	0.000 (0.000)			
R012202	RF	2	0.819 (0.049)	0.440 (0.051)	0.213 (0.019)			
R012203	RF	3	0.800 (0.050)	0.594 (0.048)	0.193 (0.018)			
R012204	RF	4	0.424 (0.011)	-0.011 (0.019)	0.000 (0.000)	1.356 (0.050)	-0.465 (0.044)	-0.891 (0.050)
R012205	RF	5	1.296 (0.072)	0.563 (0.032)	0.277 (0.013)			
R012206	RF	6	1.102 (0.031)	0.729 (0.019)	0.000 (0.000)			
R012207	RF	7	0.529 (0.034)	-0.755 (0.148)	0.237 (0.043)			
R012208	RF	8	0.868 (0.025)	-0.453 (0.024)	0.000 (0.000)			
R012209	RF	9	1.187 (0.058)	0.336 (0.034)	0.178 (0.015)			
R012210	RF	10	0.601 (0.024)	-1.550 (0.060)	0.000 (0.000)			
R012501	RJ	1	0.925 (0.224)	2.799 (0.346)	0.310 (0.008)			
R012502	RJ	2	0.850 (0.041)	-2.035 (0.107)	0.234 (0.047)			
R012503	RJ	3	0.982 (0.025)	-0.085 (0.018)	0.000 (0.000)			
R012504	RJ	4	0.663 (0.020)	-0.373 (0.026)	0.000 (0.000)			
R012505	RJ	5	1.125 (0.050)	-0.840 (0.055)	0.256 (0.026)			
R012506	RJ	6	0.765 (0.022)	-0.220 (0.023)	0.000 (0.000)			
R012507	RJ	7	1.213 (0.060)	-0.453 (0.052)	0.357 (0.023)			
R012508	RJ	8	0.979 (0.026)	-0.469 (0.021)	0.000 (0.000)			
R012509	RJ	9	0.758 (0.045)	-0.617 (0.103)	0.322 (0.035)			
R012510	RJ	10	0.929 (0.050)	-0.435 (0.071)	0.306 (0.028)			
R012511	RJ	11	0.940 (0.028)	-0.682 (0.025)	0.000 (0.000)			
R012512	RJ	12	0.363 (0.011)	0.563 (0.024)	0.000 (0.000)	0.861 (0.057)	0.226 (0.057)	-1.088 (0.070)
R012701	RG	1	1.167 (0.056)	0.002 (0.041)	0.286 (0.018)			
R012702	RG	2	0.590 (0.020)	-1.273 (0.044)	0.000 (0.000)			
R012703	RG	3	1.093 (0.029)	0.629 (0.018)	0.000 (0.000)			
R012704	RG	4	1.325 (0.063)	0.729 (0.023)	0.144 (0.010)			
R012705	RG	5	1.347 (0.045)	1.272 (0.023)	0.000 (0.000)			
R012706	RG	6	0.642 (0.025)	1.335 (0.045)	0.000 (0.000)			
R012707	RG	7	2.225 (0.102)	0.396 (0.019)	0.245 (0.010)			
R012708	RG	8	0.673 (0.018)	1.634 (0.021)	0.000 (0.000)	1.253 (0.028)	0.386 (0.036)	-1.639 (0.108)
R012709	RG	9	0.548 (0.049)	0.052 (0.147)	0.288 (0.040)			
R012710	RG	10	1.075 (0.037)	0.839 (0.023)	0.000 (0.000)			
R015701	RH	1	1.030 (0.059)	-0.674 (0.079)	0.458 (0.028)			
R015702	RH	2	0.575 (0.011)	-0.040 (0.023)	0.000 (0.000)	1.638 (0.036)	-1.638 (0.033)	
R015703	RH	3	0.680 (0.012)	0.023 (0.021)	0.000 (0.000)	1.651 (0.032)	-1.651 (0.030)	
R015704	RH	4	0.617 (0.015)	-0.214 (0.018)	0.000 (0.000)	0.394 (0.032)	-0.394 (0.028)	
R015705	RH	5	0.730 (0.017)	0.154 (0.017)	0.000 (0.000)	0.808 (0.027)	-0.808 (0.026)	
R015706	RH	6	0.921 (0.066)	1.099 (0.039)	0.192 (0.013)			
R015707	RH	7	0.511 (0.012)	0.246 (0.024)	0.000 (0.000)	1.235 (0.037)	-1.235 (0.039)	
R015708	RH	8	0.587 (0.032)	-0.194 (0.084)	0.147 (0.028)			
R015709	RH	9	0.439 (0.016)	1.089 (0.035)	0.000 (0.000)	0.400 (0.043)	-0.400 (0.057)	

APPENDIX E

TRIAL STATE ASSESSMENT REPORTING SUBGROUPS

COMPOSITE AND DERIVED COMMON BACKGROUND VARIABLES

COMPOSITE AND DERIVED REPORTING VARIABLES

APPENDIX E

REPORTING SUBGROUPS FOR THE 1994 TRIAL STATE ASSESSMENT

Results for the 1994 Trial State Assessment were reported for student subgroups defined by gender, race/ethnicity, type of location, parents' level of education, and geographical region. The following explains how each of these subgroups was derived.

DSEX (Gender)

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

DRACE7 (Race/ethnicity)

The variable DRACE7 is an imputed definition of race/ethnicity, derived from up to three sources of information. This variable is used for race/ethnicity subgroup comparisons. Two items from the student demographics questionnaire were used in the determination of derived race/ethnicity:

Demographic Item Number 2:

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

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

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

Demographic Item Number 1:

1. Which best describes you?

○ White (not Hispanic)

○ Black (not Hispanic)

○ Hispanic ("Hispanic" means someone who is Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or other Spanish or Hispanic background.)

○ Asian ("Asian" means someone who is Chinese, Japanese, Korean, Vietnamese, or other Asian background.)

○ Pacific Islander ("Pacific Islander" means someone who is from a Filipino, Hawaiian, or other Pacific Island background.)

○ American Indian or Alaskan Native ("American Indian or Alaskan Native" means someone who is from one of the American Indian tribes, or one of the original people of Alaska.)

○ Other

Students' race/ethnicity was then assigned to correspond with their selection. For students who filled in the seventh oval ("Other"), provided illegible information or information that could not be classified, or did not respond at all, race/ethnicity as provided from school records was used.

Derived race/ethnicity could not be determined for students who did not respond to background items 1 or 2 and for whom race/ethnicity was not provided by the school.

TOL8 (Type of Location)
TOL5
TOL3

The variable TOL8 is used by NAEP to provide information about the type of location in which schools are located. The variable is defined using population size information from the 1990 Census and the definitions of Metropolitan Statistical Areas (MSAs) as of February 1994. There are eight categories for TOL8:

1 Large Central City — a central city of an 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 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 a 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 or equal to 2,500 and defined as urban by the U.S. Bureau of Census
7	Rural MSA	a place within an MSA with a population of less than 2,500 and defined as rural by the U.S. Bureau of the Census
8	Rural NonMSA	a place not within an MSA with a population of less than 2,500 and defined as rural by the U.S. Bureau of the Census

The variable TOL5 collapses the information provided in the variable TOL8 to five levels:

1 Large Central City
2 Midsize Central City
3 Urban Fringe of Large City, Urban Fringe of Midsize City, and Large Town
4 Small Town
5 Rural MSA and Rural NonMSA

The variable TOL3 is used extensively in the NAEP reports. TOL3 collapses TOL8 to three levels:

1	Central City	(Large Central City and Midsize Central City) This category includes central cities of all MSAs. Central City is a geographic term and is not synonymous with "inner city."
2	Urban Fringe/Large Town	(Urban Fringe of Large City, Urban Fringe of Midsize City, and Large Town) An Urban Fringe includes all densely settled places and areas within MSAs that are classified as urban by the Bureau of the Census. A Large Town is defined as a place outside MSAs with a population greater than or equal to 25,000.
3	Rural/Small Town	(Small Town, Rural MSA, and Rural NonMSA) Rural includes all places and areas with a population of less than

2,500 that are classified as rural by the Bureau of the Census. A Small Town is defined as a place outside MSAs with a population of less than 25,000 but greater than or equal to 2,500.

PARED (Parents' education level)

The variable PARED is derived from responses to two questions, B003501 and B003601, in the student demographic questionnaire. Students were asked to indicate the extent of their mother's education (B003501—How far in high school did your mother go?) by choosing one of the following:

○ She did not finish high school.
○ She graduated from high school.
○ She had some education after high school.
○ She graduated from college.
○ I don't know.

Students were asked to provide the same information about the extent of their father's education (B003601—How far in high school did your father go?) by choosing one of the following:

○ He did not finish high school.
○ He graduated from high school.
○ He had some education after high school.
○ He graduated from college.
○ I don't know.

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

REGION (Region of the country)

States were grouped into four geographical regions—Northeast, Southeast, Central, and West—as shown in Table E-1. All 50 states and the District of Columbia are listed, with the participants in the Trial State Assessment shown in italic type. Territories were not assigned to a region. The part of Virginia that is included in the Washington, DC, metropolitan statistical area is included in the Northeast region; the remainder of the state is included in the Southeast region.

Table E-1
NAEP Geographic 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

MODAGE (Modal age)

The modal age (the age of most of the students in the grade sample) for the fourth grade students is age 9. A value of 1 for MODAGE indicates that the student is younger than the modal age; a value of 2 indicates that the student is of the modal age; a value of 3 indicates that the student is older than the modal age.

VARIABLES DERIVED FROM THE STUDENT AND TEACHER QUESTIONNAIRES

Several variables were formed from the systematic combination of response values for one or more items from either the student demographic questionnaire, the student reading background questionnaire, or the teacher questionnaire.

HOMEEN2 (Home environment——Articles [of 4] in the home)

The variable HOMEEN2 was created from the responses to student demographic items B000901 (Does your family get a newspaper regularly?), B000903 (Is there an encyclopedia in your home?), B000904 (Are there more than 25 books in your home?), and B000905 (Does your family get any magazines regularly?). The values for this variable were derived as follows:

1 0-2 types — The student responded to at least two items and answered Yes to two or fewer.

2 3 types The student answered Yes to three items.

3 4 types The student answered Yes to four items.

8 Omitted The student answered fewer than two items.

SINGLEP (How many parents live at home)

SINGLEP was created from items B005601 (Does either your mother or your stepmother live at home with you?) and B005701 (Does either your father or your stepfather live at home with you?). The values for SINGLEP were derived as follows:

1 2 parents at home The student answered Yes to both items.

2 1 parent at home The student answered Yes to B005601 and No to B005701, or Yes to B005701 and No to B005601.

3 Neither at home The student answered No to both items.

8 Omitted The student did not respond to or filled in more than one oval for one or both items.

TRUMAJ (Teacher undergraduate major)

Items T040701 and T040705 through T040710 in the teacher questionnaire (What were your undergraduate major fields of study?) were used to determine TRUMAJ as follows:

1 English/reading The teacher responded yes to T040706 or T040707 (English, reading, and/or language arts).

2 Education The teacher responded yes to T040701 (education) and No to T040706 and T040707.

3 Other Any other response.

TRGMAJ (Teacher graduate major)

Items T040801 and T040805 through T040811 in the teacher questionnaire (What were your undergraduate major fields of study?) were used to determine TRGMAJ as follows:

1 English/reading The teacher responded yes to T040807 or T040808 (English, reading, and/or language arts).

2 Education The teacher responded yes to T040801 (education) and no to T040807 and T040808.

3 Other The teacher responded yes to T040805 (other), T040809 (geography), T040810 (history), or T040811 (social studies).

4 None The teacher indicated (T040806) that he or she had had no graduate-level study.

VARIABLES DERIVED FROM READING ITEMS

BKSCOR (Booklet level score)

The booklet level score is a student-level score based on the sum of the number correct for dichotomous items plus the sum of the scores on the polytomous items, where the score for a polytomous item starts from 0 for the unacceptable category. Thus, for a 4-point extended constructed-response item, scores of "no response", "off-task", and "unsatisfactory" are assigned an item score of 0. Scores of "partial", "essential", and "extensive" are assigned item scores of 1, 2, and 3, respectively. The score is computed based on all cognitive items in an individual's assessment booklet.

LOGIT (Logit percent correct within booklet)

In order to compute the LOGIT score, a percent correct within booklet was first computed. This score was based on the ratio of the booklet score (BKSCOR) over the maximum booklet score. The percent correct score was set to .0001 if no items were answered correctly; if BKSCOR equaled the maximum booklet score, the percent correct score was set to .9999. A logit score, LOGIT, was calculate for each student by the following formula:

A logit score, LOGIT, was calculated for each student by the following formula:

$$\text{LOGITP} = \ln\left[\frac{PCTCOR}{1 - PCTCOR}\right]$$

LOGIT was then restricted to a value x, such that $-3 \leq x \leq 3$. After computing LOGIT for each student, the mean and standard deviation was calculated for each booklet as the first step in standardizing the logit scores. The standardized logit score, ZLOGIT, was then calculated for each student by the following formula:

$$\text{ZSCORE} = \left[\frac{LOGIT\ -\ mean\ logit}{standard\ deviation}\right]$$

NORMIT (Normit Gaussian score)
SCHNORM (School-level mean Gaussian score)

The normit score is a student-level Gaussian score based on the inverse normal transformation of the mid-percentile rank of a student's number-correct booklet score within that booklet. The normit scores were used to decide collapsing of variables, finalize conditioning coding, and check the results of scaling.

The number correct is based on the number of dichotomous items answered correctly plus the score obtained on extended constructed-response items. The mid-percentile rank is based on the formula:

$$\frac{CF(i)+CF(i-1)}{2N}$$

where CF(i) is the cumulative frequency at i items correct and N is the total sample size. If i = 0 then

$$\frac{CF(0)+\frac{CF(1)}{2}}{2N}$$

A school-level normit, SCHNORM, was also created; this was the mean normit across all reading booklets administered in a school. These school-level mean normit scores were used in conditioning procedures to take into account differences in school proficiency. For each school, the weighted mean of the logits for the students in that school was calculated. Each student was then assigned that mean as his or her school-level mean logit score value.

VARIABLES RELATED TO PROFICIENCY SCALING

Proficiency Score Variables

Item response theory (IRT) was used to estimate average reading proficiency for each jurisdiction and for various subpopulations, based on students' performance on the set of reading items they received. IRT provides a common scale on which performance can be reported for the nation, jurisdiction, and subpopulations, even when all students do not answer the same set of questions. This common scale makes it possible to report on relationships between students' characteristics (based on their responses to the background questions) and their overall performance in the assessment.

A scale ranging from 0 to 500 was created to report performance for each of the two content areas—Reading for Literary Experience and Reading to Gain Information. Each content-area scale was based on the distribution of student performance across all three grades assessed in the 1994 national assessment (grades 4, 8, and 12) and had a mean of 250 and a standard deviation of 50. A composite scale was created as an overall measure of students' mathematics proficiency. The composite scale for grade 4 was a weighted average of the two content area scales, where the weight for each content area was proportional to the relative importance assigned to the content area as specified in the mathematics objectives. Although the items comprising each scale were identical to those used for the national program, the item parameters for the Trial State Assessment scales were estimated from the combined data from all jurisdictions participating in the Trial State Assessment.

Scale proficiency estimates were obtained for all students assessed in the Trial State Assessment. The NAEP methods use random draws ("plausible values") from estimated proficiency distributions to compute population statistics. Plausible values are not optimal estimates of individual proficiency; instead, they serve as intermediate values to be used in estimating population characteristics. Chapter 8 provides further details on the computation and use of plausible values.

The proficiency score (plausible value) variables are provided on the student data files for each of the scales and are named as shown in Table E-2.

Table E-2
Scaling Variables for the 1994 Trial State Assessment Samples

Reading Scale	Data Variables
Reading for Literary Experience	RRPS11 to RRPS15
Reading to Gain Information	RRPS21 to RRPS25
Composite	RRPCM1 to RRPCM5

SMEANR, SMNR1 (School mean score using first plausible value)
SRANKR, SRNKR1 (School rank using first plausible value)
SRNK3R, SRK3R1 (Top, middle, bottom third using first plausible value)

A mean reading composite score (SMEANR on the student files, SMNR1 on the school files) was calculated using the first composite plausible value for each school included in the grade 4 assessment. The mean composite score was based on the values from the scaling variable RRPCM1 and was calculated using the students' sampling weights. The schools were then ordered from highest to lowest mean score (SRANKR on the student files, SRNKR1 on the school files) within a jurisdiction using school-level weights—the school with the highest mean score was given a ranking of 1 and the school with the lowest mean score was given a ranking equal to the number of schools in the jurisdiction.

These variables were then used in partitioning the schools within the national public-school comparison sample and the schools within each jurisdiction into three groups (top third, middle third, and bottom third) based on their ranking (SRNK3R on the student files, SRK3R1 on the school files).

SMEANRP, SMNR1P **(School mean score using first plausible value, public schools only)**
SRANKRP, SRNKR1P **(School rank using first plausible value, public schools only)**
SRNK3RP, SRK3R1P **(Top, middle, bottom third, using first plausible value, public schools only)**

These variables were computed in the same manner as SMEANR, SMNR1, SRANKR, SRNKR1, SRNK3R, and SRK3R1 for the subset of students who attended public schools.

SMNRx **(School mean score using plausible values 2 through 5)**
SRNKRx **(School rank using plausible values 2 through 5)**
SRK3Rx **(Top, middle, bottom third using plausible values 2 through 5)**

SMNRxP **(School mean score using plausible values 2 through 5, public schools only)**
SRNKRxP **(School rank using plausible values 2 through 5, public schools only)**
SRK3RxP **(Top, middle, bottom third, using plausible values 2 through 5, public schools only)**

School ranking results presented in the 1994 NAEP reports are based on the first plausible value. However, since there are four additional estimates of proficiency (plausible values) for each student, school ranking data were also created for those estimates. These school rank values were created using the same procedures described above, substituting proficiency variables RRPCM2 through RRPCM5 to compute the results. In the variable names, *x* stands for the plausible value 2, 3, 4, or 5. Note that these variables are included only on the school file.

QUALITY EDUCATION DATA VARIABLES (QED)

The data files contain several variables obtained from information supplied by Quality Education Data, Inc. (QED). QED maintains and updates annually lists of schools showing grade span, total enrollment, instructional dollars per pupil, and other information for each school. These data variables are retained on both the school and student files and are identified in the data layouts by "(QED)" in the SHORT LABEL field.

Most of the QED variables are defined sufficiently in the data codebooks. Explanations of others are provided below.

ORSHPT and SORSHPT are the Orshansky Percentile, an indicator of relative wealth that specifies the percentage of school-age children in a district who fall below the poverty line.

IDP and SIDP represent, at the school district level, dollars per student spent for textbooks and supplemental materials.

ADULTED and SADLTED indicate whether or not adult education courses are offered at the school site.

URBAN and SURBAN define the school's urbanicity: urban (central city); suburban (area surrounding central city, but still located within the counties constituting the metropolitan statistical area); or rural (area outside any metropolitan statistical area).

APPENDIX F

SETTING THE NAEP ACHIEVEMENT LEVELS FOR THE 1994 READING ASSESSMENT

APPENDIX F

Setting the NAEP Achievement Levels for the 1994 Reading Assessment

Mary Lyn Bourque
National Assessment Governing Board

Introduction

Since 1984, NAEP has reported the performance of students in the nation and for specific subpopulations on a 0-to-500 proficiency scale. The history and development of the scale and the anchoring procedure used to interpret specific points on that scale are described in Appendix G of *The NAEP 1992 Technical Report* (Johnson & Carlson, 1994).

The 1988 NAEP legislation[1] created an independent board, the National Assessment Governing Board (NAGB), responsible for setting policy for the NAEP program. The 1994 NAEP reauthorization[2] continued many of the Board's statutory responsibilities, including "developing appropriate student performance standards for each age and grade in each subject area to be tested under the National Assessment." Consistent with this directive, and striving to achieve one of the primary mandates of the statute "to improve the form and use of NAEP results," the Board has been developing student performance standards (called achievement levels by NAGB) on the National Assessment since 1990.

The 1990 standard-setting effort, initiated in December 1989 with the dissemination of a draft policy statement (NAGB, 1989) and culminating 22 months later in the publication of the NAGB report, *The Levels of Mathematics Achievement* (Bourque & Garrison, 1991), consisted of two phases: the main study and a replication-validation study. Although there were slight differences between the two phases, there were many common elements. Both phases used a modified (iterative/empirical) Angoff (1971) procedure for arriving at the levels; both focused on estimating performance levels based on a review of the 1990 NAEP mathematics item pool; and both phases employed policy definitions for basic, proficient, and advanced levels (NAGB, 1990) as the criteria for rating items. The 1990 process was evaluated by a number of different groups (for a discussion, see Hambleton & Bourque, 1991) who identified technical flaws in the 1990 process. These evaluations influenced the Board's decision to set the levels again in 1992,

[1] Public Law 100-297. (1988). National Assessment of Educational Progress improvement act (Article No. USC 1221). Washington, DC.

[2] Public Law 103-382. (1994). Improving America's schools act. Washington, DC.

and to not use the 1990 levels as benchmarks for progress toward the national goals during the coming decade. It is interesting to note, however, that the 1990 and 1992 processes produced remarkably similar results.

In September 1991, the Board contracted with American College Testing (ACT) to convene the panels of judges that would recommend the levels on the 1992 NAEP assessments in reading, writing, and mathematics. While the 1992 level-setting activities were not unlike those undertaken by the Board in 1990, there were significant improvements made in the process for 1992. There was a concerted effort to bring greater technical expertise to the process: the contractor selected by the Board has a national reputation for setting standards in a large number of certification and licensure exams; an internal and external advisory team monitored all the technical decisions made by the contractor throughout the process; and state assessment directors periodically provided their expertise and technical assistance at key stages in the project.

Setting achievement levels is a method for setting standards on the NAEP assessment that identify what students should know and be able to do at various points along the proficiency scale. The initial policy definitions of the achievement levels were presented to panelists along with an illustrative framework for more in-depth development and operationalization of the levels. Panelists were asked to determine descriptions/definitions of the three levels from the specific framework developed for the NAEP assessment with respect to the content and skills to be assessed. The operationalized definitions were refined throughout the level-setting process, as well as validated with a supplementary group of judges subsequent to the level-setting meetings. Panelists were also asked to develop a list of illustrative tasks associated with each of the levels, after which sample items from the NAEP item pool were identified to exemplify the full range of performance of the intervals between levels. The emphasis in operationalizing the definitions and in identifying and selecting exemplar items and papers was to represent the full range of performance from the lower level to the next higher level. The details of the implementation procedures are outlined in the remainder of this appendix.

1992 Preparation for the Reading Level-setting Meeting

It is important for the planning of any standard-setting effort to know how various process elements interact with each other. For example, panelists interact with pre-meeting materials, meeting materials (i.e., the assessment questions, rating forms, rater feedback, and so forth), each other, and the project staff. All of these elements combine to promote or degrade what has been called intrajudge consistency and interjudge consensus (Friedman & Ho, 1990).

Previous research has conceptualized the effects of two major kinds of interaction: (1) people interacting with text (Smith & Smith, 1988), and (2) people interacting with each other (Curry, 1987; Fitzpatrick, 1989). In order to assess the effects of textual and social interaction and adjust the standard-setting procedures accordingly, a pilot study was conducted as the first phase of the 1992 initiative.

Reading was chosen as the single content area to be pilot tested since it combined all of the various features found in the other NAEP assessments, including multiple-choice, and both short and extended constructed-response items. The pilot study provided the opportunity to implement and evaluate all aspects of the operational plan—background materials, meeting materials, study design, meeting logistics, staff function, and participant function.

The overall pilot was quite successful. The level-setting process worked well, and the pilot allowed the contractor to make improvements in the design before implementation activities began. For example, schedule changes were made that allowed the panelists more time to operatioralize the policy definitions before beginning the item-rating task. Also, the feedback mechanisms used to inform panelists about interjudge and intrajudge consistency data were improved for clarity and utility to the entire process.

1992 Reading Level-setting Panel

Sixty-four panelists representing 32 jurisdictions (31 states and the Virgin Islands) were selected from the 366 nominees and invited to participate in the level-setting process. They represented reading/language arts teachers at grades 4, 8, and 12, nonteacher educators, and members of the noneducator (general public) community. The group was balanced by gender, race/ethnicity, NAEP regions of the country, community type (low SES, not low SES), district size, and school type (public/nonpublic). Two panelists were unable to attend due to a family emergency and a loss of job, resulting in 62 participants, 22 at grade 4, 20 at grade 8, and 20 at grade 12.

1992 Process for Developing the Achievement Levels

The four-and-one-half-day session began with a brief overview of NAEP and NAGB, a presentation on the policy definitions of the achievement levels, a review of the NAEP reading assessment framework, and a discussion of factors that influence item difficulty. The purpose of the presentation was to focus panelists' attention on the reading framework and to emphasize the fact that panelists' work was directly related to the NAEP assessment, not to the whole domain of reading.

All panelists completed and self-scored an appropriate grade-level form of the NAEP assessment. The purpose of this exercise was to familiarize panelists with the test content and scoring protocols—as well as time constraints—before beginning to develop the preliminary operationalized descriptions of the three levels.

Working in small groups of five or six, then eventually in grade-level groups, panelists expandeo and operationalized the policy definitions of basic, proficient, and advanced in terms of specific reading skills, knowledge, and behaviors that were judged to be appropriate expectations for students in each grade, and to be in accordance with the current reading assessment framework.

The policy definitions[3] are as follows:

Basic This level, below proficient, denotes partial mastery of the knowledge and skills that are fundamental for proficient work at each grade—4, 8, and 12.

Proficient This central level represents solid academic performance for each grade tested—4, 8, and 12. Students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling.

Advanced This higher level signifies superior performance beyond proficient grade-level mastery at grades 4, 8, and 12.

The small groups were allowed to brainstorm about what student performance *should be*, using the framework and their experience in completing the NAEP assessment as guides[4]. In addition, a practice task caused panelists to examine items in the half of the item pool that they would not be rating later. A comprehensive listing of grade level descriptors was developed, and panelists were asked to identify the five or six that best described what students *should be able to do* at each of the levels. Those descriptors appearing with the greatest frequency were compiled into a discussion list for the grade-level groups. Additions, deletions, and modifications were made as a result of discussions, and the groups reached general agreement that the final list of descriptors represented what students *should be able to do* at each achievement level.

Panelists next received training in the Angoff method, which was customized to reflect the unique item formats of the particular subject area assessment. Once a conceptual consensus was reached about the characteristics of **marginally** acceptable performance at each of the three levels, practice items from the released pool were rated by the panelists according to the process defined in the contractor's plan. For multiple-choice and short constructed-response items (both of which were scored right or wrong), panelists were asked to rate each item for the expected probability of a correct response for a group of *marginally* acceptable examinees at the basic, proficient, and advanced levels. For extended constructed-response items (which were scored on a four-point rating scale using a partial credit model), panelists were asked to review a set of student response papers and select three papers, one for each achievement level, that typified *marginally* acceptable examinee performance for that level.

Following training in the Angoff method, the judges began the rating and paper selection process, inspecting and rating each dichotomously scored item in the pool for the expected

[3] NAGB revised its policy definitions on achievement levels in late 1993. The *Proficient* level now reads: this level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter. *Basic* and *Advanced* remain virtually unchanged.

[4] The panelists also reviewed about half the item pool (the half they would not be r[illegible]ater) so that the descriptors could be further modified if that was deemed appropriate.

probabilities of answering the item correctly at each level. For polytomously scored items, panelists reviewed a representative set of 24 to 28 student response papers for each item and selected the paper that best represented marginally acceptable student performance at each level. Panelists completed three rounds of item ratings and paper selections. For Round 1, panelists first answered the items related to a reading passage, then reviewed their answers using scoring keys and protocols. This process helped ensure that panelists would be thoroughly familiar with each item, including the foils and scoring rubrics, before rating the item Panelists provided item ratings/paper selections for all three achievement levels, one item at a time, for all the items related to a reading passage, then proceeded to the next reading passage and set of items, for which the process was repeated. Panelists rated items for half the items in th ir grade-level assessment; one block of exercises was common to both halves of the grade-level groups. During Round 1, panelists used their lists of descriptors and other training materials for guidance in the rating process.

Following Round 1, item response theory (IRT) was used to convert the rating results[5] for each rater to a latent ability scale, represented by the Greek letter theta (θ). This θ scale was the same scale to which the NAEP items evaluated by each panelist were calibrated. In order to provide meaningful feedback about item ratings, a special *relative scale* was constructed, which was a linear transformation of the theta scale having a mean of 75 and standard deviation of 15. Before Round 2 of the rating process, panelists were given interjudge consistency information using this relative scale. This information allowed panelists to see where their individual mean item ratings were on the scale, relative to the mean for the group and to the means for other panelists. Reasons for extreme mean ratings, including the possibility that some panelists misinterpreted the item rating task, were discussed.

Before Round 2, panelists were also given item difficulty data. This information was presented as the overall percentage of students who answered each item correctly during the actual NAEP administration, for items scored "correct" or "incorrect" (i.e., multiple-choice and short constructed-response items), and as the mean score for student responses (on a scale of 1 to 4) for the extended response items. Panelists were told that this item difficulty information should be used as a reality check. For items on which item ratings differed substantially from the item difficulty value, panelists were asked to reexamine the item to determine if they had misinterpreted the item or misjudged its difficulty. Results of the data analysis, and panelists' own evaluations, indicated that the item difficulty information was perceived as very useful but had little impact on panelists' ratings.

For Round 2, panelists reviewed the same set of items they rated in Round 1 and, using the interjudge consistency information, the item difficulty information, and the information provided prior to Round 1, they either confirmed their initial item ratings and paper selections or adjusted their ratings to reflect the additional information. About one-half of Round 1 item ratings and paper selections were adjusted during Round 2.

Prior to Round 3, panelists' ratings were reanalyzed and additional information was presented to panelists concerning intrajudge variability. For each panelist, the intrajudge

[5] Because the IRT item parameters were not available for the polytomously scored (extended constructed-response) items, these items were not included in the following discussion of results.

variability information consisted of those items that they had rated differently than items having similar difficulty, taking into consideration the panelist's aggregated item ratings. That is, the panelists' aggregated item ratings were converted to the theta (θ) scale. All items rated by the panelists were then analyzed in terms of the panelist's achievement level (θ) in comparison to actual student performance on the items. The observed item rating from each panelist was contrasted to an expected item rating. Those items with the largest differences between observed and expected ratings were identified. Panelists were given this information and asked to review each of these items and decide if their Round 2 ratings still accurately reflected their best judgments of the items. The intrajudge consistency data was to be used to flag items for reconsideration in the final round of rating.

For Round 3, panelists reviewed the same set of items they rated in Rounds 1 and 2 using both the new intrajudge variability information and the information made available during Rounds 1 and 2. In addition, panelists could discuss, within their small groups, ratings and paper selections for specific items about which they were unsure. About one-third of the item ratings were adjusted during Round 3.

1992 Process for Selecting Exemplar Items

On the final day of the achievement level-setting process, panelists reviewed items from the 1992 item pool scheduled for release to the public. The released item pool was the set from which the panelists could select items illustrative of the achievement levels for their grade. Exercises are organized in blocks, consisting of a reading passage, followed by several items, usually employing each of the three item formats, (i.e., multiple-choice, short constructed-response, and extended constructed-response). A total of 10 blocks from the 1992 exercise pool were scheduled for release: 2 blocks from the fourth-grade pool, totaling 19 items; 4 blocks from the eighth-grade pool, totaling 52 items; and 4 blocks from the twelfth-grade pool, totaling 46 items.

Panelists who had rated specific blocks of released items were asked to review those same items again to select particular ones as exemplary of each achievement level. The items were pre-assigned to each achievement level based on the final round of the judges' rating data, and using the following statistical criteria. For any given level (basic, proficient, or advanced),

(1) items having an expected p-value[6] $\geq .501$ and $\leq .750$, *at that level,* were assigned to that level;

(2) items meeting the criteria at *more than one level* were assigned to *one* level taking both the expected p-value and the appropriateness of the item for one of the levels into account; and

[6] Expected p-values were based on the average predicted performance at the cut point for each achievement level.

Because the content of items was given equal consideration in the selection process,

(3) items with expected p-values ≤.501 were assigned to levels where a specific passage had few or no items at that level.

For example, the raters' expected p-value for one of the released items might have been .366 at the basic level, .701 at the proficient level, and .932 at the advanced le\ el. This item would have been identified for review as a potential exemplar item for the proficient level. The expected p-value at the basic level was too low for consideration as a basic-level exemplar—that is, the item was judged to be too difficult, and the expected p-value at the advanced level was too high for consideration at the advanced level—that is, the item was judged to be too easy. Table F-1 shows the results of this process for each grade and level.

Panelists were asked to review the items as classified, and form an individual judgment regarding the suitability of each item to illustrate and further communicate the meaning of the levels. Each item's classification could be accepted, rejected, or reassigned, although the procedure was primarily designed to eliminate items that did not meet panelists' expectations for any reason. Items were reclassified if a strong consensus was found to hold for that change.

During the validation process, described in the next section, items were again reviewed. Those that had been selected by the originai standard-setting panel were grouped into sets of *pre-selected* items. All remaining items in the released blocks that met the statistical criteria, *but were not recommended by the original panel*, were grouped into a set identified as *additional items for review*. Exercises that had been recommended for reclassification into another achievement level category were presented in their original classification for purposes of this review. As the Table F-2 shows, 21 items were recommended as exemplars for the basic level, 17 for the proficient level, and 9 for the advanced.

1992 Process for Validating the Levels

Nineteen reading educators participated in the item selection and content validation process. Ten of the panelists were reading teachers who had participated in the original achievement level-setting process and who had been identified as outstanding panelists by grade group facilitators during this meeting, who were extensively involved with professional organizations (e.g., the International Reading Association, the National Reading Conference, or the National Council for Teachers of English), and who had outstanding service credentials. The other nine panelists represented state-level reading curriculum supervisors or assessment directors, as well as university faculty teaching in disciplines related to this subject area. To the extent possible, the group was balanced by race/ethnicity and gender.

The two-and-one-half day meeting began by briefing panelists on the purpose of the meeting and by giving them an overview of the level-setting process and results. Panelists first reviewed the operationalized descriptions of the achievement levels for qualities such as (1) within- and across-grade consistency, (2) grade-level appropriateness, and (3) utility for increasing the public's understanding of the NAEP reading results. Next, panelists reviewed the operationalized descriptions of the achievement levels for consistency with the NAGB policy definitions of basic, proficient, and advanced with the NAEP *Reading Objectives*. Working in

Table F-1

Results of First Review for Achievement-level Exemplars

Level/Status	Grade 4	Grade 8	Grade 12	All Grades
Total released	19	52	46	117
Basic				
Reviewed	4	12	18	34
Recommended	3	5	14	22
Proficient				
Reviewed	5	14	20	39
Recommended	4	12	9	25
Advanced				
Reviewed	5	6	7	18
Recommended	5	6	8	19

Table F-2

Results of Review of Additional Items for Achievement-level Exemplars

Level/Status	Grade 4	Grade 8	Grade 12	All Grades
Total items recommended	13	13	21	47
Basic				
Reviewed	3	12	12	27
Recommended	6	7	8	21
Proficient				
Reviewed	4	13	11	28
Recommended	6	3	8	17
Advanced				
Reviewed	5	8	9	22
Recommended	1	3	5	9

grade-level (4, 8, and 12) groups of 6 to 7 panelists each, then as a whole group, panelists reviewed the operationalized descriptions to provide within- and across-grade consistency, and to align the language and concepts of the descriptions more closely with the language of the NAEP *Reading Objectives*. (Both the original descriptions and the revised descriptions are included later in this appendix.) Finally, panelists suggested revisions they thought would improve the operational descriptions based on their earlier reviews.

On the final day, panelists worked in grade-level groups to review the possible exemplar items. The task was to select a set of items, for each achievement level for their grade, that would best communicate to the public the levels of reading ability and the types of skills needed to perform in reading at that level.

After selecting sets of items for their grades, the three grade-level groups met as a whole group to review item selection. During this process, cross-grade items that had been selected as exemplars for two grades (two such items were selected for grades 8 and 12) were assigned to one grade by whole-group consensus. In addition, items were evaluated by the whole group for overall quality. This process yielded 13 items as recommended exemplars for grade 4, 13 items as recommended exemplars for grade 8, and 21 items as recommended exemplars for grade 12.

Evaluation of the 1992 Levels

The 1992 achievement levels in both mathematics and reading were evaluated under a Congressional mandate by the National Academy of Education (NAE). A series of research studies were mounted by the NAE (1993a; 1993b) to look at various aspects of the validity of the levels-setting process and the levels finally adopted by the National Assessment Governing Board. Three of the studies focused specifically on the reading achievement levels, and were conducted for the Academy panel by staff at the Center for the Study of Reading at the University of Illinois at Urbana-Champaign. The first study examined the process for setting the levels in reading; the second study provided an analysis of the reading achievement levels descriptions; and the third focused on a comparison of the reading cut scores with those set by alternative means. Based on these studies the Academy's policy report concluded that the achievement levels were flawed and should be discontinued as a means of reporting NAEP data.

While the National Assessment Governing Board did not agree with the conclusions reached in the NAE studies, and while the Board's technical advisors and contractor did not believe the weight of the evidence supported the conclusions reached by the Academy (American College Testing, 1993; Cizek, 1993; Kane, 1993), the Board agreed to support further investigation into the validity of the reading achievement levels through additional studies prior to the release of the 1994 NAEP reading data, since the Board planned on using the levels to report the 1994 NAEP data.

