ED 406 453 TM 026 832

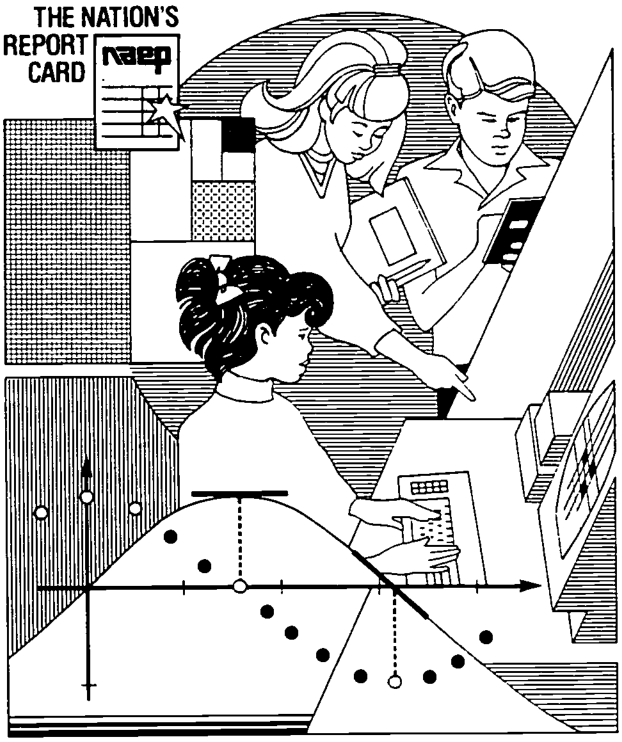
AUTHOR	Koffler, Stephen L.; And Others
TITLE	The Technical Report of NAEP's 1990 Trial State Assessment Program.
INSTITUTION	Educational Testing Service, Princeton, N.J.; National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	ISBN-0-88685-113-0; NAEP-21-ST-01
PUB DATE	Apr 91
NOTE	304p.
AVAILABLE FROM	National Assessment of Educational Progress, Educational Testing Service, Rosedale Road, Princeton, NJ 08541-0001.
PUB TYPE	Books (010) -- Reports - Evaluative/Feasibility (142) -- Statistical Data (110)
EDRS PRICE	MF01/PC13 Plus Postage.
DESCRIPTORS	*Data Analysis; *Educational Testing; Grade 8; Junior High Schools; Mathematics; Mathematics Achievement; National Surveys; *Research Design; Research Methodology; Research Reports; Sampling; Scaling; State Programs; Student Attitudes; Tables (Data); *Test Construction; *Testing Programs
IDENTIFIERS	*National Assessment of Educational Progress; *Trial State Assessment (NAEP)

ABSTRACT

This report documents the design and data analysis procedures of the Trial State Assessment Program of the National Assessment of Educational Progress (NAEP). Today the NAEP is the only survey using advanced plausible values methodology that uses a multiple imputation procedure in a psychometric context. The 1990 Trial State Assessment collected information on the mathematics knowledge, skills, understanding, and attitudes of a representative sample of more than 100,000 eighth graders in public schools in 37 states, the District of Columbia, and 2 territories. Following a Foreword by Gary W. Phillips, this report includes the following chapters: (1) "Overview: The Design, Implementation, and Analysis of the Trial State Assessment Program" (Eugene G. Johnson and Stephen L. Koffler); (2) "Developing the Objectives, Cognitive Items, Background Questions, and Assessment Instruments" (Stephen L. Koffler); (3) "Sample Design and Collection" (Jim Bethel, Keith Rust, and Jacqueline Severynse); (4) "State and School Cooperation" (Nancy Caldwell); (5) "Field Administration" (Nancy Caldwell); (6) "Processing Assessment Materials" (Dianne Smrdel, Lavonne Mohn, Linda Reynolds, and Brad Thayer); (7) "Creation of the Database and Evaluation of the Quality Control of Data Entry" (John J. Ferris, David S. Freund, and Alfred M. Rogers); (8) "Weighting Procedures and Variance Estimation" (Jim Behtel, Keith Rust, and Jacqueline Severynse); (9) "Theoretical Background and Philosophy of NAEP Scaling Procedures" (Eugene G. Johnson and Robert J. Mislevy); (10) "Data Analysis and Scaling" (John Mazzeo); and (11) "Conventions Used in Reporting the Results" (John Mazzeo). Six appendixes present further detail about test construction, reporting subgroups, and the anchoring process. (Contains 32 tables and 28 figures.) (SLD)

NATIONAL CENTER FOR EDUCATION STATISTICS

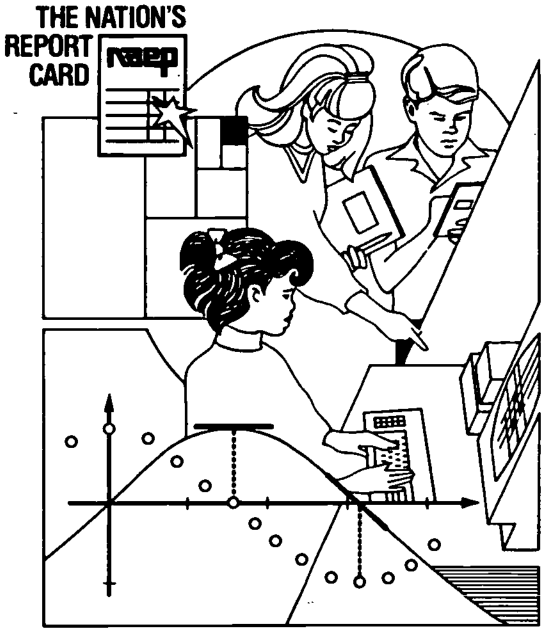
The Technical Report of NAEP's 1990 Trial State Assessment Program



Prepareo by Educational Testing Service under Contract with the National Center for Education Statistics
Office of Educational Research and Improvement • U.S. Department of Education

NATIONAL CENTER FOR EDUCATION STATISTICS

The Technical Report of NAEP's 1990 Trial State Assessment Program



Stephen L. Koffler
in collaboration with
Jim Bethel
Nancy Caldwell
David S. Freund
John J. Ferris
Eugene G. Johnson
John Mazzeo
Robert J. Mislevy
Lavonne Mohn
Linda Reynolds
Alfred M. Rogers
Keith Rust
Jacqueline Severynse
Dianne Smrdel
Brad Thayer

with a Foreword by Gary W. Phillips

Report No. 21-ST01 April 1991

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

This report, No. 21-ST-01, can be ordered from the National Assessment of Educational Progress, Educational Testing Service, Rosedale Road, Princeton, New Jersey 08541-0001.

Library of Congress, Catalog Card Number: 91-61344

ISBN: 0-88685-113-0

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.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

ACKNOWLEDGMENTS

The design, development, analysis, and reporting of the first 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 Trial State Assessment Program. It also provides insight into the rationale behind the technical decisions made about the program. The development of this 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 Trial State Assessment Program.

Special recognition is due the Council of Chief State School Officers for its considerable contributions to the program, especially its management of the National Assessment Planning Project. That project resulted in the mathematics framework and objectives for the assessment and recommendations about reporting the results of the program. In particular, we note the significant contributions of Ramsay Selden, Director of the State Education Assessment Center for the Council of Chief State School Officers and the members of the Steering Committee, Mathematics Objectives Committee, and Analysis and Reports Committee of the National Assessment Planning Project.

The Trial State Assessment was funded through the National Center for Education Statistics in the Office of Educational Research and Improvement of the U.S. Department of Education. Emerson Elliott, Acting Commissioner, NCES provided consistent support and guidance. The staff -- particularly Gary Phillips, Eugene Owen, Stephen Gorman, Maureen Treacy, and Raul Garza -- worked closely and collegially with ETS, Westat, and NCS staff and played a crucial role in all aspects of the program.

The members of the National Assessment Governing Board (NAGB) and NAGB staff also deserve credit for their advice and guidance.

We owe a great deal to the individuals who constituted the Mathematics Item Development Panel and Mathematics Scale Anchoring Panel. These people -- from school districts, colleges and universities, and State Education Agencies -- worked tirelessly to help ETS staff develop the assessment and a framework for interpreting the results.

Under the NAEP contract to Educational Testing Service, Archie Lapointe served as the project director and Ina Mullis as the deputy director. Statistical and psychometric activities for the Trial State Assessment were led by John Mazzeo, with consultation from Eugene Johnson and Donald Rock. John Barone managed the data analysis activities; Jules Goodison, the operational aspects; Walter MacDonald and Chancey Jones, test development; David Hobson, the fiscal aspects; and Stephen Koffler, state services. Kent Ashworth managed communications with the public and participating schools. Sampling and data collection activities were carried out by Westat under the supervision of Renee Slobasky, Keith Rust, Nancy Caldwell, and the late Morris Hansen. The printing, distribution, and processing of the Trial State Assessment materials were the responsibility of National Computer Systems, under the direction of John O'Neill and Lynn Zaback.

The overall responsibility for the Trial State Assessment Program lay with NAEP management at ETS. Of particular note is the contribution of Ina Mullis who deserves special recognition. She has had considerable influence on the present and future directions of the Trial State Assessment Program, and of NAEP itself.

The design and data analysis of the Trial State Assessment Program was primarily the responsibility of the NAEP research and data analysis staff, with significant contributions from the NAEP management, Westat, and NCS staffs. Particular recognition is due to Nancy Allen, Albert Beaton, Eugene Johnson, John Mazzeo, Robert Mislevy, Eiji Muraki, Donald Rock, Kentaro Yamamoto, and Rebecca Zwick.

The Data Analysis staff, under the outstanding leadership of John Barone, performed exceptionally in carrying out their responsibilities for developing the database and producing numerous data analyses. John Ferris, David Freund, Bruce Kaplan, Debra Kline, Edward Kulick, Philip Leung, Jennifer Nelson, and Alfred Rogers deserve special commendation for their innovation in developing a unique and comprehensive system and their tireless efforts in producing analyses and reports. They were ably assisted in these efforts by Drew Bowker, Laura McCamley, and Craig Pizzuti.

The staff of Westat, Inc. continued to contribute their exceptional talents in all areas of sampling design and data collection. Particular recognition is due to Renee Slobasky and Nancy Caldwell for supervising the field operations and to Keith Rust for developing and supervising the sampling design. Morris Hansen, whose untimely death in 1990 saddened us all, had considerable influence on all aspects of the Trial State Assessment Program.

Critical to the program was the contribution of National Computer Systems, Inc. which was responsible for the printing, distribution, and processing of the Trial State Assessment materials. The leadership roles of John O'Neill and Lynn Zaback are especially acknowledged.

We are especially appreciative of the contributions of the NAEP Design and Analysis Committee who provided sound advice on the technical aspects of the program.

Kent Ashworth was responsible for coordinating the cover design and final printing of this report.

Grateful appreciation is expressed to Marilyn Brennan for all of her assistance, her word processing skills, and for her patience.

Special thanks are also due to many individuals for their invaluable assistance in reviewing the reports, especially the editors who improved the text and the data analysts who checked the accuracy of the data.

FOREWORD

This technical report represents a landmark accomplishment, not only for the NAEP project, but for the broader statistical and psychometric community. It summarizes some of the most sophisticated statistical methodology used in any survey or testing program in the United States. In its 20 year history, the NAEP has employed state-of-the-art techniques such as matrix-sampling and item response theory models. Today it is the only survey using the advanced plausible values methodology which uses a multiple imputation procedure in a psychometric context.

Some of the most exciting times in the project were in the early stages in which we had to make the major technical decisions that determined the project's overall design. Now that we are through the 1990 portion of the Trial State Assessment, it is amazing at how many of those initial decisions appear to have been the proper ones to have made. Some of the initial decisions included: 1)expanding the consensus process through which objectives are determined; 2)continuing the use of focused BIB spiraling, item response theory models, and plausible values; 3)keeping the national and Trial State Assessment samples unduplicated; 4)not attempting to equate to previous trend lines; 5)doing separate stratifications and conditioning in each of the State samples; 6)making each State sample have power similar to the past regional samples (this is how the sample sizes for the State were determined); 7) equating the aggregate of the State samples to the national scale (and doing this via an augmented national sample that also was representative of the aggregate of the States); 8) using the winter half of the national sample as the national estimate against which the States would be compared; 9)limiting the State samples to public schools; and 10)using power rules to determine which subgroup comparisons were supported by sufficient sample sizes (this became the "rule of 62").

Some of our planned attempts did not work. We tried to develop an extensive opportunity-to-learn questionnaire; however, based on the field testing, the questionnaire did not produce reliable data nor did the data correlate with achievement. As a result, that questionnaire was dropped after the field test. We tried to put the estimation and higher order thinking skills items on the national scale, but determined that they had to be scaled separately. We had hoped to have the achievement level standards being set by the National Assessment Governing Board ready in time for the June 6, 1991, public release of the Trial State Assessment results; however, they were not completed in time.

NAEP continues to take on new technical challenges. Just a few years ago the statistical problems associated with State-by-State comparisons seemed insurmountable. After three years of contemplation and advice from countless people, the statistical issues appear to be solved.

As formidable as they were, those statistical issues of the past are dwarfed by the seemingly intractable problems of the future. In the future, NAEP will need new psychometric procedures for performance-type items. In 1992 the Trial State Assessment will not only have to compare more states -- in more grades and more subjects -- but will have to measure trend over time on a state-by-state basis. It will also have to report on the results in a shorter time frame than for 1990. In addition, NAEP will have to determine how to accommodate the achievement level standards being developed by the National Assessment Governing Board. The biggest technical challenge of all may be the interface of NAEP and a possible new national examination system currently being discussed around the country.

The NAEP project is not only characterized by its elegant statistical procedures; it is also noted for the dedicated professionalism of its staff. In hundreds of hours of technical advisory committees, I have not seen a single instance in which truth, honesty and reason were compromised. It is the stubborn insistence that surveys are scientific activities and the relentless quest for improved methodology that has made NAEP credible for more than 20 years.

The NAEP statistical procedures are the product of the best thinking by some of the best statisticians in the country. One of the giants on whose shoulders the project stands is Morris Hansen. He was the father of modern day survey sampling theory and was the godfather of NAEP.

Had Morris Hansen lived to see this document, he would have been very proud. We are very proud to affectionately dedicate this report to him.

Gary W. Phillips
Acting Associate Commissioner
National Center for Education Statistics

Chapter 1

OVERVIEW:

THE DESIGN, IMPLEMENTATION, AND ANALYSIS OF THE TRIAL STATE ASSESSMENT PROGRAM

Eugene G. Johnson and Stephen L. Koffler

Educational Testing Service

"The National Assessment shall develop a trial mathematics assessment survey instrument for the 8th grade and shall conduct a demonstration of the instrument in 1990 in States which wish to participate, with the purpose of determining whether such an assessment yields valid, reliable State representative data." (P.L. 100-297)

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 addition to continuing the national assessments that NAEP had conducted since its inception. As a result of the legislation, the first part of the Trial State Assessment Program was conducted in 1990. National assessments in mathematics, reading, writing, and science were conducted simultaneously in 1990 at grades four, eight and twelve. The 1990 Trial State Assessment Program collected information on the mathematics knowledge, skills, understanding, and attitudes of a representative sample of eighth-grade students in public schools in 37 states, the District of Columbia, and two territories.

Table 1-1 lists the jurisdictions that participated in the 1990 Trial State Assessment Program. The information was collected from more than 100,000 students in those jurisdictions based on a complex sample survey. The students who were assessed were administered one of seven mathematics assessment booklets also used in NAEP's 1990 national mathematics assessment. The mathematics framework and objectives established to guide the both the Trial State Assessment and national assessment were developed for NAEP through a consensus project of the Council of Chief State School Officers, funded by the National Center for Education Statistics and the National Science Foundation. The framework and objectives were also used for the 1990 national mathematics assessment. In addition, questionnaires completed by the students, their mathematics teacher, and principal or other school administrator provided an abundance of contextual data within which to interpret the mathematics results.

Table 1-1

Jurisdictions Participating in the
1990 Trial State Assessment Program

Jurisdictions			
Alabama	Guam	Minnesota	Oklahoma
Arizona	Hawaii	Montana	Oregon
Arkansas	Idaho	Nebraska	Pennsylvania
California	Illinois	New Hampshire	Rhode Island
Colorado	Indiana	New Jersey	Texas
Connecticut	Iowa	New York	Virginia
Delaware	Kentucky	New Mexico	Virgin Islands
District of Columbia	Louisiana	North Carolina	West Virginia
Florida	Maryland	North Dakota	Wisconsin
Georgia	Michigan	Ohio	Wyoming

The purpose of this report is to provide the technical information about the Trial State Assessment Program. 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 stage through the creation of the database used for analysis through the analysis and reporting. The report describes in detail the development of the cognitive and background questions, the field procedures, the creation of the database for analysis (from receipt of the assessment materials through scanning, scoring, and creation of the database), and the methods and procedures for sampling, analysis, and reporting. It does not provide the results of the assessment -- rather it provides information on how those results were derived.

Educational Testing Service (ETS) was the contractor for the 1990 NAEP program, including the Trial State Assessment Program. 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. Westat, Inc., and National Computer Systems (NCS) were subcontractors to ETS. Westat was responsible for all aspects of sampling and of field operations, while NCS was responsible for printing, distribution, and receipt of all assessment materials, and for scanning, and professional scoring

This technical report supports other reports that have been prepared for the 1990 Trial State Assessment Program, including:

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

- A *Composite Report* that provides the data for all of the 40 jurisdictions that participated in the Trial State Assessment Program as well as the results from the 1990 national mathematics assessment. There also is an accompanying *Executive Summary* that provides the highlights from the Composite Report.

- An *Almanac* for each jurisdiction, that contains a detailed breakdown of the proficiency data according to the responses to the student, teacher, and school questionnaires for the population as a whole and for important subgroups of the population. There are five sections to each Almanac:

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

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

  - ▲ *The School Questionnaire Section* provides a breakdown of the proficiency data according to the principals' (or other administrators') responses to questions in the School Characteristics and Policies questionnaire.

  - ▲ *The Subscale Section* provides a breakdown of selected questions from the questionnaires according to each of the five content-area subscales measured in the assessment.[1]

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

ORGANIZATION OF THE TECHNICAL REPORT

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

[1] The five content areas were Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions.

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

- Section 1.3 summarizes the development of the mathematics 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 and background questions to assessment booklets and describes the focused-BIB spiral design. A complete description is provided in Chapter 2.

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

- Section 1.6 summarizes the field administration procedures including securing school cooperation, training administrators, administering the assessment, and conducting quality control. Further details appear in Chapters 4 and 5.

- 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 database for analysis. Chapters 6 and 7 provide a detailed description of the process.

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

- Section 1.9 summarizes the procedures used to weight the data from the assessment and to obtain estimates of the sampling variability of subpopulation estimates. Chapter 8 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 10.

- Section 1.11 describes the item response theory subscales and the overall mathematics 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 9 with details of the scaling process in Chapter 10.

- Section 1.12 provides an overview of the linking of the scaled results from the Trial State Assessment to those from the national mathematics assessment. Details of the linking process appear in Chapter 10.

- Section 1.13 describes the reporting of the assessment results with further details supplied in Chapter 11.

- A glossary of terms and list of references are included. Finally, six appendices provide information about 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 mathematics items, the reporting subgroups, composite and derived common background and reporting variables, and a description of the mathematics scale anchoring process.

1.2 DESIGN OF THE TRIAL STATE ASSESSMENT

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

- Participation was voluntary.

- Only eighth-grade students in public schools were assessed. Students in private or parochial schools were not included in the program. A representative sample of schools was selected in each participating state and territory, and students were randomly sampled within schools.

- Mathematics was assessed at the eighth-grade level.

- The mathematics items used in the Trial State Assessment were also used in the grade 8/age 13 national assessment and contained open-ended items and items requiring scientific calculators and protractors/rulers. The total pool of mathematics items was divided into seven blocks. Each block was timed so that the student had no more than 15 minutes to complete the block.

- Background questionnaires given to the students, the students' mathematics teacher, and the principal or other administrator in the schools provided for rich contextual information. The background questionnaires for the Trial State Assessment were identical to those used in the grade 8/age 13 national assessment.

- A complex form of matrix sampling called a balanced incomplete block (BIB) spiraling design was used. With BIB spiraling, students in an assessment session received different booklets, resulting in a more efficient sample. This design also reduced student burden and provided for greater mathematics content coverage than would have been possible had every student been administered the identical set of items.

- Each assessed student was assigned a mathematics booklet that contained two five-minute background questionnaires and three of the seven 15-minute blocks containing mathematics items. There were seven different booklets assembled. The assessment time for each student was approximately 55 minutes.

- The assessments took place in the five-week period between February 2 and March 5, 1990. One-fourth of the schools in each state were assessed each week throughout the first four weeks with the fifth week being reserved for the scheduling of makeup sessions.

- Data collection, by law, was the responsibility of each participating state and jurisdiction.

- Security and uniform assessment administration were high priorities. Extensive training was conducted to assure that the administration of the assessment would be under standard, uniform procedures. Fifty percent of the assessment sessions were monitored by the contractor's staff.

1.3 DEVELOPMENT OF OBJECTIVES, ITEMS, AND BACKGROUND QUESTIONS

Similar to all previous NAEP assessments, the objectives for the Trial State Assessment were developed through a broad-based consensus process managed by the Council of Chief State School Officers (CCSSO). Educators, scholars, and citizens, representative of many diverse constituencies and points of view, designed objectives for the mathematics assessment, proposing goals they believed students should achieve in the course of their education. After careful reviews of the objectives, assessment questions were developed that were appropriate to those objectives. Representatives from State Education Agencies provided extensive input throughout the entire development process.

The framework adopted for the 1990 mathematics assessment was organized according to three mathematical abilities and five content areas. The mathematical abilities assessed were conceptual understanding, procedural knowledge, and problem solving. Content was drawn primarily from elementary and secondary school mathematics up to, but not including, calculus. The content areas assessed were Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions.

The Trial State Assessment included both multiple-choice and open-ended items. All questions underwent extensive reviews by specialists in mathematics, measurement, and bias/sensitivity, as well as reviews by representatives from State Education Agencies. The items were field tested on a representative group of students. Based on the results of the field test, items were revised or modified as necessary and then again reviewed for sensitivity, content, and editorial concerns. With the assistance of ETS/NAEP staff and outside reviewers, the mathematics Item Development Committee selected the items to include in the assessment.

Chapter 2 includes specific details about developing the objectives and items for the Trial State Assessment. The details of the professional scoring process are given in Chapter 6.

1.4 ASSESSMENT INSTRUMENTS

The assembly of cognitive items into booklets and their subsequent assignment to assessed students was determined by a balanced incomplete block (BIB) design with spiraled administration. Details of the BIB design are provided in Chapter 2.

The student assessment booklets contained five sections and included both cognitive and noncognitive items. In addition to three sections of cognitive questions, each booklet included two 5-minute sets of general and mathematics background questions designed to gather contextual information about students, their experiences in mathematics, and their attitudes toward the subject.

Besides the student assessment booklets, three other instruments provided data relating to the assessment -- an eighth-grade Mathematics Teacher Questionnaire, a School Characteristics and Policies Questionnaire, and an Excluded Student Questionnaire.

The Teacher Questionnaire was administered to the eighth-grade mathematics teachers of the students participating in the assessment. It consisted of two sections and took approximately 20 minutes to complete. The first section focused on teachers' background and experience. The second section focused on classroom information. Teachers were asked to respond to a set of questions about the classes in which students in the assessment were enrolled. Each teacher answered this set of questions for up to five different classes.

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

The Excluded 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). This questionnaire took approximately three minutes per student 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 population for the Trial State Assessment Program consisted of eighth-grade students enrolled in public schools. The representative sample of students assessed in the Trial State Assessment came from about 100 public schools in each jurisdiction, unless a jurisdiction had fewer than 100 schools in which case all or almost all schools participated. The sample in each state was designed both to produce aggregate estimates for the state, and 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 with monitoring and without monitoring. The schools were stratified by urbanicity, percentage of Black and Hispanic students enrolled, and median household income.

Thirty students selected from each school provided a sample size of approximately 3,000 students per state. The student sample size of 30 for each school was chosen to ensure at least 2,000 students participating from each state, allowing for school nonresponse, exclusion of students, inaccuracies in the measures of enrollment, and student absenteeism from the assessment.

The students within a school were sampled from a list of eighth-grade students. The decisions to exclude students from the assessment were made by school personnel, as in the national assessment. However, 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 1990 Trial State Assessment -- the construction of the school frame, the stratification process, the updating of the school frame with new schools, the actual sample selection, and the sample selection for the field test.

1.6 FIELD ADMINISTRATION

The administration for the 1990 Program and the 1989 field test involved a collaborative effort 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 1989 was to try out the items and procedures for the 1990 Program.

Each jurisdiction volunteering to participate in the 1989 field test and in the 1990 Trial State Assessment was asked to appoint a State Coordinator who became the liaison between NAEP staff and the participating schools. At the local school level, a Local 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 staff. In addition, Westat hired and trained a State Supervisor for each state. The State Supervisors were responsible for working with the State Coordinators and overseeing assessment activities. Westat also hired and trained four Quality Control Monitors in each state to monitor 50 percent of the assessment sessions. During the field test, the State Supervisors monitored all sessions.

Chapter 4 describes the procedures for obtaining cooperation from states. Chapter 5 provides details about the field activities for both the field test and 1990 program. Chapter 5 discusses the planning and preparations for the actual administration of the assessment, the training and monitoring of the assessment sessions, and a description of the responsibilities and roles of the State Coordinators, State Supervisors, Local Administrators, and Quality Control Monitors.

1.7 MATERIALS PROCESSING AND DATABASE CREATION

Upon completion of each assessment session, the school district personnel shipped the assessment booklets and forms from the field to National Computer Systems (NCS), the NAEP subcontractor for scanning and scoring, for professional scoring, entry into computer files, and checking. Then the files were sent to Educational Testing Service (ETS) for creation of the database. Careful checking assured that all data from the field were received. More than 125,000 booklets or questionnaires were received and processed. The processing of these data is detailed in Chapter 6. That chapter details the printing, distribution, receipt, processing, and final disposition of the 1990 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, innovativel; designed processing programs and a sophisticated Process Control System. This system, which is described in Chapter 6, allowed an integration of data entry and workflow management systems, including carefully planned and delineated editing, quality control, and auditing procedures.

The data transcription and editing procedures are also described in Chapter 6. These procedures resulted in the generation of 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 had to undergo a quality control process. The files were then merged into a comprehensive, integrated database. Chapter 7 describes the transcribed data files, the procedure of merging them, or bringing them together, to create the Trial State Assessment database, and the results of the quality control process.

1.8 THE TRIAL STATE ASSESSMENT DATA

Approximately 2,500 students were assessed within each state and the District of Columbia; apart from nonresponse, all eighth grade public school students were assessed in Guam and the Virgin Islands (1,617 and 1,328, respectively).

The basic information collected from the Trial State Assessment consisted of the responses of the assessed students to the 137 mathematics exercises. To limit the assessment time for each student to about one hour, a variant of matrix sampling called BIB spiraling was used to assign a subset of the full exercise pool to each student. The set of 137 items was divided into seven unique blocks, each requiring 15 minutes for completion. Each assessed student received a booklet containing three of the seven blocks according to a design, which ensured that each block was administered to a representative sample of students within each jurisdiction. The data also included responses to the background questionnaires (described in section 1.4 and Chapter 2).

The national data to which the Trial State Assessment results were compared came from a nationally representative sample of public-school students in the eighth grade. This sample was a part of the full 1990 national mathematics 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. Each age cohort sample was divided into two random half-samples, one assessed in the winter (January 8 to March 11, 1990) and the other in the spring (March 19 to May 18, 1990). Each half sample was representative of the national population of students in the age cohort.

The assessment instruments used in the Trial State Assessment were also used in the eight-grade national assessment and were administered using the identical procedures in both assessments. The time of testing for the state assessments (February 5 to March 2) occurred within the time of testing of the winter half sample of the national assessment. However, the state assessments differed from the national assessment in one important regard: Westat staff collected the data for the national assessment while, in accordance with the NAEP legislation, for the Trial State Assessment, data collection activities were the responsibility of each participating state and jurisdiction. These activities included ensuring the participation of selected schools and students, assessing students according to standardized procedures, and observing procedures for test security. To provide quality control of the Trial State Assessment, a random half of the administrations within each state was monitored.

1.9 WEIGHTING AND VARIANCE ESTIMATION

The Trial State Assessment used a complex sample design 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 taken into account was through the use of sampling weights which account for the fact that the probabilities of selection are not identical for all students. These weights also include adjustments for nonresponse of students and of schools. All population and subpopulation characteristics based on the Trial State Assessment data used the sampling weights in their estimation. Chapter 8 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 that due to sampling variability, which measures the dependence of the results on the particular sample of students actually assessed. Because of the effects of cluster selection (first schools are selected and 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 formulae will produce incorrect results. Instead, a variance estimation procedure which does take the characteristics of the sample into account was used for all analyses. This procedure, called the jackknife variance estimator, is discussed in Chapter 8.

The jackknife variance estimator 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 mathematics proficiency, such as the average mathematics proficiency of a subpopulation, do not. Because each student typically responds to relatively few items within a particular mathematics content area, there exists a nontrivial amount of imprecision in the measurement of the proficiency of any 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 Section 8.4.

1.10 PRELIMINARY DATA ANALYSIS

Immediately after receipt from NCS of the machine-readable data tapes containing students' responses, all cognitive and noncognitive items were subjected to an extensive item analysis to assure that each item represented what it was purported to measure.

Each block of cognitive items was subjected to item analysis routines, which yielded, for each item, the number of respondents, the percentage of students who selected the correct response and each incorrect response, the percentage who omitted the item, the percentage who did not reach the item, and the correlation between the item score and the block score. In addition, the item-analysis program provided summary statistics for each block, including reliability (internal consistency). These kinds of 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 anomalies that might signal unusual results or errors in creating the database.

Tables of the weighted percentages of students choosing each of the possible responses to each cognitive and background item were created and distributed to each state and jurisdiction. Additional analyses comparing the data from the monitored sessions with that from the unmonitored sessions were conducted to determine the comparability of the assessment data from the two types of administrations. Among other statistics compared were measures such as standard reliability estimates, the percentage of items attempted, and the rates of participation. Further details of the preliminary analyses conducted on the data appear in Chapter 10.

1.11 SCALING THE ASSESSMENT ITEMS

The primary analysis and reporting of the results from the Trial State Assessment used 3-parameter logistic item response theory (IRT) scale score models. Scaling models quantify a respondent's tendency to provide correct answers to the items contributing to a scale as a function of a parameter called proficiency that can be viewed as a summary measure of performance across all items entering into the scale. Chapter 9 provides an overview of the scaling model used with details provided in Chapter 10.

A series of subscales were created for the Trial State Assessment to summarize students' mathematics performance. These subscales were defined identically to those used for the scaling of the national NAEP eighth-grade mathematics data. The subscale definitions were based on the content by process area paradigm described in Chapter 2 and included five content areas: Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. Although the items comprising each subscale were identical to those used for the national program, the item parameters for the Trial State Assessment subscales were estimated from the combined data from all states and 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 the available data. To ensure equal representation in the scaling process, each state and jurisdiction was equally represented in the item calibration sample, as were the monitored and unmonitored administrations from each state and jurisdiction. Chapter 10 provides further details about the item parameter estimation.

The fit of the IRT model to the observed data was examined within each subscale by comparing the empirical item characteristic curves with the theoretic curves. In this comparison, the expected proportions of correct responses to each item for students with various levels of subscale proficiency were compared with the fitted item response curve. The expected proportions were calculated without assuming any functional form. In general, the item level results were well fit by the subscale models.

Using the item parameter estimates, subscale 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. Under the assumptions of the scaling model, these population estimates will be consistent, which would not be the case for subpopulation estimates obtained by aggregating optimal estimates of individual proficiency. Chapter 9 provides further details on the computation and use of plausible values.

In addition to the subscale plausible values, a composite of the subscales was created as a measure of overall mathematics proficiency. This composite was a weighted average of the subscale plausible values in which the weights were proportional to the relative importance assigned to each content area as specified in the mathematics objectives. The definition of the composite for the Trial State Assessment program was identical to that used for the national eighth-grade mathematics program.

The national composite scale for mathematics includes scale anchoring information as an aid to its interpretation. The anchoring process for the 1990 mathematics scale began by designating four levels on the scale (which ranges from 0 to 500): 200, 250, 300, and 350. The process then identifies items that a vast majority of students at a selected scale level can answer correctly but that most students at lower levels cannot. Such items are then reviewed by subject area specialists. The result is descriptions of student proficiency at each of the levels and a set of selected items that exemplify the interpretation. The descriptions aid in the interpretation of the results from the Trial State Assessment, after those results are linked to the national subscales. Further details of the anchoring process appear in Chapter 9.

1.12 LINKING THE TRIAL STATE RESULTS TO THE NATIONAL RESULTS.

The results from the Trial State Assessment were linked to those from the national NAEP through a linking function determined by comparing the results for the aggregate of all students assessed in the Trial State Assessment with the results for students within 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 the 37 participating states and the District of Columbia. Specifically, the SAC subsample consists of all eighth-grade students in public schools in the states and the District of Columbia who were assessed as a part of the winter administration of the national cross-sectional mathematics assessment.

A linear equating within each subscale was used to link the results of the Trial State Assessment to the national NAEP. The adequacy of linear equating was evaluated by comparing, for each subscale, the distribution of mathematics proficiency based on the aggregation of all assessed students 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 eighth grade public school students in the aggregation of the states and the District of Columbia -- the students from Guam and the Virgin Islands were not included in the equating. If a linear equating is adequate, the two distributions will have the same shape, to a close approximation, and will differ, at most, in their means and variances. This was found to be the case.

The linking was accomplished for each subscale by matching the mean and standard deviation of the subscale proficiencies across all students in the Trial State Assessment (excluding Guam and the Virgin Islands) to the corresponding subscale mean and standard deviation across all students in the SAC subsample. Further details of the linking are given in Chapter 10.

1.13 REPORTING THE TRIAL STATE ASSESSMENT RESULTS

Each state and jurisdiction that participated in the Trial State Assessment received multiple copies of a summary report providing the state's results with accompanying text and graphics, and including national and regional comparisons. These reports were generated by a computerized report-generation system in which graphic designers, statisticians, data analysts, and report writers collaborated to develop shells of the reports in advance of the analysis. The results of the data analysis were then automatically incorporated into the reports that gave, in addition to tables and graphs of the results, 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 community type. Results were presented for each subscale and for the overall mathematics composite. Results were also reported for a variety of other subpopulations based on variables derived from the student, teacher, and school questionnaires. Additionally, the estimated proportion of students who are at or above each of the four anchor points on the mathematics composite scale was presented for the key reporting categories. Standard errors were included for all statistics.

Because the demographic characteristics of the eighth-grade public-school students vary widely by state, the proportions of students in the various categories of the race/ethnicity, parental education, and type of community variables varied by state. Chapter 11 describes the rules, based on effect size and sample size considerations, that were used to establish whether a particular category contained sufficient data for reliable reporting of results for a particular state. Chapter 11 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 state and jurisdiction. This included the 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 by state.

Chapter 2

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

Stephen L. Koffler

Educational Testing Service

2.1 OVERVIEW

Similar to all previous NAEP assessments, the objectives for the Trial State Assessment were developed through a broad-based consensus process. Educators, scholars, and citizens, representative of many diverse constituencies and points of view, designed objectives for the mathematics assessment, proposing goals they believed students should achieve in the course of their education. After careful reviews of the objectives, assessment items were developed that were appropriate to those objectives. All items underwent extensive reviews by specialists in mathematics, measurement, and bias/sensitivity, as well as reviews by state representatives.

The objective and item development efforts were governed by four major considerations:

- As specified in the 1988 NAEP legislation, the objectives had to be developed through a consensus process involving subject-matter experts, school administrators, teachers, and parents, and the items had to be carefully reviewed for potential bias.

- As outlined in the ETS proposal for the administration of the NAEP contract, the development of the items had to be guided by a Mathematics Item Development Panel.

- As described in the ETS Standards for 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 specific details about developing the objectives and items for the Trial State Assessment. The chapter also describes the instruments -- the student assessment booklets, Mathematics Teacher Questionnaire, School Characteristics and Policies Questionnaire, and Excluded Student Questionnaire -- and the manner in which the items were organized into blocks to create the student booklets. Many committees worked on the development framework, objectives, and items for the Trial State Assessment. A list of the committees and consultants who participated in the 1990 development process is included in Appendix A.

2.2 CONTEXT FOR PLANNING THE 1990 MATHEMATICS ASSESSMENT[1]

Anticipating the 1988 legislation that authorized the Trial State Assessment, in mid-1987 the federal government arranged for a special grant from the National Science Foundation and the Department of Education to the Council of Chief State School Officers (CCSSO) to lay the groundwork for the Trial State Assessment.

The CCSSO established the National Assessment Planning Project to oversee the work for the Trial State Assessment. The National Assessment Planning Project, whose members included policymakers, practitioners, and citizens nominated by 18 national organizations, had two primary purposes. The first was to recommend objectives for the state-level mathematics assessment, and the second was to make suggestions for reporting state results. However, rather than focusing exclusively on the eighth-grade objectives for the Trial State Assessment, the project developed objectives for all three grades to be assessed in 1990 (fourth, eighth, and twelfth) because the assessment objectives had to be coordinated across all grades. The objectives for the Trial State Assessment Program were the same as for the eighth-grade national program.

2.3 ASSESSMENT DESIGN PRINCIPLES

A Mathematics Objectives Committee--comprising a teacher, a school administrator, mathematics education specialists from various states, mathematicians, parents, and citizens--was created by the CCSSO to recommend objectives for the assessment.

Two principles emerged during the discussions of the Mathematics Objectives Committee and became the basis for structuring the objectives and framework for the 1990 assessment. The first principle was that a national assessment, designed to provide state-level comparisons, should not be directed toward measuring only those topics and skills already in the objectives of all states or geared to the *least common denominator* of student preparation. The second principle was that the assessment should also not be used to steer instruction toward one particular pedagogical or philosophical viewpoint to the exclusion of others that are widely held.

The objectives development was also guided by several other considerations: the assessment should reflect many of the states' curricular emphases and objectives; reflect what various scholars, practitioners, and interested citizens believe should be included in the curriculum; and maintain some of the content of prior assessments to permit reporting of trends in performance. Accordingly, the committee gave attention to several frames of reference.

- States' goals and concerns, as reflected through analyses of state mathematics curriculum guides and the recommendations of state mathematics specialists.

- A report on "Issues in the Field," based on telephone interviews with leading mathematics educators, and a draft assessment framework provided by a subcommittee of the Mathematics Objectives Committee.

[1] For more details see the booklet *Mathematics Objectives 1990 Assessment* (National Assessment of Educational Progress, Princeton, N.J.: Educational Testing Service, 1988).

- The draft of the **Curriculum and Evaluation Standards for School Mathematics**, developed by the National Council of Teachers of Mathematics through intensive work by leading mathematics educators in the United States.[2]

- The design of the 1986 mathematics assessment.[3] The framework for the 1986 assessment had thirty-five cells -- seven content and five process areas. Because there were thirty-five cells, the weightings assigned to some of the cells in the 1986 framework did not result in a sufficient number of items to provide reliable measures of students' knowledge and skills. As a result, it was decided that the outline or matrix guiding the development of the 1990 assessment had to be simplified and that necessary complexity could be reflected through the designation of specific abilities and topics in each content area.

2.4 ASSESSMENT DEVELOPMENT PROCESS

The framework, objectives, and a set of sample items developed by the Mathematics Objectives Committee were distributed to the mathematics supervisor in each of the 50 State Education Agencies. These supervisors convened a panel that reviewed the draft objectives and returned comments and suggestions to the project staff. Copies of the draft were also sent to 25 mathematics educators and scholars for review. The Mathematics Objectives Committee incorporated the comments and revisions and formulated their final recommendations, which were approved by the National Assessment Planning Project Steering Committee.

The framework and objectives were then submitted to the National Center for Education Statistics (NCES), which forwarded them for review to NAEP's then-governing board, the Assessment Policy Committee (APC). The APC approved the objectives with minor provisions about the feasibility of full implementation.[4] The framework and objectives were refined by NAEP's Item Development Panel, reviewed by the Task Force on State Comparisons, and resubmitted to NCES for adoption.

2.5 FRAMEWORK FOR THE ASSESSMENT

The framework adopted for the 1990 mathematics assessment is organized according to three mathematical abilities and five content areas. The mathematical abilities assessed are conceptual understanding, procedural knowledge, and problem solving. Content is drawn primarily from elementary and secondary school mathematics up to, but not including, calculus.

[2] National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: 1987).

[3] National Assessment of Educational Progress, *Mathematics Objectives: 1985-86 Assessment* (Princeton, NJ: Educational Testing Service, 1987).

[4] This action is contained in a statement issued by the Assessment Policy Committee's Executive Committee on April 29, 1988. The recommendations were ratified by the full committee on June 18, 1988, with two stipulations: that the objectives be so weighted as to permit reporting on trends in performance; and, with regard to the use of calculator-active items and open response questions, that the assessment be developed within the resources available for its administration.

The content areas assessed are numbers and operations; measurement; geometry; data analysis, statistics, and probability; and algebra and functions.

2.6 DISTRIBUTION OF ASSESSMENT ITEMS

The assignment of percentages of the assessment items that would be devoted to each mathematical ability and content area is an important feature of the assessment design because such weighting reflects the importance or value given to each area at each grade level. For 1990, the National Assessment Planning Project was interested in creating an assessment that would be forward-thinking and could lead instruction; thus, more emphasis was given to problem solving than in previous assessments. In addition, individuals involved in the Planning Project advised that greater emphasis be given to geometry and to algebra and functions, and less to numbers and operations than in the past.

The distribution of items by mathematical ability, mathematical content area, and grade is provided in Table 2-1 and Table 2-2.

Table 2-1

Percentage Distribution of Items
by Grade and Mathematical Ability

Mathematical Ability	Grade		
	Four	Eight	Twelve
Conceptual Understanding	40%	40%	40%
Procedural Knowledge	30%	30%	30%
Problem Solving	30%	30%	30%

Table 2-2

Percentage Distribution of Items
by Grade and Mathematical Content Area

Mathematical Content Area	Grade		
	Four	Eight	Twelve
Numbers and Operations	45%	30%	25%
Measurement	20%	15%	15%
Geometry	15%	20%	20%
Data Analysis, Statistics, and Probability	10%	15%	15%
Algebra and Functions	10%	20%	25%

2.7 DEVELOPING THE COGNITIVE ITEMS

The Trial State Assessment included open-ended and multiple-choice items. The open-ended items were designed to provide an extended view of students' mathematical knowledges and skills. Building on recommendations from the CCSSO report, the NAEP Item Development Panel created open-ended items to assess objectives in the framework that are best measured using such types of items (e.g. ability to draw graphs and figures, generate informal proofs, draw figures, or generalize relationships.)

The Trial State Assessment included seven different 15-minute segments or "blocks" of multiple-choice and open-ended content items. Two of the seven blocks were designed to be answered using a calculator and one using a protractor/ruler. Because the blocks contain a variety of item types, there were no rigid criteria dictating parallel structure across blocks. The blocks were assembled three to a booklet, and each student was asked to respond to one booklet. These seven blocks (including the two requiring calculators) were balanced across seven booklets.

A carefully developed and tested series of steps were used to create the assessment items that reflected the objectives.

1. The Mathematics Item Development Panel provided guidance to the 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.

2. Item specifications were developed, and prototype items were created.

3. Item writers, both inside and outside ETS, with subject-matter expertise and skills and experience in creating items according to specifications wrote assessment items.

4. The items were reviewed and revised by NAEP/ETS staff and external reviewers.

5. Representatives from the State Education Agencies met and reviewed all items and background questionnaires (see section 2.9 for a discussion of the background questionnaires.)

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

7. Field test materials were prepared, including the materials necessary to secure OMB clearance.

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

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

10. Based on the field test analyses, items for the 1990 assessment were revised, modified and re-edited, where necessary. The items once again went through another ETS sensitivity review.

11. The Mathematics Item Development Panel selected the items to include in the 1990 assessment.

12. Items were assembled into seven different "blocks" (15-minute sections established according to statistical guidelines developed at the beginning of the process).

13. After a final review and check to ensure that each assessment booklet and each block met the overall guidelines for the assessment, the booklets were typeset and printed. 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 137 items, including 35 open-ended items. Table 2-3 provides the number of items for each content and ability group included in the Trial State Assessment. In total, seven 15-minute blocks were used. These same blocks were also used in the national mathematics assessment.

Table 2-3

Content by Ability Distribution of Items

	Content Area					
Ability	A	B	C	D	E	Total
CU	18	7	13	9	12	59
PK	15	9	4	5	8	41
PS	12	5	9	5	6	37
Total	45	21	26	19	26	137

CU = Conceptual Understanding
PK = Procedural Knowledge
PS = Problem Solving

A = Numbers and Operations
B = Measurement
C = Geometry
D = Data Analysis, Statistics, and Probability
E = Algebra and Functions

2.8 STUDENT ASSESSMENT BOOKLETS

Each student assessment booklet included three sections of cognitive mathematics items and two sections of background questions.

The assembly of mathematics items into booklets and their subsequent assignment to assessed students was determined by a *balanced incomplete block* (BIB) design with *spiraled* administration.

The first step in implementing BIB spiraling required dividing the total pool of mathematics items into blocks designed to take 15 minutes to complete. These blocks were then assembled into booklets containing two 5-minute background sections and three blocks of mathematics items according to a partially balanced incomplete block design. Thus, the overall assessment time for each student was approximately 55 minutes. The mathematics blocks were assigned to booklets in such a way that each block appeared in the same number of booklets and every pair of blocks appeared together in exactly one booklet. This is the *balanced* part oi the balanced incomplete block design. It is an *incomplete* block design because no booklet contained all items and hence there is *incomplete* data for each assessed student.

The BIB design for the 1990 national mathematics assessment (and, therefore, for the Trial State Assessment) was *focused* -- each block was paired with every other mathematics block but not with blocks from other subject areas. The *focused*-BIB design also balances the order of presentation of the blocks of items -- every block appears as the first cognitive block in one booklet, as the second block in another booklet, and as the third block in a third booklet.

The focused-BIB design used in 1990 required that seven blocks of mathematics items be assembled into seven booklets. The assessment booklets were then *spiraled* and bundled. Spiraling involves interleaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in each position in a bundle.

The final step in the BIB-spiraling procedure is 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 several students in the session received the same booklet. In the Trial State Assessment BIB-spiral design, representative and randomly equivalent samples of about 2,500 students responded to each item.

Table 2-4 provides the composition of each block of items administered in the Trial State Assessment Program. Table 2-5 provides the total number of booklets, cognitive blocks, and noncognitive blocks used for the program. Table 2-5 also provides the details of the focused-BIB design that was used with seven blocks and seven booklets. Note that these same blocks and focused-BIB design also were used for the eighth-grade national assessment.

Table 2-4

Cognitive and Noncognitive Block Information

Block	Type	Total Number of Items	Number of Multiple-Choice Items	Number of Open-ended Items	Booklets Containing Block
CA	Common Background	22	22	0	8 - 14
MB	Mathematics Background	22	22	0	8 - 14
MC	Mathematics Cognitive	23	19	4	8, 12, 14
MD	Mathematics Cognitive	21	21	0	8, 9, 13
ME	Mathematics Cognitive	16	0	16	9, 10, 14
MF	Mathematics Cognitive (Protractor/Ruler)	21	16	5	8, 10, 11
MG	Mathematics Cognitive	18	17	1	9, 11, 12
MH	Mathematics Cognitive (Calculator)	18	16	2	10, 12, 13
MI	Mathematics Cognitive (Calculator)	20	13	7	11, 13, 14

Table 2-5

Booklet Contents and Number of Booklets Administered

Booklet Number	Common Background Block	Background Block	Cognitive Blocks		
8	CA	MB	MC	MD	MF[1]
9	CA	MB	MD	ME	MG
10	CA	MB	ME	MF[1]	MH[2]
11	CA	MB	MF[1]	MG	MI[2]
12	CA	MB	MG	MH[2]	MC
13	CA	MB	MH[2]	MI[2]	MD
14	CA	MB	MI[2]	MC	ME

[1] Protractor/ruler needed for this block
[2] Calculator needed for this block

2.9 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 an iterative process that involved staff work, field testing, and review by external advisory groups. A Policy Analysis and Use 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).

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. The items were then assembled into questionnaires and underwent internal ETS review procedures to ensure fairness and quality.For the 1990 assessment, the framework focused on six educational areas: curriculum, instructional practices, teacher qualifications, educational standards and reform, school conditions, and conditions outside of the school that facilitate learning and instruction.[5]

[5] National Assessment of Educational Progress, *1990 Policy Information Framework*. (Princeton, N.J.: Educational Testing Service, 1990).

2.9.1 Student Questionnaires

In addition to the cognitive questions, the 1990 Trial State Assessment included two five-minute sets of general and mathematics background questions designed to gather contextual information about students, their experiences in mathematics, and their attitudes toward the subject.

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, school climate, academic expectations, 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.

Three categories of information were represented in the second five-minute section of mathematics background questions called the **student mathematics questionnaire** (22 questions):

- **Time Spent Studying Mathematics**: Time spent on task and mathematics coursework has been shown to be strongly related to mathematics achievement.[6] Students were asked to describe both the amount of instruction they received in mathematics and the time spent on mathematics homework.

- **Instructional Practices**: The nature of students' mathematics instruction is also thought to be related to achievement.[7] Students were asked to report their experience in using various instructional materials in the mathematics classroom, including calculators, models, and manipulatives. In addition, they were asked about the instructional practices of their mathematics teachers and the extent to which the students themselves practiced the communication of mathematical ideas--such as writing out explanations, justifications, or proofs--in their mathematics classes.

- **Attitudes Towards Mathematics**: Students' enjoyment of and confidence in their abilities in mathematics and their perceptions of the usefulness of the discipline to their present and future lives appear to be related to mathematics achievement.[8] Students were asked a series of questions about their attitudes and perceptions about mathematics, such as, do they enjoy mathematics and are they good in mathematics.