1994 Process for Validating the Levels

The methodology developed by ACT to examine the reading achievement levels descriptions required the use of reading professionals (teachers and non-teacher educators) to review the descriptions in relation to the 1992 reading item pool. Fifty-eight panelists (about 20 at each grade level) were assigned to two different task groups, A and B. Group A employed the Item Difficulty Categorization (IDC) procedure, while Group B used a Judgmental Item Categorization (JIC) procedure. The goal of both task groups was to identify any lack of congruence between the item pool and the achievement level descriptions.

The IDC procedure examined the level of support for the descriptions as evidenced by performance on the NAEP items. Items were pre-selected for each achievement level using an response probability (rp) criterion of 0.50 at the lower borderline (*can do* items). Those items not meeting the same rp criterion at the upper borderline of the level were categorized as *can't do* items, while those items meeting the rp criterion anywhere in the range (from lower borderline to upper borderline) of the achievement level were labeled *challenging* items. Panelists were trained to examine the items in each of the three categories and determine whether or not the cognitive demand of the item matched the skills and knowledge identified in the descriptions. Mismatches were identified and later resolved or accounted for through a grade level procedure involving the JIC group.

The JIC procedure asked panelists to assign items to levels based on their judgment of where it belonged given the achievement levels descriptions. Items were assigned to the lowest level of performance required to respond correctly to the item. All items were assigned to levels independently by judges in the first round. Then, working in small groups and finally in the total group, assignments were confirmed and/or moderated through a consensus process.

The final grade-level procedure brought both groups A and B together to jointly evaluate the descriptions *vis a vis* performance on the item pool. The goal of the grade level procedure was to reach general agreement on the extent of (or lack of) agreement between the descriptions and the item pool, employing somewhat different approaches to the question.

On the basis of the validation process only one recommendation was made by the panelists to improve the descriptions and bring them more in line with the performance data they had examined during the process. The general conclusion was that reference to an ability to make inferences should be included in the description of Basic level achievement at each grade level. An adjustment has been made in the 1994 descriptions to reflect that recommendation.

1994 Exemplars

The purpose of providing exemplar exercises is to provide readers with a sample of the kind of skills and knowledge that students reaching the achievement levels are likely to be able to respond to successfully. They are meant also to represent the kind of knowledge and skills embodied in the reading framework.

The selection of exemplar items for the 1994 reading assessment augment the 1992 exemplars by providing three additional passages (one for each grade level) and 13 additional exercises associated with the passages. The choice was made on the basis of criteria similar to those used in 1992, with one additional selection criterion, namely, item format. Since the percent of constructed response items increased by approximately 10% over the 1992 assessment, the choice of 1994 exemplars reflects this focus.

It should be noted that although some exemplars are associated with performance data from the 1992 and 1994 assessments (overall and conditional p-values), others have only 1992 performance estimates since they were released items in 1992 and not readministered in 1994. However, they are all reflective of the assessment framework.

Mapping the Levels onto the NAEP Scale

The process of mapping panelists' ratings to the NAEP scales used *item response theory* (IRT). IRT provided statistically sophisticated methods for determining the expected performance of examinees on particular test items in terms of an appropriate measurement scale. The same measurement scale simultaneously described the characteristics of the test items and the performance of the examinees. Once the item characteristics were set, it was possible to determine precisely how examinees were likely to perform on the test items at different points of the measurement scale.

The panelists' ratings of the NAEP test items were likewise linked, by definition, to the expected performance of examinees at the theoretical achievement level cut points. It was therefore feasible to use the IRT item characteristics to calculate the values on the measurement scale corresponding to each achievement level. This was done by averaging the item ratings over panelists for each achievement level and then simply using the item characteristics to find the corresponding achievement level cut points on the IRT measurement scale. This process was repeated for each of the NAEP reading scales within each grade (4, 8, and 12).

For the multiple-choice and short constructed-response items that were dichotomously scored, the judges each rated half of the items in the NAEP pool in terms of the expected probability that a student at a borderline achievement level would answer the item correctly, based on the judges' operationalization of the policy definitions and the factors that influence item difficulty. To assist the judges in generating consistently scaled ratings, the rating process was repeated twice, with feedback. Information on consistency among different judges and on the difficulty of each item[7] was fed back into the first repetition (Round 2), while information on consistency within each judge's set of ratings was fed back into the second repetition (Round 3). The third round of ratings permitted the judges to discuss their ratings among themselves to resolve problematic ratings. The mean final rating of the judges aggregated across multiple-choice and short constructed-response items yielded the threshold values for these items in the percent correct metric. These cut scores were then mapped onto the NAEP scale (which is defined and scored using item response theory, rather than percent correct).

For extended constructed-response items, judges were asked to select student papers that exemplified performance at the cut point of each achievement level. Then for each achievement level, the mean of the scores assigned to the selected papers was mapped onto the NAEP scale in a manner similar to that used for the items scored dichotomously.

[7] Item difficulty estimates were based on a preliminary, partial set of responses to the national assessment.

The final cut score for each achievement level was a weighted average of the cut score for the multiple-choice and short constructed-response items and the cut score for the extended constructed-response items, with the weights being proportional to the information supplied by the two classes of items. The judges' ratings, in both metrics, are shown for grade 4 in Table F-3.

Table F-3
Cut Points for Achievement Levels

Level	Mean Percent Correct, Multiple-choice and Short Constructed-response (Round 3)	Mean Paper Rating, Extended Constructed-response (Round 3)	Scale Score*	Standard Error of Scale Score**
Grade 4				
Basic	38	2.72	208	(3.6)
Proficient	62	3.14	238	(1.4)
Advanced	80	3.48	268	(6.1)

*Scale score is derived from a weighted average of the mean percents correct for multiple-choice and short constructed-response items and the mean paper ratings for extended constructed-response items after both were mapped onto the NAEP scale.

**The standard error of the scale is estimated from the difference in mean scale scores for the two equivalent subgroups of judges.

In the final stage of the mapping process, the achievement level cut points on the IRT measurement scale were combined over content areas and rescaled to the NAEP score scale. Weighted averages of the achievement level cut points were computed. The weighting constants accounted for the measurement precision of the test items evaluated by the panelists, the proportion of items belonging to each NAEP content area, and the linear NAEP scale transformations. These weighted averages produced the final cut points for the basic, proficient, and advanced achievement levels within each grade.

Figure F-1

Final Descriptions of 1992 Reading Achievement Levels

PREAMBLE

Reading for meaning involves a dynamic, complex interaction between and among the reader, the text, and the context. Readers, for example, bring to the process their prior knowledge about the topic, their reasons for reading it, their individual reading skills and strategies, and their understanding of differences in text structures.

The texts used in the reading assessment are representative of common real world reading demands. Students at grade 4 are asked to respond to literary and informational texts which differ in structure, organization, and features. Literary texts include short stories, poems, and plays that engage the reader in a variety of ways, not the least of which is reading for fun. Informational texts include selections from textbooks, magazines, encyclopedias, and other written sources whose purpose is to increase the reader's knowledge.

In addition to literary and informational texts, students at grades 8 and 12 are asked to respond to practical texts (e.g., bus schedules or directions for building a model airplane) that describe how to perform a task.

The context of the reading situation includes the purposes for reading that the reader might use in building a meaning of the text. For example, in reading for literary experience, students may want to see how the author explores or uncovers experiences, or they may be looking for vicarious experience through the story's characters. On the other hand, the student's purpose in reading informational texts may be to learn about a topic (such as the Civil War or the oceans) or to accomplish a task (such as getting somewhere, completing a form, or building something).

The assessment asks students at all three grades to build, extend, and examine text meaning from four stances or orientations:

> **Initial Understanding**—Students are asked to provide the overall or general meaning of the selection. This includes summaries, main points, or themes.
>
> **Developing Interpretation**—Students are asked to extend the ideas in the text by making inferences and connections. This includes making connections between cause and effect, analyzing the motives of characters, and drawing conclusions.
>
> **Personal Response**—Students are asked to make explicit connections between the ideas in the text and their own background knowledge and experiences. This includes comparing story characters with themselves or people they know, for example, or indicating whether they found a passage useful or interesting.
>
> **Critical Stance**—Students are asked to consider how the author crafted a text. This includes identifying stylistic devices such as mood and tone.

Figure F-1 (continued)

Final Descriptions of 1992 Reading Achievement Levels

These stances are not considered hierarchical or completely independent of each other. Rather, they provide a frame for generating questions and considering student performance at all levels. All students at all levels should be able to respond to reading selections from all of these orientations. What varies with students' developmental and achievement levels is the amount of prompting or support needed for response, the complexity of the texts to which they can respond, and the sophistication of their answers.

INTRODUCTION

The following achievement-level descriptions focus on the interaction of the reader, the text, and the context. They provide some specific examples of reading behaviors that should be familiar to most readers of this document. The specific examples are not inclusive; their purpose is to help clarify and differentiate what readers performing at each achievement level should be able to do. While a number of other reading achievement indicators exist at every level, space and efficiency preclude an exhaustive listing.

It should also be noted that the achievement levels are cumulative from basic to proficient to advanced. One level builds on the previous levels such that knowledge at the proficient level presumes mastery of the basic level, and knowledge at the advanced level presumes mastery at both the basic and proficient.

Grade 4—Basic

Fourth-grade students performing at the **basic level** *should demonstrate an understanding of the overall meaning of what they read.* When reading texts appropriate for fourth graders, *they should be able to make relatively obvious connections between the text and their own experiences*[9].

For example, when reading **literary text**, they should be able to tell what the story is generally about—providing details to support their understanding—and be able to connect aspects of the stories to their own experiences.

When reading **informational text**, basic-level fourth graders should be able to tell what the selection is generally about or identify the purpose for reading it; provide details to support their understanding; and connect ideas from the text to their background knowledge and experiences.

[9] Based on the recommendations of the 1994 reading revisi? study, the phrase "*and extend the ideas in the text by making simple inferences*" has been added here to the description of *Basic*.

Figure F-1 (continued)

Final Descriptions of 1992 Reading Achievement Levels

Grade 4—Proficient

Fourth grade students performing at the **proficient level** *should be able to demonstrate an overall understanding of the text, providing inferential as well as literal information.* When reading text appropriate to fourth grade, *they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connection between the text and what the student infers should be clear.*

For example, when reading **literary text**, proficient-level fourth graders should be able to summarize the story, draw conclusions about the characters or plot, and recognize relationships such as cause and effect.

When reading **informational text**, proficient-level students should .be able to summarize the information and identify the author's intent or purpose. They should be able to draw reasonable conclusions from the text, recognize relationships such as cause and effect or similarities and differences, and identify the meaning of the selection's key concepts.

Grade 4—Advanced

Fourth-grade students performing at the **advanced level** *should be able to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices.* When reading text appropriate to fourth grade, *they should be able to judge texts critically and, in general, give thorough answers that indicate careful thought.*

For example, when reading **literary text**, advanced-level students should be able to make generalizations about the point of the story and extend its meaning by integrating personal experiences and other readings with the ideas suggested by the text. They should be able to identify literary devices such as figurative language.

When reading **informational text**, advanced-level fourth graders should be able to explain the author's intent by using supporting material from the text. They should be able to make critical judgments of the form and content of the text and explain their judgments clearly.

Grade 8—Basic

Eighth-grade students performing at the **basic level** *should demonstrate a literal understanding of what they read and be able to make some interpretations.* When reading text appropriate to eighth grade, *they should be able to identify specific aspects of the text that reflect the overall meaning,*[10]

[10] Based on the recommendations of the 1994 reading revisit study, the phrase "*extend the ideas in the text by making simple inferences,*" has been added here to the description of *Basic*.

recognize and relate interpretations and connections among ideas in the text to personal experience, and draw conclusions based on the text.

For example, when reading **literary text**, basic-level eighth graders should be able to identify themes and make inferences and logical predictions about aspects such as plot and characters.

When reading **informative text**, they should be able to identify the main idea and the author's purpose. They should make inferences and draw conclusions supported by information in the text. They should recognize the relationships among the facts, ideas, events, and concepts of the text (e.g., cause and effect and chronological order).

When reading **practical text**, they should be able to identify the main purpose and make predictions about the relatively obvious outcomes of procedures in the text.

Grade 8—Proficient

Eighth-grade students performing at the **proficient level** *should be able to show an overall understanding of the text, including inferential as well as literal information.* When reading text appropriate to eighth grade, *they should extend the ideas in the text by making clear inferences from it, by drawing conclusions, and by making connections to their own experiences—including other reading experiences.* Proficient eighth graders *should be able to identify some of the devices authors use in composing text.*

For example, when reading **literary text,** students at the proficient level should be able to give details and examples to support themes that they identify. They should be able to use implied as well as explicit information in articulating themes; to interpret the actions, behaviors, and motives of characters; and to identify the use of literary devices such as personification and foreshadowing.

When reading **informative text**, they should be able to summarize the text using explicit and implied information and support conclusions with inferences based on the text.

When reading **practical text**, proficient-level students should be able to describe its purpose and support their views with examples and details. They should be able to judge the importance of certain steps and procedures.

Grade 8—Advanced

Eighth-grade students performing at the **advanced level** *should be able to describe the more abstract themes and ideas of the overall text.* When reading text appropriate to eighth grade, they *should be able to analyze both meaning and form and support their analyses explicitly with examples from the text; they should be able to extend text information by relating it to their experiences and to world events.* At this level, student *responses should be thorough, thoughtful, and extensive.*

Figure F-1 (continued)

Final Descriptions of 1992 Reading Achievement Levels

For example, when reading **literary text**, advanced-level eighth graders should be able to make complex, abstract summaries and theme statements. They should be able to describe the interactions of various literary elements (i.e., setting, plot, characters, and theme); to explain how the use of literary devices affects both the meaning of the text and their response to the author's style. They should be able critically to analyze and evaluate the composition of the text.

When reading **informative text**, they should be able to analyze the author's purpose and point of view. They should be able to use cultural and historical background information to develop perspectives on the text and be able to apply text information to broad issues and world situations.

When reading **practical text**, advanced-level students should be able to synthesize information that will guide their performance, apply text information to new situations, and critique the usefulness of the form and content.

Grade 12—Basic

Twelfth-grade students performing at the **basic level** *should be able to demonstrate an overall understanding and make some interpretations of the text.* When reading text appropriate to twelfth grade, they *should be able to identify and relate aspects of the text to its overall meaning,*[11] *recognize interpretations, make connections among and relate ideas in the text to their personal experiences, and draw conclusions.* They *should be able to identify elements of an author's style.*

For example, when reading **literary text**, twelfth-grade students should be able to explain the theme, support their conclusions with information from the text, and make connections between aspects of the text and their own experiences.

When reading **informational text**, basic-level twelfth graders should be able to explain the main idea or purpose of a selection and use text information to support a conclusion or make a point. They should be able to make logical connections between the ideas in the text and their own background knowledge.

When reading **practical text**, they should be able to explain its purpose and the significance of specific details or steps.

Grade 12—Proficient

Twelfth-grade students performing at the **proficient level** *should be able to show an overall understanding of the text, which includes inferential as well as literal information.* When reading text

[11] Based on the recommendations of the 1994 reading revisit study, the phrase "*extend the ideas in the text by making simple inferences,*" has been added here to the description of *Basic*.

Figure F-1 (continued)

Final Descriptions of 1992 Reading Achievement Levels

appropriate to twelfth grade, they *should be able to extend the ideas of the text by making inferences, drawing conclusions, and making connections to their own personal experiences and other readings. Connections between inferences and the text should be clear, even when implicit.* These *students should be able to analyze the author's use of literary devices.*

When reading **literary text**, proficient-level twelfth graders should be able to integrate their personal experiences with ideas in the text to draw and support conclusions. They should be able to explain the author's use of literary devices such as irony or symbolism.

When reading **informative text**, they should be able to apply text information appropriately to specific situations and integrate their background information with ideas in the text to draw and support conclusions.

When reading **practical texts**, they should be able to apply information or directions appropriately. They should be able to use personal experiences to evaluate the usefulness of text information.

Grade 12—Advanced

Twelfth-grade students performing at the **advanced level** *should be able to describe more abstract themes and ideas in the overall text.* When reading text appropriate to twelfth grade, they *should be able to analyze both the meaning and the form of the text and explicitly support their analyses with specific examples from the text.* They *should be able to extend the information from the text by relating it to their experiences and to the world.* Their *responses should be thorough, thoughtful, and extensive.*

For example, when reading **literary text**, advanced-level twelfth graders should be able to produce complex, abstract summaries and theme statements. They should be able to use cultural, historical, and personal information to develop and explain text perspectives and conclusions. They should be able to evaluate the text, applying knowledge gained from other texts.

When reading **informational text**, they should be able to analyze, synthesize, and evaluate points of view. They should be able to identify the relationship between the author's stance and elements of the text. They should be able to apply text information to new situations and to the process of forming new responses to problems or issues.

When reading **practical texts**, advanced-level twelfth graders should be able to make a critical evaluation of the usefulness of the text and apply directions from the text to new situations.

Figure F-2

Draft Descriptions of the Achievement Levels
Prepared by the Original Level-setting Panel

4th-Grade Draft Descriptions

BASIC performance in reading should include:

* Determining what a text is about
* Identifying characterizations, settings, conflicts, or plots in a story
* Supporting one's understanding of a text with appropriate details
* Explaining why one likes or dislikes a text
* Connecting material in a text to personal experiences
* Making predictions about situations beyond the confines of a text
* Demonstrating an ability to maintain a focus over the entirety of a longer text

PROFICIENT performance in reading should include:

* Summarizing a text
* Recognizing an author's intent or purpose
* Making simple inferences based on information provided in a text
* Using information from a text to draw a basic conclusion
* Determining the meaning of key concepts in the text and connecting them to the main idea
* Recognizing the progression of ideas and the cause-and-effect relationships in a text
* Using the surrounding text to assign meaning to a word or phrase

ADVANCED performance in reading should include:

* Explaining an author's intent, using supporting material from the text
* Describing the similarities and differences in characters
* Demonstrating an awareness of the use of literary devices and figurative language
* Applying inferences drawn from a text to personal experiences
* Extending the meaning of a text by integrating experiences and information outside of the text
* Making and explaining a critical judgment of a text
* Demonstrating an ability to adapt reading purpose to genre and/or writing style

8th-Grade Draft Descriptions

BASIC performance in reading should include:

* Identifying the main idea or purpose of a text using information both stated and implied
* Expressing an author's purpose, viewpoint, and/or theme

Figure F-2 (continued)

Draft Descriptions of the Achievement Levels
Prepared by the Original Level-setting Panel

* Using information from a text to draw and support conclusions
* Making inferences appropriate to the information provided in a text
* Recognizing the cause-and-effect relationships in a text
* Making logical connections from the material in a text to personal knowledge and experience

PROFICIENT performance in reading should include:

* Restating the main idea using supportive details and examples from a text
* Summarizing a text using information both stated and implied
* Making inferences from a text in order to draw valid conclusions
* Interpreting the actions, behaviors, and motives of characters
* Integrating personal knowledge and experience to enhance one's understanding of a text
* Identifying an author's use of literary devices

ADVANCED performance in reading should include:

* Describing how specific literary elements interact with each other
* Synthesizing the information in a text to obtain abstract meaning or to perform a task
* Finding new applications for information derived from a text
* Making personal and critical evaluations of a text
* Analyzing an author's purpose, viewpoint, and/or theme
* Explaining an author's use of literary devices

12th-Grade Draft Descriptions

BASIC performance in reading should include:

* Explaining the main idea of a text
* Describing the main purpose in reading a selection
* Recognizing the significance of details from a reading in order to support a conclusion or perform a task
* Applying the information gathered from reading to meet an objective or support a conclusion
* Explaining the basic elements of an author's literary devices

PROFICIENT performance in reading should include:

* Drawing conclusions from and making inferences about information from different texts and writing styles
* Integrating background information with newly acquired information to support conclusions

Figure F-2 (continued)

Draft Descriptions of the Achievement Levels
Prepared by the Original Level-setting Panel

- Applying information from a text in an appropriate manner
- Bringing personal experience and accumulated knowledge into the process of critically evaluating a text
- Explaining an author's purpose in using complex literary devices

ADVANCED performance in reading should include:

- Providing innovative elaborations from textual information
- Analyzing and evaluating different points of view by means of comparison and contrast
- Identifying the relationships between an author's or narrator's stance and the various elements of the text
- Critically evaluating a text within a specific frame of reference
- Bringing the knowledge of other texts to the process of critical evaluation
- Using cultural or historical information provided in a text to develop perspectives on other situations
- Using cultural or historical information to develop perspectives on a text

Figure F-3

Revised Draft Descriptions of the Achievement Levels
Recommended by the Follow-up Validation Panel

Revised 4th-Grade Draft Descriptions

Basic performance in reading should include:

- Determining what a story/informational text is about (i.e. topic, main idea)
- Determining the main purpose for reading a selection
- Identifying character(s), setting(s), conflict(s), or plot(s) in a story
- Supporting one's understanding of a story/informational text with appropriate details
- Explaining why one likes or dislikes what they have read [a reading]
- Connecting material from a story/informational text to personal experiences
- Making predictions about situations beyond the confines of the printed material
- Maintaining a focus over the entirety of a story/informational text

Proficient performance in reading should include:

- Summarizing a story/informational text
- Recognizing an author's intent or purpose
- Making simple inferences based on information provided in a story/informational text
- Drawing a valid conclusion from a story/informational text
- Determining the meaning of key concepts in the story/informational text and connecting them to the main idea
- Recognizing relationships in a story/informational text (time order, cause/effect, compare/contrast)

Advanced performance in reading should include:

- Explaining an author's intent, using supporting material from the story/informational text
- Describing the similarities and difference in characters, settings, and plots
- Demonstrating an awareness of the use of literary devices, such as figurative language
- Applying inferences drawn from a story/informational text to personal experiences
- Extending the meaning of a story/informational text by integrating experiences and information outside of the text
- Making and explaining a critical judgment of a story/informational text
- Demonstrating an ability to adapt reading purpose to a variety of printed material and/or writing style

Revised 8th-Grade Draft Descriptions

Basic performance in reading should include:

- Identifying the main idea, theme, or purpose of a text
- Describing the main purpose for reading a selection

Figure F-3 (continued)

Revised Draft Descriptions of the Achievement Levels
Recommended by the Follow-up Validation Panel

- Expressing an author's purpose and viewpoint
- Making inferences, predictions, and drawing conclusions that are supported by information in a text
- Recognizing the relationships among facts, ideas, events, and concepts within a text (e.g., cause and effect, chronological order, and characterization)
- Making logical connections between the text and personal knowledge
- Maintaining a focus over the entirety of a story/informational text

Proficient performance in reading should include:

- Restating the main idea, theme, or purpose of a text using supporting details and examples
- Summarizing a text using both stated and implied information
- Interpreting the actions, behaviors, and motives of characters
- Using personal knowledge and experience to enhance one's understanding of a text
- Identifying an author's use of literary devices (i.e. personification, foreshadowing, and so forth).
- Using inferences from a text in order to draw valid conclusions.

Advanced performance in reading should include:

- Describing how specific literary elements (i.e., setting, plot, characters, and themes) interact with each other
- Synthesizing the information in a text to obtain implied meaning or to perform a task
- Applying information derived from a text to new situations.
- Explaining an author's use of literary devices (i.e., irony, personification, and foreshadowing)
- Responding personally and critically to a text
- Analyzing an author's purpose and viewpoint
- Using cultural or historical information to develop perspectives on a text
- Using cultural or historical information provided in a text to develop perspectives on other situations

Revised 12th-Grade Draft Descriptions

Basic performance in reading should include:

- Explaining the main idea, theme, or purpose of a text
- Describing the main purpose for reading a selection
- Recognizing the significance of details from a reading in order to support a conclusion or perform a task

Figure F-3 (continued)

Revised Draft Descriptions of the Achievement Levels
Recommended by the Follow-up Validation Panel

- Applying the information gathered from reading to meet an objective or support a conclusion
- Identifying and explaining the basic elements of an author's literary devices
- Making logical connections between a text and personal knowledge and experience
- Maintaining a focus over the entirety of a story/informational text

Proficient performance in reading should include::

- Drawing conclusions and making inferences from different texts and writing styles
- Integrating background information with newly acquired information to support conclusions
- Applying information from a text in an appropriate manner
- Applying personal experience and accumulated knowledge to the process of critically evaluating a text
- Explaining an author's purpose in using complex literary devices (i.e. irony, symbolism)

Advanced performance in reading should include:

- All basic and proficient reading behaviors listed previously
- Prompted by information from a text, innovating in new situations and creating new answers to old situations
- Analyzing, synthesizing, and evaluating different points of view by means of comparison and contrast
- Identifying the relationships between an author's or narrator's stance and the various elements of the text
- Critically evaluating a text within a frame of reference
- Applying the knowledge of other texts to the process of critical evaluation
- Using cultural or historical information to develop perspectives on a text
- Using cultural or historical information provided in a text to develop perspectives on other situations

Figure F-4

Meeting Participants, NAEP Reading Achievement Level Setting
Original Meeting, St. Louis, Missouri, August 21 - 25, 1992

David Awbrey
Wichita Eagle
Wichita, KS

Dorothy Botham
Milwaukee Public Library
Milwaukee, WI

Anna Caballero
Attorney
Salinas, CA

Kathy Casseday
WFSP Radio Station
Kingwood, WV

Dee Ellis
Trimble Banner Newspaper
Milton, KY

Nona Smith
NAACP
New York, NY

Lillaine Speese
Oakdale Elementary School
Oroville, CA

Clifton Whetten
Retired Construction Sprvsr.
Elfrida, AZ

P. Richard Brackett
Brackett & Assoc. Motivational Marketing
Company
Brentwood, TN

Kathleen Harkey
Corporate Presentations
Nashville, TN

Patricia Oliverez
Salinas Public Library
Salinas, CA

Christine Sentz
North Milwaukee Branch Library
Milwaukee, WI

Carolyn Sullivan
Planters & Merchants Bank
Gillett, AR

Paula Abrams
City Hall
Bedford, KY

Rhonda Cantrell Dunn
Nashville Urban League
Nashville, TN

Harlon Gaskill (CPA)
Gaskill, Pharis & Pharis
Dalhart, TX

Jean McManis
Local/State Education Volunteer
State College, PA

Linda Borsum
Lakeview School District
Battlecreek, MI

Anne Kraut
Elementary Supervisor
Princeton, WV

Robert Williams
Macomb Intermediate SD
Clinton Township MI

Figure F-4 (continued)

Meeting Participants, NAEP Reading Achievement Level Setting
Original Meeting, St. Louis, Missouri, August 21 - 25, 1992

Constance Boyd
Owen J. Roberts SD
King of Prussia, PA

Mary Gonzalez
Mesa Public Schools
Mesa, AZ

James Schindler
Jordan SD
Salt Lake City, UT

Kathryn Flannery
Indiana University
Bloomington, IN

Catherine Hatala
School District of Philadelphia
Philadelphia, PA

Raymond Morgan
Old Dominion University
Virginia Beach, VA

Berton Wiser
Columbus Public School
Columbus, OH

Freda Andrews
Durham Public Schools
Durham, NC

Tim Barnes
Ashdown Public Schools
Ashdown, AR

Larry Barretto
Maplewood Elementary School
Coral Springs, FL

Gloria Darling
Conway Public Schools
Conway, AR

Nina Frederick
Marion County School System
Hackleburg, AL

Karen Fugita
Oak Grove SD
San Jose, CA

Anne Gregory
Durham Public Schools
Durham, NC

Joseph Howard
Josiah Quincy School
West Roxbury, MA

Roberta Johnson
Cleveland Public Schools
Cleveland, OH

Marcia Jolicoeur
Lisbon Falls School
Lewiston, ME

Elizabeth Litchfield
Westwood School District
Emerson, NJ

Jean Young
Houston ISD
Houston, TX

Wilma Centers
Wolfe County Middle School
Campton, KY

Eunice Coakley
Greenville School
Greenville, SC

Figure F-4 (continued)

Meeting Participants, NAEP Reading Achievement Level Setting
Original Meeting, St. Louis, Missouri, August 21 - 25, 1992

Eugenia Constantinou
Prince Georges County Schools
Silver Spring, MD

Walt Cottingham
Henderson City Schools
Zirconia, NC

Deborah Davidson
Westhampton Beach UFSD
Patchogue, NY

Julia Dominique
Department of Education USVI
Sunnyisle, VI

Patricia Gerdes
Waelder ISD
Schulenburg, TX

Leslie Leech
Elkton School
Elkton, SD

Belva Leffel
Whittier Christian Jr. High
Norwalk, CA

Harriett McAllaster
Volusia County Schools
DeLand, FL

Mary Orear
Camden-Rockport HS & MS
Rockport, ME

Judith Zinsser
Houston ISD
Houston, TX

Mary Ann Ledbetter
East Baton Rouge Parish School Board
Baton Rouge, LA

Cora Cummins
Conway Public Schools
Conway, AR

Stanley Fraundorf
Cuba City Public Schools
Cuba City, WI

Georgia Howard
Volusia County Schools
Holly Hill, FL

Roger Larsen
Campbell County SD
Gillette, WY

Judith Lusk
Norfield School District
Rockbury, VT

Donnie McQuinn
Wolfe County Board of Education
Pine Ridge, KY

Meredith Powers
Swansea School
Providence, RI

Beth Schieber
Kingfisher Schools
Okarche, OK

Carolyn Sue Wilson
Greenville, SC

Sue Zak
Cleveland Board of Education
Garfield Heights, OH

Figure F-5

Meeting Participants, NAEP Reading Achievement Level Setting
Follow-Up Validation Meeting, San Diego, California, October 9 - 11, 1992

Meredith Powers
Swansea School
Providence, RI

Roger Larsen
Campbell County SD
Gillett, WY

Beth Schieber
Kingfisher Schools
Okarche, OK

Elizabeth Litchfield
Westwood School District
Emmerson, NJ

Larry Barretto
Maplewood Elementary School
Coral Springs, FL

Anne Gregory
Durham Public Schools
Durham, NC

Debra Davidson
Westhampton Beach UFSD
Patchogue, NY

Eugenia Constantinou
Prince Georges County School
Silver Spring, MD

Eunice Coakley
Greenville School
Greenville, SC

Nancy Livingston
Brigham Young University
Salt Lake City, UT

Susan McIntyre
University Wisconsin-Eau Claire
Eau Claire, WI

Clyde Colwell
Norfolk Public School
Norfolk, VA

Jo Prather
Mississippi Department of Education
Jackson, MS

Mary Orear
Camden-Rockport HS & MS
Rockport, ME

Shelia Potter
Michigan Department of Education
Lansing, MI

Gene Jongsma
IRA Subcommittee Member
San Antonio, TX

Peggy Dutcher
Michigan Education Assessment Program
Lansing, MI

Martha Carter
Milwaukee Public Schools
Milwaukee, WI

Mark Conley
Michigan State University
Holt, MI

Figure F-6

Meeting Participants, NAEP Reading Revisit
Validation Meeting, Saint Louis, MO, October 14 - 16, 1994

Jody Alexander
Madison No. 1
Phoenix, AZ

Evelyn Alford
East Baton Rouge Public Schools
Baton Rouge, LA

Winfrey Bates
Mannsville Elementary School
Mannsville, KY

Joyce Boone
John Strange Elementary School
Indianapolis, IN

Linda Brooks
Alcorn County Public Schools
Corinth, MS

Katie Burnham
Pa Wau Lu Middle School
Gardnerville, NV

Martha Carter
Milwaukee Public Schools
Milwaukee, WI

Carol Case
Mirabean B. Lamar High School
Houston, TX

Molly Chun
Applegate Elementary School
Portland, OR

Roseine Church
Cheyenne, WY

Connie Clayton
Franklin High Schcol
Franklin, WV

David Colburn
Flathead High School
Kalispell, MT

Brenda Creel
Jessup Elementary School
Cheyenne, WY

Pam Diamond
Hellgate Middle School
Missoula, MT

Caroline Downs
Worland Middle School
Worland, WY

Esther Dunnington
Grandview High School
Grandview, MO

Sandra Forsythe
Green Valley High School
Henderson, NV

David Fredette
Westborough High School
Westborough, MA

Cynthia Freeman
Maryville High School
Maryville, TN

Rita Gallagher
Roswell, NM

Lorraine Gerhart
Elmbrook Middle School
Elm Grove, WI

Maria Valeri-Gold
Georgia State University
Atlanta, GA

Figure F-6 (continued)

Meeting Participants, NAEP Reading Revisit
Validation Meeting, Saint Louis, MO, October 14 - 16, 1994

Bill Hammond
GA Department of Education
Atlanta, GA

Sally Hellman
Las Vegas, NV

Grace Herr
West Linn High School
West Linn, OR

Sarah Herz
Coleytown Middle School
Westport, CT

Susan Hodgin
Moscow Public Schools
Moscow, ID

Beverly Hoffmaster
Berkeley Heights Elementary School
Martinsburg, WV

Roberta Horton
Custer County High School
Miles City, MT

Lory Johnson
Iowa Department of Education
Des Moines, IA

Ruth Johnson
Holmes High School
Covington, KY

Theresa Lowe
Rancho Viejo School
Yuma, AZ

Ruby Mayes
S.P. Waltrip High School
Houston, TX

Robert McKean
Havre Public Schools
Havre, MT

Pamela McNair
Lemon G. Hine Jr. High School
Washington, DC

Daniel McQuagge
Delta State University
Cleveland, MS

Cheryl Miller
Buchanan Elementary School
Baton Rouge, LA

Donna Miller
Chinook High School
Chinook, MT

Lynn Minderman
Honeoye Falls-Lima Public Schools
Honeoye Falls, NY

John Morrissey
Huntley Project Elementary School
Worden, MT

Pamela Perryman
Selah Middle School
Selah, WA

Kathleen Sanders
Los Ang :les Unified S.D.
Wilmington, CA

Helen Schotanus
NH Department of Education
Concord, NH

Terrence Smith
Verona School
Battle Creek, MI

Figure F-6 (continued)

Meeting Participants, NAEP Reading Revisit
Validation Meeting, Saint Louis, MO, October 14 - 16, 1994

Faith Stevens
Haslett Public Schools
Haslett, MI

Richard Telfer
Univ. of Wisconsin-Whitewater
Whitewater, WI

Cara Terry
Lakewood High School
St. Petersburg, FL

James Thompson
Simpson-Waverly School
Hartford, CT

Patsy Turner
Great River Co-operative
West Helena, AR

Toni Walters
Oakland University
Rochester, MI

Barbara Watson
Agricola Elementary School
Lucedale, MS

Florence Wakuya
Hawaii Department of Education
Honolulu, HI

Janet Williams
Bluewell Elementary School
Bluefield, WV

Sarah Williams
Maryville Middle School
Maryville, TN

Philip Yeaton
Concord, NH

Katie Young
Louisiana Department of Education
Baton Rouge, LA

APPENDIX G

THE EFFECT OF MONITORING ON ASSESSMENT SESSIONS IN NONPUBLIC SCHOOLS

APPENDIX G

The Effect of Monitoring on Assessment Sessions in Nonpublic Schools

Eddie H. S. Ip and Nancy L. Allen

Educational Testing Service

G.1 OVERVIEW

This appendix describes the analyses performed to look at the effect of monitoring nonpublic-school test sites in the 1994 Trial State Assessment in reading. As described in Chapter 4, a randomly selected half of the administration sessions in nonpublic schools in all jurisdictions were observed by Westat-trained quality control monitors. In the 1992 Trial State Assessment of reading, only public schools participated and were studied with regard to the effect of monitoring. Because nonpublic schools were included for the first time in the 1994 Trial State Assessment, a study was carried out on the effect of monitoring those schools.

To study the effect of monitoring, statistical analyses were performed on the 1994 Trial State Assessment reading data. The analyses carried out for the 1994 Trial State Assessment were more extensive than those done for the 1992 assessment. The first analysis, which was mainly exploratory and similar to that done for 1992, is described in section G.2. Overall statistical hypothesis testing on whether there is a monitoring effect is described in section G.3. In Section G.4, we examine the data more closely and identify jurisdictions that had considerable monitoring effects. Section G.5 is devoted to the analysis and discussion of monitoring effect in the nonpublic schools, where sample sizes might generally be small enough to become a concern. More information on technical issues in the statistical analyses can be found in Ip and Allen (1995).

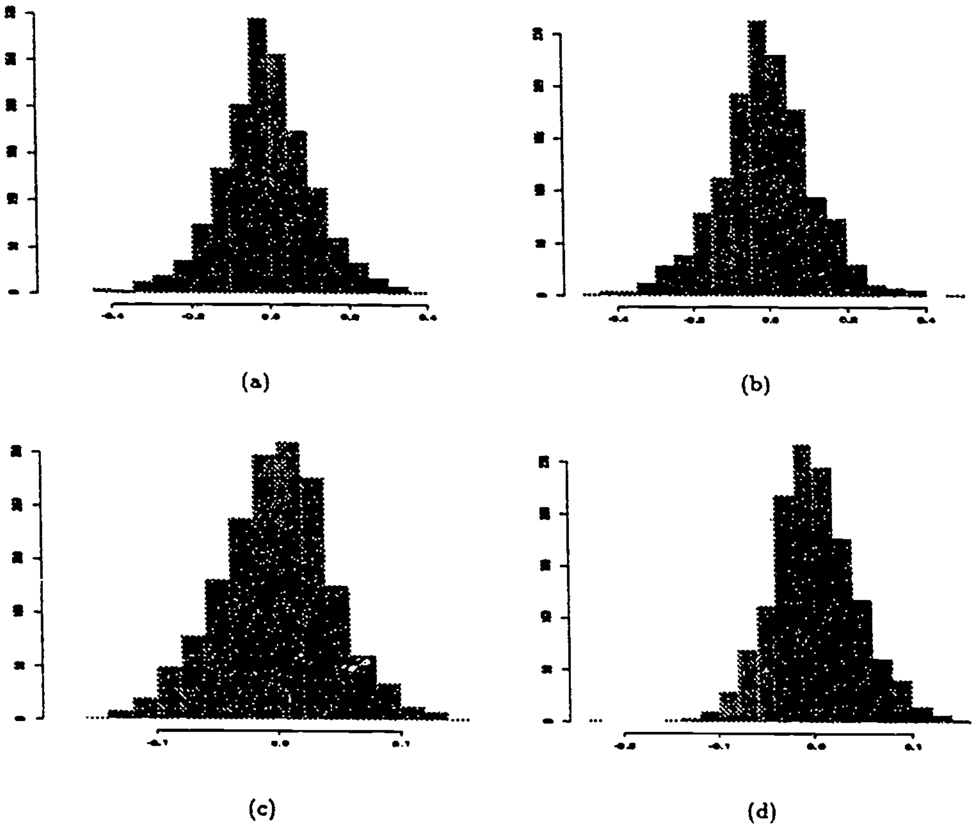
G.2 EXPLORATORY ANALYSIS

In the 1994 NAEP design for the fourth-grade reading assessment, there were two proficiency scales—Reading for Literary Experience and Reading to Gain Information. A difference between the administrative procedures of the 1994 and 1992 Trial State Assessment is that nonpublic schools were included in the 1994 assessment. In the 1994 Trial State Assessment there were 34 jurisdictions with nonpublic-school samples. However, in the nonpublic-school sessions, there were some jurisdictions that had either too low a response rate or not enough nonpublic schools to sample from. Eventually the number of jurisdictions in the nonpublic-school sessions was reduced to 23 in the analysis.

Figure G-1 contains the histograms of differences (unmonitored - monitored) in mean item scores between unmonitored and monitored schools for the public- and nonpublic-school

Figure G-1

Histograms of Differences in Mean Item Scores Between Unmonitored and Monitored Sessions*



(a)

(b)

(c)

(d)

* (a) nonpublic schools, Reading for Literary Experience scale; (b) nonpublic schools, Reading to Gain Information scale; (c) public schools, Reading for Literary Experience scale; (d) public schools, Reading to Gain Information scale.

sessions on the two proficiency scales. The distributions are all unimodal but skewed to various degrees. The range of the difference in mean item scores between unmonitored and monitored schools is larger for nonpublic schools. For public schools, the mean differences were 0.0007 and 0.0008 for the Reading for Literary Experience scale and Reading to Gain Information scale, respectively. The median differences were -0.0007 and 0.0011, respectively. For nonpublic schools, the mean differences were 0.008 and 0.01 for the Reading for Literary Experience scale and Reading to Gain Information scale, respectively. The median differences were 0.007 and 0.016, respectively.

Figures G-2 and G-3 show plots of the differences in average mean score between the unmonitored and the monitored sessions on the two scales, with nonpublic and public schools plotted separately. The symbol "1" indicates an average mean score difference for a jurisdiction on the first proficiency scale, Reading for Literary Experience and "2" indicates an average mean score difference for the second scale, Reading to Gain Information. Each point represents, for a jurisdiction, the difference between mean item scores for unmonitored and monitored sessions, averaged over items of a proficiency scale.

The difference between public- and nonpublic-school sessions is evident. In the nonpublic schools, there are more positive values than negative ones and the discrepancy between the two proficiency scales appears to be much bigger. This visualization of the data, however, does not tell in a convincing way whether there is a genuine difference between monitored and unmonitored schools. The apparent difference between the public- and nonpublic-school sessions may simply be due to a difference in the sample sizes of the students involved in the two samples. As a matter of fact, the numbers of students sampled from the nonpublic-school sessions are usually much smaller (typically a dozen to several dozen) than the numbers of students from the public-school sessions (varies from one hundred to several hundred). Since the variability from jurisdiction to jurisdiction seems considerable as seen in Figures G-2 and G-3, the issue of whether the difference is real deserves to be investigated in greater detail. The following two questions, therefore, are to be addressed in the subsequent analysis:

1. For public schools, is the difference between monitored and unmonitored sessions significant? A similar question can be asked about the nonpublic schools.

2. If there is a difference, can we identify where such a difference comes from?

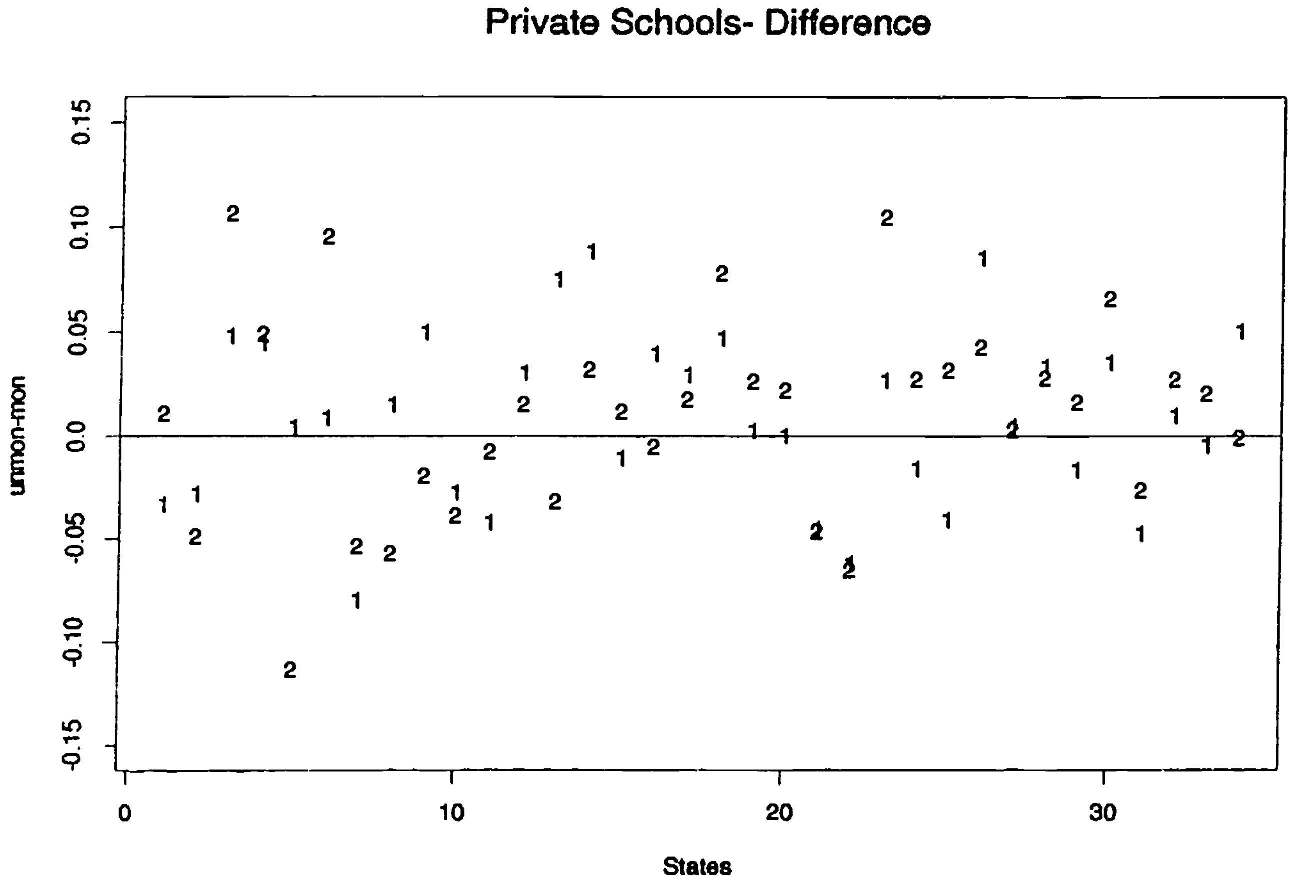
G.3 HYPOTHESIS TESTING

To facilitate discussion, we created a generic variable DIFF, which is the mean item score of unmonitored sessions minus that of monitored sessions for each item and jurisdiction. The averages of the variable DIFF over items are the values plotted in Figures G-2 and G-3.

More formally, let p_{1ijk} and p_{2ijk} denote the mean scores for students from unmonitored and monitored sessions for item i, jurisdiction j and proficiency scale k. The 1 in

Figure G-2

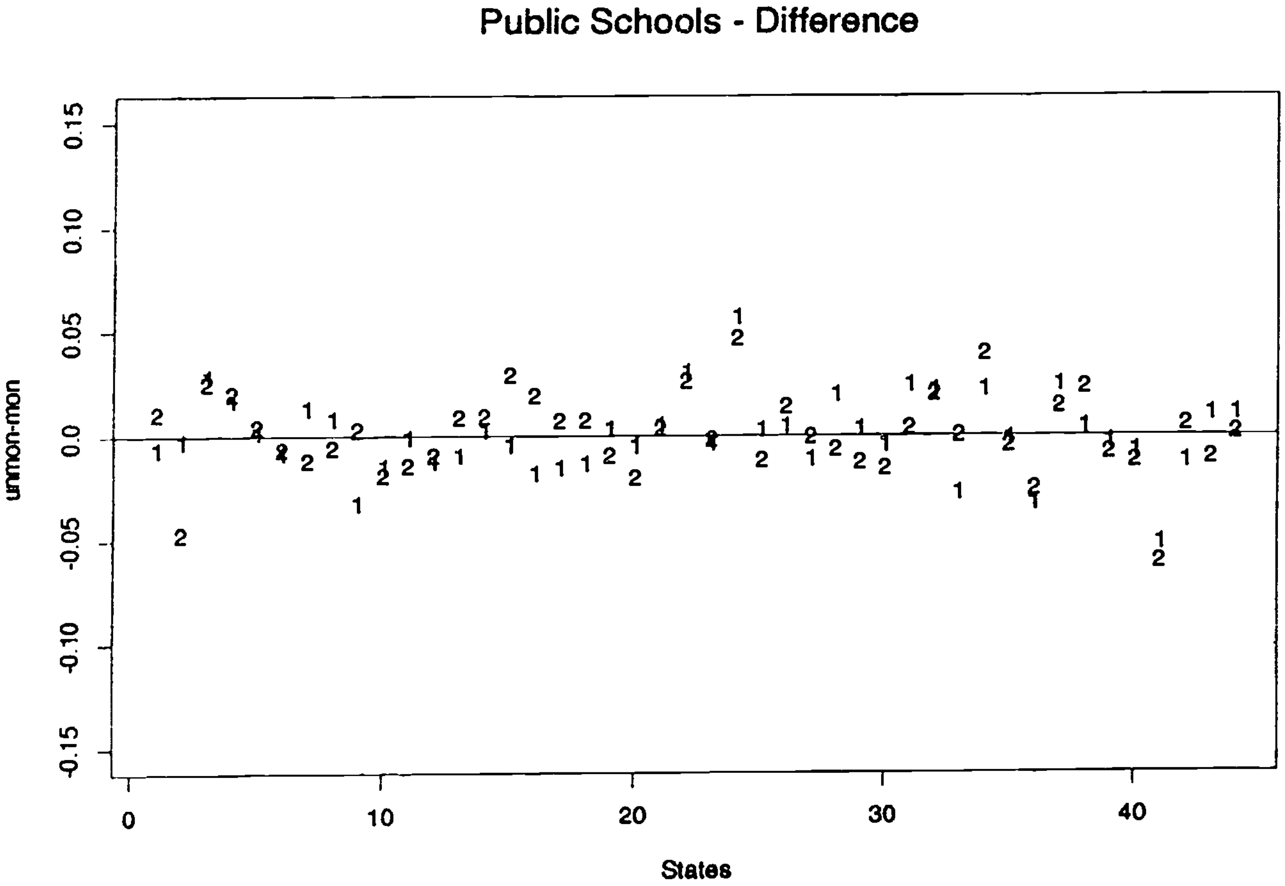
Difference in Average (Over Items) of Mean Scores of Unmonitored and Monitored Sessions in Nonpublic Schools (Unmonitored - Monitored)*



* Each point represents a jurisdiction. "1" indicates the Reading for Literary Experience scale; "2" indicates the Reading to Gain Information scale.

Figure G-3

Difference in Average (Over Items) of Mean Scores of Unmonitored and Monitored Sessions in Public Schools (Unmonitored - Monitored,*



*Each point represents a jurisdiction. "1" indicates the Reading for Literary Experience scale; "2" indicates the Reading to Gain Information scale.

the first subscript denotes unmonitored schools; the 2 denotes monitored schools. Define the variable DIFF as

$$d_{ijk} = p_{1ijk} - p_{2ijk} \;,$$

$i = 1,\ldots,I_k, j = 1,..J$ and $k = 1,2$.

Specifically, denote the average DIFF values of the items over scale k for the jth jurisdiction by $\bar{d}_{jk}$. That is,

$$\bar{d}_{jk} = \frac{1}{I_k} \sum_{i=1}^{I_k} d_{ijk} \;.$$

The vectors $\{(\bar{d}_{j1} , \bar{d}_{j2})\}$, j=1,...J of average DIFF values are assumed to be a sample from a distribution F with a mean $(\mu_1,\mu_2)^T$. The hypothesis we want to test is

$$H_{\bullet} \;:\; (\mu_1,\mu_2)^T = \mathbf{0}.$$

A function of$(\bar{d}_1 , \bar{d}_2)^T$where

$$\bar{d}_k = \frac{1}{J} \sum_{i=1}^{J} \bar{d}_{jk},$$

j=1,...,J,is a natural candidate of test statistic for testing$H_{\bullet}$. We use the test statistic $\bar{d}_1^2 + \bar{d}_2^2$. To test $H_{\bullet}$, the distribution of $\bar{d}_1^2 + \bar{d}_2^2$ under $H_{\bullet}$ needs to be known. The bootstrap method gives a nonparametric estimate of this distribution without the usual normal assumptions.

The bootstrap (Efron, 1979) is a resampling method that reuses the observed data by drawing repeatedly from the sample with replacement to produce bootstrap samples. The bootstrap samples are then used to estimate variance and bias of the statistic of interest. The basic idea of bootstrap in this application is to emulate the data generation process while still making minimal statistical assumptions. It uses the empirical distribution as a proxy for the true underlying distribution. Supposing the experiment could be replicated under the same design, we would expect to witness variabilities at two levels. The DIFF values for each item in each jurisdiction would change. If a jurisdiction is regarded as a random sample from a population, we should also find variability at the jurisdiction level. In order to account for variabilities of the data at both the item and jurisdiction levels, we propose a two-tier bootstrap method for hypothesis testing. The details of the procedures can be found in Ip and Allen (1995).

It should also be noted that due to the balanced incomplete block (BIB) design of NAEP booklets, the items cannot be regarded as independent. The clustering effect of items tends to deflate the sampling error estimated using simple normal approximation. Johnson and Rust (1992) and Johnson, Rust, and Wallace (1994) refer to such an effect as a design effect. The appropriate variance should be about twice as large as the one estimated using normal approximation. Hence, the sampling error associated with each value of DIFF, d_{ijk}, can be approximated by

$$\sigma^2_{ijk} = f_{ijk} \{ \frac{p_{1ijk}(1-p_{1ijk})}{n_{1ijk}} + \frac{p_{2ijk}(1-p_{2ijk})}{n_{2ijk}} \}, \qquad (1)$$

where n_{1ijk} and n_{2ijk} denote respectively the number of students who attempted the ith item for unmonitored and monitored sessions. The factor f_{ijk} denotes the design effect and is taken to be equal to 2.0. The sampling error given by (1) tends to be conservative for polytomous items since the error is assumed to be derived from a binomial distribution.

A bootstrap sample of $B = 300$ was used in the analysis of the NAEP 1994 Trial State Assessment data. To provide a visualization of the result, we display two-dimensional plots of the centered bootstrap samples in Figures G-4 and G-5. Each point in the plot represents a centered bivariate bootstrap sample pair. The two graphs show what one would generally expect the distribution of $(\bar{d}_1, \bar{d}_2)^T$ to look like under the null hypothesis. The "X" mark on the graph indicates the sample value of the observed data. The more deviant it is from the origin, the greater is the evidence that the null hypothesis should be rejected. For the public-school sessions, the p-value in testing against the hypothesis that $\mu = \mathbf{0}$ was 0.523. For the nonpublic-school sessions, the p-value was 0.503. These results suggest that we do not reject the hypothesis that there is a genuine difference between monitored and unmonitored sessions. A variance component analysis was also completed (see Ip & Allen, 1995) and it indicated a similar result—that the difference between monitored and unmonitored sessions was due to sampling noise.

G.4 LOOKING AT THE DATA CLOSELY

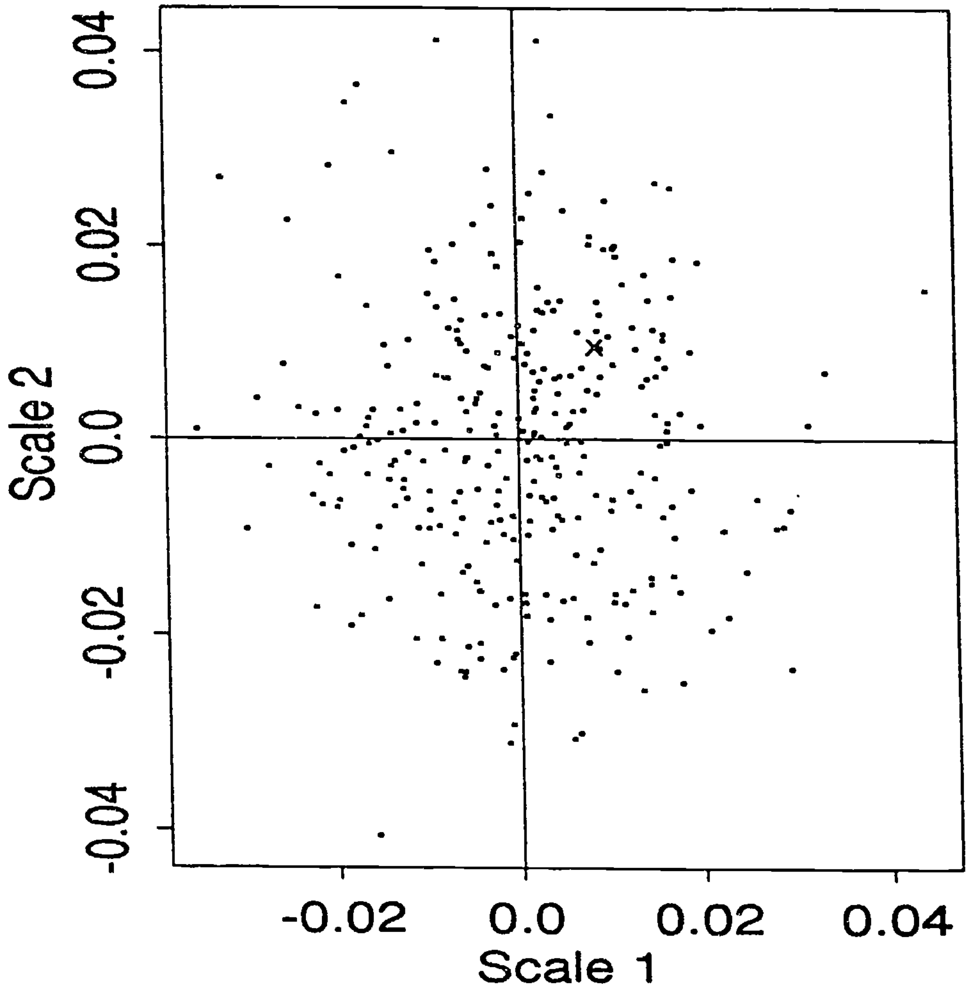
From the hypothesis testing in Section G.3, there seems to be little overall evidence that the unmonitored sessions are performing differently than the monitored sessions. However, closer examination and alternative analyses of the data show that there are some jurisdictions that might have large monitoring effects.

Table G-1 displays the list of DIFF values averaged over items for each of the jurisdictions on the two proficiency scales. Several jurisdictions either had too low a nonpublic-school participation or did not have enough nonpublic schools to sample from. They are marked by daggers and are excluded in the subsequent analyses. To help visualize the data, the points are displayed on two-dimensional plots in Figures G-6 and G-7. The effect of monitoring in public schools does look smaller.

The 2 x 2 tables in Tables G-2 and G-3 provide summaries of information contained in Table G-1. They show the number of jurisdictions that have positive and negative DIFF values on the two proficiency scales. As expected, the public schools (in Table G-3) seem to show a

Figure G-4

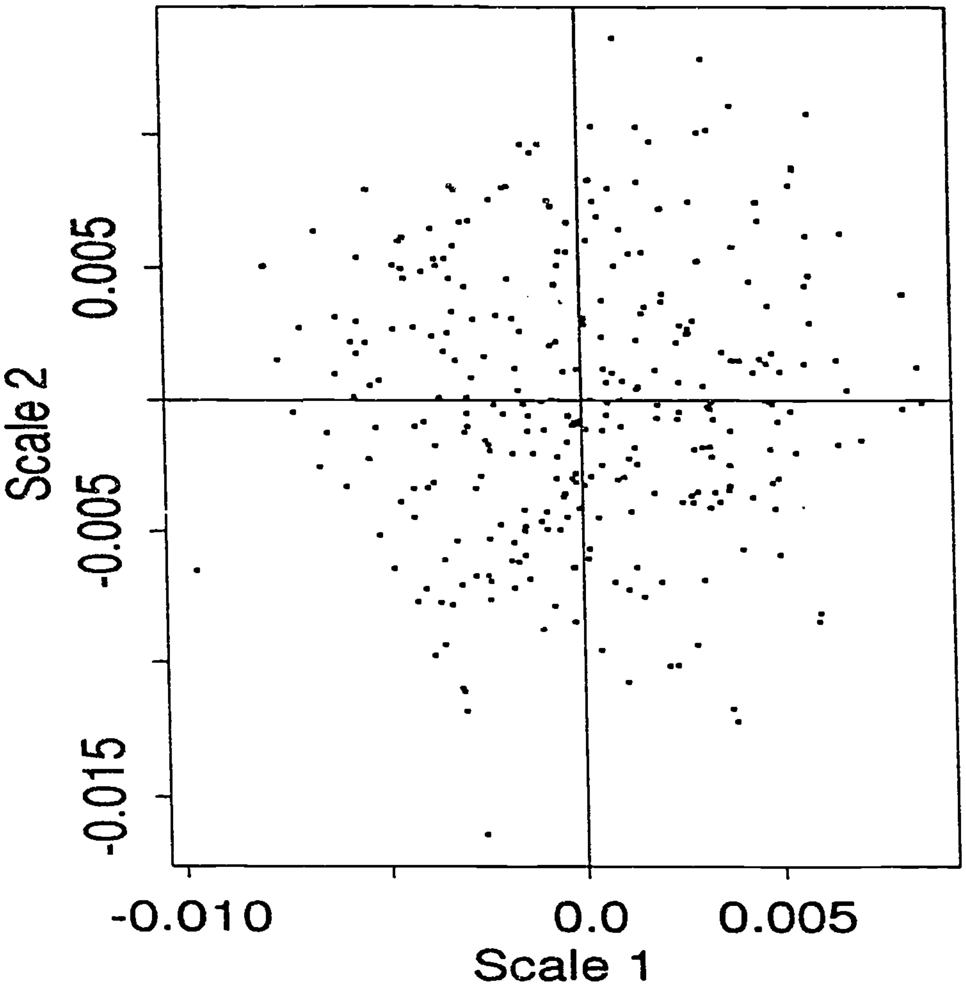
Bootstrap Results on DIFF for Nonpublic Schools*



* The horizontal scale is value for scale 1 and the vertical scale is value for scale 2. Each point represents a bivariate bootstrap sample. The sample average value of DIFF is marked by X.

Figure G-5

Bootstrap Results on DIFF for Public Schools*



* The horizontal scale is value for scale 1 and the vertical scale is value for scale 2. Each point represents a bivariate bootstrap sample. The sample average value of DIFF is marked by X. Note that the scale is different than that in Figure G-4.

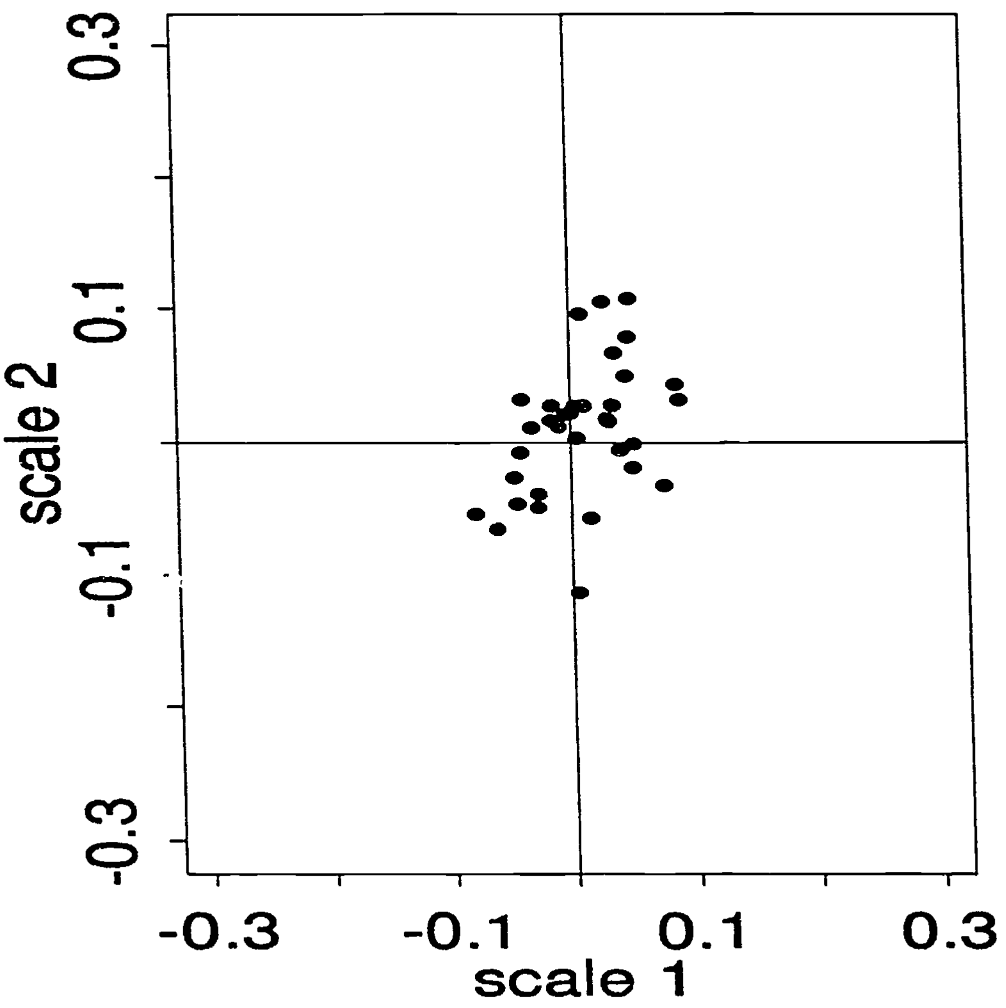
Table G-1
Table of the Variable DIFF for Nonpublic and Public Schools on Two Proficiency Scales

Jurisdiction	Nonpublic s1	Nonpublic s2	Public s1	Public s2
CO	-0.033	0.011	-0.006	0.011
PA	-0.028	-0.049	-0.002	-0.047
NM	0.049	0.107	0.029	0.025
DE	0.045	0.049	0.002	0.004
VA	0.005	-0.113	-0.008	-0.006
GU	0.009	-0.096	0.008	-0.005
IN	-0.080	-0.053	-0.015	-0.019
ME	0.015	-0.057	-0.001	-0.014
WV	0.050	-0.019	-0.009	0.009
MA	-0.027	-0.039	0.003	0.009
LA	-0.042	-0.008	-0.013	0.008
MN	0.031	0.016	0.003	-0.009
ID	0.075	-0.032	-0.004	-0.020
AL	0.089	0.032	0.006	0.003
NY*	-0.010	0.012	0.031	0.026
NJ	0.040	-0.005	-0.003	-0.001
DC*	0.030	0.018	0.057	0.047
AR	0.048	0.078	0.003	-0.011
CT	0.003	0.027	0.005	0.014
CA*	0.000	0.022	-0.011	0.000
SC*	-0.045	-0.046	0.020	-0.006
WI*	-0.062	-0.065	0.004	-0.012
MS*	0.027	0.105	-0.004	-0.015
HA	-0.016	0.027	0.025	0.004
FL*	-0.041	0.032	0.022	0.020
ND	0.086	0.043	-0.027	0.000
KY*	0.005	0.003	0.023	0.040
RI	0.034	0.028	-0.032	-0.025
NE*	-0.016	0.017	0.025	0.014
MO	0.036	0.066	0.005	0.023
IA	-0.047	-0.026	-0.002	-0.008
MD*	0.010	0.028	-0.007	-0.012
GA	-0.004	0.021	-0.051	-0.061
MT*	0.051	-0.001	-0.012	0.005

* The nonpublic-school sessions in these jurisdictions are excluded in the subsequent analysis.

Figure G-6

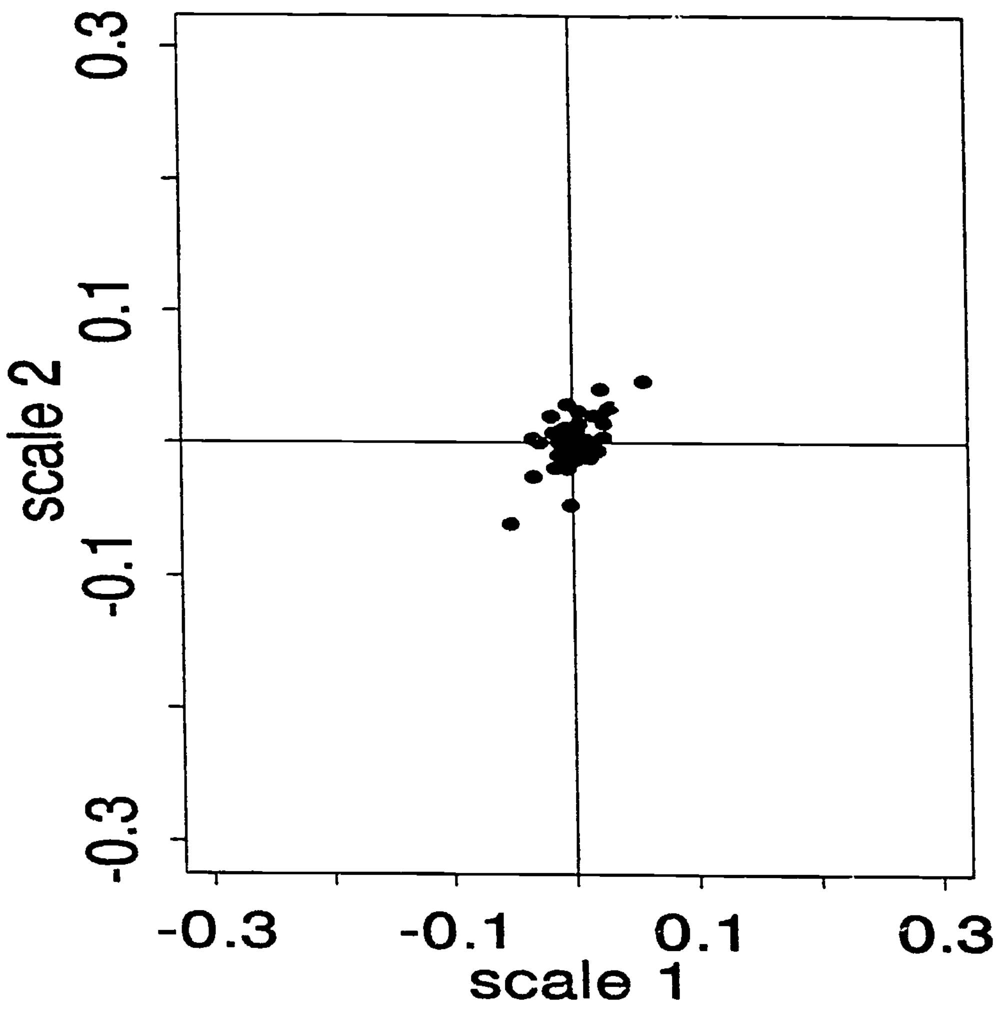
Variable DIFF for Nonpublic Schools on Two Proficiency Scales*



[*] Each point represents a jurisdiction.

Figure G-7

Variable DIFF for Public Schools on Two Proficiency Scales*



* Each pcint represents a jurisdiction.

Table G-2

Counts of Signs in Nonpublic-school Sessions*

		s_1	
		-	+
s_2	-	5	5
	+	3	10

Table G-3

Counts of Signs in Public-school Sessions*

		s_1	
		-	+
s_2	-	11	5
	+	6	12

* The number in each cell indicates the number of jurisdictions falling in that cell. Scales 1 and 2 are indicated by s_1 and s_2.

consistency in performance across scales in that larger numbers lie along the main diagonal. This indicates that if a jurisdiction has higher(or lower) score in one scale on average, it also has higher(or lower) scores on the other. Approximately the same number of jurisdictions have positive DIFF values on both scales as have negative DIFF values on both scales. On the other hand, in the nonpublic-school sessions (in Table G-2), ten jurisdictions are positive on both scales and only five are negative. The number of both positives is inconsistently high given the hypothesis that there should be about the same number of jurisdictions showing positive and negative signs on both scales. Proportionally, about the same number of jurisdictions are off the diagonal in each direction in Tables G-2 and G-3. These numbers also tend to be smaller than those on the diagonals.

Taking the different sample sizes of the public and nonpublic schools into account, we further our analysis by performing simple univariate statistical hypothesis testing on the data in Table G-1.

Using (1) as our estimate of variance, a simple z-test is conducted to test whether the average of DIFF over items in a particular scale is significantly different than zero for all jurisdictions (except those marked by daggers in the nonpublic-school sessions) listed in Table G-1. The result of this analysis for the nonpublic schools is summarized in Table G-4.

In Table G-4, we mark the jurisdictions that show significance on both scales with double asterisks, and those that show significance on either of the scales individually with a single asterisk. A total of eight values are significant in either one or both scales. Many are in the cell where unmonitored sessions are doing better than monitored sessions. This number seems to be alarmingly high given that there are only 46 tests of significance at a 0.05 level.

The picture in the public schools is somewhat more usual looking, although not totally without problems. Two jurisdictions appear to be outliers. They can be seen on the corners along the 45 degree line in Figure G-7. The District of Columbia (with means of 0.057 and 0.047 on the two scales) is extreme in being highly positive on both scales and Georgia (with means of -0.051 and -0.061 on the two scales) is extreme in being negative. Eleven p-values are less than 0.05, disregarding the District of Columbia and Georgia. Since there are only 68 tests of significance at the 0.05 level, this number of significant results also appears to be quite high.

G.5 CHECKING ON THE NONPUBLIC SCHOOLS

Given the small sample sizes in nonpublic schools, it is not clear whether differences between monitored and unmonitored sessions are real. We therefore perform a "micro-level" analysis on the nonpublic schools, especially on the ten jurisdictions in which the unmonitored school sessions are performing better than the monitored school sessions on both scales. These ten jurisdictions are (see Table G-4) : Delaware, New Mexico, Guam, Alabama, North Dakota, Missouri, Maine, Arkansas, Connecticut and Rhode Island.

Among these ten jurisdictions (referred to as the "group of ten" hereafter), Delaware, New Mexico, Guam, Alabama, North Dakota, and Missouri show significant differences in one or both scales. To cross-validate the results obtained by looking at averages of DIFF over items, we examine the composite proficiency scores of the jurisdictions classified by monitored and

Table G-4

2 x 2 Table Showing Specific Jurisdictions from Table G-2 (Nonpublic Schools)

	-	+
-	IN* MA LA IA PA	VA* ME WV ID NJ
+	CO HA GA	DE** NM* GU* AL* ND* MO* MN AR CT RI

* Significant in either one scale at the 0.05 level.
** Significant in both proficiency scales at the 0.05 level.

unmonitored status. These averages of composite proficiency scores and their standard errors for each jurisdiction (nonpublic-school sessions only) are displayed in Table G-5. The results are consistent with what we saw in the DIFF analysis of the items. All jurisdictions that show positive effects on both scales in the DIFF analysis show positive effects in the proficiency scale analysis; similar results are observed for jurisdictions with negative effects. Those jurisdictions that exhibit a significant effect in monitoring at the 0.05 level are flagged by asterisks in Table G-5. Arkansas, Guam, and North Dakota show significantly positive effect (unmonitored schools doing better) while Indiana shows significantly negative effect (monitored schools doing better), measured on a composite proficiency scale. Using mean proficiencies as the criteria, there are now fourteen jurisdictions showing positive effects and nine showing negative effects. In order to look for plausible explanations of the higher number of jurisdictions whose unmonitored schools scored higher, we examine the characteristics of the schools being sampled in the group of ten and observe two phenomena.

One finding is that many of the jurisdictions in question are performing at the low end of the scales. New Mexico, Guam, and North Dakota are the three jurisdictions that are lowest (and in that order) in mean proficiency scores in nonpublic schools. (North Dakota, however, is top in mean proficiency score in the *public*-school sessions). One possible explanation for their unmonitored schools doing better is that there is just a bigger variability in jurisdictions with lower scores.