[6] Senta Raisen and Lyle Jones, Eds., *Indicators of Precollege Education in Science and Mathematics: A Preliminary Review.* (Washington, DC: National Academy Press, 1985).

[7] Dossey, John A., Mullis, Ina V. S., Lindquist, Mary M., and Chambers, Donald L., *The Mathematics Report Card: Are We Measuring Up?* (Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service, 1988).

[8] Sheila Tobias *Succeed with Math: Every Student's Guide to Conquering Mathematics Anxiety* (New York: The College Entrance Examination Board. 1987).

2.9.2 Teacher, School, and Excluded Student Questionnaires

To supplement the information on instruction reported by students, the mathematics teachers of the eighth 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 training. The second part pertained to the procedures the teacher uses for *each class* containing an assessed student.

> **The Teacher Questionnaire, Part I: Background and Training** (34 questions) included questions pertaining to gender, race/ethnicity, years of teaching experience, certification, degrees, major and minor fields study, coursework in education, coursework in subject area, in-service training, extent of control over classroom, instruction, and curriculum, and availability of resources for classroom.

> The **Teacher Questionnaire, Part II: Classroom by Classroom Information** (35 questions) included questions on the 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, instructional emphasis given to the topics and skills 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.

> The **School Characteristics and Policies Questionnaire** (117 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 tracking, curriculum, testing practices and use, special priorities and school-wide programs, availability of resources, special services, community services, policies for parental involvement, and school-wide problems.

The **Excluded 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). This questionnaire asked about the nature of the student's exclusion and the special programs in which the student participated.

2.10 DEVELOPMENT OF FINAL FORMS

The field tests were conducted in February-March 1989 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 Mathematics Panel and Policy Panel. Using item analysis, which provides the mean percentage of correct responses for each item in the field test, committee members, ETS test development staff, and NAEP/ETS staff reviewed the materials with four objectives: to determine which items were most related to achievement; 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.

Once the committees had selected the items, all items were rechecked for content, measurement, and sensitivity concerns. In addition, another meeting of representatives from State Education Agencies was convened to review the field test results. The federal clearance process was initiated in June 1989 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 July 1989. 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

Jim Bethel, Keith Rust, and Jacqueline Severynse

Westat

3.1 OVERVIEW

The representative sample of eighth-grade students assessed in the Trial State Assessment came from about 100 public schools in each jurisdiction, unless a jurisdiction had fewer than 100 public schools with eighth-grade students in which case all or almost all such schools participated. The sample of schools in each state was selected with probability proportionate to size (pps), where the measure of size was equal to the number of students enrolled in the eighth grade in each school. The school samples were implicitly stratified based on urbanicity, percentage of minority enrollment, and household income.

Except for some schools in a few states, schools selected for the 1990 national assessment for Grade 8/Age 13 were excluded from the Trial State Assessment. Appropriate weighting adjustments were used to ensure that these exclusions did not introduce bias into estimates from the state samples. One hundred percent participation of all selected schools was the goal. Many of the schools that declined to participate were replaced in the sample by substitute selections.

The target population for the Trial State Assessment Program consisted of eighth-grade students enrolled in public schools. In general, slightly more than one hundred schools per state were selected to allow for the fact that some selected schools would not have any eligible students enrolled. Such schools arose as a result of errors in the list of schools used to compile the sampling frame. Thirty students selected from each school provided a sample size of approximately 3,000 students per state. The student sample size of 30 for each school was chosen to ensure at least 2,000 students participating from each state, accounting for school nonresponse, exclusion of students, inaccuracies in the measures of enrollment, and student absenteeism from the assessment.

The schools within each state were stratified by the following variables:

- Urbanicity (central city, suburban, other)
- Percentage of Black and Hispanic students enrolled
- Median household income

All states, except for those with 100 schools or fewer, were stratified by urbanicity and income variables. Only states with significant minority populations were stratified based on minority enrollment.

In contrast to the national assessment, which was administered by Westat field personnel, the Trial State Assessment was administered by local school or district personnel. To check on the consistency of assessment administration conditions, half the schools in the sample were monitored by Westat field staff, and half were unmonitored to permit comparisons between the two. The sample in each state was designed both to produce aggregate estimates for the state, and various 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 with monitoring and without monitoring.

The Trial State Assessment was preceded in 1989 by a field test that had three principal goals: (1) to test new items contemplated for 1990; (2)to test procedures contemplated for 1990; and (3)to give states the opportunity to observe and react to proposed procedures. Nine schools were selected for the field test from each of 28 participating states. Schools that participated in the field test had a chance of being selected for the 1990 assessment.

In this chapter, Section 3.2 documents the procedures used to select the schools for the field test. Section 3.3 describes the various aspects of selecting the sample for the 1990 Trial State Assessment, including frame construction, the stratification process, updating the school frame with new schools, and the actual sample selection. School substitution is discussed in more detail in Section 3.3.3.5. Section 3.4 discusses the school and student participation rates.

3.2 SELECTION OF SCHOOLS FOR THE 1989 FIELD TEST

3.2.1 Overview

The selection of nine pairs of schools for each participating state for the field test satisfied the principal goals for testing administration procedures and assessment items, and at the same time served the following three sampling objectives: (1) the sample of schools provided for the nation as a whole was a cross section of the types of schools likely to be encountered in the 1990 Program; (2) the schools were clustered into two geographic clusters within each state, to facilitate Westat's field operations in monitoring the assessments; and (3) pairs of schools were identified, with one of each pair included in the test, which allowed state participation in the selection of test schools, and also facilitated replacements of schools which declined to participate in the assessment. The field test did not provide a representative sample of schools from any given state in terms of demographic characteristics, urbanicity or geographic distribution. We note especially that the sampling of schools in pairs, with only one of each pair included in the test, was a special feature of the field test that was not a feature of the 1990 sample drawn to represent the state. In the 1990 sample, every reasonable effort was made to obtain the cooperation of the sample schools as initially selected, with only a very limited amount of substitution for noncooperating schools.

To achieve the three objectives listed above, we used sampling techniques that combined clustering, stratification, and matching procedures. Briefly, schools were geographically clustered -- using ZIP code and county designations -- so that a single Westat field worker could monitor up to five administrations in any given cluster within the week allotted for the assessment. This involved defining clusters that were large enough to contain schools for five administrations but also small enough geographically that the field personnel could travel from one session to another in one afternoon.

Two clusters were selected from each state, and the schools within each cluster were stratified by urbanicity, number of eighth-grade students, and percentage of minority students. Nine schools were selected from each state, with the schools being divided between the two clusters. The final step involved matching each selected school to the remaining ones in the cluster to find an alternative school of approximately similar characteristics. The within school sampling of students was carried out using the same procedures as for the 1990 Trial. These procedures are described in more detail in Chapter 10.

3.2.2 Sampling Frames

QED vs. Common Core: In September and October 1988, when the sample design for the field test was carried out, the most recently available edition of the National Center for Education Statistics (NCES) Common Core of Data (CCD) did not contain data on the number of minority students. Since we considered percentage of minority students to be a key stratification variable, we proceeded to use the most current version of the Quality Education Data, Inc., (QED) data file as the basis for the sampling frame. This file contained public schools extant during the 1987-88 school year, together with their approximate enrollments. Since it did not contain the exact eighth-grade enrollments, the enrollments were estimated as

$$\textit{Estimated Eighth-Grade Enrollment} = \frac{\textit{Total Students}}{\textit{Number of Grades}}.$$

This method is generally used with QED data for sample design and selection in national NAEP, and experience has shown it to be reasonably accurate.

Other Sources of Data: Because the QED data contains demographic data for districts (rather than for individual schools), we used data on minority enrollment from other sources wherever possible. From previous National Assessments we have accumulated demographic data on schools in most of the very large school districts. While these data are some years out of date, they are still more current than that contained in the QED files (which are derived from 1980 Census data); moreover, the data are at the school level rather than the district level.

Scope of the Survey: Only public schools with eighth-grade students were in-scope for the assessments. Because the minimum size for an assessment was 25 students, schools with fewer than 25 students were removed from the sampling frame, except in Nebraska and Montana, where the majority of schools had fewer than 25 students and where substantial percentages of students (about 18% and 14%, respectively) attended such schools. It was unnecessary to sample such small schools for the field test in each state since a fully representative sample for each state was not needed.

3.2.3 Geographic Clustering

Clustering Procedure: Geographic clusters were initially defined using three-digit-level ZIP codes. Clusters with fewer than 50 eighth grade students (of which there were only one or two) were deleted from the sampling frame. Since each cluster was to have up to five pairs of schools selected from it, the lower size limit for a cluster was set at 12 schools. Thus clusters that had fewer than 12 schools were aggregated by linking together clusters within counties. When a cluster crossed county lines, the cluster was associated with the county having the largest number of eighth-grade students in the cluster. Using computer algorithms and county maps, we defined a set of geographic clusters all of that met the requirement of containing at least 12 public schools of 25 or more eighth-grade students.

Selection of Clusters: Once defined, the clusters were selected as follows. Each school, was assigned an "urbanicity score" of one for rural, two for suburban, and three for urban, based on the classification given in the QED file, and the average urbanicity score was calculated over all the schools in each cluster. The clusters were then sorted from highest to lowest based on the average urbanicity value and a systematic sample of two clusters was selected from each state with probability proportionate to the number of eighth-grade students in the cluster. This resulted in a sample of clusters providing a broad national representation of both urban and rural areas.

3.2.4 Selection of Schools

Stratification of Schools: Within the selected clusters, the schools were stratified by broad classes of the estimated eighth grade enrollment, the estimated percentage of minority students and the percentage of students below the poverty line in the school district. The stratification was implemented by sorting so that each level of school enrollment contained each level of the minority student percentages, which contained each level of the impoverished student percentages.

Selection of Schools: Schools were selected in two stages. First, samples of four or five schools were selected systematically from each cluster. (Within a given state, five schools were taken from the larger cluster and four from the smaller one.) Once the "initial" sample of schools was selected, each selected school was matched to an unselected school in the same cluster to allow for a replacement.

Matching for Replacements: The matching procedure employed a "nearest-neighbor" rule based on the percentage of Black students, the percentage of Hispanic students, the eighth-grade enrollment, the total enrollment, and the percentage of students in the district below the poverty line. The variables were first standardized (by dividing the deviation from the mean for a given variable by the respective standard deviations), and pairwise differences were squared and summed. The result was a measure of similarity for each selected school with each unselected school in all the sample clusters. We then used a computer algorithm to pick the closest matches among the pairs of schools to arrive at four or five pairs of schools for each cluster.

3.2.5 Schools with Fewer Than 25 Students

As noted above, Montana and Nebraska both had large numbers of schools and substantial proportions of students enrolled in schools that had at most 24 eligible students. In the 1990 Trial State Assessment Program, such schools were represented in all states' samples. They were aggregated with other small or large schools to form groups of schools with at least 25 students that were sampled for the assessment. This process of aggregation involved not only having Westat personnel cluster and select the schools, but also having the states conduct small assessment sessions in different schools.

Since both aspects of this process were to be part of the 1990 Trial State Assessment, it seemed important to conduct at least a limited field test of both selection procedures and assessment methods. Because Montana and Nebraska had the largest percentage of students enrolled in small schools, these states were selected for this part of the field test.

In one of the geographic clusters in each of these states, a pair of two groups of small schools was selected instead of a pair of single schools, as done in all other states. These "groups" were formed as follows. First, all schools in the geographic cluster with between five and 24 eighth-grade students were isolated. Next, the schools were sorted by the number of eighth-grade students. Finally, groups were formed by combining the largest and smallest schools, alternatively taking one from each end of the sorted list, until a minimum of 25 students was reached. Once the groups were defined, two were selected at random. These selections did not use either the stratification or matching algorithms described above for larger schools. All students in the selected small schools were assessed.

3.2.6 Special Cases

The steps described above covered all states participating in the field test except for American Samoa, the Virgin Islands, and the District of Columbia. For American Samoa and the Virgin Islands, there were fewer than nine schools with an estimated enrollment of at least 25 eighth-grade students, so that all schools in these territories were included in the sample. In the District of Columbia, the entire district was treated as a single cluster, with the selection of schools following the remaining steps -- e.g., stratification, selection and matching of schools -- as described above.

3.3 SELECTION OF SCHOOLS AND STUDENTS FOR THE 1990 PROGRAM

3.3.1 Frame Construction

Three sources of data were combined to construct the school sampling frame:

- The NCES Common Core of Data (CCD) for 1988

- Data on school-level minority enrollment collected from school districts during the sample design phase of the 1988 and 1990 NAEP samples. The information was sought directly from those districts expected to have more than a trivial minority enrollment, but for which minority enrollment was not available from the CCD. Such districts were identified using the file of districts provided by Quality Education Data, Inc. (QED).

- 1980 Census data broken down to the ZIP code level, as provided on a file supplied by Donnelley Marketing Information Services

For the school level sample design, the frame variables used were total enrollment, eighth-grade enrollment, urbanicity, minority enrollment, and median household income.

The CCD file had complete or near complete information on both urbanicity and grade enrollments. According to the CCD documentation, urbanicity was designated as *Central City*, *Suburban*, or *Other*, depending on what type of area was served by the public school district to which the school belonged. If the district primarily served a central city of a Metropolitan Statistical Area (MSA), it was designated as *Central City*. If the district served an MSA but not primarily its central city, then it was designated as *Suburban*. Finally, a district was designated as *Other* if it did not serve an MSA.

In the few cases where the eighth-grade enrollment was missing, the school was screened to assess whether it was likely to contain an eighth grade, based primarily on the school name. If the total enrollment was less than 100, the school was assigned as having all grades between the lowest grade for which there was at least one student enrolled in 1987 and the highest grade for which there was at least one student enrolled in 1987. If the total enrollment was at least 100, the school was assigned as having all grades for which the 1987 grade enrollment was greater than zero. These schools which contained an eighth grade, as defined by this process, were included on the frame.

Minority (i.e., Black and Hispanic) enrollment was included on the 1988 CCD for all participating states except Alabama, Georgia, Idaho, Louisiana, Maryland, Montana, New Hampshire, New Mexico, Virginia, West Virginia, Wyoming, Guam, and the Virgin Islands. For Alabama, Georgia, Louisiana, Maryland, New Mexico, and Virginia, we used minority data collected from the 1987 surveys of public school districts requesting school-level minority enrollment data. If these data were not available for a district likely to have non-trivial minority enrollment, we obtained minority population data by surveying the district and obtaining school level minority enrollment. For school districts with low minority populations and for those which did not respond to the minority data survey, the 1980 Census information for percent minority in the district, obtained from the QED file, was used to approximate the percentage of minority enrollment for the school. In the other states listed (Idaho, Montana, New Hampshire, West Virginia, and Wyoming), minority data was not utilized in sample selection, since minority enrollment was very low in these five states. In the two territories all schools were included in the sample.

The other variable used in sampling was median household income. This was available from the Census file (third source above), which contained the number of households in the ZIP code area with incomes of various levels (e.g., less than $7,500, from $7,500 to $9,999, etc.) Using these data, we calculated an estimate of the median income for each ZIP code area and assigned this value to all schools having that ZIP code in their address.

In order to minimize overlap with the national NAEP school samples, in general schools selected for the national sample were excluded from the state frame. Weighting adjustments were made to account for this procedure and render unbiased estimates. These adjustments are described in Section 8.2.4. In Delaware, the District of Columbia, and Hawaii, all national schools eligible were retained on the state frame. A similar approach was used to exclude those schools having both grades 8 and 10 that were included in the school sample for the 1990 National Educational Longitudinal Study First Phase Follow-Up.

3.3.2 Stratification

States were stratified on urbanicity, percentage of minority enrollment, and household income depending on the number of eighth-grade schools within the state and the percentage of minority students within each urbanicity class:

- In states with 105 schools or less, schools were not stratified at all, since all schools in these states with at least 20 students were selected for the assessment. If a sample of smaller schools was drawn, rather than selecting them all, then this sample was not stratified. Schools in these states were called Type 1 Clusters.

- In states which *either* had 106 to 200 schools *or* a low percentage of minority students, schools were stratified by urbanicity and household income.

- In states which had more than 200 schools *and* a high percentage of minority students, schools were stratified by urbanicity, percentage of minority enrollment, and household income.

- In those states with high percentages of both Black and Hispanic students and more than 200 schools, schools were stratified on the basis of percentage of Black enrollment and percentage of Hispanic enrollment.

Urbanicity (defined more precisely in the previous section) was categorized as *Central City*, *Suburban*, and *Other*, although these classes were collapsed in some cases. If any urbanicity class had more than 10 percent Black students or 7 percent Hispanic students but not more than 20 percent of both, the schools within the state were stratified by ordering the percentage of minority enrollment within the urbanicity classes and dividing the schools into three groups with an approximately equal number of schools in each. Urbanicity strata with fewer than 10 percent Black students and 7 percent Hispanic students were not stratified by minority enrollment. Where there were high percentages of both Black and Hispanic students (i.e., more than 20 percent of each), four strata were formed:

- High Black/high Hispanic: schools above the medians for both percentage of Black students and percentage of Hispanic students.

- High Black/low Hispanic: schools above the median for percentage of Black students but below the median for percentage of Hispanic students.

- Low Black/high Hispanic: schools below the median for percentage of Black students but above the median for percentage of Hispanic students.

- Low Black/low Hispanic: schools below the medians for both percentage of Black students and percentage of Hispanic students.

Within these classes defined by urbanicity and minority enrollment, schools were sorted in serpentine order by the median household income so that bordering schools in different classes would be the most similar. For instance, within the suburban urbanicity, if the low minority class was sorted from highest median income to lowest, then the intermediate minority class was sorted from lowest median income to highest, and the highest minority class was sorted from highest median income to lowest. Table 3-1 shows the configuration of strata in each participating jurisdiction, together with the number of sampled schools.

3.3.2.1 Schools with 19 Students or Fewer

Since the target assessment size for each school was about 25 after allowance for exclusion and absenteeism, schools with 19 or fewer eighth-grade students were handled by one of two different methods, depending on the prevalence of these schools within the given state, and of the students attending them. These special procedures were adopted to provide control over the sample sizes of both schools and students and are referred to as "geographic" and "stratified" grouping. Table 3-1 gives the form of grouping used for each state.

Geographic Grouping: In states with a relatively small number of such schools (specifically, fewer than 20 percent of the schools for the state, with fewer than 1 percent of the total eighth-grade students), small schools were grouped geographically with larger ones (eighth-grade enrollment of 20 or more), and then the resulting pairs (or possibly larger groups) were initially sampled together as a single unit. These units were called Type 2 Clusters. Data for stratification were pooled between the paired schools.

Stratified Grouping: In states with larger numbers of small schools, schools were stratified into two groups, depending on whether or not their eighth-grade enrollment was 20 or more. Schools whose eighth-grade enrollment was at least 20 were referred to as Type 3A Clusters. Schools with fewer than 20 eighth-grade students were clustered into groups called Type 3B Clusters, using the following algorithm:

1. Schools were ordered from smallest to largest.

2. The largest school on the list was grouped with the smallest. If the sum of the enrollments was 20 or more, a cluster number was assigned to the two schools and they were removed from the list.

3. If the sum of the enrollments was at most 19, the next smallest school was added. Small schools continued to be added until the sum of the enrollments was at least 20, then a cluster number was assigned and the schools were removed from the list.

4. Steps 2 and 3 were repeated for each subsequent largest school.

5. If at the end of the last grouping there was a residual cluster (total enrollment less than 20) this was added to the previous cluster. The following example illustrates this point. If schools with the following enrollments remained to be clustered, the clustering would be done as shown.

Enrollments:	19, 18, 15, 10, 8, 2, 2, 1, 1
Clusters:	19+1, 18+1+2, 15+2+8+10

This approach assured that no clusters had student enrollment less than 20.

TABLE 3-1: SAMPLING AND STRATIFICATION RULES BY STATE

State	Small Schools	Type of Cluster	Original Sampled Non-Certainty Schools	Original Sampled Certainty Schools	Urbanicity	Minority
AL	Geographic	Type 2	104	2	Central	Minority
					MSA	Minority
					Other	Minority
AR	Stratified	Type 3A/3B	90	17	Central/MSA	Low Minority
					Central/MSA	Medium Minority
					Central/MSA	High Minority
					Other	Low Minority
					Other	Medium Minority
					Other	High Minority
AZ	Stratified	Type 3A/3B	76	34	Central/MSA	Low Black/Low Hispanic
					Central/MSA	Low Black/High Hispanic
					Central/MSA	High Black/Low Hispanic
					Central/MSA	High Black/High Hispani
					Other	Low Minority
					Other	Medium Minority
					Other	High Minority
CA	Stratified	Type 3A/3B	106	0	Central	Low Black/Low Hispanic
					Central	Low Black/High Hispani
					Central	High Black/Low Hispani
					Central	High Black/High Hispani
					MSA	Low Minority
					MSA	Medium Minority
					MSA	High Minority
					Other	Low Minority
					Other	Medium Minority
					Other	High Minority
CO	Stratified	Type 3A/3B	98	9	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	Low Minority
					MSA	Medium Minority
					MSA	High Minority
					Other	Low Minority
					Other	Medium Minority
					Other	High Minority

TABLE 3-1: SAMPLING AND STRATIFICATION RULES BY STATE

State	Small Schools	Type of Cluster	Original Sampled Non-Certainty Schools	Original Sampled Certainty Schools	Urbanicity	Minority
CT	Geographic	Type 2	88	20	Central	Low Black/Low Hispanic
					Central	Low Black/High Hispanic
					Central	High Black/Low Hispanic
					Central	High Black/High Hispanic
					MSA	None
					Other	None
DC	All schools taken	Type 1	0	36	–	None
DE	All schools taken	Type 1	5	32	–	None
FL	Stratified	Type 3A/3B	108	0	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	Low Black/Low Hispanic
					MSA	Low Black/High Hispanic
					MSA	High Black/Low Hispanic
					MSA	High Black/High Hispanic
					Other	Low Minority
					Other	Medium Minority
					Other	High Minority
GA	Geographic	Type 2	109	0	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	Low Minority
					MSA	Medium Minority
					MSA	High Minority
					Other	Low Minority
					Other	Medium Minority
					Other	High Minority
GU	All schools taken	Type 1	0	7	–	None
HI	All schools taken	Type 1	0	57	–	None

TABLE 3-1: SAMPLING AND STRATIFICATION RULES BY STATE

State	Small Schools	Type of Cluster	Original Sampled Non-Certainty Schools	Original Sampled Certainty Schools	Urbanicity	Minority
IA	Stratified	Type 3A/3B	101	7	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	None
					Other	None
ID	Stratified	Type 3A/3B	29	79	–	–
IL	Stratified	Type 3A/3B	107	0	Central	Low Black/Low Hispanic
					Central	Low Black/High Hispanic
					Central	High Black/Low Hispanic
					Central	High Black/High Hispanic
					MSA	Low Minority
					MSA	Medium Minority
					MSA	High Minority
					Other	None
IN	Geographic	Type 2	104	1	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	None
					Other	None
KY	Stratified	Type 3A/3B	109	3	Central/MSA	Low Minority
					Central/MSA	Medium Minority
					Central/MSA	High Minority
					Other	None
LA	Stratified	Type 3A/3B	106	2	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	Low Minority
					MSA	Medium Minority
					MSA	High Minority
					Other	Low Minority
					Other	Medium Minority
					Other	High Minority

TABLE 3-1: SAMPLING AND STRATIFICATION RULES BY STATE

State	Small Schools	Type of Cluster	Original Sampled Non-Certainty Schools	Original Sampled Certainty Schools	Urbanicity	Minority
MD	Geographic	Type 2	104	3	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	Low Minority
					MSA	Medium Minority
					MSA	High Minority
					Other	Low Minority
					Other	Medium Minority
					Other	High Minority
MI	Geographic	Type 2	105	0	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	None
					Other	None
MN	Stratified	Type 3A/3B	104	4	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	None
					Other	None
MT	Stratified	Type 3A/3B	67	57	–	–
NC	Geographic	Type 2	111	0	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	Low Minority
					MSA	Medium Minority
					MSA	High Minority
					Other	Low Minority
					Other	Medium Minority
					Other	High Minority
ND	Stratified	Type 3A/3B	53	58	Central/MSA	None
					Other	None

TABLE 3-1: SAMPLING AND STRATIFICATION RULES BY STATE

State	Small Schools	Type of Cluster	Original Sampled Non-Certainty Schools	Original Sampled Certainty Schools	Urbanicity	Minority
NE	Stratified	Type 3A/3B	79	42	Central/MSA	None
					Other	None
NH	Stratified	Type 3A/3B	22	85	Central	None
					MSA	None
					Other	None
NJ	Geographic	Type 2	111	1	Central	Low Black/Low Hispanic
					Central	Low Black/High Hispanic
					Central	High Black/Low Hispanic
					Central	High Black/High Hispani
					MSA	Low Minority
					MSA	Medium Minority
					MSA	High Minority
NM	Stratified	Type 3A/3B	25	83	Central/MSA	None
					Other	None
NY	Geographic	Type 2	105	0	Central	Black/Hispanic
					MSA	Minority
					Other	None
OH	Geographic	Type 2	105	0	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	None
					Other	None
OK	Stratified	Type 3A/3B	103	9	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	None
					Other	None