Our second finding is that in nonpublic-school sessions that show significant effects, the numbers of schools included in those sessions are generally quite small. In fact, four jurisdictions in the group of ten have fewer than ten schools, monitored and unmonitored combined, surveyed. When a few schools (even one or two) with characteristics that are atypical (for example, Department of Defense schools) are included in a monitored or unmonitored sample, that may be sufficient to make the result look unbalanced.

For example, in Alabama, which is in the group of ten, there are only 5 monitored and 4 unmonitored schools in the NAEP sample. Three unmonitored schools are large nonpublic schools with high teacher/student ratios and high percentages of parents working as professionals. On the other hand, a Department of Defense school with a high percentage of parents who were either blue-collar workers, farmers, or not regularly employed workers is present in the monitored group. Another jurisdiction from the group of ten, North Dakota, has two Bureau of Indian Affairs schools in its monitored sample (sample size 8) and only one in the unmonitored sample (sample size 6). These three schools all have mean proficiency scores below 200, while on average the other schools sampled in North Dakota have mean proficiency scores around 240. The difference in one poor-performing school might contribute to the significant difference in the unmonitored-monitored effect, given the respectively small sample sizes of 8 and 6 schools with monitored and unmonitored status. Arkansas, also from the group of ten, has only 4 monitored and 3 unmonitored schools in its sample. Two of the unmonitored schools are Catholic and the remaining nonpublic school has an extraordinarily high teacher/student ratio (1:8, while typically the ratio is 1:20). Moreover, while all unmonitored schools in the Arkansas sample have a minimum homework requirement, almost all monitored schools do not.

The two findings, namely, larger variability in low-scoring jurisdictions and small sample sizes in nonpublic-school sessions, might be sufficient to explain a large part of the discrepancy

Table G-5

Mean Proficiency Scores for Unmonitored and Monitored Sessions in Nonpublic Schools

Jurisdiction	Unmonitored	Monitored
AL	243.3 (11.5)	230.7 (7.1)
AR*	243.6 (5.4)	232.1 (4.4)
CT	229.4 (5.0)	226.7 (5.8)
DE	236.9 (2.9)	227.8 (10.5)
GA	235.7 (9.3)	232.6 (7.3)
GU*	221.5 (4.3)	210.0 (2.4)
ID	223.9 (6.2)	210.2 (34.1)
MO	243.3 (7.0)	235.3 (4.2)
MN	236.6 (4.1)	232.1 (3.4)
NJ	233.9 (5.3)	229.6 (6.5)
NM	192.0 (29.6)	178.6 (22.9)
ND*	231.3 (11.4)	208.9 (8.5)
RI	232.9 (4.0)	225.5 (7.1)
WV	241.5 (6.2)	233.1 (4.0)
CO	238.9 (4.4)	239.6 (6.3)
HI	234.8 (6.6)	235.4 (4.8)
IN*	228.5 (4.1)	243.2 (4.3)
IA	226.7 (7.4)	235.8 (4.1)
LA	223.6 (5.3)	230.5 (5.8)
ME	232.3 (8.0)	242.8 (8.6)
MA	234.6 (7.2)	241.6 (6.9)
PA	224.4 (8.5)	234.4 (4.8)
VA	233.7 (7.1)	246.9 (9.3)

* These jurisdictions show significant difference between unmonitored and monitored sessions.

in the number of jurisdictions being better or worse with respect to monitoring. However, we are not sure that this is the complete picture. For example, for the state of Delaware, which shows significance on both scales in DIFF, we did not find anything conspicuous. The characteristics of schools in the monitored sessions of Delaware match quite well with those in the unmonitored sessions. Note that Delaware does not show a significant monitoring effect when using mean proficiency scores.

There exist good statistical procedures to test whether the background variables for the monitored sessions and unmonitored sessions at the jurisdiction level are similar or not. For example, we may draw on the idea of propensity score (Rosenbaum & Rubin, 1983). Let x be the vector of covariates for a particular school and let the binary variable z indicate whether the school is monitored ($z = 1$) or unmonitored ($z = 0$). The propensity score $e(\mathrm{x})$ is defined as the conditional probability of being in the monitored session given the background variables. Specifically, $e(\mathrm{x}) = \mathrm{pr}(z = 1 \mid \mathrm{x})$. A logit transform $q(\mathrm{x}) = \log\{(1 - e(\mathrm{x}))/e(\mathrm{x})\}$ on $e(\mathrm{x})$ is often used. The logit transform $q(\mathrm{x})$ can be interpreted as the log odds against being monitored. Although propensity score applications usually arise in comparison studies where cohorts are matched so that imbalance due to background covariates in the treatment group and the control group can be adjusted, the same idea can be applied to the present situation. However, to evaluate the propensity scores $e(\mathrm{x})$, we require an estimation of the density $\mathrm{pr}(\mathrm{x} \mid z = 1)$ and $\mathrm{pr}(\mathrm{x} \mid z = 0)$. Given the small number of schools present in some jurisdictions and the fairly large number of background variables collected (about a dozen that we considered important), it is not easy to estimate such densities formally. Using simple descriptive statistics and directly examining the manageable amount of data seems to be a more sensible way in judging whether there is a difference between background variables of the monitored and unmonitored sessions. This is the approach we adopted in the preceding paragraphs.

G.6 CONCLUSION

There are two major facets to this appendix. One facet this appendix tries to address is the various statistical issues that arise from the complex structure of the NAEP data. To perform the testing of hypotheses on the effect of monitoring sessions, we propose a two-tier bootstrap method. The nonparametric testing of hypotheses provides an overall summary of the statistical analysis of data, making minimal assumptions about how the data are generated. Although no strong evidence is found against the null hypothesis that there is no effect, there are other nonstatistical issues that we need to consider. The second facet of this appendix looks at other nontechnical but decision-related issues such as specific school characteristics. We examine results on individual jurisdictions closely to gain a deeper understanding of the problem. Small sample sizes and diversity in school types are found to be factors that may lead to significant monitoring effect size. Jurisdictions at the lower end of the proficiency scale are found to exhibit larger variability. These are factors that need to be taken into consideration in future plans for administering and monitoring the assessment. Perhaps more sessions, especially those in nonpublic schools, should be monitored.

APPENDIX H

CORRECTION OF THE NAEP PROGRAM DOCUMENTATION ERROR

APPENDIX H

Correction of the NAEP Program Documentation Error

John Mazzeo and Nancy L. Allen

Educational Testing Service

In April 1995, results from the 1994 Trial State Assessment of reading were released as part of the report *1994 NAEP Reading: A First Look.* Subsequently, ETS/NAEP research scientists discovered an error in the documentation for the ETS version of the PARSCALE program, which is used to compute NAEP scale score results. The error affected how omitted responses were treated in the IRT scaling of the extended constructed-response items that received partial-credit scoring. The error affected only those items; omitted multiple-choice and omitted short constructed responses were treated appropriately.

The conventional treatment in NAEP subjects has been to treat omitted responses (blank responses to an item that are followed by valid responses to items that appear later in the test) as the lowest possible score category in the production of NAEP scale scores. In contrast, not-reached responses (blank responses that are not followed by any further student responses) are treated as missing data. As a result of the documentation error, for a number of the polytomous constructed-response items and across several subject areas, *all* blank responses (both omitted and not-reached responses) to affected items were treated as missing — an *acceptable* treatment but *not* the *conventional* option of choice for NAEP.

The error occurred because of a documentation error in the description of one of the PARSCALE control parameters, designated as POMIT. The program permits the analyst to choose two different ways of treating blank responses: (a) as missing data, and (b) as a valid response falling in the lowest score category. The documentation indicates that by setting POMIT= -1, the treatment in (a) occurs. By setting POMIT=0 or POMIT=1, the treatment in (b) is supposed to occur. The POMIT=1 setting is the program default. In reality, POMIT= 1 and POMIT= -1 operate equivalently, treating blank responses as missing data.

The error appears to have been introduced in 1992 when the programs BILOG and PARSCALE were merged to form the ETS version of PARSCALE. Verification of the accuracy of existing documentation, modifications to internal program diagnostics, and more systematic testing procedures for any and all changes to NAEP-related programs have been implemented immediately to reduce the likelihood of experiencing this kind of error in subsequent NAEP cycles.

The PARSCALE documentation error affected a number of the NAEP scales constructed since 1992. Specifically, the 1992 and 1994 national and state reading results were affected by the error. Results from these two assessments have been released to the public in a number of NAEP publications. The 1992 data are also available to the public through NCES's secondary-use data files.

It should be noted that this processing error also impacted the location of the National Assessment Governing Board (NAGB) achievement levels in reading, which were set on the 1992 scales.

NCES and ETS felt that the most technically correct plan of action would be to recalculate all affected NAEP scales, no matter how slight the change, and to issue revised results. ETS was therefore instructed by NCES to recalculate all affected scales and to work with American College Testing (ACT) in the recomputation of the achievement level cutpoints.

In recomputing the cutpoints, an additional error was discovered in the procedures used by ACT in 1992 to "map" the achievement level cutpoints onto the NAEP scale. As described in Appendix I, the procedures contained an incorrectly derived formula. ACT used revised procedures with the correct formula to map the achievement level cutpoints for the 1994 history and geography scales. However, the error in the earlier procedures did affect achievement level cutpoints for reading, which were established during the 1992 assessment. The 1992 national and state reading achievement level results were further impacted by this additional error.

A new version of the 1994 *First Look* report, containing the revised reading results, was issued by NCES in the fall of 1995. The main release of NAEP reading results, including the *Reading Report Card, Cross-state Data Compendium*, individual state reports, almanacs, technical report, and data files, originally scheduled for the end of September, took place instead in late fall.

The information documenting the original analysis of the 1992 data that appears in the *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading* is substantially correct for the revised 1992 analysis. The transformation constants for the revised analysis are provided in Table H-1. (The original values for k_1 and k_2 were 217.56 and 38.10 for reading for literary experience and 212.50 and 37.00 for reading to gain information.) The information in the other sections of the technical report for the 1994 state reading assessment describe the revised analysis of the 1994 Trial State Assessment data.

Table H-1
Transformation Constants for the 1992 Trial State Assessment in Reading

Scale	k_1	k_2
Reading for Literary Experience	216.02	36.80
Reading to Gain Information	210.95	37.40

While some *small* changes in scale score results were found, the revised numbers for reading are quite similar to the results released in 1992 and to those published in the NCES April release of the reading *First Look* report. More specifically, the revised reading results are *substantively equivalent* to the originally published 1992 results and to the results released in the *First Look*. Regarding the 1992 and 1994 national assessment data, fourth-grade results are about 1 point lower than originally reported, while twelfth-grade results are about 1 point higher. These changes are small and not substantively meaningful. The eighth-grade numbers are essentially unchanged. The revised numbers indicate the same relative distances between reporting subgroups (i.e., race/ethnicity subgroups, male, females, etc.). The significant national score decline at grade 12 is totally unaffected by the revision, as is the absence of significant changes at grades 4 and 8.

With regard to the state assessment data, all jurisdictions were affected to roughly the same degree. Thus, the revised rank ordering of state performance in both 1992 and 1994 is essentially identical to that originally published. Original and revised trend results (i.e., the change in scores between 1992 and 1994) are extremely close for all the jurisdictions. However, in four instances (for Massachusetts, New Jersey, Utah, and California), the small changes engendered by the revision are sufficient to affect the statistical significance of the change. For Massachusetts, New Jersey, and Utah, the revised decline in scores is between 0.3 and 0.5 points smaller than the originally released results — a magnitude of change that was typical across *all* participants. When rounded to an integer, the original and revised declines for these three states are of identical size. Despite this similarity, the revised results for these states are no longer statistically significant since the original results were right on the margin of statistical significance. In California, the revised decline in scores is 0.4 points larger than the originally released results and is now statistically significant.

Tables H-2 and H-3 more fully document the effect of the ETS program documentation error on the NAEP scale scores for each jurisdiction. Table H-2 contains the means, standard deviations, and percentiles for each jurisdiction before revisions were made. This information was reported in the first version of the report *1994 NAEP Reading: A First Look* released in April 1995. Table H-3 contains the same information after revisions were made. This information was reported in the revised *First Look* report, the *Reading Report Card*, and the individual state reports.

In the results for state assessment achievement levels, there is little difference in the revised and original numbers from an interpretive standpoint. As expected, correction of the ACT error generally results in lower achievement level cutpoints and, hence, slightly higher percentages above the various cutpoints. The revised achievement level results in this technical report and in the reading reports reflect the change in the formula used in setting the achievement levels.

There is one notable aspect of the revised state assessment achievement level results. Prior to the revision, only one state, Arizona, had shown a statistically significant increase from 1992 to 1994 in the percentage of students at the Advanced level. Based on the revised results, six more states — Connecticut, Florida, Kentucky, Maine, Mississippi, and Maryland — also showed a statistically significant increase at that level.

Tables H-4 and H-5 contain information about the effect of the ETS program documentation error and the incorrectly derived "mapping" formula on the achievement level results for each jurisdiction. Table H-4 contains the percentages of students at or above each achievement level and the percentage of student below the Basic level for each jurisdiction before revisions were made. These results were reported in the April 1995 version of *1994 NAEP Reading: A First Look*. Table H-5 contains the percentages for each jurisdiction after revisions were made. These results were reported in the revised *First Look* report, in the *Reading Report Card*, and the individual state reports.

Table H-2
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Weighted Means, Standard Deviations, and Percentiles
Original Results

		MEAN	STD DEV	10TH	25TH	50TH	75TH	90TH
PUBLIC SCHOOLS								
Nation	1994	213.3(1.1)	40.2(0.6)	157.9(2.4)<	188.3(1.6)	217.6(1.2)	241.9(1.2)	261.4(1.4)
	1992	215.9(1.1)	36.5(0.7)	167.6(1.7)	192.6(1.1)	218.3(1.4)	241.3(1.4)	261.2(1.9)
Alabama	1994	209.0(1.5)	38.9(1.1)	157.2(2.7)	183.7(1.7)	211.4(1.6)	236.8(2.1)	256.8(1.7)
	1992	208.3(1.7)	36.3(0.8)	160.2(2.1)	184.6(2.5)	210.3(1.7)	233.8(2.0)	253.5(1.5)
Arizona	1994	207.3(1.8)	42.5(1.1)	150.0(2.6)<	180.4(2.5)	210.8(2.0)	237.5(1.9)	259.1(2.3)
	1992	210.4(1.3)	35.0(0.8)	164.1(2.3)	187.2(2.0)	212.8(1.5)	235.5(1.1)	253.7(1.5)
Arkansas	1994	209.7(1.7)	38.6(0.9)	157.5(2.4)	185.3(2.7)	213.0(1.8)	237.5(1.4)	256.5(1.7)
	1992	211.9(1.2)	35.9(0.6)	164.4(1.9)	188.1(1.2)	214.4(1.5)	237.2(1.0)	256.2(1.6)
California	1994	198.0(1.8)	43.3(1.1)	139.0(2.3)	169.4(3.2)	202.4(2.4)	229.5(1.5)	250.2(1.5)
	1992	203.1(2.1)	41.8(1.0)	147.2(2.8)	176.2(2.7)	206.3(2.1)	232.9(3.0)	254.4(3.4)
Colorado	1994	214.5(1.3)	38.3(0.9)	163.4(2.8)<	191.6(1.5)<	218.1(1.1)	241.5(1.3)	259.9(1.4)
	1992	217.7(1.2)	32.3(0.7)	175.0(2.5)	197.9(1.3)	220.0(1.4)	240.0(1.4)	257.0(1.0)
Connecticut	1994	223.3(1.6)	39.3(1.4)	171.5(4.6)	200.5(2.1)	228.1(1.4)	250.7(1.5)	269.2(1.5)
	1992	222.9(1.3)	34.2(1.1)	176.8(2.9)	201.7(2.8)	226.2(1.1)	247.1(1.5)	263.8(1.9)
Delaware	1994	207.5(1.1)<	41.0(0.7)	152.4(2.7)<	182.4(1.5)<	211.5(1.2)	236.1(0.9)	256.9(1.6)
	1992	214.0(0.7)	35.5(0.7)	166.8(1.8)	190.4(1.6)	215.1(1.2)	238.9(0.9)	258.9(1.2)
Florida	1994	206.0(1.7)	41.5(0.9)	150.2(2.3)<	179.4(2.0)	209.4(1.8)	235.7(1.7)	256.9(2.1)
	1992	209.3(1.3)	36.4(0.9)	161.4(2.9)	185.6(1.5)	211.4(1.6)	235.6(1.4)	254.0(1.4)
Georgia	1994	208.1(2.4)	43.7(1.5)	149.6(3.7)<	180.3(2.4)	211.7(2.4)	239.2(2.4)	261.3(2.6)
	1992	213.4(1.5)	36.9(0.7)	164.0(1.8)	188.8(2.7)	215.1(1.7)	239.4(1.7)	259.3(2.3)
Hawaii	1994	201.9(1.5)	41.6(0.8)	145.5(2.6)<	175.2(1.8)	205.0(1.9)	231.4(1.6)	252.8(1.7)
	1992	204.3(1.7)	37.4(1.0)	154.3(1.7)	180.2(2.3)	206.7(1.7)	230.7(1.9)	250.0(1.7)
Idaho	1994	213.8(1.4)<	37.5(1.0)	164.4(3.0)<	191.4(2.4)<	217.0(1.8)	239.8(1.4)	258.8(1.5)
	1992	220.5(1.0)	31.3(0.8)	180.3(1.9)	200.7(1.2)	222.1(1.1)	242.1(1.1)	259.0(1.5)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-2 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Weighted Means, Standard Deviations, and Percentiles
Original Results

		MEAN	STD DEV	10TH	25TH	50TH	75TH	90TH
PUBLIC SCHOOLS								
Indiana	1994	220.7(1.3)	35.1(0.8)	174.4(1.7)	198.7(1.8)	223.6(1.4)	245.2(2.4)	262.9(1.4)
	1992	222.4(1.3)	31.7(0.7)	180.6(2.4)	202.4(2.3)	223.8(1.2)	244.4(1.2)	261.8(1.7)
Iowa	1994	223.7(1.3)	34.0(0.8)	179.0(2.4)	202.7(1.8)	226.2(1.6)	247.2(1.6)	264.7(2.4)
	1992	226.8(1.1)	31.2(0.7)	185.1(1.3)	207.0(1.1)	229.0(1.0)	248.6(1.1)	265.1(1.1)
Kentucky	1994	212.6(1.6)	37.8(0.9)	162.8(2.8)	188.4(1.8)	214.6(1.4)	239.0(1.7)	259.0(1.9)
	1992	213.6(1.3)	34.1(0.6)	167.8(3.4)	192.1(1.7)	215.9(1.5)	237.6(1.8)	255.2(1.4)
Louisiana	1994	198.0(1.3)<	38.2(0.9)	147.8(2.7)<	172.6(1.9)<	199.4(1.9)	224.9(1.9)	246.6(2.7)
	1992	204.6(1.2)	33.7(0.8)	160.7(2.0)	181.7(2.3)	205.3(1.7)	228.1(1.4)	247.2(1.5)
Maine	1994	229.1(1.3)	31.6(0.8)	187.0(1.7)	209.4(1.9)	231.4(1.1)	251.0(1.2)	267.9(1.6)
	1992	228.2(1.1)	28.7(0.6)	190.9(2.7)	209.3(1.3)	229.3(1.6)	248.3(1.1)	263.7(1.7)
Maryland	1994	210.7(1.4)	42.2(1.2)	157.1(2.0)	185.2(1.5)	214.6(1.9)	239.8(1.7)	260.3(1.3)
	1992	212.1(1.6)	37.4(1.3)	161.7(3.1)	188.7(2.4)	215.4(1.4)	238.6(1.5)	257.3(1.3)
Massachusetts	1994	224.0(1.3)<	33.6(0.8)	178.7(1.9)<	203.0(2.0)	227.0(1.4)	247.9(1.5)	264.7(1.9)
	1992	227.4(1.0)	30.3(0.5)	188.3(1.5)	208.2(1.2)	229.0(1.4)	248.5(1.0)	264.7(2.0)
Michigan	1994	213.2(2.0)	36.7(1.3)	165.3(2.7)	190.6(3.1)	215.9(2.9)	238.8(2.0)	257.2(2.2)
	1992	217.2(1.6)	33.1(0.8)	172.7(2.9)	195.5(1.6)	219.7(1.3)	240.8(1.8)	257.7(1.6)
Minnesota	1994	219.2(1.3)	37.6(1.2)	169.1(2.9)<	196.9(1.3)	223.9(1.1)	245.2(1.2)	263.3(1.5)
	1992	222.1(1.2)	32.8(0.7)	179.4(2.4)	201.1(1.4)	224.4(1.6)	245.3(0.8)	262.3(1.3)
Mississippi	1994	202.7(1.6)	38.8(0.9)	151.6(3.0)	176.5(2.9)	204.5(1.7)	229.7(1.6)	251.7(1.7)
	1992	199.8(1.3)	36.1(0.8)	152.8(2.9)	175.9(1.8)	201.0(1.8)	225.2(1.6)	246.2(1.5)
Missouri	1994	217.8(1.5)	37.5(0.9)	168.7(2.5)<	194.4(2.0)	221.1(1.9)	244.2(1.2)	263.2(1.9)
	1992	221.3(1.3)	32.6(0.8)	177.7(2.2)	200.4(1.3)	223.2(1.4)	244.1(1.4)	261.4(1.1)
Montana	1994	223.1(1.4)	33.6(0.6)	179.5(2.9)	203.0(1.8)	226.2(1.6)	246.6(1.1)	263.4(1.1)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-2 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Weighted Means, Standard Deviations, and Percentiles
Original Results

		MEAN	STD DEV	10TH	25TH	50TH	75TH	90TH
PUBLIC SCHOOLS								
Nebraska	1994	220.9(1.4)	37.4(1.0)	172.1(1.6)<	198.6(2.1)	224.6(1.8)	247.3(1.7)	264.9(1.2)
	1992	222.4(1.1)	32.4(0.9)	180.3(1.9)	202.3(1.6)	224.4(1.5)	244.6(1.4)	261.6(1.1)
New Hampshire	1994	224.3(1.5)<	34.2(0.7)	179.3(3.2)<	203.8(1.5)	227.3(1.5)	247.7(1.5)	265.3(1.4)
	1992	229.1(1.2)	30.6(0.7)	190.0(1.8)	210.0(2.2)	230.7(1.5)	249.7(1.6)	266.4(2.0)
New Jersey	1994	220.3(1.2)<	37.2(0.8)	170.5(2.0)	197.4(1.5)	223.6(1.2)	246.8(1.7)	264.9(1.5)
	1992	224.3(1.5)	34.0(0.8)	179.0(3.5)	202.5(1.6)	226.8(1.8)	248.5(1.6)	266.1(1.8)
New Mexico	1994	205.7(1.7)<	39.6(1.2)	152.6(4.6)	180.2(2.4)<	208.4(1.5)	233.4(1.2)	254.6(2.3)
	1992	211.8(1.5)	35.7(1.2)	165.7(3.1)	188.5(1.6)	213.2(1.2)	237.0(2.3)	255.8(2.4)
New York	1994	212.7(1.4)	39.4(0.9)	158.3(4.4)	188.1(2.3)	216.2(1.5)	241.0(2.1)	260.2(2.2)
	1992	215.8(1.4)	36.9(1.5)	167.1(2.6)	194.2(2.4)	219.4(1.6)	241.2(1.6)	259.5(1.4)
North Carolina	1994	215.3(1.5)	39.6(0.8)	163.4(1.5)	189.8(2.0)	218.1(1.5)	243.6(1.3)	263.3(1.4)
	1992	212.6(1.2)	37.7(0.6)	162.6(1.6)	187.8(1.3)	214.6(1.5)	239.2(1.3)	259.6(1.8)
North Dakota	1994	226.0(1.2)	32.8(1.0)	183.0(2.0)	206.2(1.8)	228.6(1.3)	248.7(1.2)	265.0(1.6)
	1992	226.9(1.2)	30.3(0.9)	188.2(3.2)	208.1(2.2)	228.7(1.7)	247.5(1.5)	263.0(1.9)
Pennsylvania	1994	216.2(1.5)<	38.7(1.0)	165.1(2.9)<	193.2(1.7)<	220.4(2.1)	243.8(1.6)	262.1(2.0)
	1992	221.9(1.3)	33.8(0.6)	177.3(3.2)	200.7(1.6)	224.5(1.6)	245.6(1.9)	263.2(1.6)
Rhode Island	1994	220.8(1.3)	35.1(0.6)	174.5(1.5)	198.8(1.7)	223.4(1.4)	245.4(2.1)	264.0(1.8)
	1992	217.8(1.8)	34.7(1.3)	171.8(3.9)	195.9(4.2)	220.3(2.1)	241.9(1.8)	260.0(1.6)
South Carolina	1994	204.6(1.4)<	38.8(0.8)	153.7(1.6)<	178.6(1.7)<	206.7(1.4)	232.3(1.3)	253.2(1.3)
	1992	210.7(1.3)	35.4(0.6)	164.6(1.6)	187.6(1.4)	211.5(1.7)	235.7(1.6)	255.6(1.9)
Tennessee	1994	213.5(1.7)	37.8(1.0)	163.1(4.3)	189.7(2.6)	216.4(2.2)	240.1(1.2)	259.1(2.4)
	1992	213.2(1.5)	34.4(0.6)	168.6(1.4)	190.5(2.0)	215.0(1.7)	237.2(1.9)	256.2(2.0)
Texas	1994	213.4(1.8)	37.8(1.1)	163.2(2.3)	190.2(3.5)	216.2(2.2)	239.4(1.6)	259.7(2.0)
	1992	213.6(1.6)	35.0(1.0)	168.4(1.9)	190.8(1.4)	215.0(1.8)	238.1(1.9)	257.1(2.5)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-2 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Weighted Means, Standard Deviations, and Percentiles
Original Results

		MEAN	STD DEV	10TH	25TH	50TH	75TH	90TH
PUBLIC SCHOOLS								
Utah	1994	218.2(1.2)<	36.4(0.8)	171.4(2.2)<	197.1(1.2)	221.9(1.1)	243.2(1.4)	260.4(1.3)
	1992	221.6(1.2)	31.5(0.8)	180.0(2.0)	201.2(1.5)	223.9(1.4)	243.9(0.9)	260.1(1.2)
Virginia	1994	214.5(1.4)<	37.0(0.7)	165.3(3.5)<	190.2(1.7)<	216.0(1.8)<	240.3(2.6)	261.4(2.0)
	1992	222.0(1.4)	33.7(0.8)	177.2(2.5)	200.3(1.7)	223.8(1.6)	245.5(1.2)	263.0(2.0)
Washington	1994	213.7(1.4)	38.1(1.0)	162.7(2.7)	190.6(1.8)	217.3(1.3)	240.2(1.1)	259.5(2.2)
West Virginia	1994	214.1(1.1)	36.4(0.6)	166.1(2.4)	191.1(1.4)	216.3(1.0)	239.7(1.6)	258.9(1.1)
	1992	216.6(1.3)	33.9(0.7)	172.1(2.6)	195.2(1.6)	218.3(1.5)	239.7(1.2)	258.5(1.7)
Wisconsin	1994	225.0(1.1)	31.3(0.6)	184.3(1.9)	205.0(1.2)	227.1(1.6)	246.7(1.7)	262.9(1.6)
	1992	224.9(1.0)	30.8(0.7)	184.6(2.1)	204.4(1.3)	226.6(1.1)	246.4(0.9)	263.1(0.8)
Wyoming	1994	222.0(1.2)	31.3(0.6)	180.1(1.9)	201.8(2.0)	224.4(1.6)	244.1(0.9)	260.2(1.6)
	1992	224.3(1.2)	31.1(0.6)	183.0(2.0)	205.2(1.9)	226.6(1.3)	245.8(1.5)	262.2(1.1)
DoDEA Overseas	1994	218.5(0.9)	33.0(0.6)	175.8(1.3)	197.3(1.3)	220.4(1.1)	241.4(1.3)	259.4(1.4)
Guam	1994	182.5(1.2)	41.5(1.0)	128.2(1.5)	155.5(1.7)	184.8(1.6)	211.7(0.9)	234.3(1.1)
	1992	182.7(1.4)	41.9(0.9)	127.0(2.2)	155.0(2.3)	185.6(1.5)	212.5(1.4)	234.8(1.6)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-3
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Weighted Means, Standard Deviations, and Percen'iles
Revised Results

		MEAN	STD DEV	10TH	25TH	50TH	75TH	90TH
PUBLIC SCHOOLS								
Nation	1994	212.3(1.1)	40.9(0.6)	155.9(2.1)<	186.9(1.5)<	216.7(1.2)	241.4(1.2)	261.2(1.5)
	1992	214.8(1.0)	35.6(0.6)	167.6(1.9)	192.0(1.0)	217.2(1.7)	239.7(1.3)	259.1(2.3)
Alabama	1994	208.0(1.5)	39.6(1.1)	155.2(2.8)	182.2(1.8)	210.4(2.3)	236.3(2.3)	256.6(1.8)
	1992	207.4(1.7)	35.4(0.8)	160.5(1.8)	184.1(2.5)	209.3(1.5)	232.3(1.8)	251.5(1.2)
Arizona	1994	206.2(1.9)	43.2(1.1)	147.9(2.5)<	178.8(2.7)	209.8(2.1)	236.9(1.8)	259.0(2.0)>
	1992	209.4(1.2)	34.2(0.8)	164.2(2.0)	186.7(2.0)	211.7(1.4)	233.9(1.1)	251.7(1.5)
Arkansas	1994	208.6(1.7)	39.3(1.0)	155.5(2.8)<	183.8(2.8)	211.9(1.9)	236.9(1.2)	256.3(1.6)
	1992	210.9(1.2)	35.0(0.6)	164.6(2.0)	187.7(1.4)	213.4(1.5)	235.5(1.1)	254.2(1.6)
California	1994	196.8(1.8)<	44.1(1.1)	136.6(2.7)<	167.5(3.4)	201.2(2.2)	228.8(1.5)	249.9(1.5)
	1992	202.2(2.0)	40.9(1.0)	147.7(3.1)	176.0(2.8)	205.5(2.1)	231.4(2.4)	252.4(2.6)
Colorado	1994	213.5(1.3)	39.0(1.0)	161.5(3.2)<	190.2(1.5)<	217.2(1.4)	241.0(1.5)	259.8(1.3)>
	1992	216.5(1.1)	31.6(0.7)	174.7(2.1)	197.1(1.6)	218.8(1.2)	238.3(1.3)	254.9(1.0)
Connecticut	1994	222.4(1.6)	40.0(1.4)	169.6(4.2)	199.1(2.6)	227.2(1.3)	250.3(1.6)	269.2(1.5)>
	1992	221.6(1.3)	33.4(1.1)	176.6(3.0)	200.9(2.4)	224.9(1.2)	245.3(1.2)	261.7(1.7)
Delaware	1994	206.4(1.1)<	41.7(0.7)	150.5(3.0)<	180.9(2.0)<	210.5(1.4)	235.4(0.9)	256.6(1.2)
	1992	212.9(0.6)	34.7(0.7)	166.8(2.2)	189.8(1.2)	214.0(1.2)	237.2(1.0)	256.9(1.4)
Florida	1994	204.9(1.7)	42.2(1.0)	148.1(2.2)<	177.8(2.0)<	208.3(1.6)	235.1(2.1)	256.6(2.0)
	1992	208.3(1.2)	35.6(0.9)	161.4(3.5)	185.1(1.6)	210.4(1.4)	234.1(1.3)	252.0(1.6)
Georgia	1994	206.8(2.4)	44.5(1.5)	148.0(3.9)<	178.5(2.6)<	210.3(3.1)	238.9(2.0)	260.2(2.3)
	1992	212.3(1.5)	36.1(0.7)	164.1(2.1)	188.3(2.3)	214.0(1.5)	237.8(1.7)	257.2(2.1)
Hawaii	1994	200.9(1.7)	42.5(1.0)	143.6(3.4)<	173.5(2.0)	204.2(2.1)	230.9(1.6)	252.8(1.5)
	1992	203.4(1.7)	36.6(0.9)	154.6(1.7)	180.0(2.5)	205.8(1.3)	229.3(1.9)	248.2(1.8)
Idaho	1994	212.8(1.4)<	38.2(1.0)	162.6(3.4)<	189.9(1.7)<	216.0(1.7)	239.2(1.7)	258.7(1.6)
	1992	219.3(0.9)	30.5(0.8)	180.1(1.8)	199.9(1.1)	220.9(1.0)	240.3(1.1)	256.9(1.5)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-3 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Weighted Means, Standard Deviations, and Percentiles
Revised Results

		MEAN	STD DEV	10TH	25TH	50TH	75TH	90TH
PUBLIC SCHOOLS								
Indiana	1994	219.8(1.3)	35.7(0.8)	172.6(1.5)<	197.4(1.9)	222.7(1.4)	244.8(1.8)	262.8(1.5)
	1992	221.1(1.3)	31.0(0.7)	180.3(2.4)	201.5(1.7)	222.6(1.2)	242.6(1.1)	259.7(1.5)
Iowa	1994	222.9(1.3)	34.6(0.8)	177.4(2.7)	201.5(1.6)	225.4(1.7)	246.8(1.7)	264.6(2.2)
	1992	225.5(1.1)	30.5(0.7)	184.7(1.6)	206.1(1.2)	227.6(0.9)	246.8(1.3)	262.9(1.1)
Kentucky	1994	211.6(1.6)	38.5(0.9)	160.8(3.2)	186.9(1.9)	213.6(1.4)	238.4(1.9)	258.8(2.2)
	1992	212.5(1.3)	33.3(0.6)	167.8(3.2)	191.5(2.0)	214.8(1.4)	235.9(1.7)	253.1(1.5)
Louisiana	1994	196.8(1.3)<	38.8(1.0)	145.7(2.8)<	170.8(1.9)<	198.1(1.8)<	224.1(1.6)	246.2(2.6)
	1992	203.7(1.2)	32.9(0.8)	160.9(2.5)	181.3(2.2)	204.4(1.2)	226.6(1.3)	245.4(1.5)
Maine	1994	228.4(1.3)	32.2(0.8)	185.5(2.2)	208.3(1.7)	230.6(1.1)	250.7(1.4)	267.9(1.4)>
	1992	226.8(1.1)	28.1(0.6)	190.3(2.1)	208.3(1.1)	228.0(1.1)	246.5(1.1)	261.5(1.9)
Maryland	1994	209.6(1.5)	42.9(1.3)	155.1(1.8)	183.8(1.8)	213.7(1.9)	239.3(1.3)	260.1(1.5)
	1992	211.0(1.6)	36.6(1.2)	161.9(2.8)	188.2(2.7)	214.2(1.4)	237.0(1.4)	255.1(1.3)
Massachusetts	1994	223.2(1.3)	34.2(0.8)	177.1(2.3)<	201.8(3.0)	226.2(1.6)	247.5(1.6)	264.7(2.1)
	1992	226.0(0.9)	29.6(0.5)	187.9(1.7)	207.2(1.6)	227.6(1.3)	246.7(0.9)	262.5(2.2)
Michigan	1994	212.2(2.0)	37.4(1.4)	163.3(3.2)	189.1(2.9)	214.9(2.5)	238.1(1.9)	257.1(2.0)
	1992	216.0(1.5)	32.4(0.7)	172.6(2.3)	194.8(1.8)	218.5(1.4)	239.2(2.0)	255.6(1.6)
Minnesota	1994	218.3(1.4)	38.2(1.2)	167.3(3.1)<	195.5(1.2)<	223.0(1.3)	244.8(1.3)	263.3(1.1)
	1992	220.9(1.2)	32.1(0.7)	179.1(2.0)	200.3(1.4)	223.1(1.6)	243.5(0.8)	260.1(1.0)
Mississippi	1994	201.5(1.6)	39.5(0.9)	149.5(3.5)	174.9(2.3)	203.4(1.8)	229.0(2.3)	251.4(1.7)>
	1992	199.1(1.3)	35.3(0.8)	153.1(2.1)	175.8(1.6)	200.3(1.6)	223.9(1.6)	244.4(1.5)
Missouri	1994	216.9(1.5)	38.2(0.9)	166.9(2.5)<	193.1(2.3)	220.2(1.9)	243.8(1.0)	263.1(1.8)
	1992	220.0(1.2)	31.8(0.8)	177.5(1.9)	199.7(1.3)	221.8(1.7)	242.3(1.7)	259.3(1.2)
Montana	1994	222.3(1.4)	34.2(0.7)	177.9(3.3)	201.8(1.9)	225.4(1.4)	246.2(1.0)	263.3(1.1)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-3 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Weighted Means, Standard Deviations, and Percentiles
Revised Results

		MEAN	STD DEV	10TH	25TH	50TH	75TH	90TH
PUBLIC SCHOOLS								
Nebraska	1994	220.0(1.5)	38.1(1.0)	170.2(1.6)<	197.3(1.9)	223.8(1.8)	246.9(1.9)	264.9(1.3)>
	1992	221.2(1.1)	31.7(0.9)	179.9(1.7)	201.5(1.6)	223.0(1.1)	242.9(1.5)	259.4(1.0)
New Hampshire	1994	223.5(1.5)<	34.9(0.7)	177.6(3.0)<	202.6(1.5)	226.6(1.5)	247.3(1.5)	265.3(1.3)
	1992	227.7(1.2)	29.9(0.7)	189.4(1.9)	208.9(2.1)	229.2(1.3;	247.9(1.2)	264.3(1.9)
New Jersey	1994	219.4(1.2)	37.9(0.9)	168.8(2.2)<	196.1(1.6)	222.8(1.0)	246.3(1.4)	264.9(1.9)
	1992	223.0(1.4)	33.3(0.8)	178.8(3.1)	201.8(1.7)	225.4(1.7)	246.7(1.6)	264.0(1.7)
New Mexico	1994	204.5(1.7)<	40.2(1.2)	150.6(4.3)<	178.7(2.5)<	207.3(1.6)	232.7(1.0)	254.4(2.6)
	1992	210.7(1.5)	34.9(1.2)	165.7(2.8)	188.0(1.4)	212.1(1.3)	235.4(2.2)	253.8(2.1)
New York	1994	211.7(1.4)	40.1(0.9)	156.5(4.2)	186.6(1.9)	215.2(1.3)	240.4(1.8)	260.0(2.0)
	1992	214.6(1.4)	36.1(1.5)	167.1(2.5)	193.6(2.4)	218.2(1.3)	239.6(1.1)	257.5(1.5)
North Carolina	1994	214.4(1.5)	40.2(0.8)	161.6(2.2)	188.3(1.9)	217.2(1.5)	243.2(1.3)>	263.1(1.7)
	1992	211.5(1.1)	36.8(0.6)	162.8(2.1)	187.4(1.2)	213.6(1.4)	237.7(1.6)	257.5(2.1)
North Dakota	1994	225.2(1.2)	33.4(1.0)	181.3(1.7)	205.0(1.4)	227.9(1.1)	248.4(1.2)	265.0(1.6)
	1992	225.5(1.1)	29.6(0.9)	187.7(2.7)	207.1(2.0)	227.3(1.6)	245.7(1.3)	260.9(2.0)
Pennsylvania	1994	215.3(1.6)<	39.5(1.1)	163.1(3.4)<	191.8(1.8)<	219.5(2.1)	243.4(1.6)	262.1(2.1)
	1992	220.7(1.3)	33.1(0.6)	177.0(2.2)	199.8(1.4)	223.2(1.7)	243.8(1.7)	261.0(1.3)
Rhode Island	1994	219.9(1.3)	35.8(0.6)	172.7(1.8)	197.6(1.6)	222.5(1.4)	244.9(2.5)	263.9(2.8)
	1992	216.6(1.8)	34.0(1.3)	171.7(3.9)	195.3(4.0)	219.1(1.8)	240.3(1.7)	257.9(1.6)
South Carolina	1994	203.4(1.4)<	39.5(0.8)	151.6(1.8)<	176.9(1.9)<	205.6(1.5)	231.6(1.2)	253.0(1.4)
	1992	209.7(1.3)	34.6(0.5)	164.6(1.8)	187.0(1.7)	210.4(1.5)	234.2(1.3)	253.5(2.0)
Tennessee	1994	212.5(1.7)	38.5(1.1)	161.2(5.0)	188.2(2.4)	215.4(2.0)	239.6(1.3)	258.9(2.3)
	1992	212.1(1.4)	33.6(0.6)	168.5(1.3)	189.9(1.9)	213.9(1.4)	235.7(1.9)	254.2(1.5)
Texas	1994	212.4(1.9)	38.5(1.1)	161.2(2.7)	188.8(3.5)	215.2(2.3)	238.9(1.8)	259.5(1.7)
	1992	212.5(1.6)	34.2(1.0)	168.4(2.5)	190.2(1.8)	213.9(1.7)	236.5(2.2)	255.3(2.0)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-3 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Weighted Means, Standard Deviations, and Percentiles
Revised Results

		MEAN	STD DEV	10TH	25TH	50TH	75TH	90TH
PUBLIC SCHOOLS								
Utah	1994	217.3(1.3)	37.1(0.8)	169.6(1.9)<	195.8(1.4)	221.1(1.0)	242.8(1.3)	260.3(1.5)
	1992	220.4(1.1)	30.8(0.8)	179.7(2.2)	200.4(1.7)	222.5(1.2)	242.1(0.9)	258.1(1.2)
Virginia	1994	213.5(1.5)<	37.7(0.7)	163.4(3.6)<	188.7(2.0)<	215.0(2.0)<	239.8(2.8)	261.3(2.2)
	1992	220.8(1.4)	32.9(0.8)	177.1(2.6)	199.6(1.8)	222.5(1.6)	243.7(1.3)	260.8(1.9)
Washington	1994	212.7(1.5)	38.8(1.0)	160.8(2.7)	189.2(1.8)	216.4(1.3)	239.7(1.2)	259.4(2.3)
West Virginia	1994	213.1(1.1)	37.1(0.6)	164.2(2.0)<	189.7(1.2)	215.3(1.0)	239.3(1.5)	258.7(1.4)
	1992	215.5(1.3)	33.2(0.7)	172.0(2.2)	194.5(1.7)	217.1(1.5)	238.1(1.3)	256.5(1.6)
Wisconsin	1994	224.2(1.1)	31.9(0.6)	182.8(2.0)	203.8(1.4)	226.4(1.4)	246.2(1.7)	262.8(1.3)
	1992	223.6(1.0)	30.1(0.6)	184.2(2.0)	203.6(1.2)	225.2(1.1)	244.6(0.8)	260.9(0.9)
Wyoming	1994	221.2(1.2)	31.9(0.6)	178.6(2.0)	200.6(2.1)	223.6(1.6)	243.8(1.2)	260.2(1.5)
	1992	223.0(1.1)	30.4(0.6)	182.6(2.5)	204.2(1.5)	225.2(1.0)	244.0(1.5)	260.1(0.9)
DoDEA Overseas	1994	217.6(0.9)	33.6(0.6)	174.2(1.3)	195.9(1.4)	219.5(0.9)	240.9(1.2)	259.3(1.7)
Guam	1994	181.0(1.2)	42.2(1.0)	125.8(1.3)	153.5(1.8)	183.3(1.6)	210.6(1.4)	233.6(1.2)
	1992	182.3(1.4)	40.9(0.9)	128.0(2.4)	155.3(3.0)	185.1(1.6)	211.4(1.3)	233.1(2.0)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-4
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Percent of Students by Achievement Levels
Original Results

		N	WEIGHTED PCT [CV]	MEAN	ADVANCED	PRFCIENT	BASIC	< BASIC
PUBLIC SCHOOLS								
Nation	1994	6030	100.0(0.0) [1%]	213.3(1.1)	4.1(0.5)	24.0(1.1)	55.8(1.2)	44.2(1.2)
	1992	5045	100.0(0.0) [1%]	215.9(1.1)	4.0(0.6)	23.5(1.2)	56.9(1.2)	43.1(1.2)
Alabama	1994	2646	93.2(1.2) [2%]	209.0(1.5)	3.2(0.5)	19.5(1.3)	49.4(1.6)	50.6(1.6)
	1992	2571	100.0(0.0) [4%]	208.3(1.7)	2.2(0.4)	17.0(1.3)	47.9(2.1)	52.1(2.1)
Arizona	1994	2651	100.0(0.0) [3%]	207.3(1.8)	3.9(0.6)>	20.6(1.4)	48.9(1.8)	51.1(1.8)
	1992	2677	100.0(0.0) [2%]	210.4(1.3)	2.1(0.4)	17.7(1.1)	51.0(1.7)	49.0(1.7)
Arkansas	1994	2535	94.7(0.9) [3%]	209.7(1.7)	2.7(0.6)	19.9(1.4)	51.0(1.9)	49.0(1.9)
	1992	2589	100.0(0.0) [4%]	211.9(1.2)	2.6(0.4)	19.6(1.3)	52.5(1.6)	47.5(1.6)
California	1994	2252	91.1(1.1) [3%]	198.0(1.8)	2.1(0.4)	14.3(1.1)	41.4(2.0)	58.6(2.0)
	1992	2365	100.0(0.0) [3%]	203.1(2.1)	2.7(0.5)	17.1(1.6)	44.8(2.3)	55.2(2.3)
Colorado	1994	2730	94.2(1.5) [3%]	214.5(1.3)	3.7(0.5)	23.5(1.4)	56.4(1.6)	43.6(1.6)
	1992	2897	10^.0(0.0) [3%]	217.7(1.2)	2.7(0.4)	21.7(1.4)	60.3(1.6)	39.7(1.6)
Connecticut	1994	2578	89.1(1.0) [2%]	223.3(1.6)	6.7(0.8)	33.2(1.7)	65.6(1.7)	34.4(1.7)
	1992	2514	100.0(0.0) [3%]	222.9(1.3)	4.6(0.9)	29.6(1.4)	65.7(1.9)	34.3(1.9)
Delaware	1994	2239	82.2(1.4) [0%]	207.5(1.1)<	3.1(0.5)	19.4(1.0)	49.6(1.3)	50.4(1.3)
	1992	2048	100.0(0.0) [0%]	214.0(0.7)	3.3(0.4)	21.3(1.3)	53.9(1.3)	46.1(1.3)
Florida	1994	2666	89.8(0.9) [2%]	206.0(1.7)	3.2(0.5)	19.1(1.4)	47.4(1.8)	52.6(1.8)
	1992	2767	100.0(0.0) [3%]	209.3(1.3)	2.2(0.4)	18.1(1.1)	49.3(1.6)	50.7(1.6)
Georgia	1994	2765	93.2(0.9) [4%]	208.1(2.4)	4.5(0.9)	22.0(2.0)	49.8(2.4)	50.2(2.4)
	1992	2712	100.0(0.0) [3%]	213.4(1.5)	3.7(0.5)	22.1(1.5)	53.4(1.8)	46.6(1.8)
Hawaii	1994	2732	87.5(0.9) [2%]	201.9(1.5)	2.5(0.4)	15.7(1.1)	43.6(1.6)	56.4(1.6)
	1992	2642	100.0(0.0) [2%]	204.3(1.7)	1.7(0.3)	14.5(1.4)	44.4(2.0)	55.6(2.0)
Idaho	1994	2598	96.3(0.4) [.4]	213.8(1.4)<	3.2(0.5)	21.8(1.3)	55.3(1.6)<	44.7(1.6)>
	1992	2674	100.0(0.0) [3%]	220.5(1.0)	3.0(0.5)	23.9(1.3)	63.0(1.3)	37.0(1.3)

> INDICATES SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-4 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Percent of Students by Achievement Levels
Original Results

		N	WEIGHTED PCT [CV]	MEAN	ADVANCEO	PRFCIENT	BASIC	< BASIC
PUBLIC SCHOOLS								
Indiana	1994	2655	92.7(1.0) [2%]	220.7(1.3)	4.4(0.6)	27.4(1.4)	62.8(1.8)	37.2(1.8)
	1992	2535	100.0(0.0) [3%]	222.4(1.3)	3.9(0.7)	26.7(1.4)	64.3(1.7)	35.7(1.7)
Iowa	1994	2759	88.1(2.0) [3%]	223.7(1.3)	5.0(0.8)	29.5(1.5)	66.1(1.6)	33.9(1.6)
	1992	2756	100.0(0.0) [4%]	226.8(1.1)	4.8(0.6)	31.5(1.5)	70.0(1.4)	30.0(1.4)
Kentucky	1994	2758	90.3(1.2) [2%]	212.6(1.6)	3.6(0.6)	21.7(1.4)	53.2(1.7)	46.8(1.7)
	1992	2752	100.0(0.0) [4%]	213.6(1.3)	2.3(0.5)	19.4(1.4)	54.7(1.8)	45.3(1.8)
Louisiana	1994	2713	83.9(1.0) [3%]	198.0(1.3)<	1.4(0.4)	12.1(0.9)	37.6(1.5)	62.4(1.5)
	1992	2848	100.0(0.0) [3%]	204.6(1.2)	1.3(0.3)	12.7(1.0)	42.4(1.7)	57.6(1.7)
Maine	1994	2436	97.0(0.8) [4%]	229.1(1.3)	5.9(0.7)	34.8(1.5)	72.5(1.6)	27.5(1.6)
	1992	1916	100.0(0.0) [4%]	228.2(1.1)	4.3(0.7)	31.5(1.7)	72.1(1.4)	27.9(1.4)
Maryland	1994	2555	90.1(1.1) [3%]	210.7(1.4)	4.2(0.6)	22.3(1.5)	52.4(1.7)	47.6(1.7)
	1992	2786	100.0(0.0) [3%]	212.1(1.6)	3.0(0.5)	20.7(1.1)	53.4(1.8)	46.6(1.8)
Massachusetts	1994	2517	90.5(0.8) [3%]	224.0(1.3)<	4.7(0.7)	30.5(1.7)	66.8(1.6)	33.2(1.6)
	1992	2545	100.0(0.0) [3%]	227.4(1.0)	4.5(0.6)	31.8(1.4)	71.0(1.4)	29.0(1.4)
Michigan	1994	2142	100.0(0.0) [4%]	213.2(2.0)	3.3(0.5)	20.6(1.4)	54.0(2.2)	46.0(2.2)
	1992	2437	100.0(0.0) [4%]	217.2(1.6)	2.6(0.5)	22.5(1.9)	59.0(1.9)	41.0(1.9)
Minnesota	1994	2655	88.0(0.8) [3%]	219.2(1.3)	4.2(0.7)	27.5(1.4)	62.4(1.5)	37.6(1.5)
	1992	2589	100.0(0.0) [4%]	222.1(1.2)	3.8(0.5)	27.6(1.4)	64.8(1.7)	35.2(1.7)
Mississippi	1994	2762	92.6(1.3) [3%]	202.7(1.6)	2.2(0.5)	14.9(1.1)>	42.1(1.7)	57.9(1.7)
	1992	2657	100.0(0.0) [3%]	199.8(1.3)	1.2(0.3)	11.6(0.7)	37.9(1.8)	62.1(1.8)
Missouri	1994	2670	87.8(1.2) [4%]	217.8(1.5)	4.4(0.7)	26.3(1.5)	59.5(1.9)	40.5(1.9)
	1992	2562	100.0(0.0) [5%]	221.3(1.3)	3.9(0.4)	26.2(1.5)	63.2(1.5)	36.8(1.5)
Montana	1994	2501	93.1(1.3) [4%]	223.1(1.4)	4.1(0.6)	29.1(1.5)	66.2(1.7)	33.8(1.7)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-4 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Percent of Students by Achievement Levels
Original Results

PUBLIC SCHOOLS		N	WEIGHTED PCT [CV]	MEAN	ADVANCED	PRFCIENT	BASIC	< BASIC
Nebraska	1994	2395	89.2(0.8) [4%]	220.9(1.4)	5.3(0.7)	29.5(1.7)	63.1(1.6)	36.9(1.6)
	1992	2364	100.0(0.0) [3%]	222.4(1.1)	4.0(0.7)	26.8(1.6)	65.0(1.5)	35.0(1.5)
New Hampshire	1994	2197	100.0(0.0) [4%]	224.3(1.5)<	5.1(0.7)	30.3(1.6)	67.3(1.8)	32.7(1.8)
	1992	2239	100.0(0.0) [4%]	229.1(1.2)	5.7(0.7)	33.7(1.5)	73.0(1.9)	27.0(1.9)
New Jersey	1994	2509	85.6(1.2) [3%]	220.3(1.2)<	5.4(0.7)	28.8(1.5)	62.2(1.5)	37.8(1.5)
	1992	2239	100.0(0.0) [4%]	224.3(1.5)	5.5(0.9)	30.9(1.7)	66.3(1.9)	33.7(1.9)
New Mexico	1994	2635	90.6(2.3) [3%]	205.7(1.7)<	2.8(0.4)	17.2(1.4)	46.0(1.7)	54.0(1.7)
	1992	2305	100.0(0.0) [5%]	211.8(1.5)	2.8(0.6)	19.6(1.6)	51.3(1.7)	48.7(1.7)
New York	1994	2495	83.4(1.0) [3%]	212.7(1.4)	3.9(0.7)	23.1(1.3)	54.3(1.6)	45.7(1.6)
	1992	2285	100.0(0.0) [2%]	215.8(1.4)	3.4(0.5)	23.4(1.1)	57.9(1.4)	42.1(1.4)
North Carolina	1994	2833	100.0(0.0) [2%]	215.3(1.5)	4.9(0.8)	25.7(1.6)	56.2(1.6)	43.8(1.6)
	1992	2883	100.0(0.0) [3%]	212.6(1.2)	3.8(0.5)	21.8(1.2)	52.9(1.4)	47.1(1.4)
North Dakota	1994	2544	89.5(1.8) [4%]	226.0(1.2)	4.9(0.5)	31.7(1.7)	69.7(1.5)	30.3(1.5)
	1992	2158	100.0(0.0) [4%]	226.9(1.2)	4.1(0.6)	30.6(1.5)	71.2(1.9)	28.8(1.9)
Pennsylvania	1994	2290	83.3(1.3) [3%]	216.2(1.5)<	4.3(0.6)	25.8(1.6)	58.5(1.7)	41.5(1.7)
	1992	2805	100.0(0.0) [4%]	221.9(1.3)	4.4(0.6)	27.8(1.5)	64.3(1.9)	35.7(1.9)
Rhode Island	1994	2342	88.2(1.2) [4%]	220.8(1.3)	4.6(0.6)	27.4(1.6)	62.7(1.5)	37.3(1.5)
	1992	2347	100.0(0.0) [4%]	217.8(1.8)	3.4(0.5)	23.9(1.7)	59.5(2.1)	40.5(2.1)
South Carolina	1994	2707	94.8(1.0) [3%]	204.6(1.4)<	2.6(0.5)	16.3(1.1)	44.4(1.4)	55.6(1.4)
	1992	2758	100.0(0.0) [3%]	210.7(1.3)	2.5(0.6)	18.8(1.2)	49.4(1.8)	50.6(1.8)
Tennessee	1994	1998	100.0(0.0) [3%]	213.5(1.7)	3.6(0.7)	22.1(1.4)	54.6(2.1)	45.4(2.1)
	1992	2734	100.0(0.0) [3%]	213.2(1.5)	2.5(0.5)	19.9(1.4)	53.1(1.7)	46.9(1.7)
Texas	1994	2454	100.0(0.0) [3%]	213.4(1.8)	3.7(0.7)	21.8(1.6)	54.3(2.3)	45.7(2.3)
	1992	2571	100.0(0.0) [4%]	213.6(1.6)	2.9(0.5)	20.1(1.7)	53.2(2.0)	46.8(2.0)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-4 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Percent of Students by Achievement Levels
Original Results

		N	WEIGHTED PCT [CV]	MEAN	ADVANCED	PRFCIENT	BASIC	< BASIC
PUBLIC SCHOOLS								
Utah	1994	2733	100.0(0.0) [2%]	218.2(1.2)<	3.8(0.5)	25.3(1.7)	60.9(1.6)	39.1(1.6)
	1992	2829	100.0(0.0) [2%]	221.6(1.2)	3.3(0.5)	26.0(1.3)	63.7(1.5)	36.3(1.5)
Virginia	1994	2719	94.1(1.1) [3%]	214.5(1.4)<	4.5(0.7)	23.1(1.6)	54.2(1.9)<	45.8(1.9)>
	1992	2786	100.0(0.0) [3%]	222.0(1.4)	4.7(0.8)	27.6(1.5)	63.6(1.8)	36.4(1.8)
Washington	1994	2737	100.0(0.0) [3%]	213.7(1.4)	3.4(0.5)	22.5(1.2)	55.5(1.6)	44.5(1.6)
West Virginia	1994	2757	95.0(0.9) [3%]	214.1(1.1)	3.3(0.4)	22.2(1.3)	54.6(1.5)	45.4(1.5)
	1992	2733	100.0(0.0) [4%]	216.6(1.3)	3.4(0.5)	21.8(1.3)	57.5(1.5)	42.5(1.5)
Wisconsin	1994	2331	83.6(1.4) [3%]	225.0(1.1)	4.0(0.5)	29.6(1.3)	68.1(1.7)	31.9(1.7)
	1992	2712	100.0(0.0) [4%]	224.9(1.0)	4.3(0.5)	29.2(1.1)	67.3(1.3)	32.7(1.3)
Wyoming	1994	2699	100.0(0.0) [3%]	222.0(1.2)	3.1(0.5)	26.4(1.3)	65.1(1.7)	34.9(1.7)
	1992	2775	100.0(0.0) [3%]	224.3(1.2)	3.7(0.5)	28.4(1.7)	67.9(1.5)	32.1(1.5)
DoDEA Overseas	1994	2413	100.0(0.0) [2%]	218.5(0.9)	3.3(0.6)	23.4(1.0)	59.7(1.3)	40.3(1.3)
Guam	1994	2203	85.3(0.1) [1%]	182.5(1.2)	0.6(0.2)	6.3(0.8)	24.6(1.0)	75.4(1.0)
	1992	2029	100.0(0.0) [0%]	182.7(1.4)	0.5(0.2)	6.3(0.7)	25.3(1.2)	74.7(1.2)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-5
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Percent of Students by Achievement Levels
Revised Results

		N	WEIGHTED PCT [CV]	MEAN	ADVANCED	PRFCIENT	BASIC	< BASIC
PUBLIC SCHOOLS								
Nation	1994	6030	100.0(0.0) [1%]	212.3(1.1)	6.8(0.7)	28.1(1.2)	58.6(1.1)	41.4(1.1)
	1992	5045	100.0(0.0) [1%]	214.8(1.0)	5.7(0.6)	26.6(1.3)	60.0(1.1)	40.0(1.1)
Alabama	1994	2646	93.2(1.2) [2%]	208.0(1.5)	5.2(0.7)	23.4(1.3)	52.3(1.6)	47.7(1.6)
	1992	2571	100.0(0.0) [4%]	207.4(1.7)	3.3(0.4)	19.6(1.5)	51.5(2.1)	48.5(2.1)
Arizona	1994	2651	100.0(0.0) [3%]	206.2(1.9)	6.2(0.8)>	24.1(1.5)	51.8(1.9)	48.2(1.9)
	1992	2677	100.0(0.0) [2%]	209.4(1.2)	3.1(0.4)	20.8(1.2)	54.3(1.8)	45.7(1.8)
Arkansas	1994	2535	94.7(0.9) [3%]	208.6(1.7)	4.7(0.6)	23.9(1.4)	53.6(1.8)	46.4(1.8)
	1992	2589	100.0(0.0) [4%]	210.9(1.2)	3.7(0.6)	22.7(1.2)	55.8(1.5)	44.2(1.5)
California	1994	2252	91.1(1.1) [3%]	196.8(1.8)<	3.4(0.5)	17.5(1.3)	44.3(2.0)	55.7(2.0)
	1992	2365	100.0(0.0) [3%]	202.2(2.0)	3.6(0.7)	19.5(1.7)	47.6(2.2)	52.4(2.2)
Colorado	1994	2730	94.2(1.5) [3%]	213.5(1.3)	5.8(0.7)	28.2(1.5)	59.3(1.4)	40.7(1.4)
	1992	2897	100.0(0.0) [3%]	216.5(1.1)	3.9(0.6)	25.3(1.4)	63.7(1.6)	36.3(1.6)
Connecticut	1994	2578	89.1(1.0) [2%]	222.4(1.6)	10.8(1.1)>	38.2(1.6)	68.1(1.7)	31.9(1.7)
	1992	2514	100.0(0.0) [3%]	221.6(1.3)	6.3(1.0)	33.7(1.4)	68.8(1.7)	31.2(1.7)
Delaware	1994	2239	82.2(1.4) [0%]	206.4(1.1)<	5.0(0.8)	23.0(1.1)	52.3(1.3)<	47.7(1.3)>
	1992	2048	100.0(0.0) [0%]	212.9(0.6)	4.7(0.5)	24.2(1.1)	57.5(1.2)	42.5(1.2)
Florida	1994	2666	89.8(0.9) [2%]	204.9(1.7)	5.1(0.6)>	22.6(1.5)	50.3(1.8)	49.7(1.8)
	1992	2767	100.0(0.0) [3%]	208.3(1.2)	3.2(0.4)	21.2(1.1)	52.6(1.6)	47.4(1.6)
Georgia	1994	2765	93.2(0.9) [4%]	206.8(2.4)	6.7(1.0)	25.6(2.0)	52.1(2.3)	47.9(2.3)
	1992	2712	100.0(0.0) [3%]	212.3(1.5)	5.2(0.8)	24.8(1.5)	56.9(1.7)	43.1(1.7)
Hawaii	1994	2732	87.5(0.9) [2%]	200.9(1.7)	4.1(0.5)	19.4(1.4)	46.4(1.8)	53.6(1.8)
	1992	2642	100.0(0.0) [2%]	203.4(1.7)	2.6(0.5)	17.0(1.5)	47.7(1.9)	52.3(1.9)
Idaho	1994	2598	96.3(0.4) [2%]	212.8(1.4)<	5.5(0.8)	26.2(1.4)	58.5(1.5)<	41.5(1.5)>
	1992	2674	100.0(0.0) [3%]	219.3(0.9)	4.4(0.7)	27.8(1.2)	66.5(1.3)	33.5(1.3)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-5 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Percent of Students by Achievement Levels
Revised Results

		N	WEIGHTED PCT [CV]	MEAN	ADVANCED	PRFCIENT	BASIC	< BASIC
PUBLIC SCHOOLS								
Indiana	1994	2655	92.7(1.0) [2%]	219.8(1.3)	7.2(0.8)	32.7(1.5)	65.7(1.6)	34.3(1.6)
	1992	2535	100.0(0.0) [3%]	221.1(1.3)	5.6(0.9)	30.3(1.5)	67.9(1.6)	32.1(1.6)
Iowa	1994	2759	88.1(2.0) [3%]	222.9(1.3)	8.1(1.0)	34.7(1.5)	69.1(1.6)	30.9(1.6)
	1992	2756	100.0(0.0) [4%]	225.5(1.1)	7.0(0.7)	35.9(1.6)	73.2(1.4)	26.8(1.4)
Kentucky	1994	2758	90.3(1.2) [2%]	211.6(1.6)	5.8(0.8)>	25.7(1.9)	56.3(1.6)	43.7(1.6)
	1992	2752	100.0(0.0) [4%]	212.5(1.3)	3.4(0.5)	22.8(1.6)	58.0(1.7)	42.0(1.7)
Louisiana	1994	2713	83.9(1.0) [3%]	196.8(1.3)<	2.4(0.5)	14.7(1.2)	40.3(1.5)<	59.7(1.5)>
	1992	2848	100.0(0.0) [3%]	203.7(1.2)	2.0(0.4)	15.2(1.1)	45.9(1.6)	54.1(1.6)
Maine	1994	2436	97.0(0.8) [4%]	228.4(1.3)	10.0(1.0)>	40.7(1.5)	75.2(1.6)	24.8(1.6)
	1992	1916	100.0(0.0) [4%]	226.8(1.1)	6.3(0.8)	36.0(1.7)	75.4(1.4)	24.6(1.4)
Maryland	1994	2555	90.1(1.1) [3%]	209.6(1.5)	6.5(0.7)>	26.2(1.4)	55.3(1.6)	44.7(1.6)
	1992	2786	100.0(0.0) [3%]	211.0(1.6)	4.3(0.6)	24.0(1.2)	56.6(1.8)	43.4(1.8)
Massachusetts	1994	2517	90.5(0.8) [3%]	223.2(1.3)	8.0(1.0)	35.8(1.7)	69.3(1.5)<	30.7(1.5)>
	1992	2545	100.0(0.0) [3%]	226.0(0.9)	6.6(0.8)	35.9(1.5)	74.1(1.3)	25.9(1.3)
Michigan	1994	2142	100.0(0.0) [4%]	212.2(2.0)	5.3(0.7)	25.3(1.6)	57.4(2.3)	42.6(2.3)
	1992	2437	100.0(0.0) [4%]	216.0(1.5)	4.0(0.6)	26.3(2.0)	62.1(1.9)	37.9(1.9)
Minnesota	1994	2655	88.0(0.8) [3%]	218.3(1.4)	7.3(0.7)	32.5(1.4)	65.1(1.5)	34.9(1.5)
	1992	2589	100.0(0.0) [4%]	220.9(1.2)	5.6(0.7)	31.4(1.5)	68.0(1.7)	32.0(1.7)
Mississippi	1994	2762	92.6(1.3) [3%]	201.5(1.6)	3.9(0.6)>	17.9(1.3)>	45.3(1.7)	54.7(1.7)
	1992	2657	100.0(0.0) [3%]	199.1(1.3)	1.9(0.4)	13.5(0.9)	41.3(1.7)	58.7(1.7)
Missouri	1994	2670	87.8(1.2) [4%]	216.9(1.5)	7.4(0.9)	30.7(1.6)	62.1(1.8)	37.9(1.8)
	1992	2562	100.0(0.0) [5%]	220.0(1.2)	5.6(0.7)	29.9(1.5)	66.6(1.5)	33.4(1.5)
Montana	1994	2501	93.1(1.3) [4%]	222.3(1.4)	7.2(0.7)	34.9(1.5)	69.3(1.7)	30.7(1.7)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-5 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Percent of Students by Achievement Levels
Revised Results

		N	WEIGHTED PCT [CV]	MEAN	ADVANCED	PRFCIENT	BASIC	< BASIC
PUBLIC SCHOOLS								
Nebraska	1994	2395	89.2(0.8) [4%]	220.0(1.5)	8.2(0.9)	34.1(1.8)	66.0(1.6)	34.0(1.6)
	1992	2364	100.0(0.0) [3%]	221.2(1.1)	5.5(0.7)	30.7(1.5)	68.4(1.5)	31.6(1.5)
New Hampshire	1994	2197	100.0(0.0) [4%]	223.5(1.5)<	8.5(1.0)	36.1(1.6)	70.4(1.9)	29.6(1.9)
	1992	2239	100.0(0.0) [4%]	227.7(1.2)	7.9(1.1)	38.1(1.6)	75.9(1.8)	24.1(1.8)
New Jersey	1994	2509	85.6(1.2) [3%]	219.4(1.2)	8.2(0.8)	33.2(1.6)	64.9(1.5)	35.1(1.5)
	1992	2239	100.0(0.0) [4%]	223.0(1.4)	7.7(1.0)	34.7(1.8)	69.1(1.8)	30.9(1.8)
New Mexico	1994	2635	90.6(2.3) [3%]	204.5(1.7)<	4.5(0.5)	20.6(1.5)	49.3(1.6)	50.7(1.6)
	1992	2305	100.0(0.0) [5%]	210.7(1.5)	3.7(0.7)	22.7(1.7)	54.6(1.7)	45.4(1.7)
New York	1994	2495	83.4(1.0) [3%]	211.7(1.4)	6.2(0.8)	27.4(1.5)	57.2(1.7)	42.8(1.7)
	1992	2285	100.0(0.0) [2%]	214.6(1.4)	4.8(0.6)	26.7(1.3)	61.4(1.4)	38.6(1.4)
North Carolina	1994	2833	100.0(0.0) [2%]	214.4(1.5)	7.7(0.8)	29.5(1.7)	59.0(1.5)	41.0(1.5)
	1992	2883	100.0(0.0) [3%]	211.5(1.1)	5.2(0.7)	24.7(1.3)	56.0(1.4)	44.0(1.4)
North Dakota	1994	2544	89.5(1.8) [4%]	225.2(1.2)	8.2(0.8)	37.5(1.5)	72.5(1.4)	27.5(1.4)
	1992	2158	100.0(0.0) [4%]	225.5(1.1)	6.0(0.8)	34.9(1.5)	74.1(1.8)	25.9(1.8)
Pennsylvania	1994	2290	83.3(1.3) [3%]	215.3(1.6)<	6.8(0.8)	30.3(1.3)	61.1(1.6)<	38.9(1.6)>
	1992	2805	100.0(0.0) [4%]	220.7(1.3)	5.9(0.8)	31.9(1.7)	67.5(1.7)	32.5(1.7)
Rhode Island	1994	2342	88.2(1.2) [4%]	219.9(1.3)	7.9(1.0)	32.1(1.4)	65.4(1.6)	34.6(1.6)
	1992	2347	100.0(0.0) [4%]	216.6(1.8)	5.1(0.7)	27.6(1.7)	62.6(2.2)	37.4(2.2)
South Caroline	1994	2707	94.8(1.0) [3%]	203.4(1.4)<	4.1(0.6)	19.8(1.3)	47.5(1.5)	52.5(1.5)
	1992	2758	100.0(0.0) [3%]	209.7(1.3)	3.7(0.7)	21.5(1.4)	53.0(1.9)	47.0(1.9)
Tennessee	1994	1998	100.0(0.0) [3%]	212.5(1.7)	6.0(0.9)	26.6(1.5)	57.6(2.1)	42.4(2.1)
	1992	2734	100.0(0.0) [3%]	212.1(1.4)	3.9(0.7)	23.1(1.5)	56.6(1.7)	43.4(1.7)
Texas	1994	2454	100.0(0.0) [3%]	212.4(1.9)	5.9(0.8)	25.9(1.8)	57.6(2.3)	42.4(2.3)
	1992	2571	100.0(0.0) [4%]	212.5(1.6)	4.1(0.7)	23.6(1.8)	56.6(2.0)	43.4(2.0)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

Table H-5 (continued)
NAEP 1992 and 1994 Trial State Reading Assessments
Grade 4 Weighted Percentages and Composite Proficiency Means
Percent of Students by Achievement Levels
Revised Results

		N	WEIGHTED PCT [CV]	MEAN	ADVANCED	PRFCIENT	BASIC	< BASIC
PUBLIC SCHOOLS								
Utah	1994	2733	100.0(0.0) [2%]	217.3(1.3)	6.2(0.8)	30.0(1.6)	64.0(1.6)	36.0(1.6)
	1992	2829	100.0(0.0) [2%]	220.4(1.1)	4.8(0.6)	30.0(1.6)	67.3(1.6)	32.7(1.6)
Virginia	1994	2719	94.1(1.1) [3%]	213.5(1.5)<	7.0(0.7)	26.4(1.7)	57.3(1.8)<	42.7(1.8)>
	1992	2786	100.0(0.0) [3%]	220.8(1.4)	6.3(1.0)	31.4(1.6)	66.7(1.8)	33.3(1.8)
Washington	1994	2737	100.0(0.0) [3%]	212.7(1.5)	5.8(0.7)	26.6(1.2)	58.7(1.6)	41.3(1.6)
West Virginia	1994	2757	95.0(0.9) [3%]	213.1(1.1)	5.6(0.6)	26.2(1.4)	57.7(1.4)	42.3(1.4)
	1992	2733	100.0(0.0) [4%]	215.5(1.3)	4.6(0.7)	25.2(1.4)	61.1(1.4)	38.9(1.4)
Wisconsin	1994	2331	83.6(1.4) [3%]	224.2(1.1)	7.0(0.7)	35.3(1.6)	71.0(1.6)	29.0(1.6)
	1992	2712	100.0(0.0) [4%]	223.6(1.0)	6.0(0.6)	33.4(1.3)	70.7(1.3)	29.3(1.3)
Wyoming	1994	2699	100.0(0.0) [3%]	221.2(1.2)	5.5(0.6)	31.8(1.4)	68.2(1.7)	31.8(1.7)
	1992	2775	100.0(0.0) [3%]	223.0(1.1)	5.4(0.6)	32.7(1.5)	71.0(1.6)	29.0(1.6)
DoDEA Overseas	1994	2413	100.0(0.0) [2%]	217.6(0.9)	5.7(0.7)	28.0(1.1)	62.7(1.5)	37.3(1.5)
Guam	1994	2203	85.3(0.1) [1%]	181.0(1.2)	1.2(0.3)	8.2(0.8)	27.4(1.1)	72.6(1.1)
	1992	2029	100.0(0.0) [0%]	182.3(1.4)	0.8(0.3)	7.7(0.8)	27.8(1.2)	72.2(1.2)

> INDICATES A SIGNIFICANT INCREASE (OR DECREASE "<") BETWEEN 1992 AND 1994

APPENDIX I

THE INFORMATION WEIGHTING ERROR

APPENDIX I

The Information Weighting Error[1]

Susan C. Loomis, Luz Bay, and Wen-Hung Chen

American College Testing

The Error

In the process of recomputing the reading cutscores set in 1992 for the three achievement levels, an error in the information weighting function was detected. The error affected data for all achievement levels set in 1992: reading and mathematics. The Muraki information weighting function published in 1993 was used in the 1994 programs to compute achievement levels, so only 1992 levels are affected.

The procedures used for 1992 were printed and reported in numerous places. No one had detected an error. The psychometrician who developed the programs for the 1994 process used Muraki's information weighting function because he found it to be more straightforward than the 1992 procedure.

The 1992 equation[2] is as follows:

$$I_j(\theta) = D^2 a_j^2 \sum_{c=1}^{m_j} P_{jc}(\theta)[1 - P_{jc}(\theta)] \ . \tag{1}$$

The 1994 equation[3] is as follows:

$$I_j(\theta) = D^2 a_j^2 \sum_{c=1}^{m_j} [T_c - \bar{T}_j(\theta)]^2 P_{jc}(\theta) \ , \tag{2}$$

[1] This appendix was reviewed by Mark Reckase and Alan Nicewander of American College Testing.

[2] Luecht, Richard M. (April, 1993). *Using IRT to improve the standard setting process for dichotomous and polytomous items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA. Notations for this equation were modified to correspond to those of equation 2. The reader will need to refer to the articles for a complete explanation of the equations.

[3] Muraki, Eiji (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17(4)*, 351-362.

where

$$\bar{T}_j(\theta) = \sum_{c=1}^{m_j} T_c P_{jc}(\theta) .$$

Analysis of the Error: Magnitude

The differences in achievement levels reported for 1992 and 1994 and the corrected achievement levels are due both to the error in item parameters and to the error in information weights. The cutscores and percentages of students scoring at or above each for each achievement level are reported in Tables I-1 and I-2. Data in Table I-1 are the previously reported (incorrect) data, and data in Table I-2 are the corrected data.