TABLE 3-1: SAMPLING AND STRATIFICATION RULES BY STATE

State	Small Schools	Type of Cluster	Original Sampled Non-Certainty Schools	Original Sampled Certainty Schools	Urbanicity	Minority
OR	Stratified	Type 3A/3B	101	8	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	None
					Other	None
PA	Geographic	Type 2	106	0	Central	Minority
					MSA	None
					Other	None
RI	All schools taken	Type 1	0	52	–	None
TX	Stratified	Type 3A/3B	107	0	Central	Low Black/Low Hispanic
					Central	Low Black/High Hispanic
					Central	High Black/Low Hispanic
					Central	High Black/High Hispanic
					MSA	Low Black/Low Hispanic
					MSA	Low Black/High Hispanic
					MSA	High Black/Low Hispanic
					MSA	High Black/High Hispanic
					Other	Low Black/Low Hispanic
					Other	Low Black/High Hispanic
					Other	High Black/Low Hispanic
					Other	High Black/High Hispanic
VA	Geographic	Type 2	104	2	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	Medium Minority
					MSA	High Minority
					Other	Low Minority
					Other	Medium Minority
					Other	High Minority
VI	All schools taken	Type 1	0	6	–	None

TABLE 3-1: SAMPLING AND STRATIFICATION RULES BY STATE

State	Small Schools	Type of Cluster	Original Sampled Non-Certainty Schools	Original Sampled Certainty Schools	Urbanicity	Minority
WI	Stratified	Type 3A/3B	108	1	Central	Low Minority
					Central	Medium Minority
					Central	High Minority
					MSA	None
					Other	None
WV	Stratified	Type 3A/3B	102	5	Central	None
					MSA	None
					Other	None
WY	All > 20 taken < 20 Stratified		11	58	–	None

.E 3-1: SCHOOL PARTICIPATION STATUS BY STATE

	NON-CERTAINTY UNITS			CERTAINTY UNITS				
State	Inscope Participating	Inscope Non-Part.	Out of Scope	Inscope Participating	Inscope Non-Part.	Out of Scope	Substitutes Offered	Participating Substitutes
AL	86	13	5	1	1	0	14	11
AR	90	0	0	17	0	0	0	0
AZ	67	3	6	33	0	1	1	0
CA	98	6	2	0	0	0	0	0
CO	96	0	2	9	0	0	0	0
CT	84	0	4	19	0	1	0	0
DC	0	0	0	36	0	0	0	0
DE	0	0	5	30	0	2	0	0
FL	100	2	6	0	0	0	2	0
GA	106	0	3	0	0	0	0	0
GU	0	0	0	6	0	1	0	0
HI	0	0	0	52	1	4	1	0
IA	85	9	7	7	0	0	9	0
ID	26	2	1	75	3	1	5	0
IL	82	23	2	0	0	0	23	19
IN	91	12	1	1	0	0	12	5
KY	101	0	8	3	0	0	0	0
LA	97	0	9	2	0	0	0	0
MD	102	0	2	3	0	0	0	0
MI	90	11	4	0	0	0	11	8
MN	90	11	3	4	0	0	11	3
MT	49	10	8	51	6	0	16	0
NC	106	0	5	0	0	0	0	0
ND	46	4	3	52	4	2	8	8
NE	55	16	8	39	3	0	19	9
NH	21	1	0	73	9	3	10	4
NJ	105	3	3	1	0	0	3	1
NM	24	0	1	82	0	1	0	0
NY	90	15	0	0	0	0	0	0

T E 3-1: SCHOOL PARTICIPATION STATUS BY STATE

	NON-CERTAINTY UNITS			CERTAINTY UNITS				
State	Inscope Participating	Inscope Non-Part.	Out of Scope	Inscope Participating	Inscope Non-Part.	Out of Scope	Substitutes Offered	Participating Substitutes
OH	99	4	2	0	0	0	4	2
OK	76	27	0	9	0	0	27	23
OR	98	0	3	8	0	0	0	0
PA	92	10	4	0	0	0	6	3
RI	0	0	0	49	3	0	2	2
TX	91	12	4	0	0	0	12	9
VA	102	1	1	2	0	0	1	0
VI	0	0	0	6	0	0	0	0
WI	104	1	3	1	0	0	1	0
WV	96	0	6	5	0	0	0	0
WY	11	0	0	58	0	0	0	0

The number of Type 3B Clusters selected was proportionate to the number of students that attended schools with 20 or fewer students, up to a maximum of ten clusters. This maximum was imposed to keep the size of the sample of small schools to within reasonable bounds. These Type 3B clusters were not stratified on urbanicity, minority enrollment, or income, but were selected systematically with probability proportionate to the total eighth-grade enrollment in the cluster. Type 3A Clusters were stratified on urbanicity, minority enrollment, and income as discussed above.

3.3.3 Selection of School Sample

3.3.3.1 States With Geographic Clustering of Small Schools (Type 2 Clusters)

In states with 200 or fewer schools, clusters were sorted by urbanicity and median income. In states with more than 200 schools, clusters were sorted by urbanicity, minority strata (which varied by state and urbanicity level), and median income. After the removal of certainty schools (those with selection probability greater than 1), a systematic sample of clusters was then selected with probability proportionate to total eighth-grade enrollment, to provide a total sample of 105 clusters.

Following the selection of clusters, there was some thinning of small schools. The purpose of thinning was to give students in small schools (enrollment less than 20) approximately the same chance of selection as those from larger schools, and to control the sample size of schools to be close to the desired number of 105. All small schools in a cluster were discarded from the sample with probability 30/X, where X denotes the total enrollment for all schools in the cluster to which the small schools belonged. Otherwise, the small schools were retained in the sample.

3.3.3.2 States With Stratification of Small Schools (Type 3A and 3B Clusters)

For all states, the percentage of eighth grade students in the state who attended small schools (i.e., schools with 19 or fewer students) was determined. The sample design for the selection of *small schools* was the same for all such states, except Montana, North Dakota, and Nebraska. In every state the percent of students in small schools, p, was rounded to the nearest integer that was at least one, with this integer being called k. Montana, North Dakota, and Nebraska were exceptions where the values of k which were in excess of 10 in each case were reduced to 10 to keep the sample size of small schools to within reasonable bounds. A random sample of k clusters of small schools was selected.

The sample selection of *large schools* (i.e., schools with 20 or more students) varied by state. In states with 105 or fewer schools, all large schools were selected. In states with 106 to 200 schools, after the large schools were sorted by urbanicity and median income, and certainty selections were removed, a systematic sample of schools was selected with probability proportionate to total eighth-grade enrollment, such that the total sample size of large schools, including certainty selections, was $(105 - k)$. The exceptions were Montana, North Dakota, and Nebraska, where the total sample size for large schools was set at 90 in each case. Once again, the special exception for these states was designed to limit the total number of schools selected.

In states with more than 200 schools, the large schools were sorted by urbanicity, minority strata, and median income, and the certainty schools were removed. Then a systematic sample of schools was selected with probability proportionate to total eighth-grade enrollment such that the total sample of large schools, including certainties, was (105 - *k*).

After the sample of schools was selected, weighted tabulations were produced to verify that it was representative of the population. The number of clusters sampled, along with the estimated and actual counts of clusters and students, were listed by urbanicity level and minority level for each state. The differences between the actual and estimated numbers of clusters were also calculated. For example, Table 3-2 shows the difference between actual and estimated counts of clusters for the state of New York. The differences shown in Table 3-2 were very similar to what happened in most states, with the exceptions being that the percentage difference between the number of estimated versus actual clusters was higher when the number of sampled clusters was small.

3.3.3.3 Designating Schools to be Monitored

The objective in assigning each school to be monitored or otherwise, was to produce two equivalent half-samples. This was achieved by pairing similar clusters, and randomly designating one pair member to be monitored, independently from pair to pair. This was achieved using the following procedure.

The sampled clusters were sorted in the order in which they were selected. The basic algorithm for designating schools to be monitored was based on the following steps:

1. A random number, x, was generated uniformly between zero and one. If x>.50, then the first cluster was designated to be monitored, otherwise the second was designated.

2. Step one was repeated for each succeeding pair in the sort order.

3. If there was an even number of clusters, then steps one and two were sufficient for designating schools for monitoring. Otherwise, a random number x was generated for the last cluster and that cluster was designated for monitoring if x > .50.

The above steps were followed exactly for Type 1, 2, and 3A clusters. For Type 3B clusters, the algorithm was applied to schools within clusters, with random sort order within cluster.

This algorithm was followed for all schools in all states except for those states where schools were not stratified. For these states, pairing was done on the basis of school size because these states in general showed little variation with regard to urbanicity and race/ethnicity, and there was no household income data readily available.

TABLE 3-2: **DIFFERENCES BETWEEN ACTUAL AND ESTIMATED COUNTS**
NEW YORK STATE

Minority Strata	Urbanicity Strata	# Clusters Sampled	# Actual Clusters	# Estimated Clusters	Cluster Difference	# Actual Students	# Estimated Students	Student Difference
Low Black/Low Hispanic		8	55	37	-0.32421	12,916	12,509	-0.031497
Low Black/High Hispanic		14	75	71	-0.05721	21,745	21,891	0.006717
High Black/Low Hispanic		9	73	94	0.28461	13,149	14,073	0.070258
High Black/High Hispanic		10	57	59	0.02784	15,656	15,636	-0.001247
Low Minority		9	117	80	-0.31394	14,217	14,073	-0.010141
Medium Minority		24	190	222	0.16780	36,690	37,528	0.022828
High Minority		19	141	125	-0.11389	30,256	29,709	-0.018069
None		12	224	199	-0.11175	19,554	18,764	-0.040413
	Central City	41	260	260	0.00093	63,466	64,110	0.010140
	Suburban	52	448	427	-0.04667	81,163	81,310	0.001807
	Other	12	224	199	-0.11175	19,554	18,764	-0.040413
STATE TOTAL			1,864	1,773		328,366	328,366	

3.3.3.4 Updating the School Sample with New Schools

In sampling for the Trial State Assessment, some districts had new schools that were not listed on the sampling frame, either because these schools were completely new or because they had been formed by some combination of old schools. In either case, to provide a mechanism for allowing these new schools into the sample, after the initial sample was selected, all districts in which schools were sampled were contacted and provided with the list of schools from the sampling frame for that district. The district was then asked to provide an updated list containing any schools not listed on the frame which were operating in the 1989-1990 school year and which contained the eligible grade. A sample of new schools was then selected from the lists provided. In order that all schools in each participating state had a chance of being selected for the Trial State Assessment, schools on the updated list were sampled and, if selected, were asked to participate in the program. The determination as to how many new schools were selected and how the data from selected schools were weighted is discussed below. Since a self-weighting sample of students was desired, the required sample size of new schools depended on the method used to weight the data estimation. Table 3-3 shows the number of new schools sampled per state.

In previous national assessments, unconditional selection probabilities were determined for those districts already having at least one school selected, and then used to determine the sample selection probabilities for new schools from such districts. For the Trial State Assessment, a somewhat simpler approach that involved multiplicity weighting was used. In the following discussion, the multiplicity weighting approach is described, along with the implications for sampling and weighting new schools.

The number of schools selected in a district was determined so that, after correcting for the multiple-selection probabilities of the district, the chance of a student's being selected for assessment was approximately the same for all students in the state, regardless of whether the student attended a new school or an older one.

Estimating the sample size of new schools was relatively straightforward. To obtain the new school sample size for each district, the sum of eighth-grade enrollments in new schools in that district was multiplied by a factor derived from data on the district's sampled schools. This factor, which was calculated for every district that had a sampled school, was represented by the following equation:

$$\begin{aligned} \mathit{Factor} &= \left(\frac{\mathit{District\ Total\ of\ Sampled\ School\ Weights}}{\mathit{Number\ of\ Schools\ in\ District}} \right) \\ &\ast \left(\frac{3000}{\mathit{State\ Total\ Grade\ Eight\ Enrollment} \ast 30} \right) \end{aligned}$$

The derivation of this factor is given in Chapter 8, Section 8.2.2.

TABLE 3-3: NUMBER OF SCHOOLS BY STATE

State	Total # Schools	Total # 8th Graders	# NELS Schools Dropped From Frame	# of Clusters	# Schools Thinned	Original # of Sampled Schools after Thinning	# Sampled New Schools Added	# Out of Scope Schools	# Regular Session Substitutes Provided	# Double Session Substitutes Provided
AL	533	54,912	5	105	5	106	0	5	14	0
AR	351	33,068	3	105	-	107	0	0	0	0
AZ	321	39,862	0	105	-	107	3	7	3	0
CA	1,667	294,111	7	105	-	106	0	2	0	0
CO	316	37,660	0	105	-	107	0	2	0	0
CT	225	31,144	3	105	7	107	1	5	0	0
DC	36	5,710	0	1	-	36	0	0	0	0
DE	47	6,284	1	33	-	36	1	7	0	0
FL	577	121,984	3	105	-	108	1	7	2	0
GA	400	80,060	3	105	5	105	4	3	0	0
GU	7	1,920	0	1	-	7	0	1	0	0
HI	55	10,865	0	1	-	55	2	4	1	0
IA	451	31,072	2	105	-	108	0	7	9	0
ID	153	15,513	0	105	-	107	1	2	5	0
IL	1,423	118,317	4	105	-	107	0	2	23	0
IN	455	68,134	4	105	8	105	0	1	12	0
KY	518	44,386	2	105	-	107	5	8	0	0
LA	440	52,229	3	105	-	106	2	9	0	0
MD	228	45,481	0	105	6	107	0	2	0	0
MI	787	106,403	6	105	8	105	0	4	11	0
MN	466	48,190	1	105	-	107	1	3	11	0
MT	293	10,113	0	100	-	123	1	8	16	0
NC	552	80,332	1	105	5	109	2	5	0	0
ND	286	8,553	2	100	-	111	0	5	8	0
NE	611	19,418	1	97	-	120	1	8	17	2
NH	127	11,582	1	101	-	127	0	3	6	4
NJ	659	71,771	2	105	3	110	2	3	3	0
NM	149	18,801	1	105	-	108	0	2	0	0
NY	969	172,041	19	105	5	105	0	0	0	0
OH	875	126,271	9	105	6	105	0	2	4	0
OK	632	40,814	0	105	-	112	0	0	27	0
OR	355	33,195	0	105	-	109	1	4	0	0
PA	726	115,038	11	105	1	105	1	4	6	0

TABLE 3-3: NUMBER OF SCHOOLS BY STATE

State	Total # Schools	Total # 8th Graders	# NELS Schools Dropped From Frame	# of Clusters	# Schools Thinned	Original # of Sampled Schools after Thinning	# Sampled New Schools Added	# Out of Scope Schools	# Regular Session Substitutes Provided	# Double Session Substitutes Provided
RI	52	9,543	1	1	-	52	0	0	1	2
TX	1,592	220,661	2	105	-	107	0	4	12	0
VA	324	67,459	5	105	1	105	1	1	1	0
VI	6	1,626	0	1	-	6	0	0	0	0
WI	524	50,737	2	105	-	107	2	3	1	0
WV	274	25,858	1	105	-	106	1	6	0	0
WY	89	7,103	0	63	-	69	0	0	0	0
Total	18,551	2,338,221	105	3544	60	3822	33	139	193	8

Once the sample size for new schools was estimated, the sample was selected in the following manner. The new schools were ordered from highest to lowest eighth-grade enrollment. The certainty units, if any, were selected, and the sample size and total eighth-grade enrollment was decremented by the number of certainty units and the eighth-grade enrollment of certainties, respectively. A systematic sample was then selected of the remaining new schools. It should be noted that in most cases the expected sample size for new schools was less than one, and, depending on the random start selected, no schools were selected. Section 8.2.2 gives further details of the method of sampling new schools.

3.3.3.5 School Substitution

A substitute school was selected for each selected school containing eligible students, for which school non-participation was established by the state coordinator as of November 10, 1989.[1] The process of selecting a substitute for a school involved identifying the most similar school in terms of the following characteristics: urbanicity, percentage of Black enrollment, percentage of Hispanic enrollment, eighth-grade enrollment, and median income.

To identify candidates for substitution, a set of schools were found which provided reasonable matches with regard to eighth-grade enrollment and percentage of Black and Hispanic enrollment. From among this set a match was selected considering all five characteristics. Schools in the national assessment sample or the 1990 National Education Longitudinal Study First Phase Follow-up were avoided in the selection of substitutes, where possible. Furthermore, the substitute was selected from the same district, wherever possible, to avoid placing the burden of replacing a refusing school from one district on another district. This was often not possible, however, because in the majority of cases, the decision not to participate was made at the district level.

In a few cases where no suitable substitute could be found among those schools not sampled (most often because all or most schools were included in the original sample), a school already in the sample conducted a double session, of which one session served as a substitute for students in the refusing school. The same criteria were applied in selecting the schools that conducted double sessions, i.e., a reasonable match was found based on eighth-grade enrollment, percentage of Black and Hispanic enrollment, median income, and urbanicity.

Table 3-3 includes information about the number of substitutes provided in each state. Of the forty jurisdictions participating, fourteen were provided with at least one substitute. Among states receiving no substitutes, the majority had 100 percent participation from the original sample, but in a few cases refusals did occur, following the November 10 deadline. The number of substitutes provided to a state ranged from 1 to 23, with 110 substitutes being provided in total (102 substitutes of schools not originally selected, and eight double session substitutes). Some states did not attempt to solicit participation from the substitute schools provided, as they considered the timing too late to seek cooperation from schools not previously notified about the assessment.

[1] Appendix B contains a summary of the participation rate data -- including the number of substitute schools provided and the number of substitute schools that participated in the Trial State Assessment -- for each jurisdiction. including the

3.3.4 Student Sample Selection

For all schools in each state, a student sample size of 30 was drawn from each selected school per state, expect for states with fewer than 100 schools (Type 1 Clusters). In these states either 60 or 100 students were sampled in the larger-certainty schools, depending on the size of the state and the size of the school.

In November 1989, school officials were asked to forward a list of the names of all of the eighth-grade students enrolled in the school to a central location (usually the State Department of Education). Schools were not asked to list students in any particular order, but were asked to implement checks to ensure that all eighth-grade students were listed. Based on the total number of students on this list, called the Student Listing Form (SLF), sample line numbers were generated for student sample selection. To generate these line numbers, the person responsible for drawing the sample (typically, the State Supervisor) went to the State Department of Education and entered the following into a calculator that had been programmed with the sampling algorithm: the number of students on the SLF, the state identity, and the sample size if it was different from 30. The calculator generated a random start which was used to systematically select the 30 (or more if necessary) line numbers. To compensate for new enrollees not on the SLF, extra line numbers were generated for a supplemental student sample of new students. All students were selected in those schools with 35 or fewer eighth-grade enrollees. This sample design was intended to give each student within the state approximately the same probability of selection.

After the student sample was selected, the administrator at each school excluded students who were incapable of taking the assessment (i.e., students who either had an Individualized Education Plan or who were Limited English Proficient).

When the assessment was conducted in a given school, a count was made of the number of non-excluded 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 all students who were absent from the initial session were asked to attend.

3.4 SCHOOL AND STUDENT PARTICIPATION RATES

The levels of school participation varied considerably across the forty participating jurisdictions. Prior to substitution, weighted response rates (for which each school was weighted inversely to its selection probability) ranged from 73 percent to 100 percent. The two states with relative low initial response rates obtained good cooperation from their substitute schools, so that, after substitution, the lowest response rate was 85 percent.

Student participation rates were uniformly high, except in one state where parental consent requirements keep this rate at 80 percent.

Table 3-4 provides the school and student participation rates, and the rate of student exclusion, for each of the forty participating jurisdictions. Appendix B contains the derivations of the weighted participation rates and the NCES guidelines for levels of school and student participation.

Table 3-4: PARTICIPATION AND EXCLUSION RATES

STATE	FINAL SCHOOL RESPONSE RATE	FINAL STUDENT RESPONSE RATE	EXCLUSION RATE
Alabama	97.1%	95.3%	5.74%
Arkansas	100.0%	95.2%	8.03%
Arizona	97.2%	93.0%	5.36%
California	93.9%	92.7%	8.23%
Colorado	100.0%	94.2%	4.92%
Connecticut	100.0%	94.8%	7.11%
Delaware	100.0%	92.9%	4.50%
District of Columbia	100.0%	87.8%	5.99%
Florida	98.0%	92.2%	7.01%
Georgia	100.0%	94.2%	3.81%
Guam	100.0%	93.0%	4.13%
Hawaii	100.0%	92.7%	5.04%
Idaho	97.1%	96.0%	2.54%
Illinois	96.2%	95.5%	5.64%
Indiana	93.6%	94.9%	5.25%
Iowa	91.3%	95.9%	3.94%
Kentucky	100.0%	95.2%	5.24%
Louisiana	100.0%	94.3%	4.49%
Maryland	100.0%	94.0%	5.00%
Michigan	97.1%	94.7%	4.59%
Minnesota	93.3%	95.4%	3.24%
Montana	90.3%	96.1%	2.44%
Nebraska	94.3%	95.1%	3.13%
New Hampshire	96.9%	94.9%	4.71%
New Jersey	98.2%	94.3%	7.66%
New Mexico	100.0%	94.0%	7.32%
New York	85.5%	92.8%	6.98%
North Carolina	100.0%	94.6%	3.44%
North Dakota	100.0%	96.3%	2.91%
Ohio	98.1%	95.2%	5.83%
Oklahoma	98.5%	80.0%	5.71%
Oregon	100.0%	93.4%	3.04%
Pennsylvania	93.3%	94.4%	5.51%
Rhode Island	97.2%	93.4%	7.08%
Texas	97.2%	95.7%	6.65%
Virginia	99.0%	93.8%	5.74%
Virgin Islands	100.0%	93.1%	3.26%
West Virginia	100.0%	94.1%	5.92%
Wisconsin	99.1%	94.2%	4.69%
Wyoming	100.0%	95.8%	3.81%
AVERAGE RATE	97.6%	93.9%	5.1%

Chapter 4

STATE & SCHOOL COOPERATION

Nancy Caldwell

Westat

4.1 OVERVIEW

This chapter describes the process of selecting the schools and the schools' cooperation rates for both the 1989 field test and 1990 Trial State Assessment Program.

4.2 THE 1989 FIELD TEST

In preparation for the 1990 Program, a field test of the forms, procedures, and booklet items was held in early 1989. The 1989 field test was designed to give states an opportunity to learn about their responsibilities and rehearse the data collection procedures envisioned for 1990 -- with a much smaller set of schools and students than would be included in 1990.

The 1989 field test was conducted from February 6 to March 3, 1989 in 23 states, the District of Columbia, and three U.S. Territories volunteering to participate. Table 4-1 lists the participating states and territories.

The intent of the field test was to give states and the contractors practice and experience with the proposed materials and procedures. For these purposes, it was decided that a sample of nine schools in each jurisdiction would be sufficient except for Guam and the Virgin Islands in which all of their schools were asked to participate in the field test because of the small number of schools in these territories. For each state and the District of Columbia, Westat, the sampling and field operations contractor, selected nine pairs of schools clustered in two geographical locations. Because the sample for each state and territory was so small, no attempt was made to make it representative at the jurisdiction level; it was representative only at the national level.

Each jurisdiction volunteering to participate in the 1989 State Field Test was asked to appoint a State Coordinator. In general, the State Coordinator was the liaison between NAEP/Westat staff and the participating schools. The State Coordinator was sent a list of the schools and requested to obtain the cooperation of one school from each pair. In a few rare instances, the State Coordinator could not obtain cooperation from either school in the pair. For these situations, Westat provided a substitute school(s) which were asked to participate.

Table 4-1

Jurisdictions Participating
in the 1989 Field Test

Jurisdictions		
American Samoa	Michigan	Oregon
Arkansas	Minnesota	Pennsylvania
District of Columbia	Mississippi	South Carolina
Florida	Montana	Texas
Guam	Nebraska	Virginia
Illinois	Nevada	Virgin Islands
Kentucky	New Mexico	West Virginia
Louisiana	Ohio	Wisconsin
Maryland	Oklahoma	Wyoming

All jurisdictions were able to elicit the cooperation of the requisite number of schools. A total of 237 schools agreed to participate. Of these 237 schools, 233 actually conducted assessments. In the other four schools, last minute refusals, weather, and scheduling problems led to the cancellation of the assessment.

4.3 THE 1990 TRIAL STATE ASSESSMENT

Thirty-seven states, the District of Columbia, and two territories volunteered for the 1990 Trial State Assessment. A thirty-eighth state, South Carolina, agreed to participate, but had to withdraw because of the damage done in the state by Hurricane Hugo. Table 4-2 lists the jurisdictions participating in the 1990 program.

A stratified random sample of approximately 100 schools was selected by Westat for each jurisdiction having that number of schools with eighth grade. Chapter 3 contains detailed information about the selection of schools.

Lists of the sampled schools were sent to the State Coordinators in July 1989, along with instructions describing their responsibilities in the assessment. Additional materials, such as forms to be used to list eligible students, NAEP reports, and descriptions of the schools' role in the assessment were sent by ETS and Westat in the fall. State Coordinators were requested to determine the cooperation of the sampled schools and to notify Westat by November 1. Substitute schools were selected for those schools that had refused by that date (see section 3.3.3.5 for more details about school substitution).