The maximum difference in cutscores originally reported and the corrected cutscores is found for grade 4 at the Advanced level:

((original cutscore = 275) - (corrected cutscore = 268)) = 7 points.

The differences attributable to each error appear to be rather small in most cases.

Table I-3 reports the differences in cutscores due to the two errors, examined one at a time. Relative to the *correct* data, the information weighting error generally resulted in a higher composite cutscore, while the recoding error resulting in incorrect item parameters generally resulted in a lower composite cutscore.

Figures I-1, I-2, and I-3 show comparisons of percentages of students who scored at or above each achievement level in 1994.[4] The center bar on each of these graphs shows the percentage of students who scored at or above the achievement level in 1994 using both correct item parameters and correct information weights. The bar on the left shows the percentage of students who would have scored at or above the achievement level in 1994 computed with the *correct* item parameters and *incorrect* information weights. The bar on the right shows the percentage computed with the *incorrect* item parameters and the *correct* information weights. These graphs show that the effect of each error was about the same, with respect to the distribution of student scores relative to the cutscores for achievement levels. The greatest difference due to the error in item parameters only is seen for grade 12 at the Advanced level. The greatest difference due to the error in information weights only is seen at the Advanced level for both grades 4 and 8.

[4] Distribution data were not recomputed for the 1992 data using incorrect item parameters and corrected information weights, so these comparisons cannot be presented for 1994.

Table I-1

Reading Cutpoints and Percents At or Above as Reported

Grade		Basic	Proficient	Advanced
4	Cutpoint	212	243	275
	% ≥92 Dist	59.0	25.3	4.5
	% ≥94 Dist	57.7	25.4	4.6
8	Cutpoint	244	283	328
	% ≥92 Dist	68.8	27.5	2.1
	% ≥94 Dist	69.1	27.8	1.9
12	Cutpoint	269	304	348
	% ≥92 Dist	75.2	37.0	3.2
	% ≥94 Dist	70.3	33.6	3.5

Table I-2

Corrected Reading Cutpoints and Percents At or Above

Grade		Basic	Proficient	Advanced
4	Cutpoint	208	238	268
	% ≥92 Dist	62.1	28.6	6.4
	% ≥94 Dist	60.5	29.6	7.4
8	Cutpoint	243	281	323
	% ≥92 Dist	69.5	29.2	2.9
	% ≥94 Dist	69.4	29.5	2.8
12	Cutpoint	265	302	346
	% ≥92 Dist	79.7	40.2	3.9
	% ≥94 Dist	74.5	36.3	4.2

Table I-3

Composite NAEP Scale Cutpoint Differences in Reading
Due to Errors

Achievement Level Cutpoint	Information Weighting*	Item Parameters**
Grade 4		
Basic	3	5
Proficient	3	2
Advanced	5	-1
Grade 8		
Basic	1	1
Proficient	1	-1
Advanced	5	-4
Grade 12		
Basic	0	-1
Proficient	1	-3
Advanced	2	-9

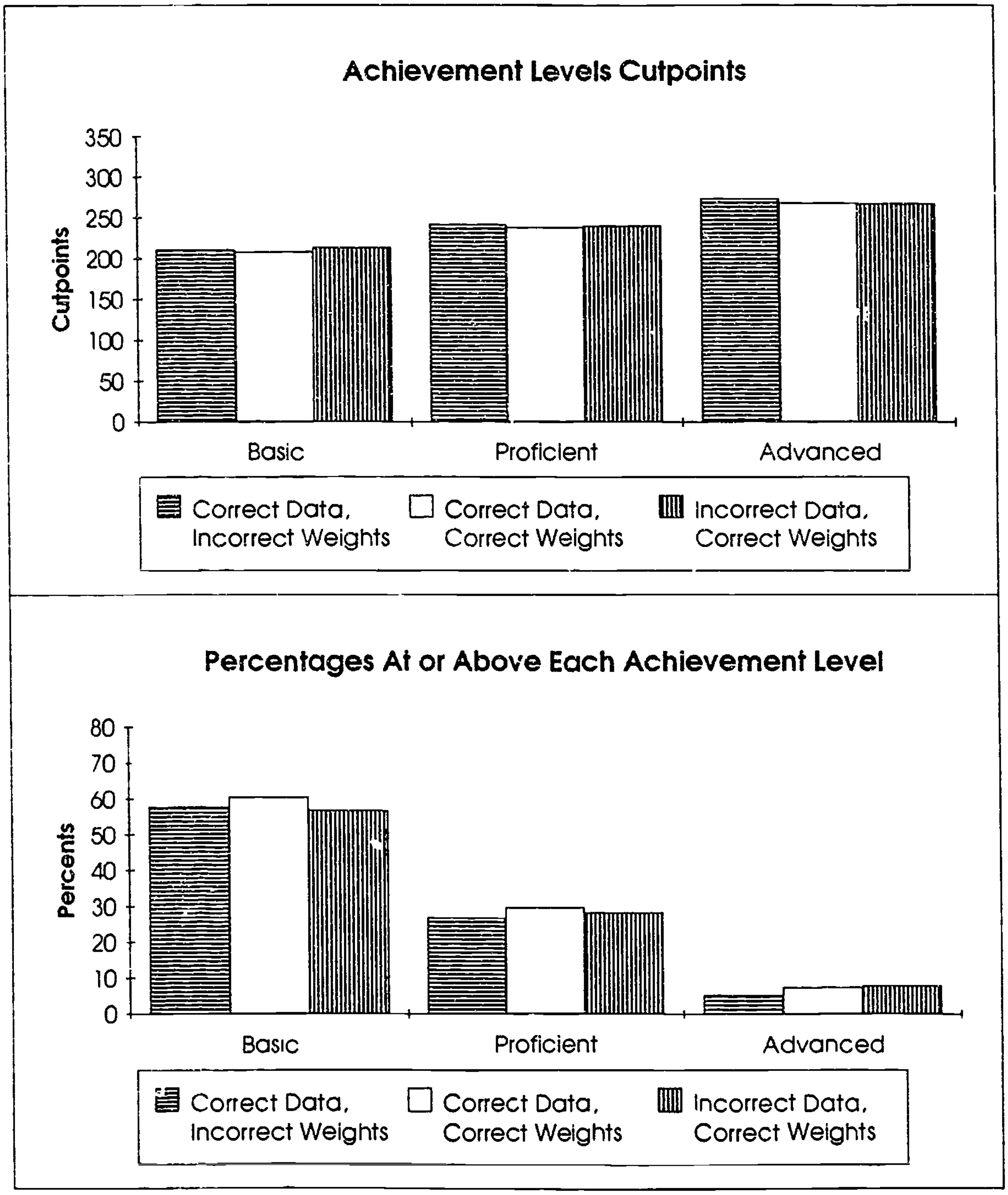
* Difference = Incorrect - Correct, based on correct item parameters. If the recoding of data had been correct, the cutpoints would have been in error by these amounts, due to the incorrect information weighting function.

** Difference = Incorrect - Correct, based on correct information weights. If the correct information weighting function had been used, the cutpoints would have been in error by these amounts due to the recoding error resulting in incorrect item parameters.

Figure I-1

NAEP Reading Achievement Levels: Cutpoints and 1994 Distribution Data

Grade 4



Achievement Levels Cutpoints

Percentages At or Above Each Achievement Level

Figure I-2

NAEP Reading Achievement Levels: Cutpoints and 1994 Distribution Data

Grade 8

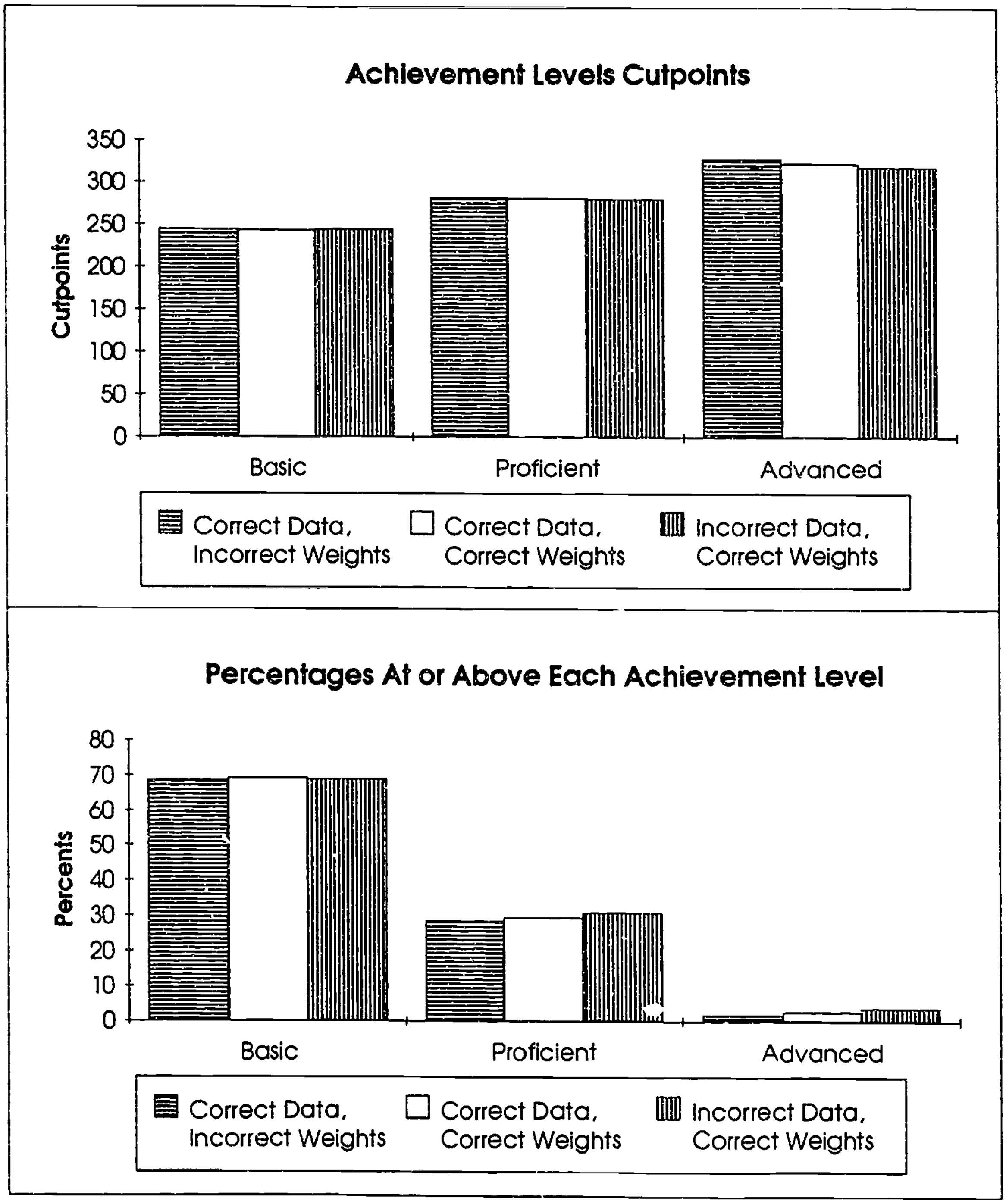
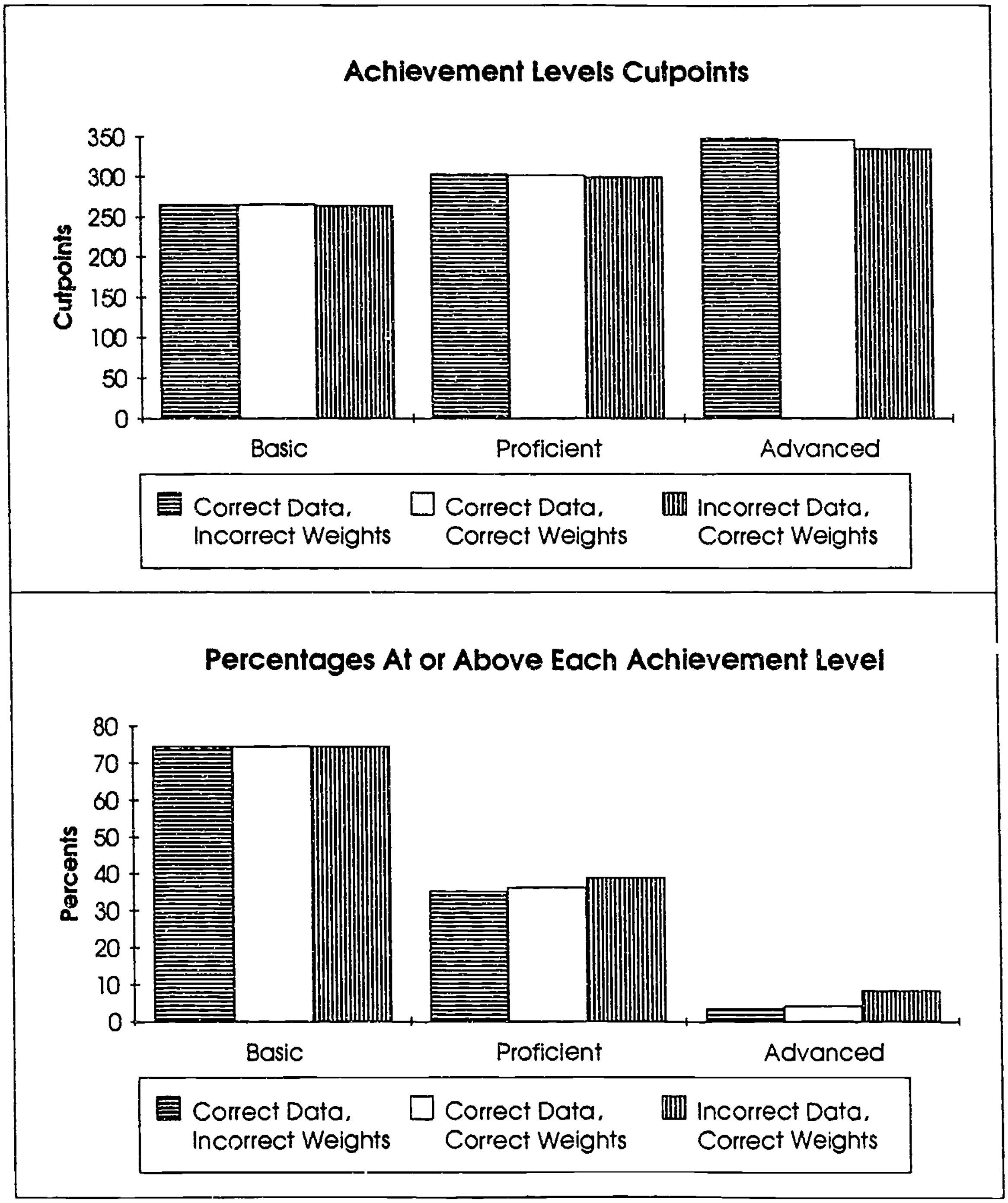


Figure I-3

NAEP Reading Achievement Levels: Cutpoints and 1994 Distribution Data

Grade 12



Achievement Levels Cutpoints

Percentages At or Above Each Achievement Level

Analysis of the Error: The Information Weighting Functions

Various analyses were conducted to determine what, if any, general conclusions could be drawn to help inform users of NAEP achievement levels data about the factors related to differences in cutscores due to the information weighting error.

Item ratings are collected from two groups of panelists at each grade level. These groups are called item rating groups, and panelists are assigned to an item rating group so that the two are as equivalent as possible in terms of panelist type (teacher, career educator, or general public; gender; race/ethnicity; and region of residence). These item rating groups rate slightly over half of all items at their grade level. Item rating pools are developed so that the items in each are as equivalent as possible in terms of item difficulty, item format (multiple choice, short constructed response, and extended constructed response), test time for the block, and so forth. Item blocks remain intact for the item rating pools. At least one block (a "common block") is rated by all panel members, i.e., both item rating groups, in the grade group.

Item ratings are placed on the NAEP scale by computing a theta value for the dichotomous items and for the polytomous ıtems in each subscale for each rating group. Information weights are applied for the polytomous items at the subscale level before computing the subscale score for both dichotomous and polytomous items.

Information weighting functions were graphed for the polytomous items in each subscale. Subscales contain items from various blocks. Some subscales contain few polytomous items. The number of items included in the graphs is further reduced by the fact that subscale rating data by item type (polytomous or dichotomous) is by item rating group (A or B). Thus, some subscale scores for a rating group are based on ratings for only one or two polytomous items.

Figures I-4 through I-7 show the "old," i.e., incorrect, information weighting function used for computing the cutscores in 1992 and the "new," i.e., correct, information weighting function used for computing cutscores for grade 4. The graphs in Figure I-4 show information weighting functions for which the greatest differences in the two (old and new) were found (grade 4, group A, subscale 1, at the Advanced level). Figures I-8 through I-11 are graphs comparing the information weighting functions for some subscales for grades 8 and 12. Note in Figures I-9 and I-10, there are negative differences, i.e., the "new" weights exceed the "old" weights, in the areas where information is maximum.

Table I-4 presents the information weights computed for each rating group and each subscale for the Reading NAEP achievement levels. Those data show that there is no *consistent* pattern of error caused by the incorrect information function. The corrected cutscores are consistently neither higher nor lower as a result of this error, but the impact of the error is generally to estimate a higher cutscore for the polytomous items.

Analysis of the Effect of Item Discrimination. Figures I-12 through I-15 show graphs for the correct ("new") and the incorrect ("old") information functions holding other parameters constant while varying the item discrimination parameter in the generalized partial credit IRT ...odel. In general, the differences between correct and incorrect weights increase as item discrimination increases.

Figure I-4

Comparison Between the "Old" and "New" Information Functions for 1992 Reading, Grade 4, Group A, Subscale 1

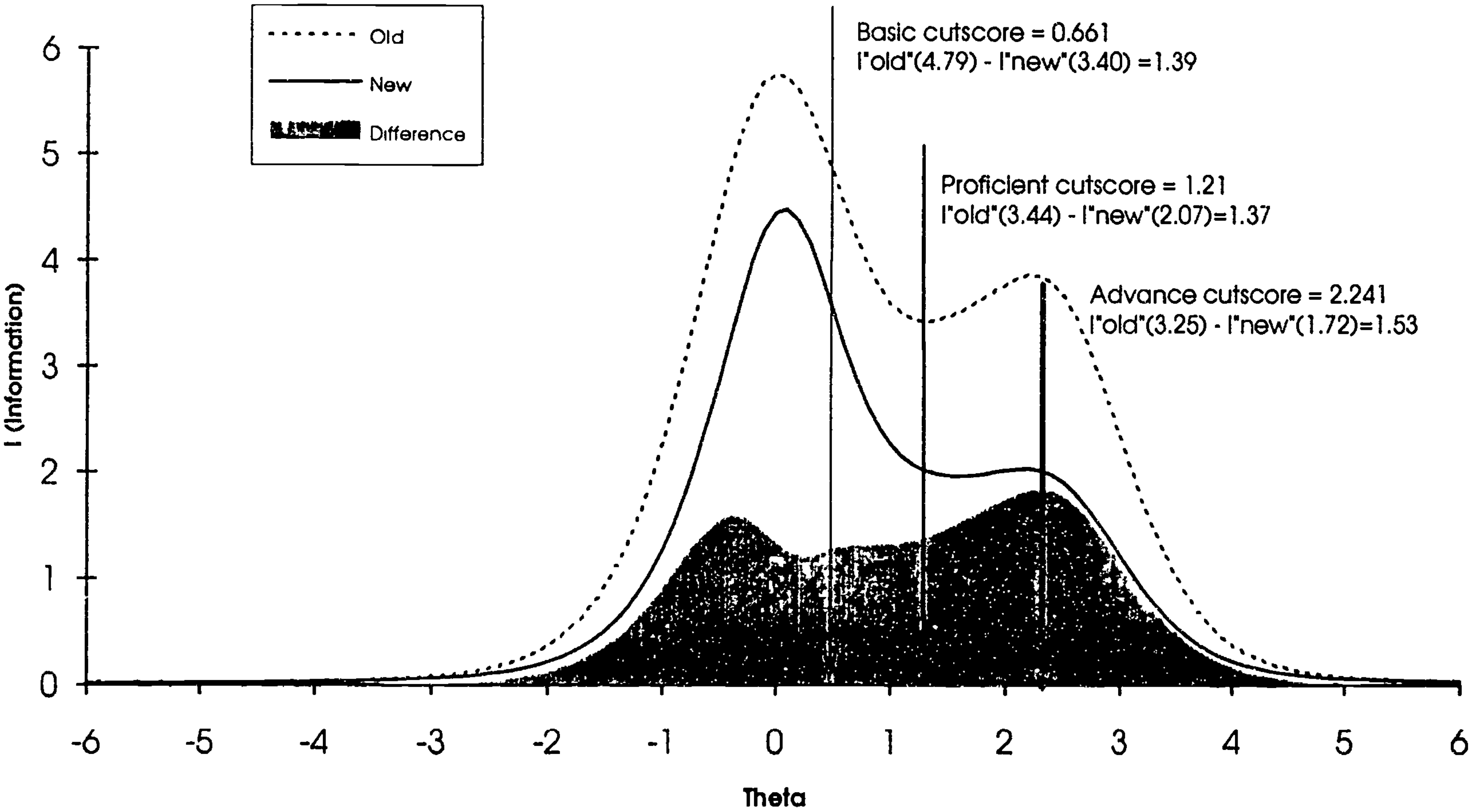


Figure I-5

Comparison Between the "Old" and "New" Information Functions for 1992 Reading, Grade 4, Group A, Subscale 2

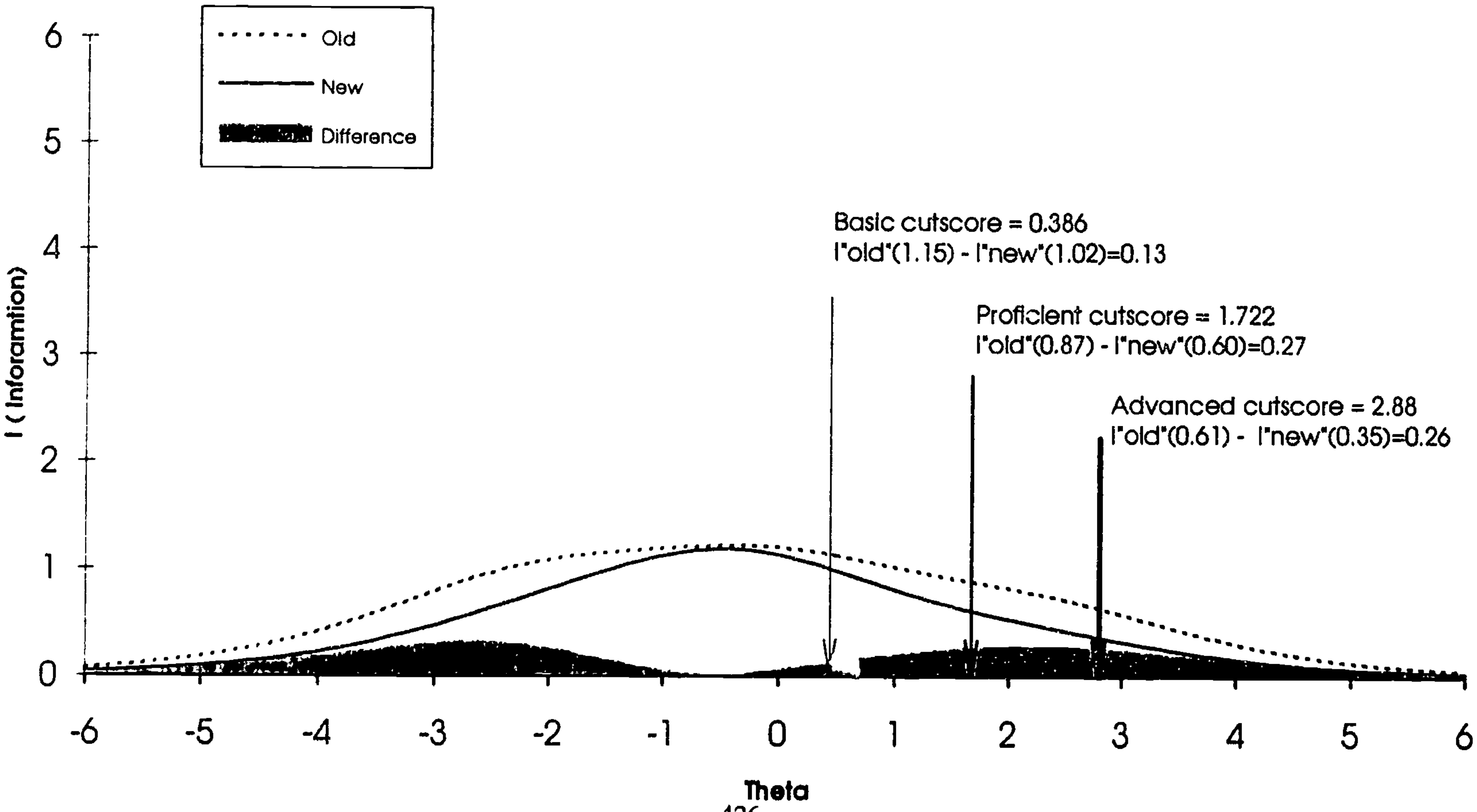
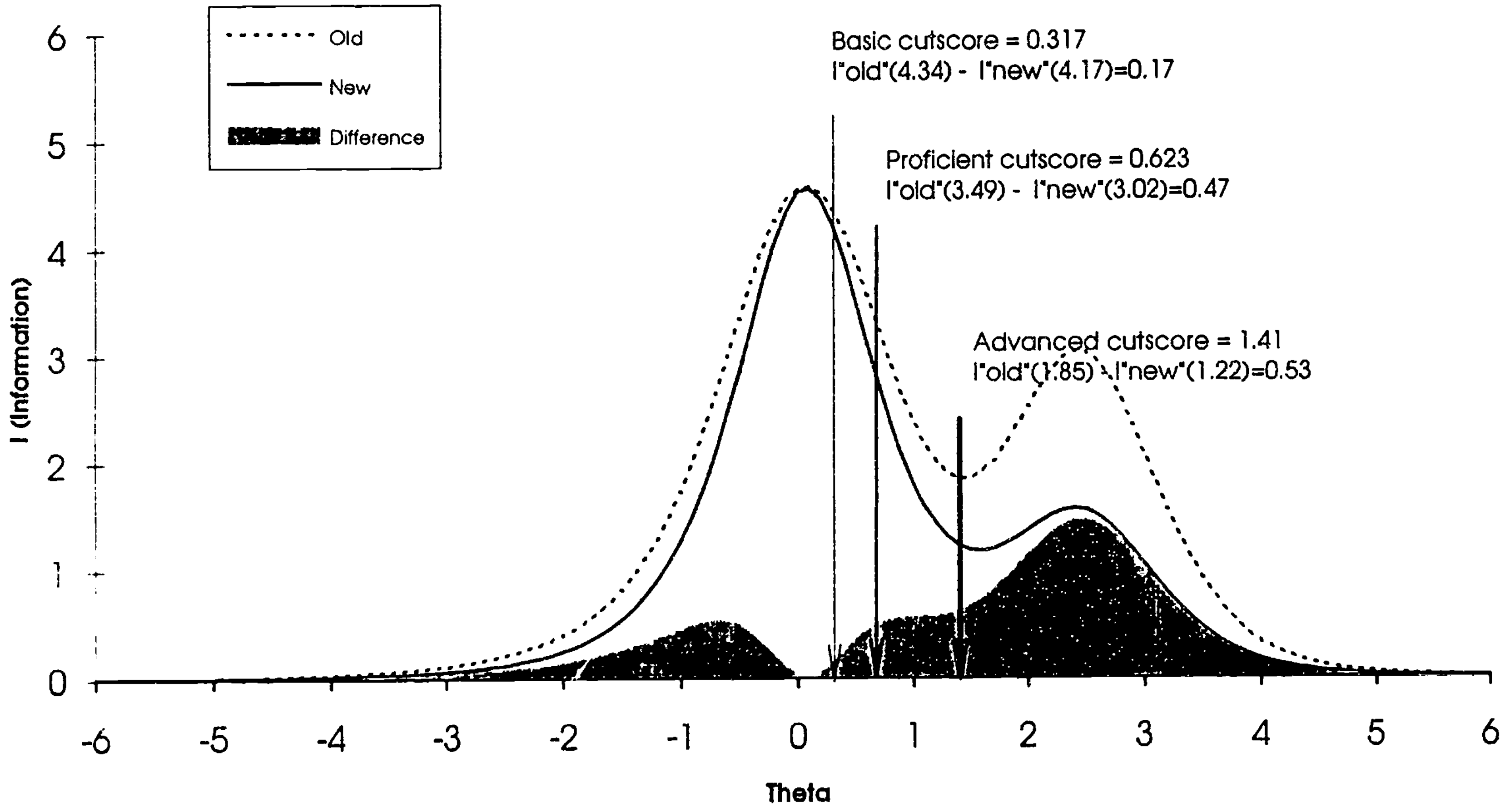


Figure I-6

Comparison Between the "Old" and "New" Information Functions for 1992 Reading, Grade 4, Group B, Subscale 1



Basic cutscore = 0.317
I"old"(4.34) - I"new"(4.17)=0.17

Proficient cutscore = 0.623
I"old"(3.49) - I"new"(3.02)=0.47

Advanced cutscore = 1.41
I"old"(1.85) - I"new"(1.22)=0.53

Figure I-7

Comparison Between the "Old" and "New" Information Functions for 1992 Reading, Grade 4, Group B, Subscale 2

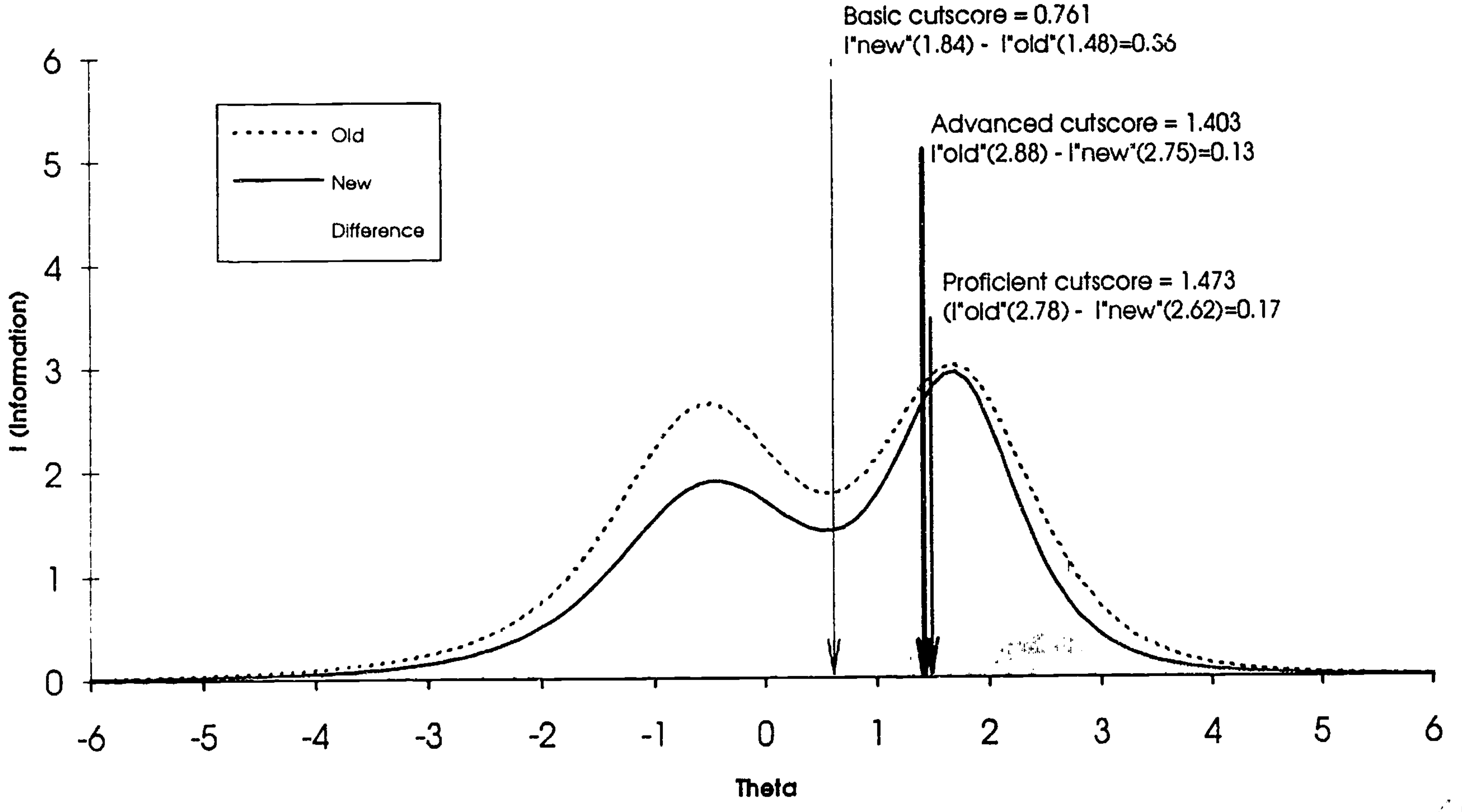


Figure I-8

Comparison Between the "Old" and "New" Information Functions for 1992 Reading, Grade 8, Group A, Subscale 2

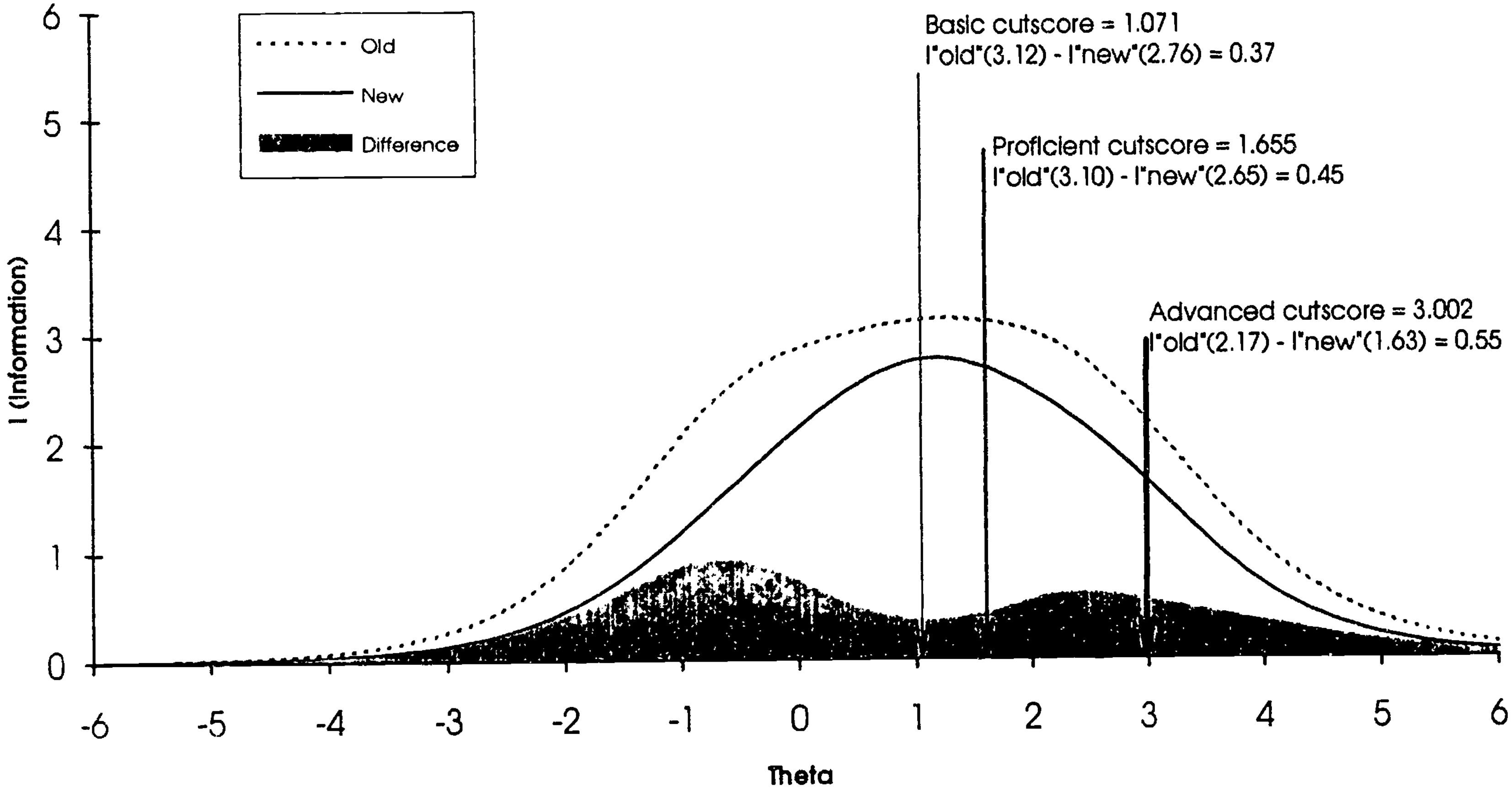
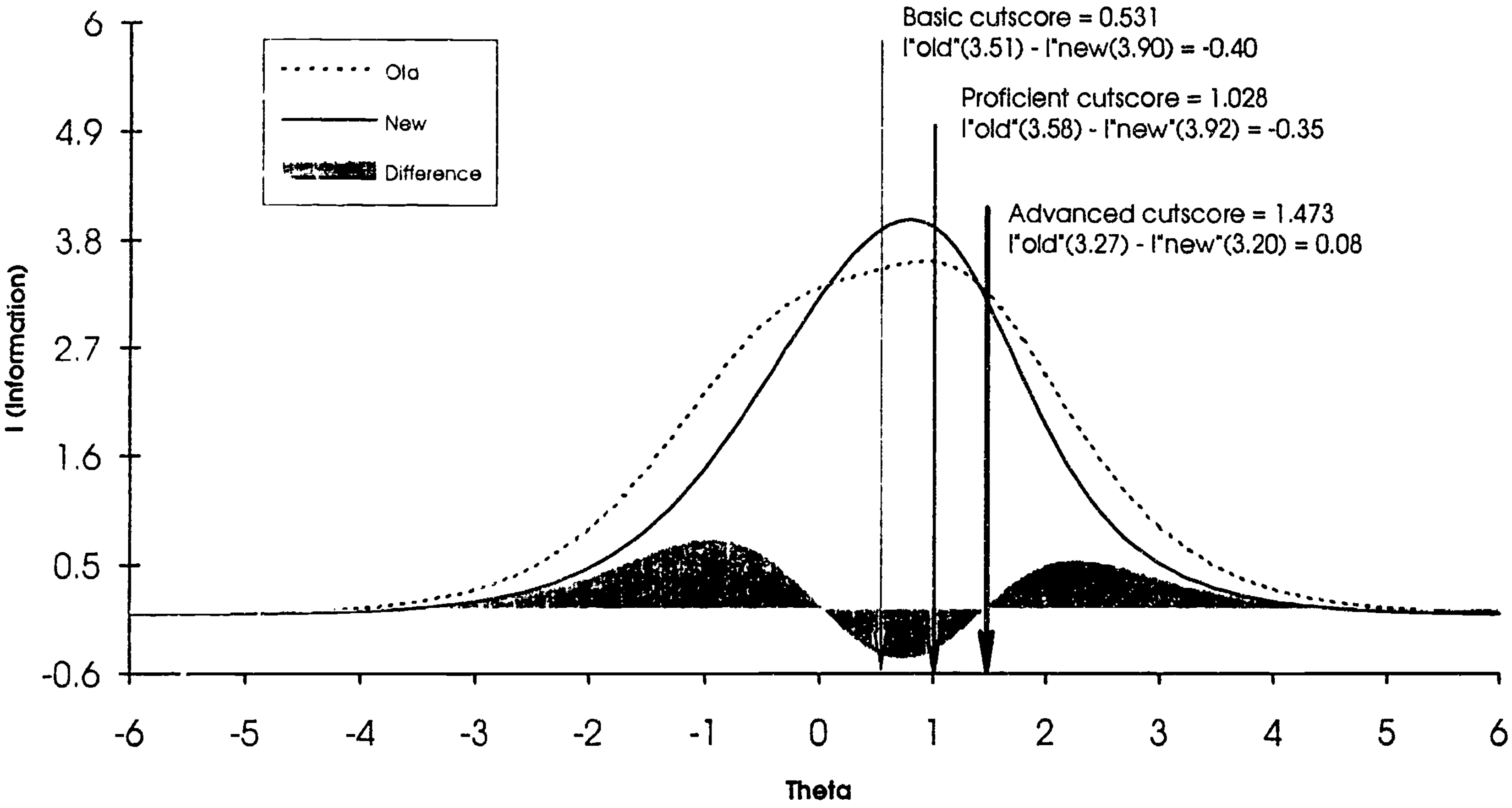