Table 4-2

Jurisdictions Participating
in the 1990 Trial State Assessment Program

Jurisdictions			
Alabama	Guam	Minnesota	Oklahoma
Arizona	Hawaii	Montana	Oregon
Arkansas	Idaho	Nebraska	Pennsylvania
California	Illinois	New Hampshire	Rhode Island
Colorado	Indiana	New Jersey	Texas
Connecticut	Iowa	New York	Virginia
Delaware	Kentucky	New Mexico	Virgin Islands
District of Columbia	Louisiana	North Carolina	West Virginia
Florida	Maryland	North Dakota	Wisconsin
Georgia	Michigan	Ohio	Wyoming

Table 4-3 provides the results of the State Coordinators' efforts to gain the cooperation of the selected schools. In summary, 94 percent of schools in the original sample that could have participated in the assessment did so. An additional 100 substitute schools also agreed to participate. Thus, the participation rate, after substitution for refusals, was 97 percent.

Table 4-3

Cooperation Rates

	Number	Percent
Schools in the original sample	3853	
Schools that were out of scope (closed, no eighth grade, not a regular school)	137	
Schools "eligible" to participate	3716	100.0%
Original schools participating	3496	94.1%
Nonparticipating schools (school or district refusal)	220	5.9%
Participating substitute schools	100	
Total participating schools	3596	96.8%

Chapter 5

FIELD ADMINISTRATION

Nancy Caldwell

Westat

5.1 OVERVIEW

The data collection for the 1990 Trial State Assessment and its 1989 field test involved a collaborative effort between the participating states and the NAEP contractors, especially Westat, the field administration contractor. Westat's responsibilities included:

- Selecting the sample of schools and students for each participating state;
- Developing the administration procedures and manuals;
- Training the state personnel who would conduct the assessments; and
- Conducting an extensive quality assurance program.

Each jurisdiction volunteering to participate in the 1989 field test and in the 1990 Program was asked to appoint a State Coordinator. In general, the State 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 eighth-grade students from the schools;
- Coordinate the flow of information between the schools and the NAEP contractors;
- Provide space for the State Supervisor to use when sampling;
- Notify Local Administrators about training and send them their manuals;
- Send the lists of sampled students to the schools.

At the local school level, a Local 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 staff. The Local 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;
- Notifying sampled students and their teachers;
- Administering the assessment session;
- Completing assessment forms; and
- Preparing the assessment materials for shipment.

Westat hired and trained a State Supervisor for each state. The State Supervisors were responsible for working with the State Coordinators and overseeing assessment activities. Their primary tasks were:

- To make sure the arrangements for the assessments were set and Local administrators identified;
- To schedule and conduct the Local Administrators training sessions;
- To select the sample of students to be assessed; and
- To coordinate the monitoring of the assessment sessions and makeup sessions.

For the Trial State Assessment, Westat hired and trained four Quality Control Monitors in each state to monitor 50 percent of the assessment sessions. During the field test, the State Supervisors monitored all sessions.

5.2 THE 1989 FIELD TEST

5.2.1 The Field Test Schedule

The schedule of activities for the 1989 field test was as follows:

November 11, 1988	Field Test materials including List of Selected Schools mailed to State Coordinators by NAEP/Westat.
Nov. 11-Dec. 9, 1988	State Coordinators secured cooperation of schools, obtained names of Local Administrators, and determined

eighth-grade enrollment in participating schools. As soon as cooperation was secured, State Coordinators mailed Summary of School Tasks, Student Listing forms, and Principal Questionnaires to participating schools.

December 9, 1988

Meeting in Washington, D.C. for the State Coordinators from states participating in the field test, reviewed and commented on the Local Administrator's Manual and discuss procedures for the field test.

December 16, 1988

Final date for State Coordinators to return one copy of the list of Selected Schools to Westat. The lists were updated to identify the participating schools, Local Administrators, and eighth-grade enrollment.

January 5-7, 1989

Training of State Supervisors.

January 6, 1989

Final date for schools to send State Coordinators the Principal Questionnaires and lists of all students enrolled in the eighth grade.

January 9-13, 1989

State Coordinators reviewed lists of enrolled eighth-grade students for completeness. State Supervisors telephoned State Coordinators to inform them of sites for Local Administrator training and to discuss assessment schedules and arrangements for sampling. State Coordinators notified Local Administrators of training sites.

January 16-27, 1989

State Supervisors visited state offices and selected student samples.

Feb. 6-March 3, 1989

All field tests in a state were conducted during a specific week within this four-week period.

Approximately 10-14
days before the field
test assessment week:

The State Coordinator sent the lists of sampled students and Local Administrator's Manuals to the schools;

Local Administrators were trained by NAEP contractor staff; and

Schools received assessment materials from NAEP/ETS.

March 6-10, 1989

Makeup sessions were conducted in schools when it was not possible to schedule an assessment during the four-week assessment period.

5.2.2 Preparation for the Field Test

Westat home office staff coordinated preliminary field test activities with the State Coordinators. As the assessment period approached, Westat hired and trained a State Supervisor to work within each participating state.

State Supervisors called the State Coordinators in early January 1989 to verify information on the participating schools, assessment dates, Local Administrator training dates, and to make plans to visit the State Coordinators' offices to draw the sample of students.

Working from lists of eighth-grade students sent to the State Coordinator's office, the State Supervisor drew a random sample of 30 students per school in late January. If a school had 35 or fewer eighth-grade students, all students were included in the sample.

After sampling, the supervisor copied the names of the sampled students onto Administration Schedules (rosters). These rosters were left with the State Coordinator to be sent to the schools approximately two weeks before the field test was scheduled.

A six-hour training session for Local Administrators was conducted by experienced NAEP/Westat trainers in each state about two weeks before the state's field test week. In a few states, two sessions were held to avoid requiring Local Administrators to travel long distances or staying overnight.

After receiving the Administration Schedule, the Local Administrator followed NAEP procedures to: select a sample of newly enrolled students, review the Administration Schedule, and identify students who could not be assessed according to NAEP criteria. NAEP/NCS sent Assessment booklets and questionnaires to the school two weeks to ten days before the scheduled field test administration. The questionnaires were to be distributed to the appropriate school staff and collected before the day of the administration.

A week before the administration, the supervisor called the Local Administrator to ensure that assessment materials had arrived from NAEP/NCS and that the Local Administrator had begun the preliminary activities such as distribution of questionnaires and sampling of newly enrolled students. The supervisor also verified the assessment date and time and answered any questions the Local Administrator had.

5.2.3 Assessment Sessions

On the day of the field test assessment, the State supervisor brought the calculators, protractor-rulers, and the timer; observed the session; and mailed the school's assessment materials to NAEP/NCS.

State supervisors observed all field test sessions to evaluate procedures and materials. An Observation Form was used to record information about the major events related to the assessment. The State supervisor was to arrive at the school one hour before the assessment was scheduled to observe the opening of the bundle of assessment booklets. He or she also reviewed the updating of the Administration Schedule by the Local Administrator, the exclusion of students from the assessment, and the sampling of newly enrolled students.

During the assessment session, the supervisor observed and recorded the accuracy with which the Local Administrator followed the prepared script for administering the session and timed the booklet sections. The State Supervisors were instructed that their role was strictly one of an observer, allowing the Local Administrator to conduct the session in his or her own way. The supervisor was to intervene only if he or she felt the deviations from procedures would jeopardize the assessment. The Observation Form noted points in the assessment where the supervisor might need to intervene.

After the assessment was completed, the supervisor asked the students whether they had any difficulty with the assessment items. Comments were noted in the Observation Form to assist the test developers in preparing the 1990 assessment items.

The supervisor observed the Local Administrator's record-keeping after the assessment and reviewed the completed Administration Schedule and School Worksheet. The Local Administrator packed the box of materials and gave it to the supervisor to mail to NAEP/NCS.

If more than four students were absent, a makeup session was required. In this case, the Local Administrator and State Supervisor discussed a mutually acceptable date for the makeup session. When a makeup session was required, the State Supervisor took the assessment materials from the school and, on the appropriate day, brought the materials needed for the makeup session.

After all activities in the school were completed, the supervisor asked the Local Administrator a series of questions from the Observation Form. This gave him or her an opportunity to react to the assessment and provide suggestions for improving the training, procedures, or materials.

5.2.4 Results of the Field Test

Two hundred and thirty-three schools in 27 states and territories participated in the 1989 Field Test. Overall, 5,987 (92%) of the students who should have been assessed, were assessed.

As a result of the monitoring and suggestions from State Coordinators and Local Administrators, modifications were made in the training program, assessment materials, and procedures for 1990.

5.3 THE 1990 TRIAL STATE ASSESSMENT PROGRAM

5.3.1 Schedule of Mailings from NAEP Contractors

The schedule of mailings for the 1990 program was as follows:

Early June 1989	First letter sent to Chief State School Officer concerning NAEP. Copies of this letter also sent to the State Test Director and the State coordinator.
Mid-June 1989	Second letter containing a list of districts with selected schools for the Trial State Assessment and the National Assessment sent to Chief State School Officer. Copies of this letter also sent to the State Test Director and the State Coordinator.
End of June 1989	State Coordinator sent the list of sampled schools and, by separate mailing, a set of reports summarizing results from previous assessments. Copy of the cover letter sent to the State Test Director in each State Education Agency.
Mid-July 1989	Districts with selected schools received a set of NAEP reports.
Early August 1989	Second mailing sent to districts with sampled schools. The mailing included: (a) the list of selected schools for the district, and (b) a list of all schools in the district with selected grades to be used to update the sample frame. (Explanation for updating the sample frame can be found in Chapter 3.)
End of August 1989	State Coordinators sent copies of the Student Listing Forms and the Principal Questionnaires for distribution to participating schools.
Mid-October 1989	Each participating school sent a set of NAEP reports.

5.3.2 Schedule of Activities

The schedule of activities for the 1990 Trial State Assessment was as follows:

July - October 1989	State Coordinators obtained the cooperation of selected schools.
Mid-September 1989	State Supervisor contacted State Coordinator to discuss status of participating schools and schedule of assessments.

October 1989	Meeting of State Coordinators to review materials and procedures for the assessment.
Oct. - Nov. 1989	A limited number of substitutions selected for refusing schools.
December 1, 1989	Final date to have assessments scheduled and for schools to send State Coordinator the list of all students enrolled in the eighth grade and completed Principal Questionnaire.
December 4-13, 1989	State Supervisor visited State Coordinator to select the sample of students. State Supervisor left Administration Schedules with the names of the sampled students for each school and provided information on the times and places of Local Administrators' training and copies of the Manual for Local Administrators.
Dec. 4, 1989 - Jan. 8, 1990	State Coordinator notified Local Administrators of the date and time of training and sends each a copy of the Manual for Local Administrators.
Jan. 15 - Feb. 2, 1990	Local Administrators were trained.
Jan. 22 - Feb. 23, 1990	State Coordinator mailed the Administration Schedule to the school two weeks before the scheduled assessment date.
Feb. 5 - March 2, 1990	Assessments conducted.
March 5 - March 9, 1990	Makeup sessions held if unable to be scheduled during the four-week assessment period.

5.3.3 Preparations for the Trial State Assessment

The focal point of the schedule for the Trial State Assessment was the period between February 5 and March 2, 1990, 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.

Lists of selected schools and other NAEP materials were sent to State Coordinators during the summer so that they could begin contacting districts and schools to secure cooperation and develop a preliminary schedule.

The State Supervisors selected to work on this project were recruited and hired during the summer of 1989. All applicants who had not worked previously for Westat were interviewed in person by Home Office staff or experienced Westat supervisors. After a thorough check of references, the hiring decisions were made, and the newly-selected supervisors were invited to attend the training session held in the Washington, D.C. area between September 15-19, 1989.

Each State Supervisor's first assignment following training was to contact the State Coordinator. This contact was to serve as an introduction to the State Coordinator, since the two would be working together, (or at least be in regular contact), for the next several months. The contact was also used to determine the status of the notification of the school districts and schools concerning the assessment.

Between the conclusion of training and the end of the calendar year, there were several remaining critical dates. The first of these was October 31. By that date State Coordinators were to have determined which schools would participate in the assessment. This information was to be relayed to the State Supervisors who notified the Westat Home Office as to the participation status of each of the originally-sampled schools within their state. After that date, replacements for schools refusing to participate were selected for the study and the State Coordinator was informed of these selections.

Also by that date, the supervisors were to have developed a training schedule for the sessions to be conducted for the Local Administrators. This schedule was developed in coordination with the State Coordinator. Although it was not always possible to adhere to these guidelines, two major criteria were set for the scheduling of these training sessions. First, the training session was to be located so that no participant would have to travel more than two hours one-way in order to attend. Second, the training sessions were to be scheduled no more than two to three weeks before the dates of the attendees' scheduled assessments so that the training information would still be recent for the Local Administrators.

During the following month and a half (November 1 - December 13), several more tasks were completed by State Supervisors. One of these tasks was to recruit and hire four Quality Control Monitors (QCMs) for each state. It was the QCMs' job to observe the sessions designated to be "monitored," complete an observation report on each session, and to intervene when the correct procedures were not followed. In each state, half of the sessions were designated to be monitored. This information was known only to the contractor staff; it was not recorded on any of the listings provided to the State Coordinator.

Further required tasks were to find out from the State Coordinator and to report to Westat the names of the school/district employees who would be serving as Local Administrators and to confirm the assessment schedule.

The task of developing the assessment schedule was handled differently among the states. In all cases, some input from the State Supervisor was necessary because only he or she knew the schools that were to be monitored. In some states, the supervisor developed the entire schedule. In others, the State Coordinator proposed a schedule, and the Supervisor made only the alterations necessary to ensure that all the designated sessions would be monitored.

The final responsibility of the State Supervisors before the end of 1989 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 eligible students, (all of the school's eighth graders), to the State Coordinator by December 1. The sample selection activities were conducted in the State Coordinators' offices to maintain the confidentiality of the students.

Using a preprogrammed calculator, the supervisors selected a sample of 30 students per school except in states with fewer than 100 eighth-grade schools. In such cases larger student samples were required from participating schools. After the sample was selected, the supervisor completed an Administration Schedule for the session and sent it to the school so that the Local Administrator would know the identity of the students to be assessed. Instructions for sampling any new students who had enrolled at the school since the time that the list was created was also included in the same package.

Activities for the State Supervisors resumed immediately at the beginning of January 1990 when the second part of the State Supervisor training was conducted. The training took place between January 3 and January 7, 1990. The first three days focused on preparing the Supervisors for the training they would be conducting for the Local Administrators. During the final two days, the QCMs also were at the training. The emphasis for the QCMs was on their responsibilities, primarily targeting the completion of the Quality Control Forms that were filled out for each assessment.

Shortly after their training, the Supervisors began conducting their training sessions for Local Administrators (LA). In order to ensure uniformity in the training sessions, Westat developed a highly structured program involving a script for trainers, a video, and a training example to be completed by the trainees. The Supervisors were directed to read the script verbatim as they proceeded through the training. The importance of ensuring that all trainees received all of the same information dictated that such a script be used. 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 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 LA's. The portions of the video depicting the actual assessment were taped in a classroom with students in attendance so that a close representation of an actual assessment could be provided. The video was divided into five sections, with breaks for review by the trainer and practice for the trainees.

The final component of the presentation was the "Training Example for Local Administrators." This consisted of a set of exercises keyed to each part of the training package. A portion of the video and part of the script were presented, and then exercises related to that material were completed before going on to the next subject.

Prior to attending the training session, each Local Administrator received the Manual for Local Administrators provided by NAEP/Westat to the State Coordinator. This manual was a comprehensive document which explained every step of assessment procedures.

The entire training session generally ran for about five to six hours. In all, a total of 356 LA training sessions were held, (an average of nine per state), with 3,463 LA's being trained.

5.3.4 Monitoring and the Assessment

The QCM's attended several LA training sessions to assist the supervisor and to become thoroughly familiar with the LA's responsibilities. The Observation Forms, which the QCM's were to complete, began with a pre-assessment telephone call five days before the scheduled assessment. Most of the questions asked in the pre-assessment call were designed to gauge the LA's preparedness for the session. Questions were asked about whether the Administration Schedules had come from the State Coordinator and whether they had received the testing materials sent out by National Computer Systems. The LA's were also asked about some of the tasks they were to complete with the objective of finalizing as much of the preparations as possible before the actual assessment date.

Each of the LA's were also asked to provide directions to the school. Even though this information would only be of use in the half-sample of schools scheduled to be observed, it was asked of everyone. This was done so that no one could conclude that their school would or would not be visited.

For the sessions that were not observed, the final task completed by the QCM's was a call to the LA three days after the assessment to complete the Observation Form. Questions asked sought to obtain the LA's impressions of how the session went, (and how well the training had prepared them for it), and to ensure that all the post-assessment activities had been completed.

For the schools that were observed, the QCM's job was more involved. The QCM's were to arrive at the schools one hour before the assessment to observe the opening of the bundle of booklets. They also reviewed the procedures used in sampling new students and observe all the assessment activities, indicating whether they were performed correctly, with or without prompting.

One of the safeguards built into the design of the Trial State Assessment to ensure the confidentiality of assessment items was the packaging of the actual test booklets. These booklets were mailed out in "shrink-wrapped" bundles so that they could not be viewed before the assessment. One of the major directions given in training, and reinforced during the pre-assessment call, was that these bundles were not to be opened until the QCM arrived or 45 minutes before the assessment was to begin. This procedure was implemented as a security safeguard and seemed to work well, as there were only 29 instances (or 1.5 percent) in the 1924 monitored sessions where the bundle had been opened prior to the designated time.

The final requirement for conducting the Trial State Assessment was that a makeup session had to be held for every session where four or more sampled students were absent. If the original session was monitored, the makeup also was monitored. Therefore, the scheduling of the makeup had to be coordinated between the LA and the QCM. Makeups were required for 16 percent of the sessions and accounted for an additional 1,663 students being assessed, (or an average of slightly more than three students per makeup session). This raised the response rate from 92 to 94 percent. Appendix B provides participation rate data for each state and territory that participated in the Trial State Assessment.

5.3.5 Participation Rates

The results of the assessment concerning participation rates are given in Table 5-1. The original number of students sampled by the State Supervisors totaled 115,545. The sample size increased by 3,313 to a total of 118,858 after the supplemental samples were drawn, (from students newly enrolled since the original lists were created).

Table 5-1

Participation Rates

Category	Number
Number of students sampled	118,858
Original sample	115,545
Supplemental sample	3,313
Number of students withdrawn	5,714
Number of students excluded	5,835
Number of students to be assessed	107,309
Number of students assessed	100,849
Initial sessions	99,184
Makeup sessions	1,665

From this sample, a number of students were removed. The first group to be deleted were 5,714 students who had withdrawn from their schools since the time that the original Student Listing Forms were prepared. The next group of students who were removed from the sample were those who were to be excluded from the assessments. There were two reasons that students could be excluded, according to NAEP criteria. The first criteria related to students with an Individualized Education Plan (IEP). Only those students who were mainstreamed in less than 50 percent of their academic subjects and/or were judged incapable of participating meaningfully in the assessment were removed from the sample. The other reason students could be excluded was if they were Limited English Proficient (LEP). If a student was a native speaker of a language other than English, had been enrolled in an English-speaking school for less than two years, and was judged to be incapable of taking part in the assessment, he or she could also be excluded from the assessment.

Of the 9,448 students in the sample identified as having an IEP, 4,978 were excluded. Of the 1,639 students identified as LEP, 846 were excluded. Thus, approximately half of each group were judged to be incapable of participating meaningfully in the assessment and were not assessed.

These exclusions left 107,309 students to be assessed.[1] Of that total, 100,849 students were actually assessed -- 99,184 in initial sessions and an additional 1,665 in makeups. Appendix B provides the participation rate data for each of the jurisdictions in the Trial State Assessment Program.

5.3.6 Results of the Observations

During the assessments, the QCM's were to note instances when the LA's deviated from the prescribed procedures and whether any of these deviations were serious enough to warrant their intervention. During the observed sessions, there were no serious breaches of the procedures or major problems that would question the validity of any assessments.

The activity where the LA's most often deviated from the prescribed procedures was in reading the full script which introduced the assessment and provided the directions. The QCM's noted that there were major deviations in some portion of the script in 13 percent of the sessions. Samples of these deviations include skipping portions of the script, adding substantial comments and forgetting to distribute protractors to the students. The QCM 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 LA's. As previously reported, over 98 percent of the LA's had not opened their bundle of test booklets prior to the designated time. In 97 percent of the observed assessments, the timing was done accurately, so that all the students had the allotted time to complete each section. In 99 percent of the sessions that were monitored, student questions were responded to properly, or there were no questions. There were more problems distributing and collecting calculators, yet this procedure was performed correctly in 90 percent of the sessions.

After the assessment, LA's were asked how they thought the assessment went and whether they had any suggestions or comments. Eighty-nine percent of the LA's thought the assessment went very well. Comments about the assessment materials and procedures were generally favorable. The timers used, the amount of record keeping and paperwork, and the procedures for using calculators received the most negative comments. About 6 percent of LA's commented on one or more of these items.

In addition to gathering data about the assessments from the Observation Forms, Westat held a debriefing after the testing period with nearly all of the State Supervisors. This session produced many suggestions that will be applied to the future Trial State Assessments. A meeting was also held with most of the State Coordinators in attendance to gather their impressions and to report on plans for future assessments. The State Coordinators were also sent a questionnaire so that specific data could be collected from all the states about the entire assessment process. Overall, 35 State Coordinators said that they thought the assessments went well (22 states) or very well (13 states). Like the Local Administrators, State Coordinators also criticized the timer and calculator procedures. A variety of other suggestions were received about specific assessment materials and procedures. The input from these sources will be evaluated and applied to the planning of future assessments. In September 1990, each participating state and territory received a summary of its participation rate data, local administrators' reactions to training, and data collection activities.

[1] Eighty students were excluded for other reasons, such as, temporary physical disabilities, homebound, etc.

Chapter 6

PROCESSING ASSESSMENT MATERIALS

Dianne Smrdel, Lavonne Mohn, Linda Reynolds, and Brad Thayer

National Computer Systems

6.1 OVERVIEW

This chapter describes the printing, distribution, receipt, processing and final disposition of the Trial State Assessment materials. The scope of the effort required to process the materials is evidenced by the following:

- 107,337 assessment booklets and 20,225 questionnaires were received and processed.
- Approximately 2.2 million double-sided pages from test booklets and questionnaires were optically scanned.
- 1,840,100 student responses from 35 open-ended items were professionally scored.
- 3,538 questionnaires were manually key-entered and verified.

The processing included:

- Using the NCS Process Control System (PCS) and Workflow Management Systems (WFM) to track, audit, edit, and resolve characters of information.
- Utilizing an accuracy check system and selecting and comparing a quality control sample of characters of transcribed data to the actual responses in assessment booklets.
- Creating and distributing 4,019 bundles of assessment booklets to multiple sites.

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 allowed an integration of data entry and workflow management systems, including carefully planned and delineated editing, quality control, and auditing procedures.

The magnitude of the effort is made apparent considering that the activities described in this chapter were completed concurrently with the processing of the national assessments, that all of processing activities were completed within 10 weeks, and that an accuracy rate of fewer than three errors for every 10,000 characters of information was achieved.

6.2 PROCESS CONTROL SYSTEM (PCS)

NCS developed a Process Control System (PCS) consisting of numerous specialized programs and processes to accommodate the unique demands of concurrent assessment processing and a unified ETS/NCS system integration. The PCS was necessary to maintain control of all shipments of materials to the field, of all receipt from the field, and of any work in process. The system is a unique combination of several reporting systems currently in use at NCS, along with some application specific processes. These systems are the Workflow Management System, the Bundle Assembly Quality Control System, the Outbound Mail Management System, and the On-line Inventory Control system. Data was collected from all these systems and recorded in the file called the "NAEP Process Control System", while some information was entered directly into the PCS system.

6.3 WORKFLOW MANAGEMENT SYSTEM

The functions of the Workflow Management (WFM) system are to keep track of where the production work is and where it should be and to collect data for status reporting, forecasting, and other ancillary subsystems. The primary purpose of the WFM system is used to analyze the current work load by project across all work stations.

To a large extent, the data processing and control systems are determined by the type of assessment booklets and answer documents processed. For the Trial State Assessment only scannable student documents were used. In addition, three questionnaires were administered to collect data about school characteristics, teachers associated with sampled students, and students excluded from the assessment. The Excluded Student and Teacher Questionnaires were scannable documents; the School Questionnaire was a key-entry document.

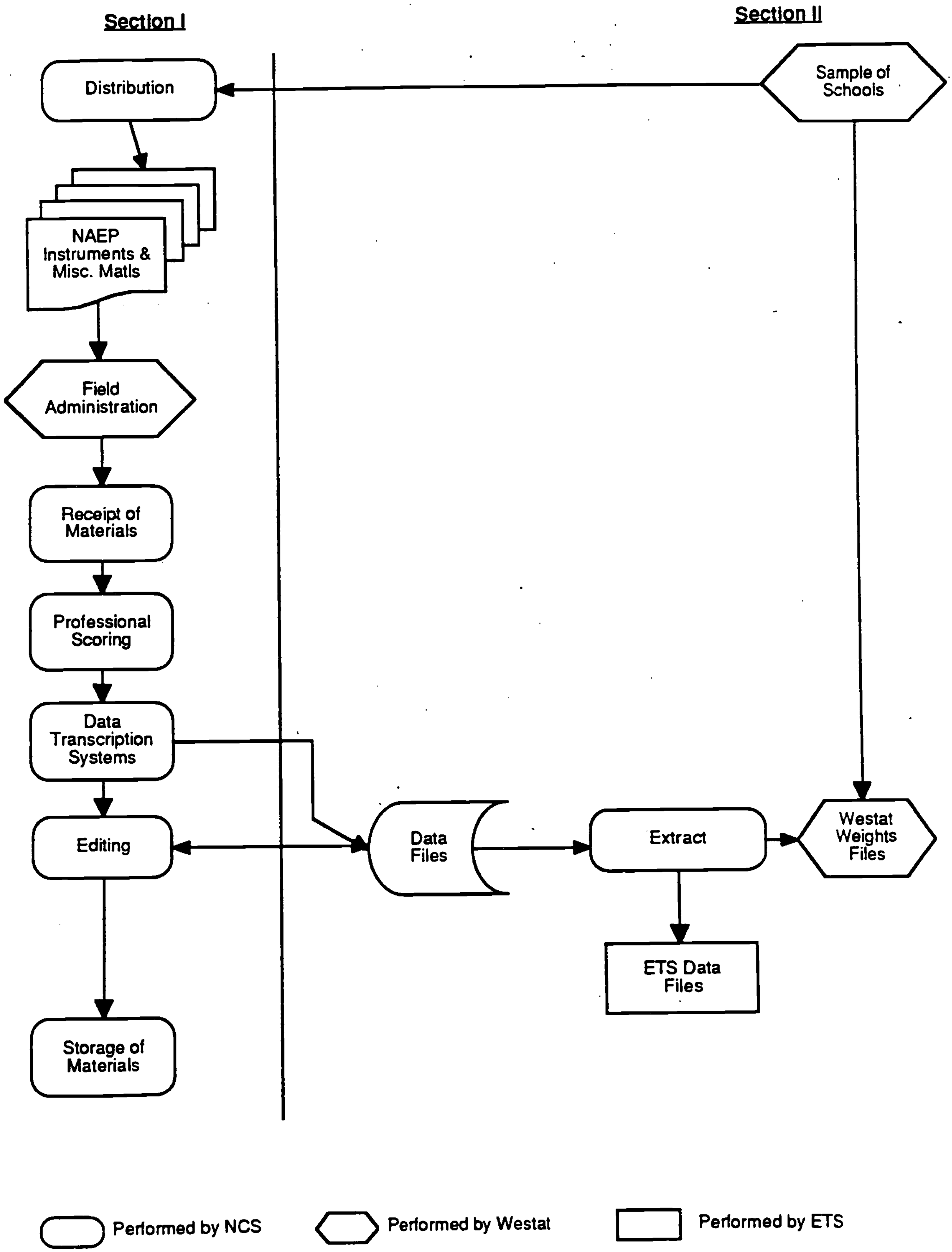
6.4 PROCESS FLOW OF NAEP MATERIALS AND DATABASE CREATION

Figure 6-1 is a flow diagram that shows the conceptual framework of processes that were applied to the Trial State Assessment materials, as well as to the national NAEP materials.

Section I of Figure 6-1 depicts the flow of NAEP printed materials. Information from the administration schedule and packing list was used to control the processing of materials. The figure follows the path of each assessment instrument -- Student Test Booklets, School Characteristics and Policies Questionnaires, Teacher Questionnaires, Excluded Student Questionnaires, Packing List, and Administration Schedules -- as they were tracked through the appropriate processes that resulted in the final integrated NAEP database.

Figure 6-1

Data Flow Overview



The remainder of this chapter provides an overview of the materials processing activities as shown in Section I of Figure 6-1 and detailed in Figure 6-2. Section II of Figure 6-1 depicts the evolution of the NAEP/NCS database from the transcribed data to the final files, provided to Westat for creation of weights and to ETS for analysis and reporting.

The 1990 NAEP data collection resulted in six classes of data files (student, school, teacher, excluded student, sampling weight, and item information files). The structure and internal data format of the 1990 NAEP database was a continuation of the integrated design originally developed by ETS in 1983.

6.5 MATERIALS DISTRIBUTION

In past assessments, a unique number was imprinted on the book by the vendor and then coded in scannable form on the cover by the student or test administrator. The use of bar code technology in document control was a major innovation introduced by NCS in the 1990 assessments. To reduce effort and error in transcribing the book ID, NCS implemented the use of bar code numbers. Bar codes were applied to the front cover of each document. The bar code consisted of the two digit book number, a five digit sequential number, and a check digit.

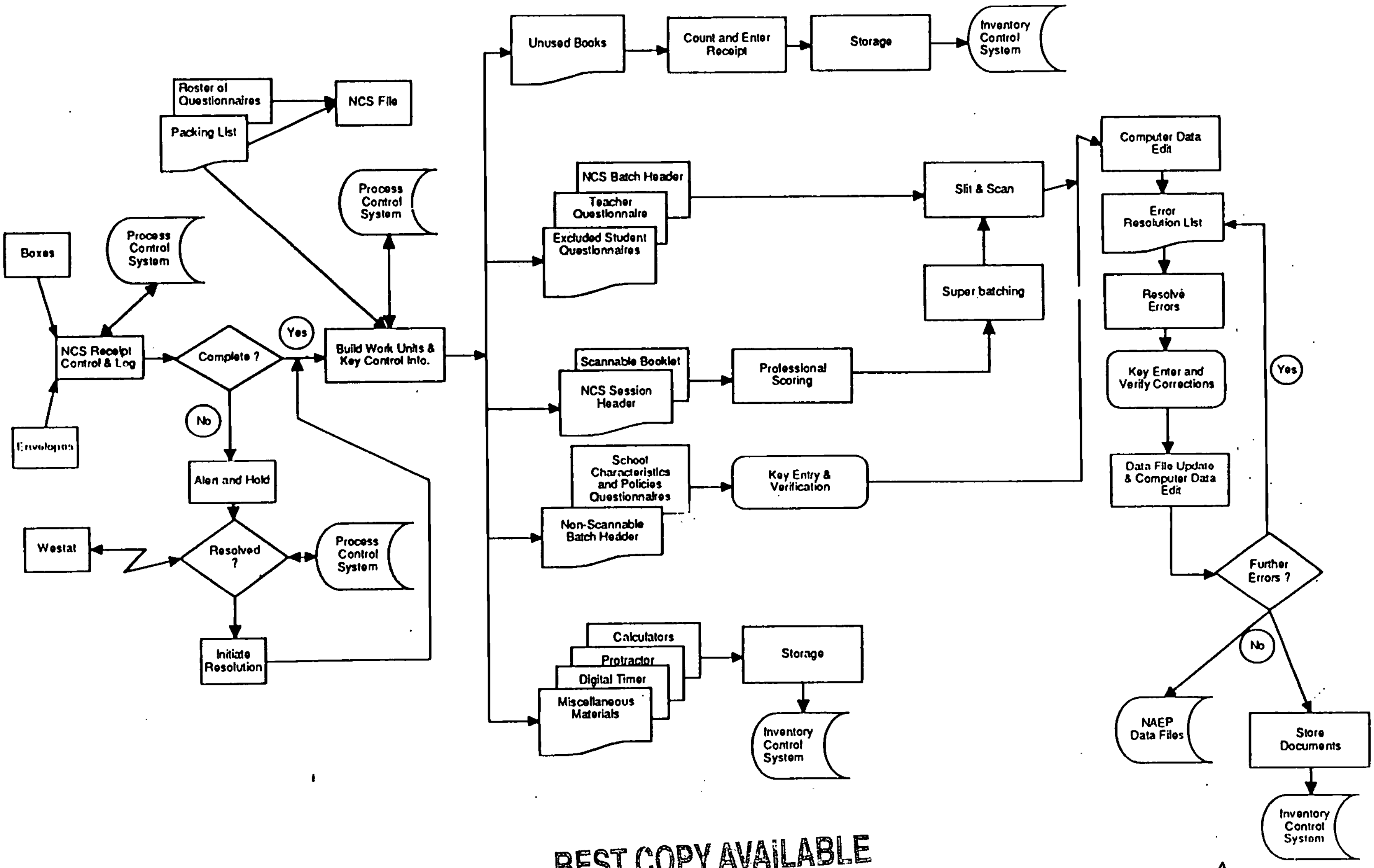
The books were spiraled into seven unique bundles consisting of 35 books in a set pattern. A bundle header sheet was attached to each bundle that indicated the assessment type, the bundle type, the bundle number, and a list of the book types to be included in the bundle.

The bundle numbers on the bundle header sheet were created to identify the type of bundle. All bundles were then passed under a scanner programmed to interpret this type of bar code and the file of scanned barcodes was transferred from the scanner to the mainframe. A computer program compared the bundle type expected to the one actually scanned after the header and verified that there were 35 booklets in each bundle. Any discrepancies were printed on an error listing forwarded to the Packaging Department. In that department, the error was corrected, and the bundle was again read into the system for another quality control check. This process was repeated until all bundles were correct.

The bundles were shrink-wrapped in clear plastic, strapped in each direction, and a bright label placed over the cross of the straps that read "Do Not Open Until 45 Minutes Before Testing". Following this, bundles were ready for assignment and distribution.

The timing of shipment of these materials to the participating schools was critical since the shipments needed to be in the school at least one week, but not more than two weeks, prior to testing. Also of concern was the fact that calculators were in limited supply, and the shipments for assessments occurring during the last two weeks could not be completed until shipments from the first week's assessments were returned.

Figure 6-2
NAEP Trial State Assessment Materials Processing Flow



91

91A

Each school conducted at least one session and in some instances more than one. In either case, each session was treated as an independent shipment. Each shipment contained the same quantities of materials which permitted pre-assembly of materials for distribution:

1	Bundle of 35 assessment booklets
15	Scientific calculators
15	Protractors
1	Digital timer
1	Pad of appointment cards
1	Return postage paid label
1	Post-it note pad
1	Shipping tape.
5	Excluded Student Questionnaires
5	Mathematics Teacher Questionnaires
1	School Characteristics and Policies Questionnaire
1	Roster of Questionnaires
1	Calculator Poster
2	Assessment Notifications
1	Pre-addressed envelope
1	Pre-addressed box

Shipments were sent according to the week of assessment. Some schools found they needed extra quantities of materials, (i.e., more than five excluded students or more than five mathematics teachers), and calls were received requesting these additional materials.

Aiding in the security of the shipments was the decision to send all shipments, whenever possible, via Airborne. NCS is connected to the Airborne system via computer link thus expediting tracing of any misdirected shipments. This system provides data on date and time of delivery as well as the name of the person who signed for the shipment. All shipments were recorded in the Airborne Libra system. If a shipment needed to be sent via UPS or US Postal Service, this information was also recorded in a similar manner and transferred to the mainframe.

6.6 PROCESSING ASSESSMENT MATERIAL

The materials from each session were to be returned to NCS in the same box in which they was originally mailed. It was the responsibility of the Local Administrator in the unmonitored schools and the Quality Control Monitor in the monitored schools to re-package the items in the proper order, complete all paper work and return the shipment via the U.S. Postal Service, utilizing the postage paid label provided.

With approximately 4,000 individual shipments arriving over a four week period, it was necessary to devise a system that would quickly acknowledge receipt of a school's material. The label used to ship the materials to the school or administrator contained a bar code which indicated the school number and the project number. When the shipment arrived at NCS, the bar code was read and the shipment forwarded to the receiving area. The file was then transferred to the mainframe via a PC link and a computer program was used to apply the shipment receipt date to the appropriate school within the PCS system. This provided current status of shipments received regardless of any processing backlog. Any shipment that was not received within seven days of the scheduled assessment was flagged on a report that was issued to Westat. The status of the administration was checked and in some cases a trace was initiated on the shipment.

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, regardless of whether it was monitored or unmonitored.

The materials were checked against the packing list (see Figure 6-3) to verify that all materials were returned. If any discrepancies were found, an alert was issued. If all assessment instruments were returned, then processing continued. Quantities of scientific calculators were in short supply; therefore, during the first two weeks of the assessment, calculators were taken from the incoming shipments and returned to the packaging area to be included in other shipments for the last weeks of testing.

Each booklet and Excluded Student Questionnaire was verified against the administration schedule. This included verification of all counts of booklets returned and the matching of information on the front cover of the booklets to that on the administration schedule. If any discrepancy was discovered, an alert was issued.

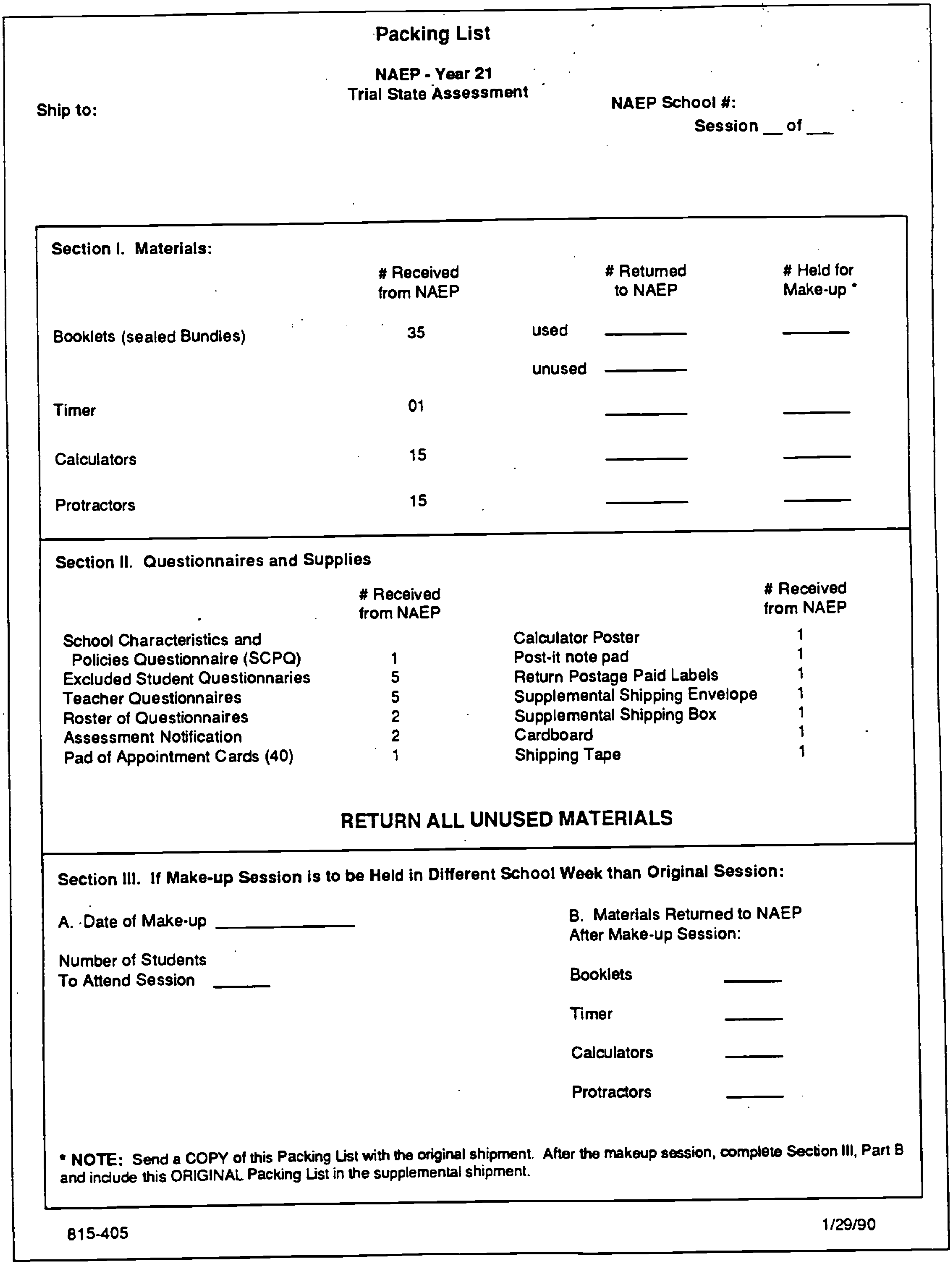
After the contents of the shipment had been identified and verified, the packing list information was entered into the PCS. That information included school number, session number, counts of the number of used and unused booklets, and makeup status. If a makeup session was expected, an information alert was issued to facilitate tracking. The control counts were used for verification of processing counts.

If quantities and individual information matched, the booklets were organized into work units and batched by session. Each session was assigned a unique batch number composed of a session identifier and the school number. The batch number was entered on the Workflow Management System, facilitating the internal tracking of the session and allowing departmental resource planning. A scannable session header was coded and placed on top of the stack of documents. All student documents were forwarded to professional scoring, other documents to a key entry activity, others directly to a machine scanning function, and others to appropriate record filing systems.

The Excluded Student Questionnaires and Teacher Questionnaires were compared to the Roster of Questionnaires and the administration schedule to verify demographic information. 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 School Characteristics and Policies Questionnaire, a key-entry document, was compared to the Roster of Questionnaires and the school number was verified to match all other materials in the shipment. As with the other questionnaires, this document may not have been returned with the shipment and could also be returned in the supplemental envelope. There was no additional effort made to collect any unreturned questionnaires.

Figure 6-3



Packing List

NAEP - Year 21
Trial State Assessment

Ship to:

NAEP School #:
Session __ of ___

Section I. Materials:

	# Received from NAEP		# Returned to NAEP	# Held for Make-up *
Booklets (sealed Bundles)	35	used	______	______
		unused	______	
Timer	01		______	______
Calculators	15		______	______
Protractors	15		______	______

Section II. Questionnaires and Supplies

	# Received from NAEP		# Received from NAEP
School Characteristics and Policies Questionnaire (SCPQ)	1	Calculator Poster	1
Excluded Student Questionnaries	5	Post-it note pad	1
Teacher Questionnaires	5	Return Postage Paid Labels	1
Roster of Questionnaires	2	Supplemental Shipping Envelope	1
Assessment Notification	2	Supplemental Shipping Box	1
Pad of Appointment Cards (40)	1	Cardboard	1
		Shipping Tape	1

RETURN ALL UNUSED MATERIALS

Section III. If Make-up Session is to be Held in Different School Week than Original Session:

A. Date of Make-up ______________

Number of Students
To Attend Session ______

B. Materials Returned to NAEP
After Make-up Session:

Booklets ______

Timer ______

Calculators ______

Protractors ______

* **NOTE:** Send a COPY of this Packing List with the original shipment. After the makeup session, complete Section III, Part B and include this ORIGINAL Packing List in the supplemental shipment.

815-405

1/29/90

Unlike the national assessment, absent students were not accounted for by use of an Absent Student Form. Absent students were assigned a test booklet. To indicate an absence, the "A" bubble in the Administration Code column on the front cover of the booklet was gridded. The booklet was then processed with assessed booklets to maintain session integrity.

The packing list (Figure 6-3) was used by the schools to account for all materials received from and returned to NCS. Any discrepancies in quantities received or returned to NCS were indicated. Also indicated was whether a makeup session was to be held, the date of scheduled makeup, number of students involved, and quantities of materials being held for later return.

The administration schedule contains the demographic characteristics of the students selected for the assessment. This information included the sex, race/ethnicity, birth date, and IEP/LEP indicators. The booklet number of the student selected was recorded on the administration schedule during the assessment process, and the demographic information was transferred to the booklet covers either by the students or Local Administrator.

The demographics of the sampled students who did not participate in the assessment (exclusions and absentees) were provided to Westat to be used to adjust the sampling weights of the students who did participate. The excluded student information was obtained from the Excluded Student Questionnaire. The absent student information was taken from the front cover of the booklet that was assigned prior to the start of the assessment. This procedure eliminated the need for an additional form for absent students.

Counts contained on the Administration Schedule of students added to or removed from the sample were entered into the file. This information was used by Westat to produce participation statistics for the states and included number of students originally selected for the assessment, supplemental students, withdrawn, excluded and absent students, and the number of students actually assessed.

For the Rosters of Questionnaires, two numbers were entered for each type of questionnaire: number of questionnaires expected and number actually received. The Packing List, Administration Schedule, and Roster of Questionnaires were forwarded to the operations coordinator and filed by school within state for future reference. If any questionnaires remained outstanding, the roster remained on file in the receipt area for check-in when they arrived.

6.7 PROFESSIONAL SCORING

The student assessment booklets were forwarded to the professional scoring area once all appropriate materials were received from each school. Like the national assessments, the Trial State Assessment included open-ended items. Open-ended and multiple choice items were administered in scannable assessment booklets that were identical to the mathematics booklets used in the eighth-grade national assessment. Scores for the open-ended items in these booklets were gridded in ovals at the bottom of the pages on which the items appeared.

The scoring of the Trial State Assessment was conducted simultaneously with the scoring of the mathematics portion of the national program and the same readers scored the open-ended questions from both programs. The readers for the Trial State Assessment were organized into five teams of twelve readers and one team leader. The five team leaders reviewed discrepancies between readers and reviewed decisions regularly so that all readers scored each item similarly.

6.7.1 Description of Scoring

Each open-ended item had a unique scoring guide that identified the range of possible scores for the item and defined the criteria to be used in evaluating students' responses.

Eighteen items were categorized as right/wrong, while 17 items included categories of specific correct and incorrect responses. For the items scored as "right/wrong", a correct response was coded a score of "8" and an incorrect response was coded a score of "1". Items with two correct responses were given a score of "7" for the second correct response.

Various types of incorrect responses were also tracked with separate score points. The incorrect responses were assigned a score point from "1" to "5" to capture information on the specific types of errors students were making.

6.7.2 Training

The readers had to be trained to make certain that they would reliably score the open-ended items. The purpose of the training, which was conducted during a one-week period, was to familiarize the group with the scoring guides and to reach a high level of agreement among the readers.

Before the training program began, the team leaders worked with ETS mathematics test development staff to prepare training sets (sets of sample responses to accompany the scoring guides). Training involved explaining each item and its scoring guide to the readers and discussing responses that were representative of the various score points in the guide. The training was conducted by ETS mathematics test development specialists with assistance from the five team leaders. Following the explanations, the readers scored and discussed 5 to 20 carefully selected "practice papers" for each item, depending on the complexity of the item. Then, each reader practiced by scoring all the open-ended items in each of approximately 12 bundles of booklets, with an average of 27 booklets per bundle. During this practice, discussion sessions were held to review responses that received a wide range of scores.

Once the practice session was completed, the formal scoring process began. During the scoring, notes on various items were compiled for the readers for their reference and guidance. In addition, short training sessions were conducted when the team leaders determined by reviewing discrepancies that certain items were causing difficulties for the scorers. The team leaders also consulted with individual readers as the scoring progressed. When a reader's score was judged to be discrepant with that of another reader, the supervisor discussed the response and its score with that reader.

Twenty percent of the responses to the open-ended items were scored by a second reader to obtain statistics on inter-reader reliability. Each item was read twice at least 9,200 times. The average inter-reader reliability for all 35 open-ended items was 97%. The reliability was 95% or greater for 29 or the 35 items. For just two of the items was the reliability less than 90% (86.8% and 89.3). The reliability information was used to monitor the capabilities of particular readers and the uniformity across readers about each task.

It is important to note that all reliability scoring was truly "blind", uninfluenced by any score already given, because the reliability scoring occurred first. Further, the reliability scoring was recorded on a separate reliability scoring sheet, which expedited the scoring process by eliminating the need to mask the scores assigned by the primary reader.

6.8 DATA TRANSCRIPTION SYSTEMS

The transcription of the student response data into machine-readable form was achieved through the use of three separate systems: data entry, (scanning or key entry), validation, (pre-edit), and resolution.

6.8.1 Data Entry

The data entry process is the first time that booklet level data were input to the computer system. One of two methods was used to transcribe data to a computerized form. The data on scannable documents was collected using NCS optical scanning equipment while data on non-scannable material was keyed through an interactive on-line system. In both cases, the data were edited and questionable data were resolved before further processing.

To ensure data integrity, edit rules were applied to each scanned data field. This procedure validated each field and reported all problems for subsequent resolution. After each field was examined and corrected, the edit rules were re-applied for final verification.

6.8.2 Scanning

After the professional scoring, the scannable documents (the student booklets, Excluded Student Questionnaires, and Teacher Questionnaires) were transported to a slitting area where the folded and stapled spine was removed from each document. Scanning operations were performed by NCS's HPS Optical Scanning equipment. The optical scanning devices and software used at NCS permits a complete mix of NAEP scannable materials to be scanned with no special grouping requirements. However, for manageability and tracking purposes, student documents, Excluded Student Questionnaires and Teacher Questionnaires were batched separately. In addition to the capture of scannable responses, the bar code identification numbers used to maintain process control were also decoded and transcribed to the NAEP computerized data file.

The scanning program is a table driven software process utilizing standard routines and application-specific tables that identify and define the documents and formats to be processed. When a booklet cover is scanned, the program uses the booklet number to determine the sequence of pages and the formats to be processed. By reading the booklet cover, the program recognizes which pages would follow and in what order.