Figure I-9

Comparison Between the "Old" and "New" Information Functions for 1992 Reading, Grade 8, Group B, Subscale 2



Basic cutscore = 0.531
I"old"(3.51) - I"new(3.90) = -0.40

Proficient cutscore = 1.028
I"old"(3.58) - I"new"(3.92) = -0.35

Advanced cutscore = 1.473
I"old"(3.27) - I"new"(3.20) = 0.08

Figure I-10

Comparison Between the "Old" and "New" Information Functions for 1992 Reading, Grade 12, Group A, Subscale 2

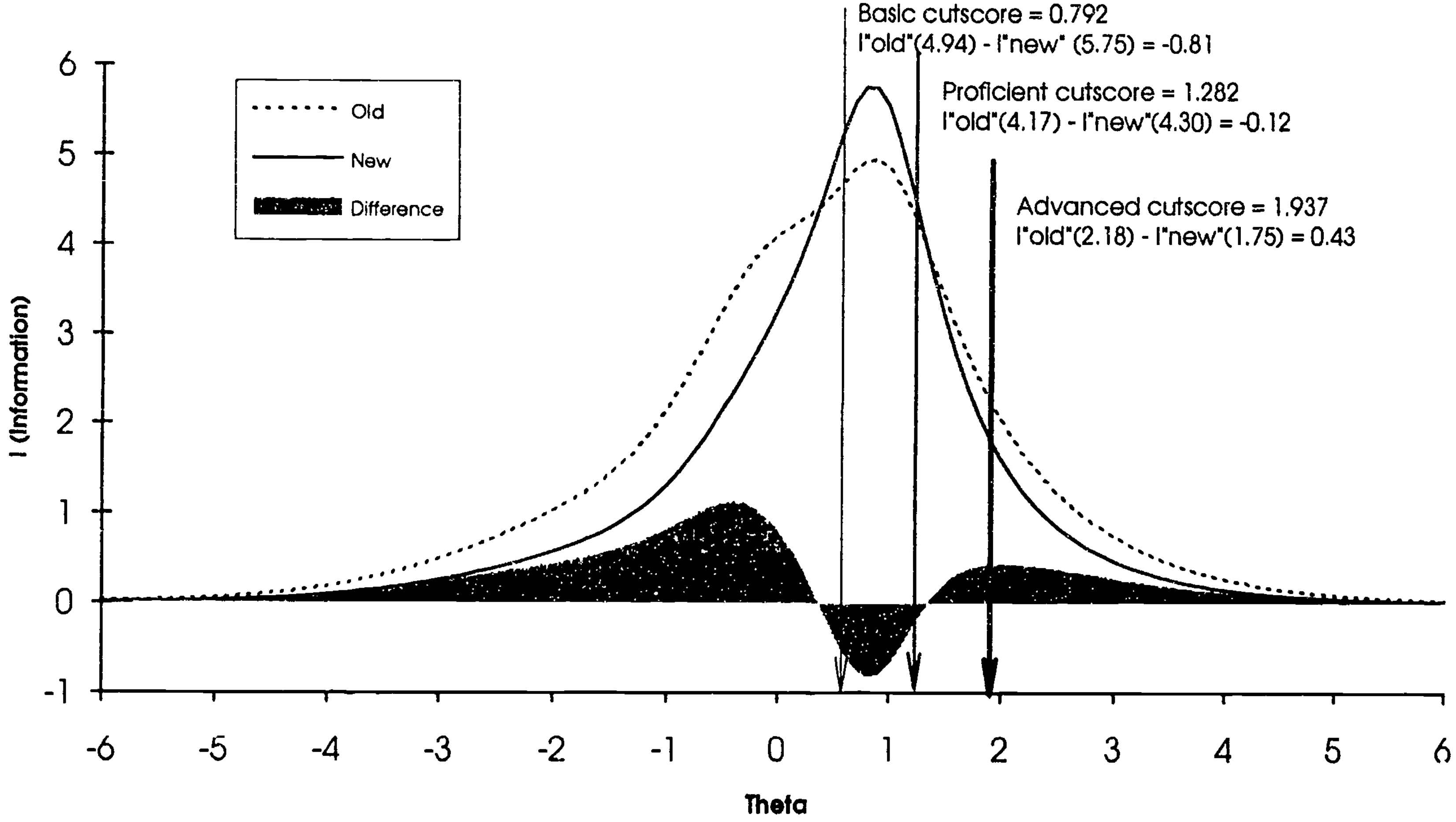


Figure I-11

Comparison Between the "Old" and "New" Information Functions for 1992 Reading, Grade 12, Group B, Subscale 3

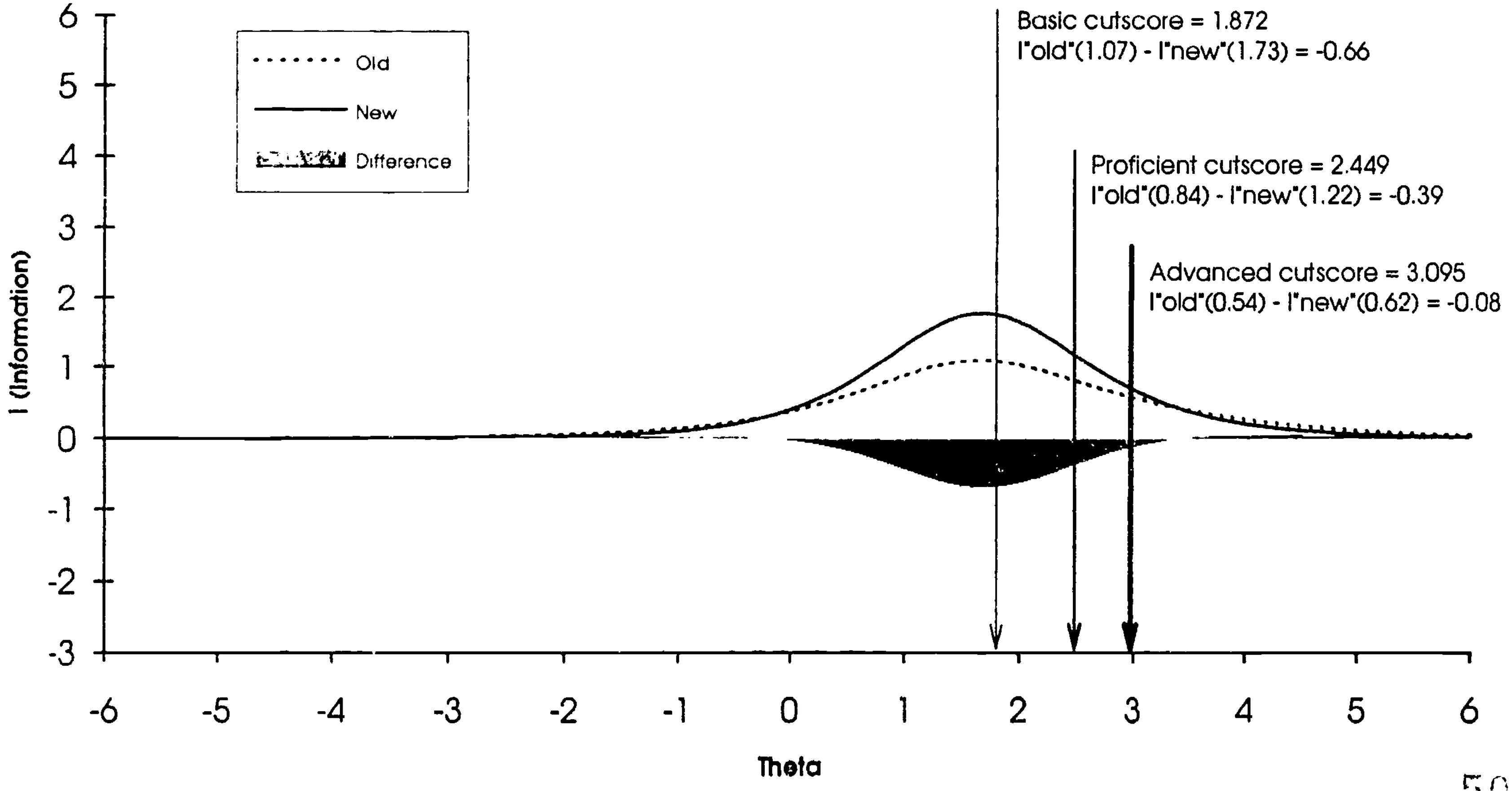


Table I-4

Comparison Between the "Old" and "New" Information Weights for 1992 Reading Achievement Levels Based on Corrected Item Parameters

Grade	Rating Group	Sub-scale	Basic			Proficient			Advanced		
			Old	New	Diff	Old	New	Diff	Old	New	Diff
4	A	1	4.79	3.40	1.39	3.44	2.07	1.37	3.25	1.72	1.53
		2	1.15	1.02	0.13	0.87	0.60	0.27	0.61	0.35	0.26
	B	1	4.34	4.17	0.17	3.49	3.02	0.47	1.85	1.22	0.53
		2	1.84	1.48	0.36	2.88	2.75	0.13	2.78	2.62	0.17
8	A	1	4.34	3.73	0.61	4.15	3.37	0.78	3.30	1.94	1.36
		2	3.12	2.76	0.37	3.10	2.65	0.45	2.17	1.63	0.55
		3	0.77	0.81	-0.04	0.85	1.08	-0.24	0.68	0.72	-0.04
	B	1	1.68	1.55	0.13	1.15	0.70	0.46	0.81	0.42	0.40
		2	3.51	3.90	-0.40	3.58	3.92	-0.35	3.27	3.20	0.08
		3	0.66	0.77	-0.11	0.66	0.77	-0.11	0.59	0.82	-0.23
12	A	1	1.85	1.38	0.47	1.65	1.22	0.43	1.99	1.13	0.86
		2	4.94	5.75	-0.81	4.17	4.30	-0.12	2.18	1.75	0.43
		3	0.81	1.05	-0.24	0.53	0.57	-0.04	0.47	0.48	-0.01
	B*	2	6.31	5.19	1.13	3.82	2.35	1.47	1.51	0.84	0.67
		3	1.07	1.73	-0.66	0.84	1.22	-0.39	0.54	0.62	-0.08

* No polytomous items in subscale 1 were rated by panelists in Group B.

Figure I-12

Comparison Between the "Old" and "New" Information Functions Using Hypothetical Item Parameters a=0.2, b=0, d0=0, d1=2, d2=0, and d3=-2, in Case of Four Response Categories

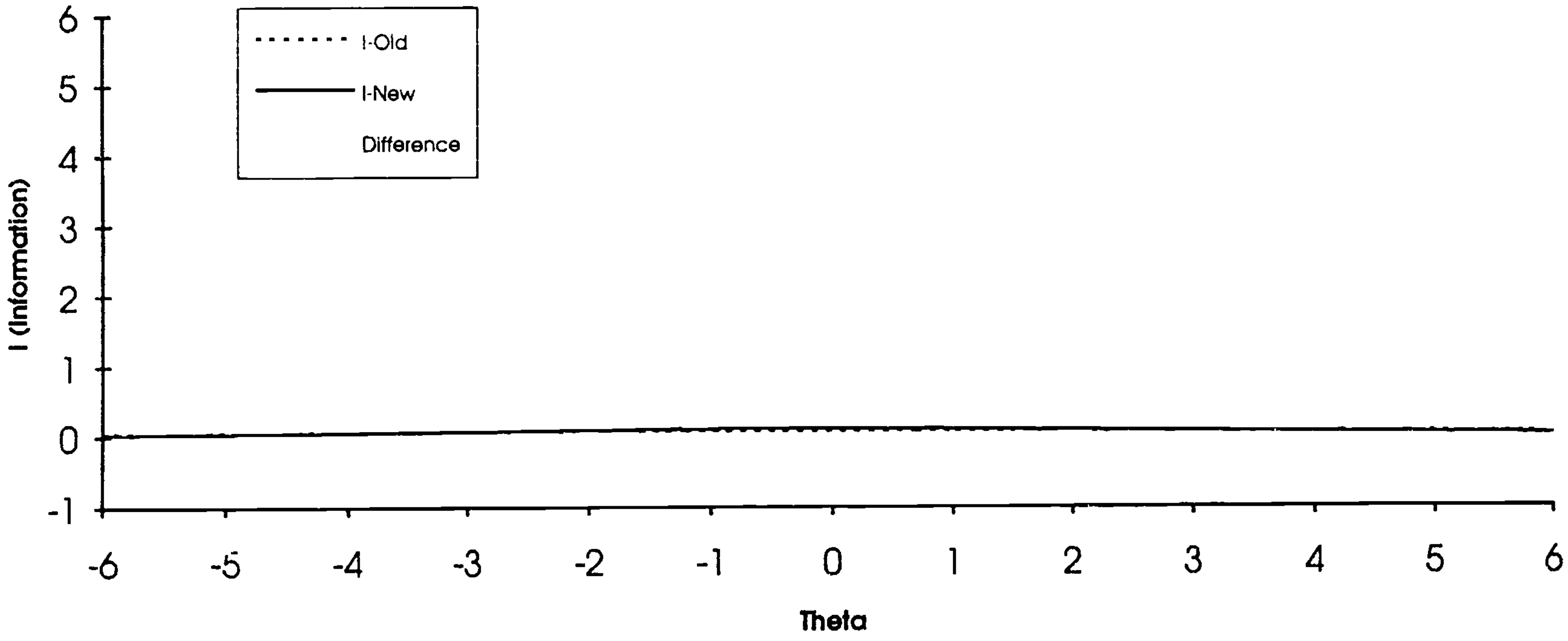


Figure I-13

Comparison Between the "Old" and "New" Information Functions Using Hypothetical Item Parameters a=0.5, b=0, d0=0, d1=2, d2=0, and d3=-2, in Case of Four Response Categories

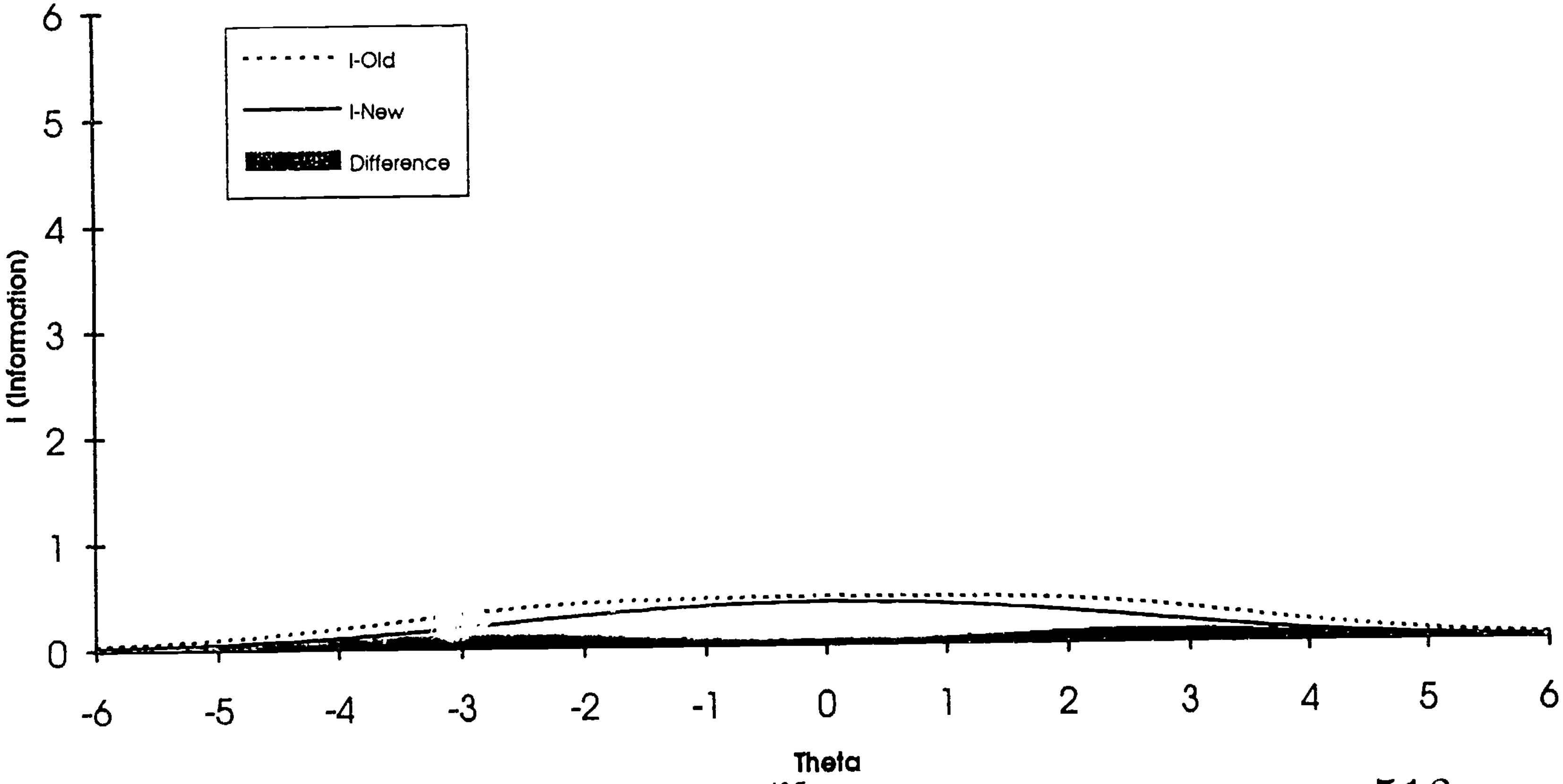
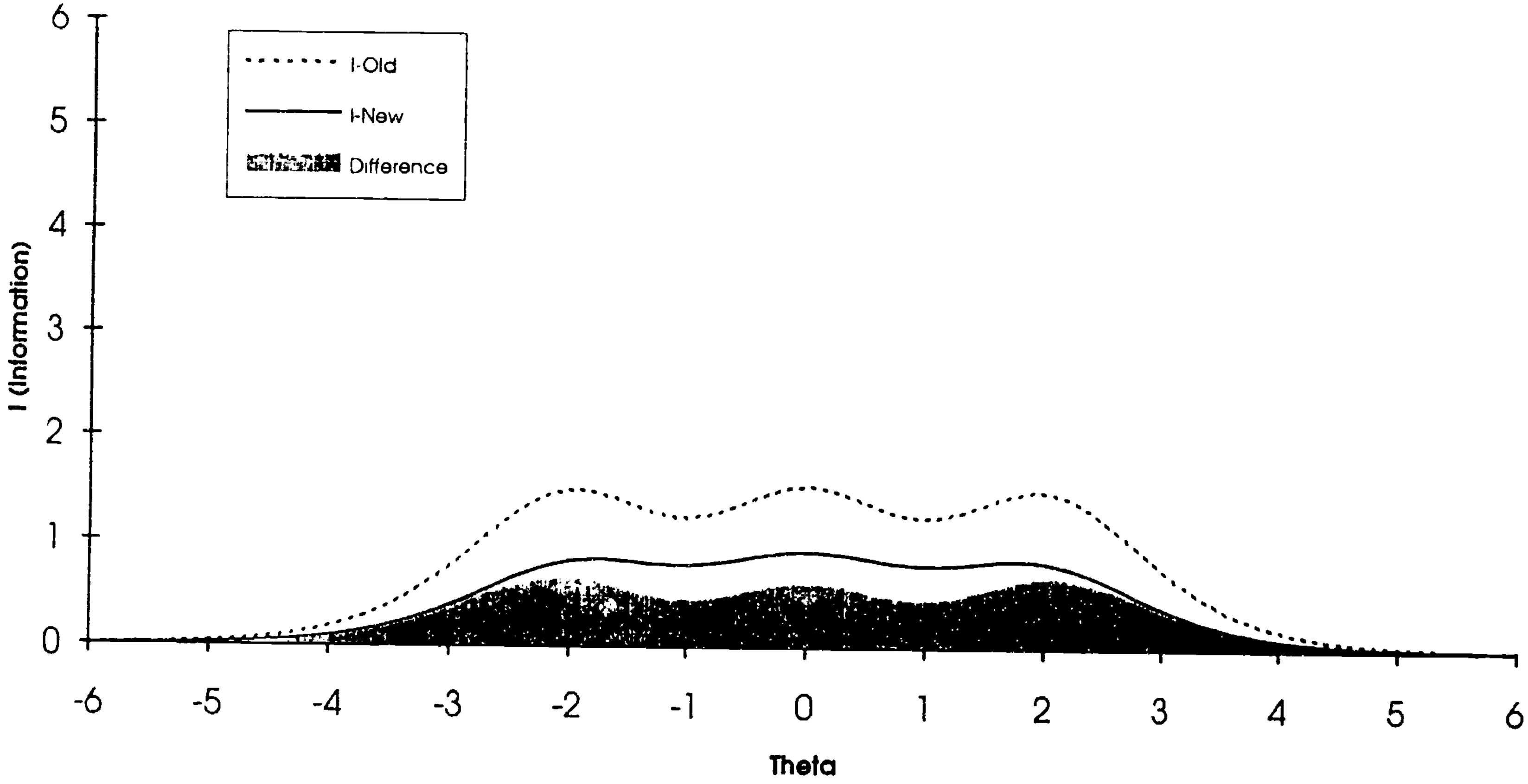


Figure I-14

Comparison Between the "Old" and "New" Information Functions Using Hypothetical Item Parameters a=1, b=0, d0=0, d1=2, d2=0, and d3=-2, in Case of Four Response Categories



Analysis of the Effect of Location Parameters. Figures I-15, I-16, and I-17 show the correct and incorrect information weighting functions for varying location parameters. The location parameters only shift the distribution of information, and that is the case for both the correct and incorrect information weighting functions. The amount of difference between the two is unchanged; only the locations change.

Analysis of the Effect of the Threshold Parameters. If the threshold parameters are close, in terms of the locations of ICCs, the correct information function will have a high peak. In the area of maximum information, i.e., around the peak of the distribution, the incorrect information function underestimates information.

As can be seen in Figures I-15 and I-15a, when the threshold parameters are relatively far apart, the information is relatively low and the distribution is multimodal. When the threshold parameters are closer (see Figures I-18 and I-18a) maximum information is higher and the distribution tends to be more unimodal. As the threshold parameters move even closer, the difference between the correct and incorrect information functions decreases. (See Table I-5.) As the threshold parameters become even closer (see Figures I-19, I-19a, I-20, and I-20a), the rate at which the incorrect weighting function increments at the peak of the distribution is slower than that for the correct function. This results in a negative difference between the two functions in the area of maximum information. (See Figures I-20 and I-20a.)

Figure I-10 is an example of how the location of the threshold parameters can impact the weights of the polytomous cutpoints. The Basic cutpoint falls where the incorrect function weights are less than the correct function weights, but the Advanced cutpoint falls where the incorrect function weights are greater than the correct function weights. Thus, the impact of the incorrect information function is seen to be inconsistent and to depend upon the location of the cutpoint.

Analysis of the Error: Conclusions

Three general conclusions can be drawn from our analyses.

1. Generally, the incorrect information weighting function results in a higher information weight. This is not a consistent pattern, however, because the incorrect function can result in lower information weights than the correct function.

2. The difference between the correct and incorrect weights increases as item discrimination increases.

3. The impact of the incorrect information weighting function on the cutscores is not consistent. The impact depends upon the location of the cutscore and the relative weight of the dichotomous items.

Figure I-15

Comparison Between the "Old" and "New" Information Functions Using Hypothetical Item Parameters a=1.5, b=0, d0=0, d1=2, d2=0, and d3=-2, in Case of Four Response Categories

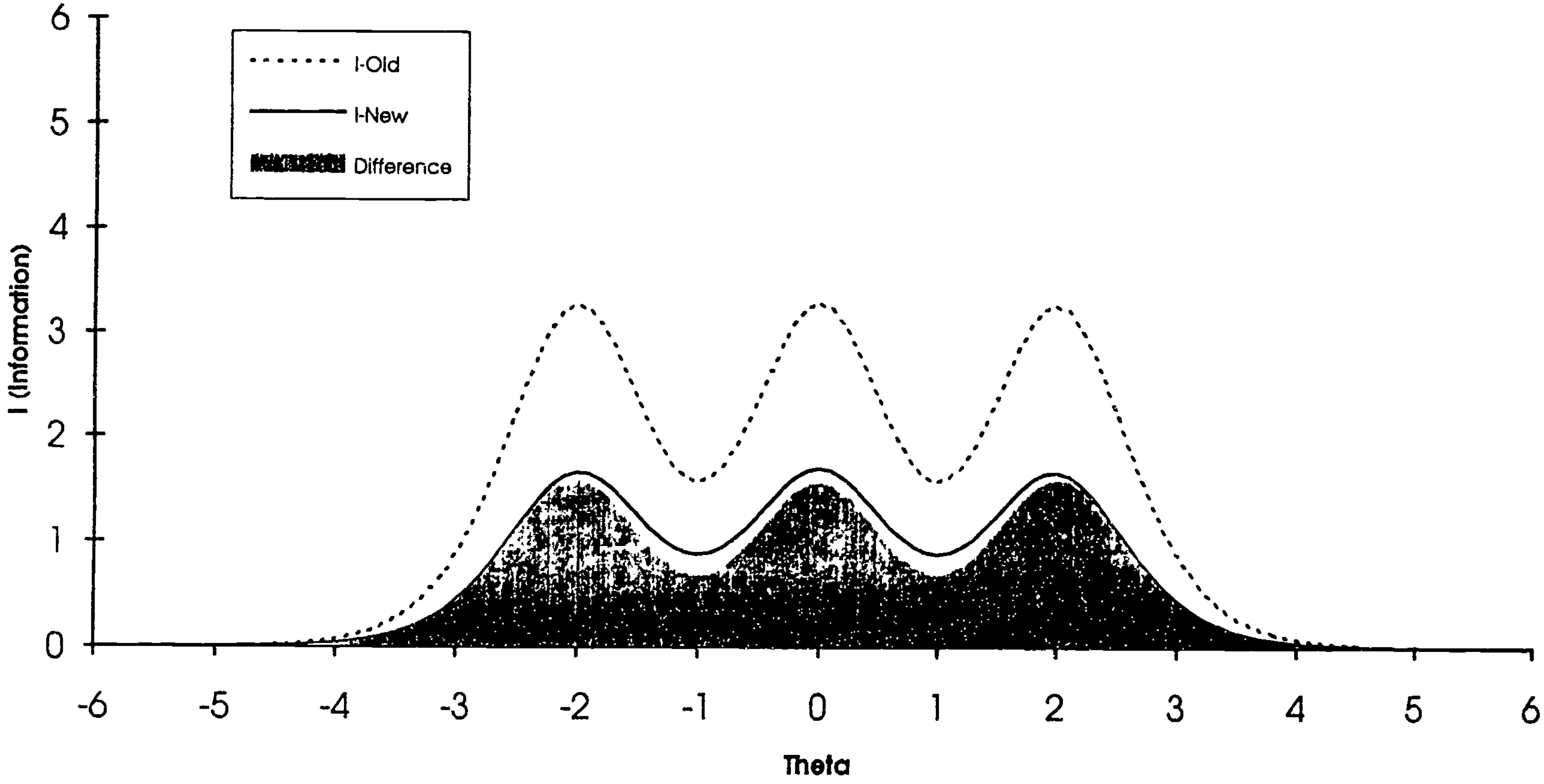


Figure I-15a

ICC of a Polytomous Item With a=1.5, b=0, d0=0,d1=2, d2=0, d3=-2

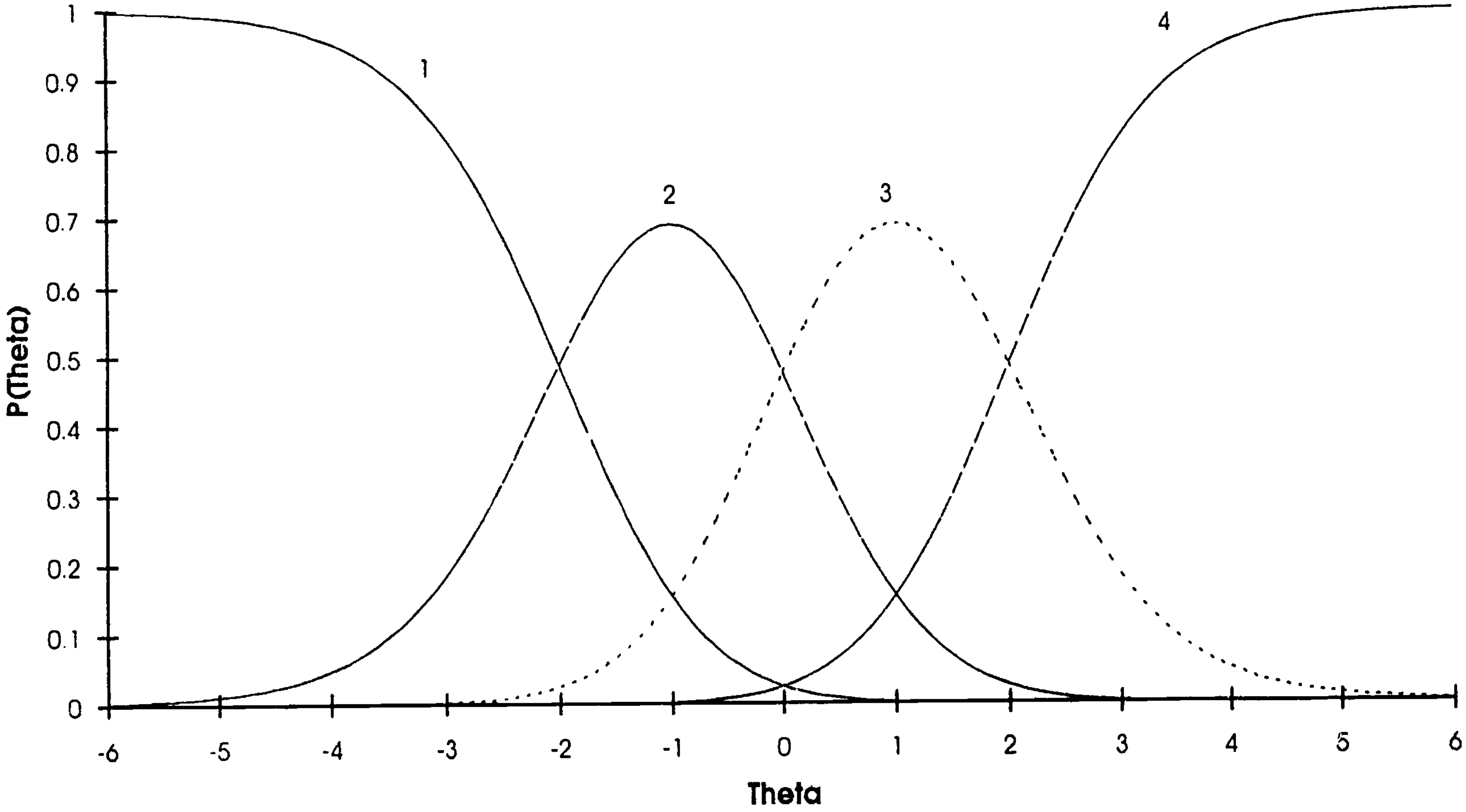


Figure I-16

Comparison Between the "Old" and "New" Information Functions Using Hypothetical Item Parameters a=1.5, b=-1, d0=0, d1=2, d2=0, and d3=-2, in Case of Four Response Categories

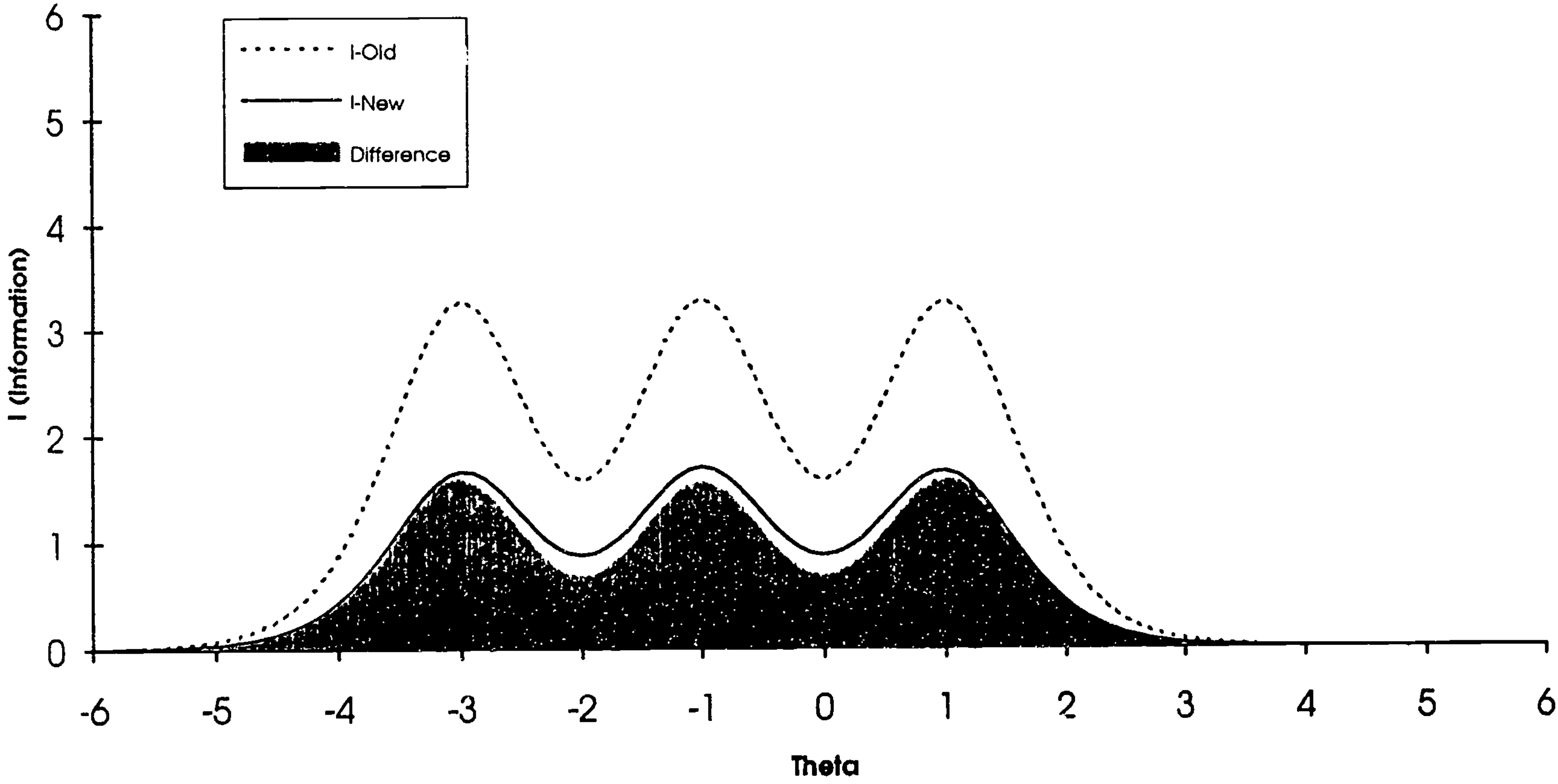


Figure I-17

Comparison Between the "Old" and "New" Information Functions Using Hypothetical Item Parameters a=1.5, b=1, d0=0, d1=2, d2=0, and d3=-2, in Case of Four Response Categories