The scanning program wrote four types of data records into the data set: 1)a batch header record containing information coded onto the batch header sheet by receipt processing staff; 2)a session header record containing information coded onto the session batch header sheet by receipt processing staff; 3)a data record containing all of the translated marked ovals from all pages in a booklet; and 4)a dummy data record, serving as a place holder in the file for a booklet with an unreadable cover sheet. The document code is written in the same location on all records to distinguish them by type.

The following coding rules were used:

- The data values from the booklet covers and scorer identification fields were coded as numeric data.
- Unmarked fields were coded as blanks and processing staff were alerted to missing or uncoded critical data.
- Fields that had multiple marks were coded as asterisks (*).
- The data values for the item responses and scores were returned as numeric codes.
- The multiple-choice, single response format items were assigned codes depending on the position of the response alternative; that is, the first choice was assigned the code "1", the second "2", and so forth.
- The circle-all-that-apply items were given as many data fields as response alternatives; the marked choices are coded as "1" and the unmarked choices as blanks.
- The fields from unreadable pages were coded "X" as a flag for resolution staff to correct.

6.8.3 Key Entry

The School Characteristics and Policies Questionnaire and professional scoring reliability scoring sheets were non-scannable documents. The Falcon system, an on-line data entry system designed to replace most methods of data input such as keypunch, key-to-disk, and many of the microcomputer data entry systems, was used to capture the data from this questionnaire. The same facility was also used to make corrections to the scannable documents.

6.9 DATA VALIDATION

The data entry and resolution system used for the Trial State Assessment Program was also used for the national assessment program. The system is able to process materials from three age groups simultaneously, three assessment types, one absent form, and five questionnaires submitted to the system from scannable and non-scannable media. The use of batch identification codes -- comprising the school and session codes as well as the batch sequence numbers for suspect record identification -- facilitated the management of the system and correction of incorrectly gridded or keyed information.

As the program processed each data record, it first read the booklet number and checked it against the batch session code for appropriate session type. Any mismatch was recorded on the error log and processing continued. The booklet number was compared against the first two digits of the student identification number. If they disagreed, because of improper bar coding, a message was written to the error log. The remaining booklet cover fields were then read and validated for the correct range of values. The school codes had to be identical to those on the PCS record and grade code had to be an eight. 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.

Document definition files describe each document as a series of blocks which were described as a series of items. The blocks in a document were traversed in the order that they appear on the document. Each block's fields were validated during this process. If a document contained suspect fields, the cover information was recorded on the edit log with a description of the suspect data. Some fields, (i.e., AGE or DOB), required special types of edits. These fields were identified in the document definition fields, and a subroutine was invoked to handle these cases.

The scorer identification fields were processed at this point and certain checks were made. If a booklet contained any open-ended items, the first scorer field had to be filled in on the session header. If a booklet was part of the reliability sample, the second scorer ID, along with the scores, was pulled from the file of key entered reliability sheets for that batch.

The program next 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 the number of fields to be processed for each block and the field type, (alphabetic or numeric), the field width in the data record, and the valid range of values for each field. The routine 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 and processing continued with the next field. No action was taken on blank-filled fields for multiple-choice items since that code indicated a non-response. The field was validated for range of response, recording anything outside of that range to the edit log. The item type code was used by the program to make a further distinction among open-ended item scores and other numeric data fields. If the data field was an open-ended item, the routine validated the score range and did not permit a blank field. If a document contained open-ended items and no score was indicated, or if an open-ended item was not scored, the disparity was noted in the edit log. If the item type indicated a secondary scoring and it was non-blank, the routine checked for the presence of a second scorer code. Moving the translated and edited data field into the output buffer was the last task performed in this phase of processing.

The completed string of data was written to the data file when the entire document has been processed. Then, when the next session header record was encountered, the program repeated the same set of processes for that session. The program closed the data set and generated an edit listing when it encountered the end of a file.

Accuracy checks were performed on each batch processed. Every five hundredth document of each book form was printed in its entirety, with a minimum of one document type per batch. This record was checked, item by item, with the source document for errors.

6.10 EDITING

Quality procedures and software throughout the system ensure that the NAEP data are correct. The initial editing that took place during the receipt control process included verification of the schools and sessions. Receipt control personnel checked that all student documents on the Administration Schedule were undamaged and assembled correctly. The machine edits performed during data capture verified that each sheet of each document was present and that each field had an appropriate value. All batches entered into the system, whether key-entered or machine scanned, were edited for errors.

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

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

During data entry, counts of documents processed by type were generated. These counts were checked against the packing list counts entered into the PCS during the receiving process. The number of assessed and absent students processed had to match the number of used books indicated on the PCS.

The second phase of data editing utilized an experienced editing staff using a predetermined set of rules to review the field errors and record corrections to be made to the student data file. The same computerized edit list used in phase one was used to perform this function.

The editing staff made corrections using the edit log prepared by the computer and the actual source document listed on the edit log. The corrections were identified by batch sequence numbers and field name for suspect record and field identification. The edit log indicated the current composition of the field. This particular piece of information was then visually checked against the NAEP source document by the editing staff for double grids, erasures, smudge marks or omitted items were flagged. Then for each flagged item, one of the following took place:

- *Correctable Error* -- If the error could be corrected by the editing staff, according to the editing specifications, the corrections were indicated on the edit listing.
- *Field Correctable* -- 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 indicated on the edit listing.
- *Non-correctable Error* -- If an error suspect was found to be correct as stated, and no alteration was possible according to source documents and specifications, the programs were tailored to allow this information to be accepted into the data record and no corrective action was taken.

These corrections were noted on the edit list and when the entire batch of sessions was resolved, the list was forwarded to the key entry staff. The corrections were entered and verified via the Falcon system. When all corrections were entered and verified for a batch, an extract program was run to pull the correction records to a mainframe data set.

The post edit program was initiated next. This program applied the corrections to the specified records and once again applied the error criteria to all records. If there were further errors, another edit list 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 school and sessions. This allowed for an accumulation process to accurately measure the number of documents processed for a session within a school and the number of documents processed by form. The posting of booklet ID's also ensured that a booklet ID was not processed more than once. These data allowed the progress of the assessment to be monitored and reported on the status report.

As one final quality control check, ETS identified a random sample of each booklet type from the master student file. The designated documents were located, removed from storage and forwarded to ETS for quality control (see Chapter 7). On completion of quality control processing, the booklets were returned to NCS for return to storage.

6.11 QUESTIONNAIRES

The questionnaires were either received with the session shipment or in a later shipment. Once the questionnaires were verified with the roster, they were accumulated by the receiving clerks. The School Characteristics and Policies Questionnaires were batched in groups of 25 and transcribed through the NAEP key entry system. The Teacher Questionnaires and Excluded Student Questionnaires were batched and sent to scanning at regular intervals. In order to assure that all documents for a session were being processed and in order to deliver all data at the same time, every effort was made to keep current on all forms.

The key entry programs were similar to those used for entry of student documents in the national assessment program. All documents, regardless of method of entry, were run through the process of error identification and resolution.

6.12 STORAGE OF DOCUMENTS

Once the editing process had been successfully completed on the batches, they were sent to the NCS warehouse for storage. The storage location of all documents was recorded on the inventory control system and stored for later retrieval. Unused materials were sent to temporary storage until the completion of the assessment and acceptance of the data tape, at which time they were destroyed.

Chapter 7

CREATION OF THE DATABASE AND EVALUATION OF THE QUALITY CONTROL OF DATA ENTRY

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

Educational Testing Service

7.1 OVERVIEW

The data transcription and editing procedures described in Chapter 6 resulted in the generation of 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. These files were then merged into a comprehensive, integrated database. To evaluate the effectiveness of the quality control of the data entry process, the final integrated database was sampled, and the data were verified in detail against the original instruments received from the field.

This chapter begins with a description of the transcribed data files and the procedure of merging them, or bringing them together, to create the Trial State Assessment database. The last section presents the results of the quality control evaluation.

7.2 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: one file for each of the three questionnaires (teacher, school, and excluded student) and one for the student response data. The process of deriving sample weights produced an additional three files of sampling weights which were produced by Westat -- one for students, schools, and excluded students. (See Chapter 8 for a discussion of the sampling weights.). These seven files were the ones needed for the analysis of the Trial State Assessment data. Before data analyses could be performed, these files had to be integrated into a coherent and comprehensive database.

The Trial State Assessment database consisted of three files: student, school, and excluded student files. Each record on the student file contained a student's responses to the particular assessment booklet the student took and the information from the questionnaire that the student's mathematics teacher completed. It was not necessary to have a separate teacher file since teacher response data can only be reported at the student level. The school file was separate and could be linked to the student file through the state and school codes. The excluded student file was also separate.

The student data file was created in two steps. First, the student response data was merged with the student weights file. The resulting file was then merged with the teacher response data. In both steps, the assessment booklet 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 variables, supplied by Westat, which included demographic information about the schools collected from the principal's questionnaire. The state and school codes were used as the matching criteria. Since some schools did not return a questionnaires and/or were missing principal's questionnaire data, some of the records in the school file contained only school identifying information and weights information.

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

When the three files had been created -- student, school, and excluded student --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 each state 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 computer standard errors.

7.3 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 reporting programs, will provide correct parameters for processing the data fields and consistent labeling 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 items and professionally scored 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., mathematics content and process areas) for the data fields. For a particular classification field, a nonblank value indicates the code of the subcategory within the classification categories for the data field. Having 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 concurrent to 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.

7.4 QUALITY CONTROL EVALUATION

Seven assessment booklets, numbered 8 through 14, were administered as part of the Trial State Assessment Program. Table 7-1 provides the total number of each booklet for which data were scanned into data files.

Table 7-1

Number of Assessment Booklets Scanned

	Booklet Number							
	8	9	10	11	12	13	14	Total
Total Booklets Scanned	14,417	14,404	14,355	14,402	14,405	14,442	14,419	100,844

The number of students assessed in each of the 40 participating jurisdictions varied: In 38 of the jurisdictions an average of 2,576 students were assessed (an average of 368 per booklet), in one jurisdiction 1,617 students were assessed (231 per booklet), and in one jurisdiction 1,326 students were assessed (189 per booklet).

The purpose of the data entry quality control procedure is to gauge the accuracy of the scanning and scoring processes. The procedure involves examining the actual responses made in the student booklets and comparing those with the responses recorded in the final database that is used for analysis and reporting. Because it was desirable for the quality control evaluation to sample at least one of each of the seven booklets from each state, it was decided to sample exactly one of each, for a total of 280 booklets (seven booklets in each of the 40 jurisdictions). Variations across jurisdictions in the number of students per booklet resulted in a sampling rate that ranged from 1/186 to 1/411, with an average rate of 280/100,844 or about 1/360. This rate was comparable to the rate of 1/400 used in similar quality control evaluations.

The quality control evaluation detected 11 errors in this collection of booklets: a single error made by one of the professional scorers who scored a response as missing when there was a response; two instances of multiple responses which were not detected; and eight instances of erasures that were recorded instead of ignored. The usual quality control analysis based on the binomial theorem permits the inference described in Table 7-2.

Table 7-2

Inference from the Quality Control Evaluation

Subsample	Entry Type	Different Booklets Sampled	Number of Booklets Sampled	Characters Sampled	Number of Errors	Observed Rate	99.8% Confidence Limit
Student Data	Scanned	7	280	40,600	11	.0003	.0006

This error rate is quite respectable and should not interfere with the validity of any data analyses. However, as always, there was some indication that it could be improved. This error rate was somewhat higher than the error rate observed for scanned booklets from previous national assessments.

Chapter 8

WEIGHTING PROCEDURES AND VARIANCE ESTIMATION

Jim Bethel, Keith Rust, and Jacqueline Severynse

Westat

8.1 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 were 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, without regard to whether the student was in a monitored or unmonitored session. This weight is known as the full, or overall, sample weight. A second weight, known as the comparison weight, was derived from the full sample weight for use in making comparisons, within and across states, in the performance of students who were assessed in monitored sessions and unmonitored sessions.

The full sample weight contained three components. First, a base weight -- the inverse of the overall probability of selection of the sampled student -- was established. This base weight incorporated the probability of selection of the student's school, and of the student within school, and accounts for the impact of procedures used to keep to a minimum the overlap of the school sample with both the national assessment eighth-grade school sample, and the sample of schools involved in the National Educational Longitudinal Study First Phase Follow-up. The base weight was then adjusted for two sources of non-participation -- school level and student level. These weighting adjustments seek to reduce the potential for bias from such non-participation 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 (or a make-up) as scheduled. The details of how these weighting steps were implemented are given in Sections 8.2 through 8.4.

The comparison weights were obtained from these full sample weights using an additional adjustment procedure. This adjustment, described in Section 8.5, resulted in the distributions of weighted counts for various student characteristics that were very similar for the students from the monitored sessions and the unmonitored sessions. The characteristics involved were known from previous NAEP studies to be related to eighth-grade mathematics proficiency. The purpose of this adjustment was to decrease the variance of comparisons between the results from monitored sessions and those from unmonitored sessions within each state. These adjustments of the comparison weights were not considered necessary or appropriate for the full-sample weights, since it was highly likely that these adjustments would actually act to increase the variance of estimates made across the monitored and unmonitored students, by adding random variability to the sampling weights (comparison and full sample). Thus, for this reason, two estimation weights were provided for each assessed student.

In addition to estimation weights, a set of replicate weights was provided for each student. These replicate weights are used in estimating sampling errors of estimates obtained from the data, using the jackknife repeated replication (or jackknifing) method. Full details of the method of using these replicate weights to estimate sampling errors are contained in the 1988 and 1990 National Assessment Technical Reports. Section 8.6 of this chapter describes how the sets of replicate weights were generated for the Trial State Assessment data. The methods of deriving these weights were aimed at reflecting appropriately the features of the sample design in each state so that when the jackknife variance estimation procedure was implemented as intended, approximately unbiased estimates of sampling variance would result.

Similar to the estimation weights, two sets of replicate weights were derived for each student. The first set is referred to as the overall replicate weights, and correspond to the full sample weight. They are used for estimating the sampling errors of estimates derived using the full sample weights. These weights are designed to reflect the method of sampling schools, and account for the type of stratification used and whether or not the student's school was included in the sample with certainty. The method of sampling students within schools is also reflected, implicitly in the case of non-certainty schools and explicitly for schools included with certainty. These overall replicate weights also reflect the impact on sampling errors of the school- and student-level nonresponse adjustments applied to the full sample weights.

The second set of replicate weights, known as the comparison replicate weights, are for use in estimating sampling errors of estimates obtained using the comparison weights. These replicate weights differ from the overall replicate weights in two ways. First, in addition to reflecting features of the sample design and weighting procedures, they reflect the impact on sampling error of the raking procedure (described in Section 8.5) used to equate weighted distributions from the monitored and unmonitored half samples in each state. Second, in those states where some or all schools were selected into the sample with certainty, the comparison weights reflect the fact that such certainty selections were assigned to be monitored or unmonitored at random. Thus, these certainty schools contribute a school level component of variance to the comparison of monitored and unmonitored assessments, which is appropriately reflected in the comparison replicate weights. The details on the formation of replicate groups and the assignment of replicate weights are given in Section 8.6.

One additional survey weighting component was used in the analysis of the Trial State Assessment data. This was a special weighting component which was applied to a subset of the national sample data for use in equating the National and Trial State Assessment eighth-grade mathematics assessments. This adjustment, a raking procedure similar to that described in Section 8.5, was used to bring the weighted distributions for certain characteristics from the national subset closely in line with the distributions given by the aggregate weighted state sample data. This special adjustment to the national data subsample is described in the 1990 Technical Report for the national program that includes a description of weighting procedures for the national survey data.

8.2 WEIGHTING PROCEDURES

The base weight assigned to a school was the reciprocal of the probability of selection of that school. The base weight reflected the actual probability used to select the school from the frame. It also included two factors that reflected the impact on the selection probability of the avoidance of school sample overlap with both the national NAEP samples and the 1990 National Educational Longitudinal Study first follow-up of tenth-grade students. Schools that substituted for a refusing school were assigned the weight of the refusing school, unless of course the substitute refused. For schools that conducted double sessions because they were substitutes for a refusing school, half of the students were assigned the school base weight of the participating school and half were assigned the weight of the refusing school. These half samples were chosen at random, with each half-sample constituting a simple random sub-sample of the full sample of students from the schools. The exact procedures for deriving school base weights are given below.

The student base weight was a product of the base weight of the school in which the student was enrolled and the within-school student weight, where the student weight was given as:

$$\textit{Student Weight} = \frac{\textit{Actual Eighth-Grade Enrollment}}{\textit{Sample Size}},$$

reflecting the within-school student probability of selection.

8.2.1 Calculating the School Cluster Weights and School Base Weights

As described in Section 3.3.3, schools were selected in clusters, so as to sample small schools appropriately. For all certainty clusters (including all schools in Type 1 Clusters) the school base weights were one. For the remaining clusters, the formulas for the cluster weights are given below. In all of the formulas, "Total Eighth-Grade Enrollment" was for noncertainty clusters only, and n was the number of non-certainty clusters selected.

For Cluster Type 2 States (Noncertainty):

$$\textit{Cluster Weight} = \frac{\textit{Total Eighth-Grade Enrollment for State}}{\textit{Total Eighth-Grade Enrollment for Cluster}} * \frac{1}{n}$$

The base weight for a school with eighth-grade enrollment of 20 or more was equal to the cluster weight for the cluster containing the school. For small schools, subject to thinning (see Section 3.3.3.1), the school base weight was obtained from the cluster weight by multiplying by the factor x/30, where x denotes the total eighth-grade enrollment for the cluster.

For Cluster Type 3 States, Type 3A Clusters (Noncertainty):

$$\textit{Cluster Weight} = \frac{\textit{Total Eighth-Grade Enrollment for 3A's}}{\textit{Total Eighth-Grade Enrollment for Cluster}} * \frac{1}{\textit{Number of 3A Clusters}}$$

The school base weight is equal to the cluster weight for the cluster containing the school (there is one school per cluster in Type 3A clusters).

For Cluster Type 3 States, Type 3B Clusters (Noncertainty):

$$\textit{Cluster Weight} = \frac{\textit{Total Eighth-Grade Enrollment for 3B's}}{\textit{Total Eighth-Grade Enrollment for Cluster}} * \frac{1}{\textit{Number of 3B Clusters}}$$

The school base weight is equal to the cluster weight for the cluster containing the school.

8.2.2 Calculating the Expected Sample Sizes for New Schools

This section provides a theoretical justification for the methods used to sample and weight new schools. For a given state, suppose that there are M school districts, of which the first m had at least one school in the initial sample. Within the i-th district, let $K_2^{(i)}$ denote the total number of schools, with the first $K_1^{(i)}$ appearing on the original frame. Thus units on the frame are indexed as:

μ_{ij} = j-th *old* school in the i-th district for $1 \le j \le K_1^{(i)}$

= $(j - K_1^{(i)})$-th *new* school in the i-th district for $K_1^{(i)} + 1 \le j \le K_2^{(i)}$

Let y_{ij} be a characteristic of interest, such as an estimated average or total test score for the students in the school. (By convention, y_{ij} will be defined to be zero when the school is closed, does not have the appropriate grade, etc.) Consider the problem of estimating the quantity

$$Y = \sum_{i=1}^{M} \sum_{j=1}^{K_2^{(i)}} y_{ij}.$$

Dividing the total Y into old and new schools, it can be written as

$$Y = Y_1 + Y_2$$

where

$$Y_1 = \sum_{i=1}^{M} \sum_{j=1}^{K_1^{(i)}} y_{ij}$$

$$Y_2 = \sum_{i=1}^{M} \sum_{j=K_1^{(i)}+1}^{K_2^{(i)}} y_{ij}$$

Assuming that the y_{ij} have selection probabilities π_{ij}, the Horvitz estimator of Y_1 is

$$\hat{Y}_1 = \sum_{i=1}^{m} \sum_{j\epsilon S} \frac{y_{ij}}{\pi_{ij}}$$

where the second summation is over the schools in the sample ($j \epsilon$ S). (Notice that $j \epsilon$ S necessarily implies $j \leq K_1^{(i)}$.) The problem now is to estimate Y_2.

For the Trial State Assessment, the within-school sample size was fixed at $n = 30$ (except in the smallest states where nearly all of the schools were selected). Denote the number of new schools to be selected in the i-th district as k_i^*, the number of students in the ij-th school as N_{ij}, the number of students in new schools in the i-th district as N_i, and the overall sampling fraction for students as f.

Assume that new schools will be selected with probability proportionate to size, so that the j-th new school will have a selection probability of $k_i^* N_{ij}/N_i$. Now, in order that the sample of *new* schools be self-weighting with respect to the original sample of *old* schools, we must have

$$\textit{School Weight} = \sum_{j\epsilon S} \frac{1}{\pi_{ij}} \frac{1}{K_1^{(i)}} \frac{N_{ij}}{30} \frac{N_i}{N_{ij}} \frac{1}{k_i^*} = \frac{1}{f}$$

or, rearranging terms,

$$k_i^* = \sum_{j\epsilon s} \frac{1}{\pi_{ij}} \frac{f}{K_1^{(i)}} \frac{N_i}{30}$$

It should be noted that k_i^* was actually the expected sample size, since, in general, it was fractional. Systematic sampling was used to select new schools from within each district. Thus, the actual sample size for a given district was either the integer immediately above or below k_i^*.

8.2.3 Weighting New Schools

As described in Chapter 3, new schools were sampled from the updated sampling frame list from each district. Assume for simplicity of exposition that this procedure consisted of selection with probability 1 (this assumption will be removed later). In this case, *all* the measured characteristics, y_{ij}, for new schools entered the sample if *any* of the old schools in the district were selected. Thus, let

$$\xi_{ij} \quad = \quad \begin{cases} 1 & \textit{if } u_{ij} \textit{ was in the sample of old schools} \\ 0 & \textit{otherwise} \end{cases}$$

and

$$t_{ij} = \sum_{j=K_1^{(i)}+1}^{K_2^{(i)}} y_{ij} \,.$$

Note that the term $t_{ij} > 0$ if $\xi_{ij} = 1$ for some $j \leq K_1^{(i)}$, i.e., if at least one old school was selected in the sample. Consider the estimator given by

$$\hat{Y}_2 = \sum_{i=1}^{m} t_{ij} \sum_{j \epsilon S} \frac{c_{ij}}{\pi_{ij}}$$

$$= \sum_{i=1}^{M} t_{ij} \sum_{j=1}^{K_1^{(i)}} \frac{c_{ij}}{\pi_{ij}} \xi_{ij} \,.$$

where the c_{ij} are constants to be specified below.

Note that $E(\xi_{ij}) = \pi_{ij}$. Under the condition that

$$\sum_{j=1}^{K_1^{(i)}} c_{ij} = 1$$

for each $i = 1, 2, ..., M$, it was easily seen that $\hat{Y}_2$ was an unbiased estimate of Y_2:

$$E(\hat{Y}_2) = E\left(\sum_{i=1}^{M} t_{ij} \sum_{j=1}^{K_1^{(i)}} \frac{c_{ij}}{\pi_{ij}} \xi_{ij}\right)$$

$$= \sum_{i=1}^{M} t_{ij} \sum_{j=1}^{K_1^{(i)}} \frac{c_{ij}}{\pi_{ij}} E(\xi_{ij})$$

$$= \sum_{i=1}^{M} t_{ij}$$

$$= \sum_{i=1}^{M} \sum_{j=K_1^{(i)}+1}^{K_2^{(i)}} y_{ij} = Y_2 \, .$$

Thus an appropriate weight to y_{ij} was

$$w_{ij} = \sum_{j \epsilon S} \frac{c_{ij}}{\pi_{ij}}$$

provided that the c_{ij} were defined to sum to 1 over the original schools in the district. The method used was to take

$$c_{ij} = \frac{1}{K_1^{(i)}}$$

(One alternative would have been to let c_{ij} be the percentage of the students in th *i*-th district that attend the *j*-th school.)

In the derivation presented above, it was assumed that all new schools would be taken. When subsampling was done among the new schools, the effect was to replace the term t_{ij} with

$$t'_{ij} = \sum_{j=K_1^{(i)}+1}^{K_2^{(i)}} u'_{ij} \frac{y_{ij}}{\pi'_{ij}}$$

where

$$u'_{ij} = \begin{cases} 1 & \textit{if } u_{ij} \textit{ was in the sample of new schools} \\ 0 & \textit{otherwise} \end{cases}$$

and where π'_{ij} represented the selection probability of the *j*-th new school in the *i*-th district. By conditioning on the selection of districts, it was easily seen that the theory presented above can be adapted to cover the more general case.

8.2.4 Adjusting for NAEP National Sample and National Longitudinal Study Selection Probabilities

This procedure reflected the probability that the school was excluded from the state frame as a result of drawing the national NAEP samples. Adjustments were made to school base weights to account for the school's possible inclusion in the national assessment. The adjustment, given below, was multiplied by the inverse of a school's probability of being selected from the frame to produce the adjusted weight.

$$\textit{Adjustment} = \begin{cases} 1/(1 - P_N) & \textit{if } P_N < 0.5 \\ 2 & \textit{if } P_N \geq 0.5 \end{cases}$$

where P_N is the National Selection Probability, conditional on the national sample of geographic PSU's. This adjustment procedure reflected the procedure used to exclude schools from the state frame. If $P_N < 0.5$, and the school was selected for the national sample, then it was excluded from the state frame. If $P_N \geq 0.5$, and the school was selected for the national sample, it was retained on the state frame with probability of $\frac{(P_N - 0.5)}{P_N}$, independently for each such school in the state.