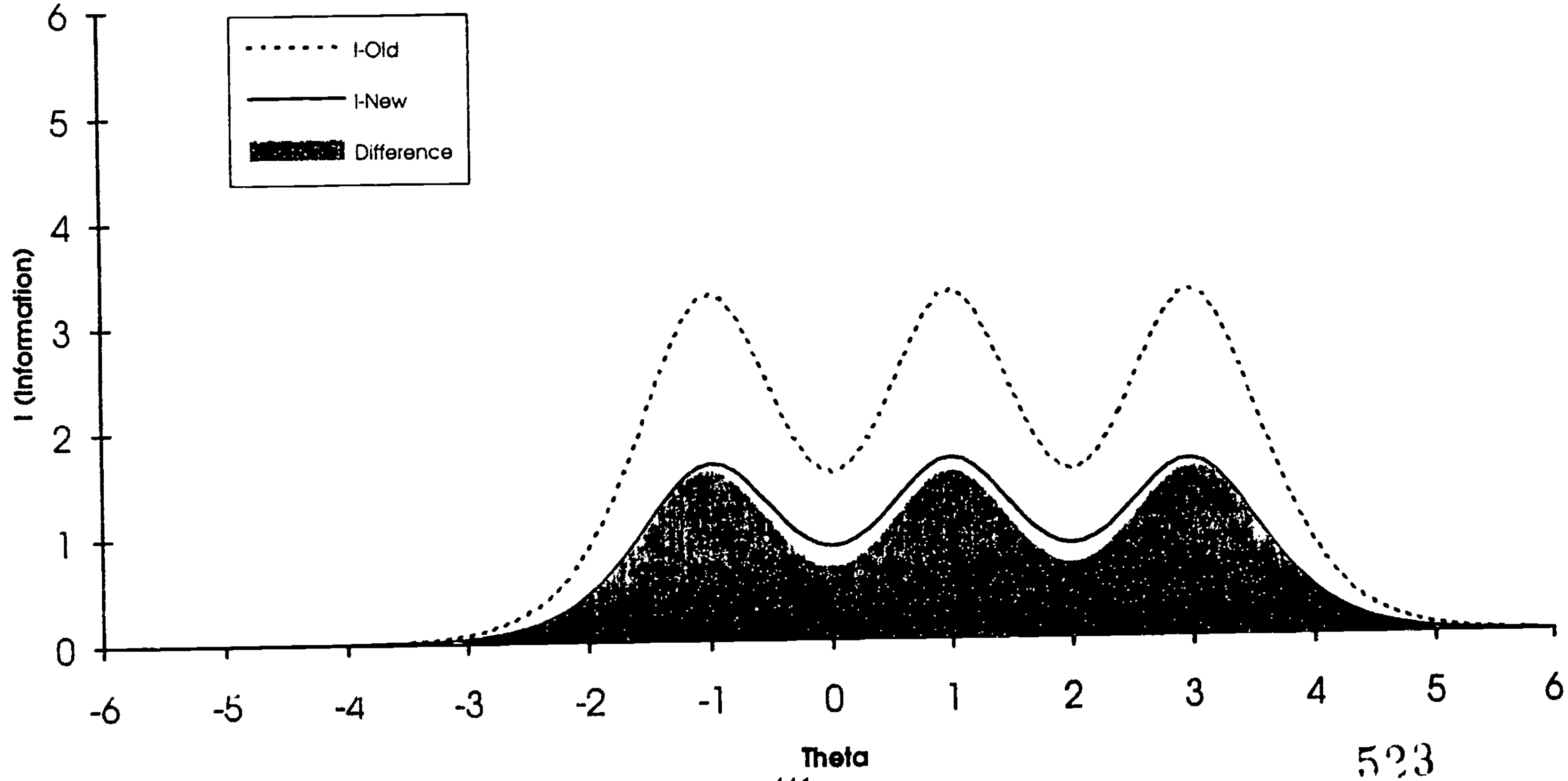


Figure I-18

Comparison Between the "Old" and "New" Information Functions
Using Hypothetical Item Parameters a=1.5, b=0, d0=0, d1=1, d2=0, and d3=-1, in
Case of Four Response Categories

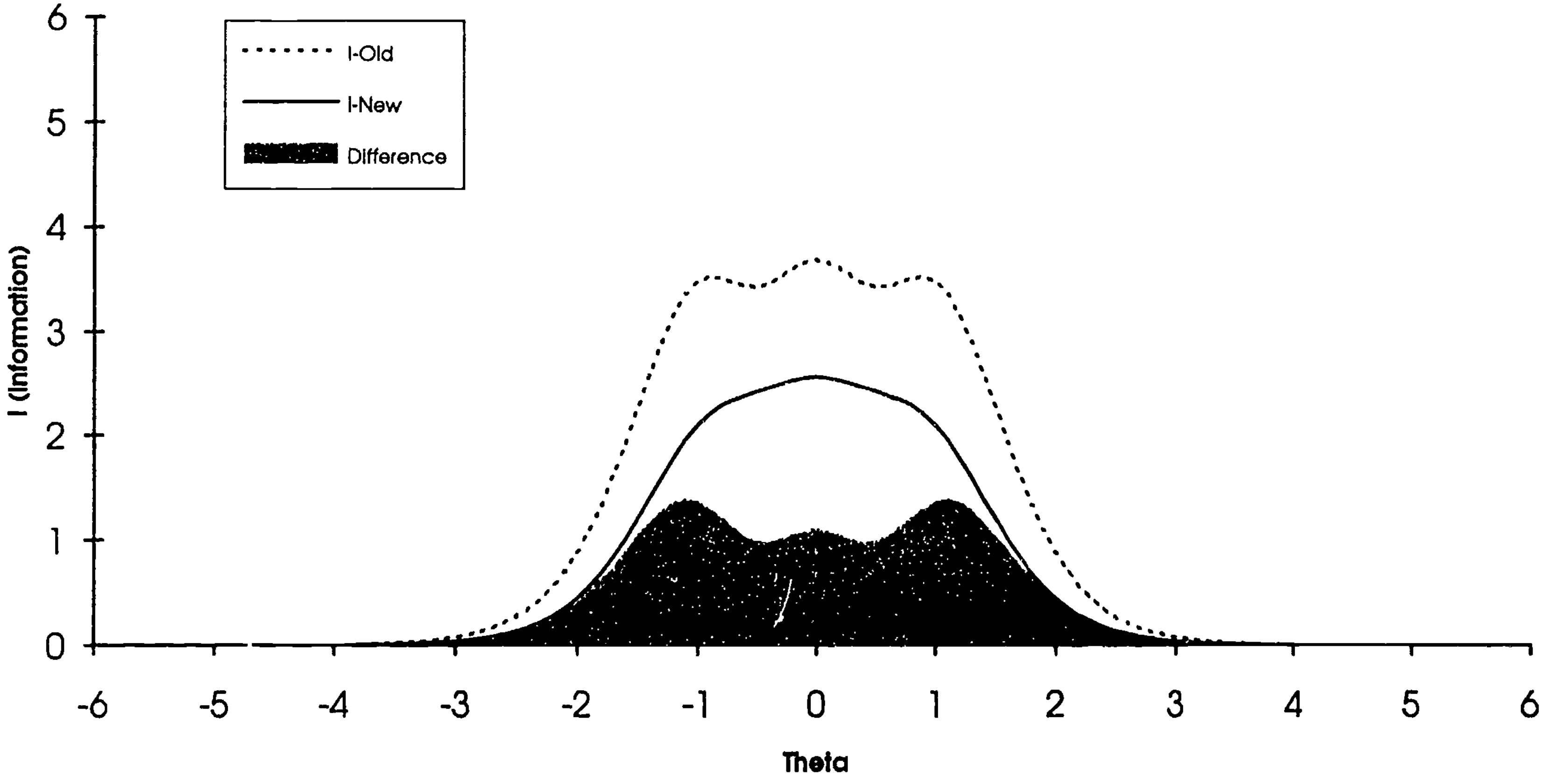


Figure I-18a

ICC of a Polytomous Item With a=1.5, b=0, d0=0,d1=1, d2=0, d3=-1

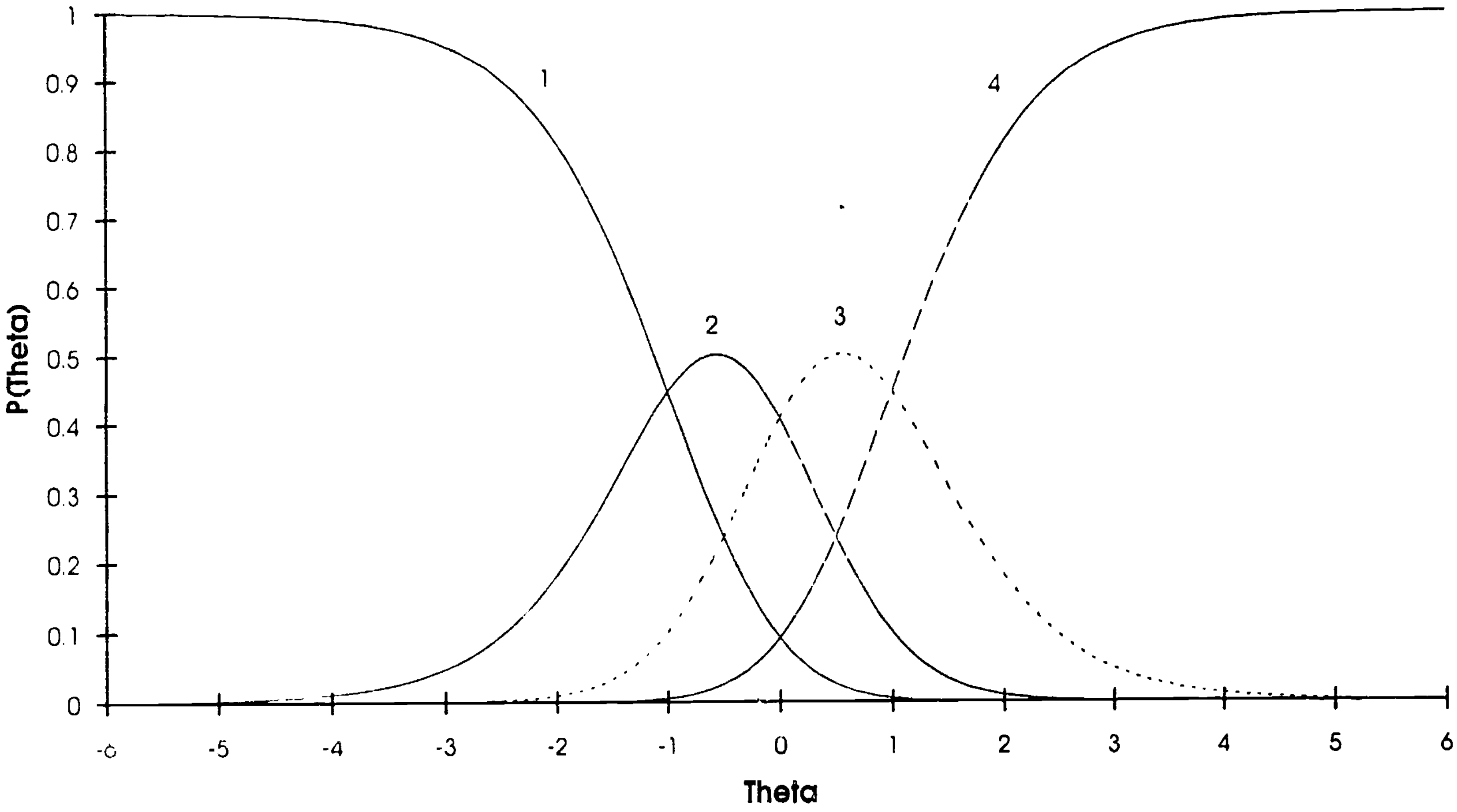


Table I-5

Item Parameters Used for Figures I-12 through I-20 and the Sum of the Difference Between the Old and the New Information Weighting Functions

Figure	a	b	d_0	d_1	d_2	d_3	Sum
12	0.2	0	0	2	0	-2	1.37
13	0.5	0	0	2	0	-2	9.67
14	1.0	0	0	2	0	-2	38.52
15	1.5	0	0	2	0	-2	70.47
16	1.5	-1	0	2	0	-2	70.46
17	1.5	1	0	2	0	-2	70.40
18	1.5	0	0	1	0	-1	46.30
19	1.5	0	0	.52	0	-.52	22.40
20	1.5	0	0	.1	0	-.1	21.46

Figure I-19

Comparison Between the "Old" and "New" Information Functions Using Hypothetical Item Parameters a=1.5, b=0, d0=0, d1=0.52, d2=0, and d3=-0.52, in Case of Four Response Categories

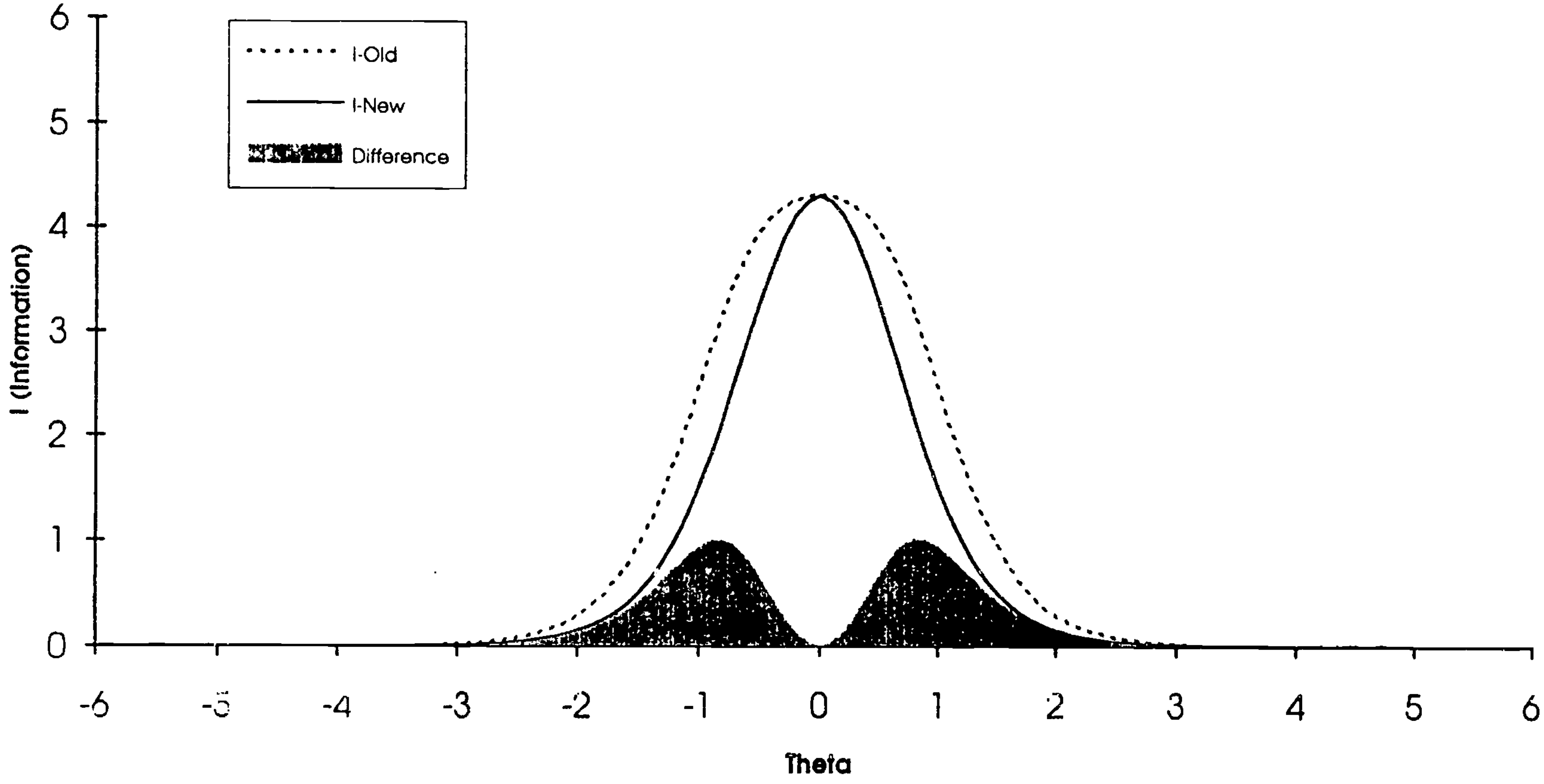


Figure I-19a

Item Characteristic Curves of a=1.5, b=0, d0=0, d1=0.52, d2=0, and d3=-0.52

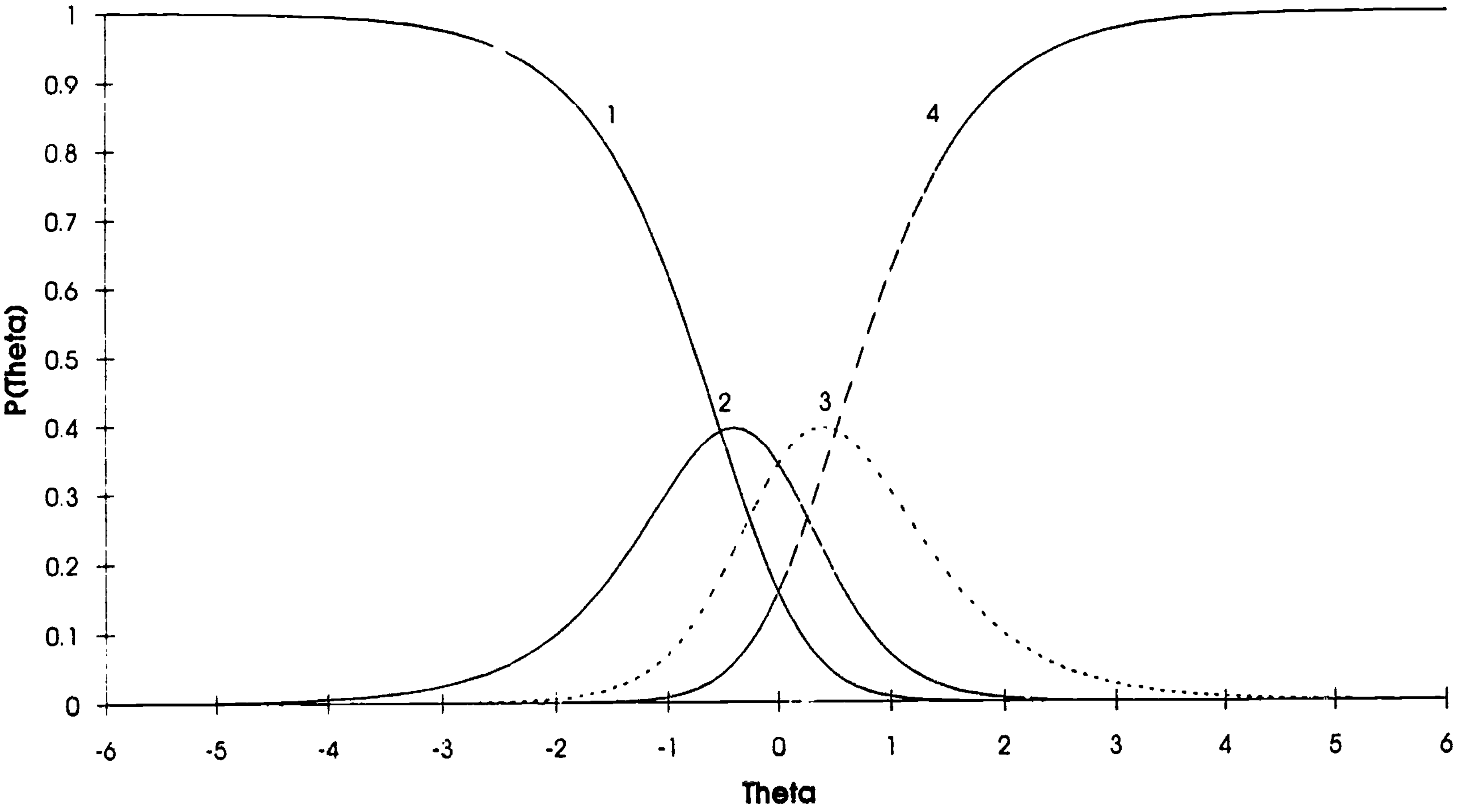


Figure I-20

Comparison Between the "Old" and "New" Information Functions Using Hypothetical Item Parameters a=1.5, b=0, d0=0, d1=0.1, d2=0, and d3=-0.1, in Case of Four Response Categories

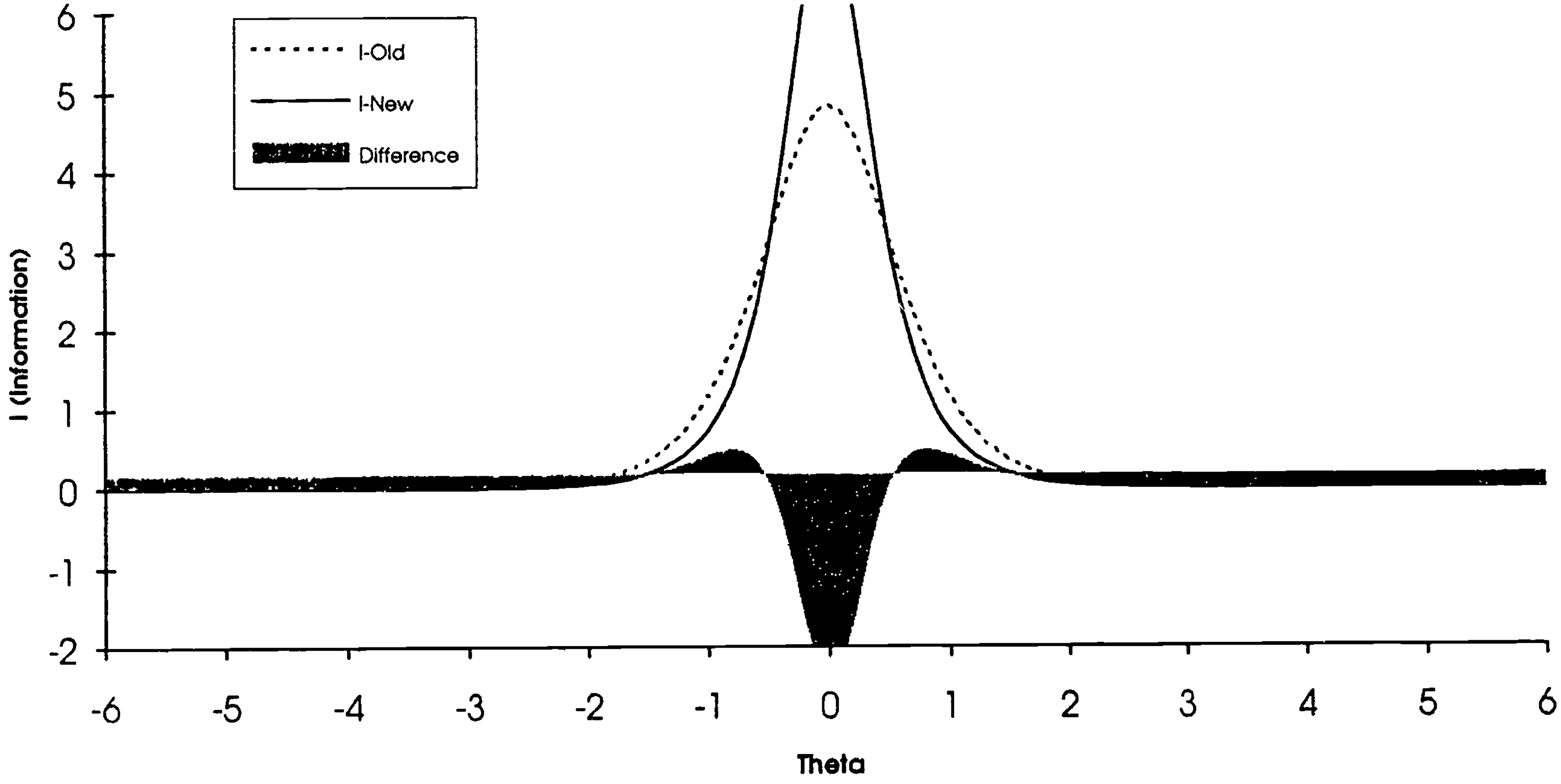
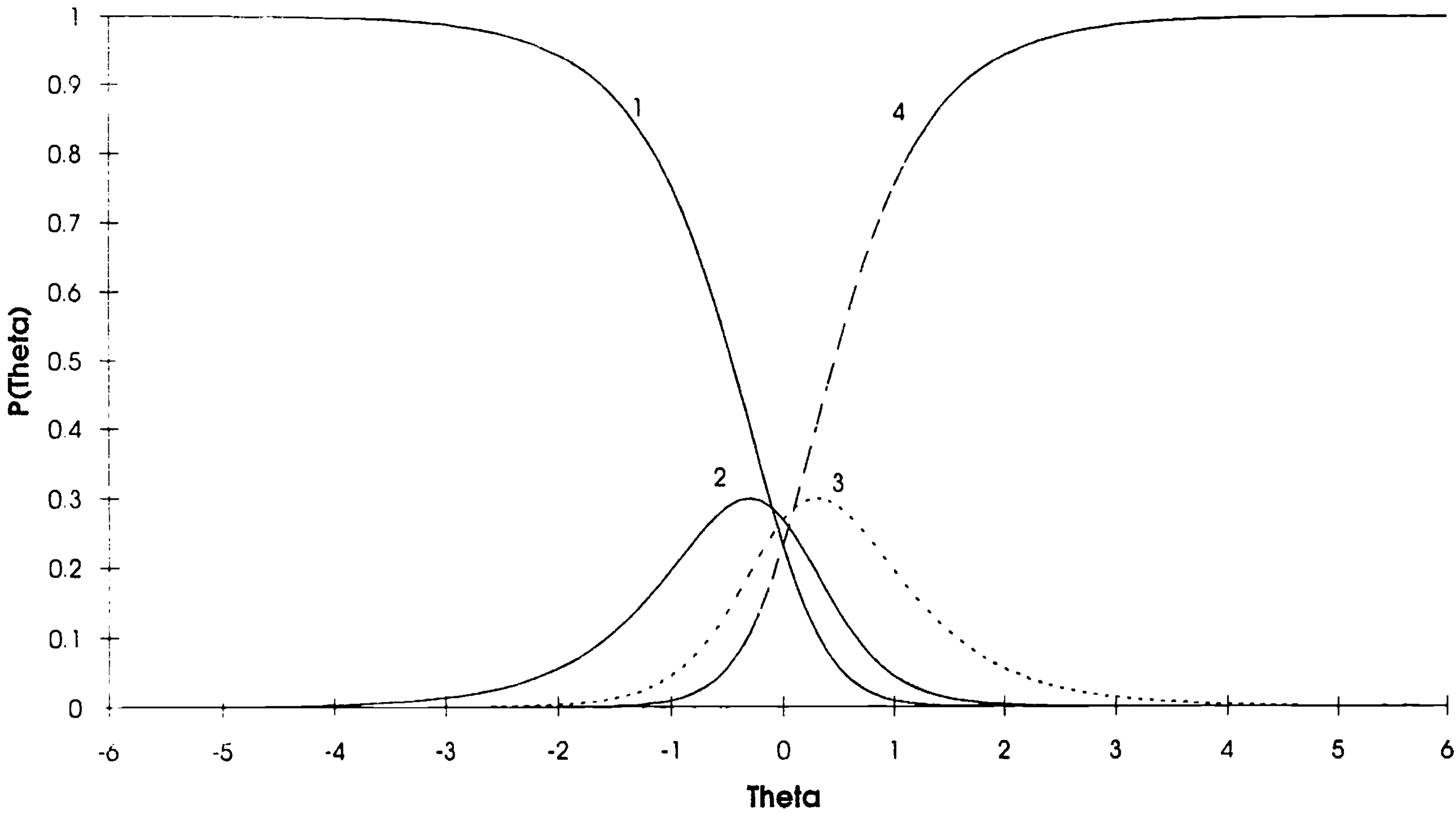


Figure I-20a

Item Characteristic Curves of a=1.5, b=0, d0=0, d1=0.1, d2=0, and d3=-0.1



REFERENCES CITED IN TEXT

REFERENCES CITED IN TEXT

Allen, N. L., & Donoghue, J. R. (1994, April). *Differential item functioning based on complex samples of dichotomous and polytomous items.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Allen, N. L., & Donoghue, J. R. (in press). Anplying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement.*

Allen, N. L., Mazzeo J., Isham, S. P., Fong, Y. F., & Bowker, D. W. (1994). Data analysis and scaling for the 1992 trial state assessment in reading. In E. G. Johnson, J. Mazzeo, & D. Kline, *Technical report of the NAEP 1992 trial state assessment program in reading.* Washington, DC: National Center for Education Statistics.

American College Testing. (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing: A technical report on reliability and validity.* Iowa City, IA: Author.

Andersen, E. B. (1980). Comparing latent distributions. *Psychometrika, 45*, 121-134.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508 - 600). Washington, DC: American Council on Education.

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement, 26(2)*, 163-175.

Beaton, A. E., & Johnson, E. G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15*, 9-38.

Beaton, A. E., & Zwick, R. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly.* (No. 17-TR-21) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.

Bourque, M. L., & Garrison, H. H. (1991). *The levels of mathematics achievement. Vol. I, national and state summaries.* Washington, DC: National Assessment Governing Board.

Brennan, R. L. (in press). Standard setting from the perspective of generalizability theory. In *Proceedings of the Joint Conference on Standard Setting for Large-scale Assessments.* Washington, DC: U. S. Government Printing Office.

Cizek, G. (1993). *Reactions to National Academy report, "Setting performance standards for student achievement"*. Washington, DC: National Assessment Governing Board.

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley & Sons.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220.

Curry, L. (1987, April). *Group decision process in setting cut-off scores.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Educational Testing Service (1987). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Educational Testing Service (1992). *Innovations and ingenuity: A foundation for the future. Application for cooperative agreement for NAEP*. CFDA Number: 84.999E. Princeton, NJ: Author.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Staistics, 7*, 1-26.

Engelen, R. J. H. (1987). *Semiparametric estimation in the Rasch model.* Research Report 87-1. Twente, the Netherlands: Department of Education, University of Twente.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research, 59*, 315-328.

Friedman, C. B., & Ho, K. T. (1990, April). *Interjudge consensus and intrajudge consistency: Is it possible to have both on standard setting?* Paper presented at the annual meeting of the National Council for Measurement in Education, Boston, MA.

Hambleton, R. K., & Bourque, M. L. (1991). *The levels of mathematics achievement. Vol. II, technical report.* Washington, DC: National Assessment Governing Board.

Hoijtink, H. (1991). *Estimating the parameters of linear models with a latent dependent variable by nonparametric maximum likelihood.* Research Bulletin HB-91-1040-EX. Groningen, The Netherlands: Psychological Institute, University of Groningen.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Ip, E. H. S., & Allen, N. L. (1995). *A study on the effect of monitoring sessions in the 1994 NAEP Trial State Assessment.* Technical report. Princeton, NJ: Educational Testing Service.

Jerry, L. (1995). *The NAEP computer-generated reporting system for the 1994 Trial State Assessment.* Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., & Allen, N. L. (1992). *The NAEP 1990 technical report* (No. 21-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., & Carlson, J. E. (1994). *The NAEP 1992 technical report* (No. 23-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., & Freund, D. S. (1994, April). *Scoring of performance items using image processing: Comparability study with paper-and-pencil approach.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17,* 175-190.

Johnson, E. G., Rust, K. F., & Wallace, L. (1994). Weighting procedures and estimation of sampling variance. In E. Johnson & J. Carlson (Eds.), *The NAEP 1992 technical report.* Princeton, NJ: Educational Testing Service, National Center for Education Statistics.

Kane, M. (1993). *Comments on the NAE Evaluation of the NAGB achievement levels.* Washington, DC: National Assessment Governing Board.

Keyfitz, N. (1951). Sampling with probability proportional to size; adjustment for changes in probabilities. *Journal of the American Statistical Association*, *46*, 105-109.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association, 73*, 805-811.

Langer, J. A. (1989). *The process of understanding literature.* (Technical report No. 2.1.) Albany: State University of New York, Center for the Learning and Teaching of Literature.

Langer, J. A. (1990). The process of understanding: Reading for literary and informative purposes. *Research in the Teaching of English, 24,* 229-257.

Lindsey, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86*, 96-107.

Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data. *American Statistician, 37*, 218-220.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley & Sons.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Mazzeo, J. (1991). Data analysis and scaling. In S. L. Koffler, *The technical report of NAEP's 1990 Trial State Assessment program* (No. ST-21-01). Washington, DC: National Center for Education Statistics.

Mazzeo, J., Chang, H., Kulick, E., Fong, Y. F., & Grima, A. (1993). Data analysis and scaling for the 1992 Trial State Assessment in mathematics. In E. G. Johnson, J. Mazzeo, & D. L. Kline, *Technical report of the NAEP 1992 Trial State Assessment program in mathematics.* Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Mazzeo, J., Johnson, E. G., Bowker, D., & Fong, Y. F. (1992). *The use of collateral information in proficiency estimation for the Trial State Assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R. J. (1990). Scaling procedures. In E.G. Johnson and R. Zwick, *Focusing the new design: The NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993-997.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17(2)*, 131-154.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A.E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (No. 15-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Mislevy, R. J., & Stocking, M. L. (1987). *A consumer's guide to LOGIST and BILOG.* (ETS Research Report 87-43). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Wu, P-K. (1988). *Inferring examinee ability when some item responses are missing* (ETS Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2)*, 159-176.

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data.* Chicago, IL: Scientific Software, Inc.

National Academy of Education. (1993a). *Setting performance standards for student achievement.* Stanford, CA: Author.

National Academy of Education. (1993b). *The trial state assessment: prospects and realities.* Stanford, CA: Author.

National Assessment Governing Board (1989). *Setting achievement goals on NAEP, a draft policy statement*. Washington, DC: Author.

Petersen, N. (1988). *DIF procedures for use in statistical analysis.* Internal memorandum.

Rogers, A. M. (1991). *NAEP-MGROUP: Enhanced version of Sheehan's software for the estimation of group effects in multivariate models* [Computer program]. Princeton, NJ: Educational Testing Service.

Rogers, A. M., Barone, J. L., & Kline, D. L. (1995). *A guide to the NAEP item information database*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Rogers, A. M., Kline, D. L., Barone, J. L., Mychajlowycz, A. W., & Forer, D. C. (1989). *A primer for the NAEP item information database.* Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Rosenbaum, P. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rubin, D. B. (1991). EM and beyond. *Psychometrika, 56*, 241-254.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Rust, K. R., & Johnson, E. G. (1992). Sampling and weighting in the national assessment. *Journal of Educational Statistics, 17(2)*, 111-129.

Sheehan, K. M. (1985). *M-GROUP: Estimation of group effects in multivariate models* [Computer program] Princeton, NJ: Educational Testing Service.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement*, *25*, 259-274.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician*, *40*, 106-108.

Stone, C. A., Ankenmann, R. D., Lane, S., & Liu, M. (1993). *Scaling QUASAR's performance assessments.* Paper presented at the annual meeting of the American Educational Research Association.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, *82*, 528-550.

Thissen D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Thomas, N. (1992). *Higher order asymptotic corrections applied in an EM algorithm for estimating educational proficiencies*. Unpublished manuscript.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing Co.

Westat, Inc. (1995). *Report on data collection activities for the 1994 National Assessment of Educational Progress.* Rockville, MD: Author.

Wainer, H. (1974). The suspended rootogram and other visual displays: An empirical validation. *The American Statistician, 28(4)*, 143-145.

Wingersky, M., Kaplan, B. A., & Beaton, A. E. (1987). Joint estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report*. (No 15-TR-20) Princeton, NJ: National Association of Educational Progress, Educational Testing Service.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17(2)*, 155-173.

Zieky, M. (1993). Practical questions in the use of DIF statistics. In P. W. Holland & H. Wair.er (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zwinderman, A. H. (1991). Logistic regression Rasch models. *Psychometrika, 56*, 589-600.

UNITED STATES
DEPARTMENT OF EDUCATION
WASHINGTON, DC 20208-5653

NCES 96–116