The exceptions were schools in Delaware, the District of Columbia, and Hawaii. These schools received no adjustment because there were so few schools eligible for the Trial State Assessment in these states that all the eligible schools were included in the state sample, regardless of whether they were included in the national sample. In the other states where all schools were selected (Rhode Island, Wyoming, Guam, U.S. Virgin Islands), no geographic PSU was selected for the national sample, so that the issue of control of overlap of the national and state samples did not arise.

Schools with both grade 8 and grade 10, previously selected for the National Educational Longitudinal Study (NELS) and surveyed in 1990 for the NELS First Phase Follow-up, were also excluded for the Trial State Assessment wherever possible. The procedures used were similar to those used to exclude national NAEP sample schools. Weights were adjusted for these exclusions using the formulas shown above.

8.3 ADJUSTMENT OF BASE WEIGHTS FOR NONRESPONSE

The base weight for a student was adjusted by two nonresponse factors: one to adjust for non-participating schools for which no substitute participated, and one to adjust for students who were invited to the assessment but did not appear either in the scheduled session or in a makeup session.

8.3.1 School Level Nonresponse Adjustments

Nonresponse classes were created based on urbanicity and minority strata. In states where no minority stratification was used, nonresponse classes were created based on median household income. The procedure for creating income classes was as follows. Three classes of schools were formed for each urbanicity stratum so that (1) each class had approximately the same number of sample schools and (2) the classes were ranked -- the school with the highest median income in the first class had a lower median income than the school with the lowest median income in the second class, and so forth. This was done using only the schools in the sample (including new schools), sorting them by median income and then dividing the schools into three groups with equal numbers of schools. In carrying this out, all schools were used in all states except Montana, North Dakota, Nebraska and Oklahoma. In these states only large schools were used to form income strata. In creating the nonresponse adjustments, urbanicity stratum and minority/income stratum were used as the primary and secondary variables, respectively, for creating adjustment classes. In all states except Montana, North Dakota, Nebraska, and Oklahoma, all in-scope schools were sorted into the nonresponse classes described above and used in the nonresponse adjustments. In these four states, Type 3B clusters were treated as a separate nonresponse class.

The original strata are shown for each state in Table 8-1. For example, Alabama (AL) had nine strata, three levels of urbanicity (Central City = 1, Metropolitan = 2, Other = 3) and three levels of minority (Low = 1, Medium = 2, High = 3). These strata formed the initial nonresponse adjustment classes. The classes varied from one state to another. For example, New York had Black/Hispanic stratification in *Central City*, minority stratification in *Metropolitan* and income stratification (i.e., no minority stratification) in *Other*. Note that in a number of states and territories there was no school level non-participation so that the school nonresponse adjustment for all schools was 1.0.

8.3.2 Certainty Schools

It was determined, nonresponse class by nonresponse class, whether or not all certainty schools participated. If all certainties in a given class participated, then nonresponse adjustments were made only to noncertainty schools because the certainty schools were not part of the randomized selection process. However, if at least one certainty did not participate, then the nonresponse adjustment for that class was made using both certainty and noncertainty schools. Even though the certainties were still not part of the randomized selection process, the nonresponding certainty needed a weight adjustment so that the full set of certainty schools would be represented appropriately in the estimates.

8.3.3 Preliminary Evaluation of Nonresponse Classes

The objectives in forming the nonresponse classes were to create as many classes as possible, as homogeneous as possible, but such that the resulting nonresponse adjustment factors were not subject to large random variations resulting from sampling error. The procedures discussed below were established with the aim of striking the necessary balance between these objectives.

The schools were sorted into nonresponse classes and the following counts and ratios were listed for each initial nonresponse class:

- Total in-scope schools from the original sample
- Participating in-scope schools from the sample (both original and substitutes)
- Total in-scope schools from the original sample divided by participating in-scope schools from the sample

TABLE 8-1 : INITIAL NONRESPONSE ADJUSTMENT CLASSES

State	Urbanicity	Minority
AL	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority
AR	Central City/Suburban	Low Minority
	Central City/Suburban	Medium Minority
	Central City/Suburban	High Minority
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority
AZ	Central City/Suburban	Low Black/Low Hispanic
	Central City/Suburban	Low Black/High Hispanic
	Central City/Suburban	High Black/Low Hispanic
	Central City/Suburban	High Black/High Hispanic
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority
CA	Central City	Low Black/Low Hispanic
	Central City	Low Black/High Hispanic
	Central City	High Black/Low Hispanic
	Central City	High Black/High Hispanic
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority
CO	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority

117

TABLE 8-1: INITIAL NONRESPONSE ADJUSTMENT CLASSES

State	Urbanicity	Minority
CT	Central City	Low Black/Low Hispanic
	Central City	Low Black/High Hispanic
	Central City	High Black/Low Hispanic
	Central City	High Black/High Hispanic
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
DC	–	Low Median Income
	–	Medium Median Income
	–	High Median Income
DE	–	Low Median Income
	–	Medium Median Income
	–	High Median Income
FL	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Black/Low Hispanic
	Suburban	Low Black/High Hispanic
	Suburban	High Black/Low Hispanic
	Suburban	High Black/High Hispanic
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority
GA	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority
GU	–	Low Median Income
	–	Medium Median Income
	–	High Median Income

TABLE 8-1: INITIAL NONRESPONSE ADJUSTMENT CLASSES

State	Urbanicity	Minority
HI	–	Low Median Income
	–	Medium Median Income
	–	High Median Income
IA	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
ID	–	Low Median Income
	–	Medium Median Income
	–	High Median Income
IL	Central City	Low Black/Low Hispanic
	Central City	Low Black/High Hispanic
	Central City	High Black/Low Hispanic
	Central City	High Black/High Hispanic
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
IN	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
KY	Central City/Suburban	Low Minority
	Central City/Suburban	Medium Minority
	Central City/Suburban	High Minority
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income

TABLE 8-1 : INITIAL NONRESPONSE ADJUSTMENT CLASSES

State	Urbanicity	Minority
LA	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority
MD	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority
MI	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
MN	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
MT	–	Low Median Income
	–	Medium Median Income
	–	High Median Income

TABLE 8-1: INITIAL NONRESPONSE ADJUSTMENT CLASSES

State	Urbanicity	Minority
NC	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority
ND	Central City/Suburban	Low Median Income
	Central City/Suburban	Medium Median Income
	Central City/Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
NE	Central City/Suburban	Low Median Income
	Central City/Suburban	Medium Median Income
	Central City/Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
NH	Central City	Low Median Income
	Central City	Medium Median Income
	Central City	High Median Income
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
NJ	Central City	Low Black/Low Hispanic
	Central City	Low Black/High Hispanic
	Central City	High Black/Low Hispanic
	Central City	High Black/High Hispanic
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority

TABLE 8-1: INITIAL NONRESPONSE ADJUSTMENT CLASSES

State	Urbanicity	Minority
NM	Central City/Suburban	Low Median Income
	Central City/Suburban	Medium Median Income
	Central City/Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
NY	Central City	Low Black/Low Hispanic
	Central City	Low Black/High Hispanic
	Central City	High Black/Low Hispanic
	Central City	High Black/High Hispanic
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
OH	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
OK	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income

TABLE 8-1 : INITIAL NONRESPONSE ADJUSTMENT CLASSES

State	Urbanicity	Minority
OR	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
PA	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
RI	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
TX	Central City	Low Black/Low Hispanic
	Central City	Low Black/High Hispanic
	Central City	High Black/Low Hispanic
	Central City	High Black/High Hispanic
	Suburban	Low Black/Low Hispanic
	Suburban	Low Black/High Hispanic
	Suburban	High Black/Low Hispanic
	Suburban	High Black/High Hispanic
	Other	Low Black/Low Hispanic
	Other	Low Black/High Hispanic
	Other	High Black/Low Hispanic
	Other	High Black/High Hispanic

TABLE 8-1 : INITIAL NONRESPONSE ADJUSTMENT CLASSES

State	Urbanicity	Minority
VA	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Minority
	Suburban	Medium Minority
	Suburban	High Minority
	Other	Low Minority
	Other	Medium Minority
	Other	High Minority
VI	–	Low Median Income
	–	Medium Median Income
	–	High Median Income
WI	Central City	Low Minority
	Central City	Medium Minority
	Central City	High Minority
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
WV	Central City	Low Median Income
	Central City	Medium Median Income
	Central City	High Median Income
	Suburban	Low Median Income
	Suburban	Medium Median Income
	Suburban	High Median Income
	Other	Low Median Income
	Other	Medium Median Income
	Other	High Median Income
WY	–	Low Median Income
	–	Medium Median Income
	–	High Median Income

The following guidelines were established for reviewing these counts and ratios and determining what collapsing should be done. Within an initial nonresponse class, if the ratio of inscope schools to participating schools was less than 1.35, with at least six participating schools in the class, there was no need to collapse the particular cell. If any nonresponse class had fewer than 6 schools or a ratio greater than 1.35, it was collapsed with another class such that the new class met these conditions. The order of variables to be collapsed (from most desirable to least desirable) was income strata or minority strata, followed by urbanicity strata. The exceptions occurred in cases where minority classes within an urbanicity stratum varied considerably as to the relative sizes of the minority population. In such a case we collapsed over urbanicity first to keep the classes as homogeneous as possible with regard to race/ethnicity.

Some preliminary collapsing was required in six states: Iowa, Minnesota, Montana, Nebraska, Oklahoma, and Pennsylvania. All the collapsing was over either income or minority strata, depending on the nonresponse class (with no collapsing of urbanicity strata). The majority of classes had income variables.

8.3.4 Adjustment Factors

The school nonresponse adjustment factor for the *j*-th school in the *h*-th class was

$$f_{hj}^{(1)} = (1 - \gamma_j) + \gamma_j \frac{\sum_{l \in C_h} W_l m_l \gamma_l}{\sum_{l \in C_h} W_l m_l \gamma_l \delta_l}$$

where

C_h = the set of schools in class *h*

W_l = Base weight of the *l*-th school

m_l = Frame grade enrollment (G08) for the *l*-th school

$$\delta_i = \begin{cases} 1 & \text{if } l\text{-th school participated} \\ 0 & \text{otherwise} \end{cases}$$

$$\gamma_l = \begin{cases} 0 & \text{if } l\text{-th school was a certainty and all certainties in class } h \text{ participated} \\ 1 & \text{if } l\text{-th school was a certainty and at least one certainty in class } h \text{ did not participate} \\ 1 & \text{if } l\text{-th school was not a certainty.} \end{cases}$$

A school was said to have participated if

(1) it was selected for the sample -- from the original frame or from the lists of new schools provided by participating school districts -- and student assessment data were obtained from the school, OR

(2) the school refused but was assigned a regular substitute and student assessment data were obtained from that substitute, OR

(3) the school refused but was assigned a double-session substitute and the substitute was verified to have provided student assessment data from both sessions.

Both the numerator and denominator contained only in-scope schools. Notice that the numerator and the denominator of this adjustment included certainties if at least one certainty in the class did not participate. Otherwise, certainties were excluded from both the numerator and denominator and the adjustment factor was set to one for such certainties.

Once these adjustments were calculated, the following were listed for each class of urbanicity and minority or income:

- Adjustment factor (f_{hj})
- Numerator and denominator of the adjustment factor

We reviewed these factors and considered whether any values of f_{hj} exceeded 1.35. Since this did not occur, no additional collapsing of classes was needed. In summary, the only collapsing done was in the preliminary evaluation of nonresponse classes, as described in the last paragraph of section 8.3.3.

8.3.5 Student Level Nonresponse Adjustments

The variables for adjusting student nonresponse were: urbanicity, percentage of minority students or median income, whether or not the student was age 13 or younger, and in addition whether the session which the student was to attend was monitored or unmonitorerd. The definition of "13 or younger" was "born on or after 10/1/75". To determine whether the nonresponse classes need collapsing, we reviewed each nonresponse class and examined the numbers of assessed students, the combined total of the assessed and absent students, and the ratio of the latter to the former. Excluded students were processed separately, with each nonresponse class showing the number of excluded students with completed questionnaires, the total number of excluded students, and the ratio of the latter to the former.

The following guidelines were established for collapsing nonresponse cells. Any cell with fewer than 20 assessed students was collapsed regardless of the adjustment factor. The order in which variables were collapsed, from most preferable to least preferable, was by monitor status, urbanicity, minority/income status, and lastly, over age 13. The exception was the case where it was not possible to collapse urbanicity because the two urbanicity variables had different minority strata. Then the order was by monitor status, minority/income status, urbanicity, and over age 13. If a cell had between 20 and 30 assessed students, the ratio of invited to assessed had to be no larger than 1.5 to avoid collapsing. For more than 30 assessed students, the ratio had to be no larger than 2.0. We continued collapsing until all these rules held true. Based on these guidelines, most states had some amount of collapsing, almost all over monitor status. The reason for the numerous collapses was more often due to the small number of assessed students within the cell, rather than to a high ratio of invited to assessed.

In the double session schools, the students who were associated with their own school retained the base weight and the nonresponse adjustment factors of the school selected. To the students who were associated with the non-participating school that was being substituted for, we assigned the base weight and the school nonresponse adjustment factors of that non-participating school for which the substitute session was performed.

8.3.6 Assessed Student Nonresponse Adjustments

We made separate nonresponse adjustments for the assessed students from monitored and unmonitored schools. The nonresponse classes were denoted as A_{h1} and A_{h2}, respectively. Within hi-th nonresponse adjustment class A_{hi}, the assessed student nonresponse adjustment was calculated as:

$$f_{hi}^{(2)} = \frac{\sum_{m \epsilon A_{hi}} W_m^{(l)}}{\sum_{m \epsilon A_{hi}} W_m^{(l)} \delta_m}$$

where

$$\delta_m = \begin{cases} 1 & \textit{if m-th student was assessed} \\ 0 & \textit{otherwise.} \end{cases}$$

and $W_m^{(1)} = f_{kl}^{(1)} W_l$ for student m from school l, in school nonresponse class k, and the sums were across all invited students within the class.

8.3.7 Excluded Student Nonresponse Adjustments

For excluded students the same basic procedures as described above for assessed students were used, except that 1)the numerator and denominator contained excluded rather than assessed students; 2)we made no distinction between the monitored and the unmonitored schools; and 3) there were no student age classes. Specifically, the excluded students nonresponse adjustments were calculated as:

$$f_h^{(3)} = \frac{\sum_{m \epsilon A_h} W_m^{(l)}}{\sum_{m \epsilon A_h} W_m^{(l)} \lambda_m}$$

where

$$\lambda_m = \begin{cases} 1 & \textit{if excluded student form was available for the } m\textit{-th student} \\ 0 & \textit{otherwise} \end{cases}$$

and $W_m^{(1)} = f_{kl}^{(1)} W_l$ for student *m* from school *l*, in school nonresponse class *k*, and the sums are over all excluded students within the class.

8.3.8 Student Nonresponse Adjustments for the Territories

The definitions of the nonresponse adjustment classes for the territories were somewhat different from the definitions for the other participants. The class A here represented an individual school or a pair of schools, (there were only six schools in each territory). Within each nonresponse adjustment class A, we defined A_{h1} and A_{h2} in the same way as in the general case.

8.4 VARIATION IN WEIGHTS

After completion of the weighting steps reflecting probabilities of selection and adjustments for nonresponse, an analysis was conducted in each state of the distribution of the resulting student weights. The analysis was intended both as a part of the process of checking that the various weight components had been derived and combined appropriately in each state and of examining the likely impact of the variability in sample weights within a state on the level of precision of 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 "final" (i.e., school and student-nonresponse-adjusted) weights, both for the approximately 2,500 assessed students in each state and for subgroups defined by age, sex, race, level of urbanicity, 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 weight distribution and the presence of outliers, where the weights were several standard deviations away from the mean weight.

The coefficient of variation was considered because of the impact of variable weights on the effective sample size of the particular sample. Assuming that the value of the variables for individual students were uncorrelated with the weights for the students, the use of data where the weights have coefficient of variation of C percent, gives sampling variances for estimated averages or aggregates of approximately $\{1 + \left(\frac{C}{100}\right)^2\}$ times as great as does an unweighted sample. Outliers, in terms of numbers of standard deviations from the mean weight, were considered because the presence of such an outlier was a likely indication of the existence of an error in the weighting procedure and because it was likely that a relatively few outlying 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 (see Table 8-2). 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. The principle exceptions were those states where a number of schools were selected with certainty, but the sample size of students within such schools was limited to 30 (Montana, Nebraska, New Mexico, North Dakota). The result was that the proportion of sample students who were from larger schools was somewhat small relative to the general eighth grade population, and therefore such students received relatively large weights.

Since in the states with large coefficient of variation there was a strong relationship between the weight the student received and the type of school (in particular the size of the school enrollment) that they attended, any process to trim the weights of students in such schools, and thus reduce the coefficient of variation in the weights and also the sample variance of estimates, had the strong potential to introduce non-negligible bias into the survey estimates. It was thus decided that no trimming of survey weights should be carried out in such cases.

In six states there were a group of noticeably outlying weights contributed by students from a single school (three schools in one state), which arose because the eighth-grade enrollment assumed at the time of sample selection of the school proved to be less than the actual enrollment. Since sample size within schools was limited (in general) to 30 students, this resulted in noticeably large weights for students from these schools. An evaluation was made of the impact of trimming these largest weights to a level comparable with the highest other weights found in the state. Such a procedure produced some reduction in the size of the coefficient of variation. It was sufficiently modest in each case, however, that we judged that the potential for the introduction of bias through trimming, when combined with the considerable effort that would be required to implement an appropriate trimming procedure, was such that it was preferable not to apply any trimming to the weights in these states.

The analyses conducted confirmed that weight components had been calculated and combined correctly, and it was concluded that weight trimming should not be undertaken in any case.

TABLE 8-2: COEFFICIENTS OF VARIATION

STATE	N	CV	WEIGHTING EFFECT	CLUSTERING EFFECT	DESIGN EFFECT
Alabama	2531	25.313%	1.06	3.50	3.72
Arizona	2558	25.555%	1.07	3.63	3.87
Arkansas	2669	26.315%	1.07	2.30	2.46
California	2424	24.587%	1.06	3.35	3.55
Colorado	2675	16.553%	1.03	2.98	3.06
Connecticut	2672	23.551%	1.06	2.73	2.88
Delaware	2110	58.786%	1.35	0.71	0.96
District of Columbia	2135	29.770%	1.09	1.09	1.19
Florida	2534	21.538%	1.05	3.15	3.30
Georgia	2766	27.780%	1.08	3.74	4.03
Guam	1617	2.240%	1.00	0.50	0.50
Hawaii	2551	28.810%	1.08	0.64	0.69
Idaho	2716	60.900%	1.37	1.35	1.85
Illinois	2683	29.519%	1.09	6.42	6.98
Indiana	2568	21.610%	1.05	3.33	3.49
Iowa	2474	31.720%	1.10	3.01	3.31
Kentucky	2680	37.265%	1.14	3.25	3.70
Louisiana	2572	23.369%	1.05	3.92	4.13
Maryland	2794	15.810%	1.02	4.22	4.33
Michigan	2587	23.825%	1.06	3.04	3.21
Minnesota	2586	32.043%	1.10	2.23	2.46
Montana	2486	84.758%	1.72	1.36	2.34
Nebraska	2519	47.718%	1.23	1.90	2.33
New Hampshire	2568	75.080%	1.56	1.34	2.09
New Jersey	2710	39.187%	1.15	2.19	2.53
New Mexico	2643	56.705%	1.32	1.42	1.88
New York	2303	29.448%	1.09	3.00	3.26
North Carolina	2843	15.970%	1.03	2.64	2.71
North Dakota	2485	98.688%	1.97	2.07	4.08
Ohio	2673	19.180%	1.04	2.78	2.88
Oklahoma	2222	33.319%	1.11	3.53	3.92
Oregon	2709	15.158%	1.02	2.95	3.02
Pennsylvania	2534	27.685%	1.08	6.03	6.49
Rhode Island	2675	32.809%	1.11	0.52	0.58
Texas	2565	25.361%	1.06	3.94	4.19
Virginia	2661	22.733%	1.05	4.74	4.98
Virgin Islands	1328	5.718%	1.00	0.50	0.50
West Virginia	2601	25.900%	1.07	2.36	2.52
Wisconsin	2750	21.070%	1.04	5.01	5.23
Wyoming	2701	33.623%	1.11	1.24	1.38

8.5 RAKING OF WEIGHTS FOR COMPARING MONITORED AND UNMONITORED SESSIONS

The monitored and the unmonitored schools comprised two random half samples. In order to compare the results for them with greater precision, a form of post-stratification to certain characteristics of the whole sample was employed on each of these half samples. This procedure, called raking, was intended to reduce variance of estimates of differences between monitored and unmonitored sessions, by controlling for sampling variability in student characteristics known to be related to mathematics proficiency, but unrelated to the monitoring process.

The post-stratification was carried out with respect to three sets (A, B, and C) of student classifications:

A Race/Ethnicity by Parents' Education
B Age by Sex
C Type of Mathematics Course Taken by Self-Reported Mathematics Ability

We obtained weighted and unweighted counts from the whole sample in the marginal distributions A, B, and C. This enabled us to decide whether there was a need to collapse any marginal cells within a state, and the choice of cells to collapse.

Table 8-3 gives the preliminary structure of the set A. There were twelve different patterns for the structure of A. For example, in Indiana there were six cells. The first one consisted of the any category of parents' education for the Blacks, American Indians, and Hispanics. The next five consisted of race/ethnicity other than Black, American Indian, and Hispanic and parents' education being unknown, less than high school, high school, more than high school but no college, and college.

Table 8-4 gives the structure of the set B. It consisted of male and female students of age less than or equal to the appropriate age for the eighth grade and age older than the appropriate age for the eighth grade. The definition of appropriate age was defined as "born on or after 10/1/75".

Table 8-5 gives the preliminary structure of the set C. The collapsing was done by the following rule: in the first round collapse (if needed) 1 and 2, 3a and 3b, and 4a and 4b. If this was not sufficient then we collapsed 1, 2, 3a, and 3b on one side , and 4a, 4b, and 5 on the other.

The rule for collapsing was to join the adjacent cells, i.e., the most similar cells, for any cell with fewer than 75 students. In all of the states we collapsed the cells from set C of "Agree", Algebra and "Agree," Not Algebra. In many states we also collapsed the cells of Race/Parents' Education equal to Other/<High School and Other/High School.

TABLE 8-3 : PRELIMINARY STRUCTURE OF RACE/ETHNICITY BY PARENT'S EDUCATION

States	Race	Parent's Education
IN, OK, RI, WI	Black, American Indian	All
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College
CO, MI, OH, PA	Black, American Indian	Less than High School or Unknown
	Black, American Indian	High School and higher
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College
AL, AR, DE, NC, VA	Black, Amr Ind, Hispanic	Less than High School or Unknown
	Black, Amr Ind, Hispanic	High School
	Black, Amr Ind, Hispanic	More than High School or College
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College
GA, LA, MD	Black, Amr Ind, Hispanic	Unknown
	Black, Amr Ind, Hispanic	Less than High School
	Black, Amr Ind, Hispanic	High School
	Black, Amr Ind, Hispanic	More than High School
	Black, Amr Ind, Hispanic	College
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College
CA	Black, American Indian	All
	Hispanic	Unknown
	Hispanic	Less than High School
	Hispanic	High School
	Hispanic	More than High School
	Hispanic	College
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College

TABLE 8-3 : PRELIMINARY STRUCTURE OF RACE/ETHNICITY BY PARENT'S EDUCATION

States	Race	Parent's Education
CT, NV	Black, American Indian	All
	Hispanic	All
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College
TX	Black, American Indian	Less than High School or Unknown
	Black, American Indian	High School and higher
	Hispanic	Unknown
	Hispanic	Less than High School
	Hispanic	High School
	Hispanic	More than High School
	Hispanic	College
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College
NJ, NY	Black, American Indian	Less than High School or Unknown
	Black, American Indian	High School and higher
	Hispanic	Less than High School or Unknown
	Hispanic	High School and higher
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College
FL, IL	Black, American Indian	Less than High School or Unknown
	Black, American Indian	High School
	Black, American Indian	More than High School or College
	Hispanic	All
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College

TABLE 8-3: PRELIMINARY STRUCTURE OF RACE/ETHNICITY BY PARENT'S EDUCATION

States	Race	Parent's Education
AZ, NM	Hispanic	Unknown
	Hispanic	Less than High School
	Hispanic	High School
	Hispanic	More than High School
	Hispanic	College
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College
HI	Asian, etc.	Unknown
	Asian, etc.	Less than High School
	Asian, etc.	High School
	Asian, etc.	More than High School
	Asian, etc.	College
	Other	Unknown
	Other	Less than High School
	Other	High School
	Other	More than High School
	Other	College
DC, IA, ID, KY, MN, MT, ND, NE, NH, OR, WV, WY, AS, GU, VI	All	Unknown
	All	Less than High School
	All	High School
	All	More than High School
	All	College

Table 8-4

Preliminary Structure of Age by Sex

Age	Sex
≤ Modal Grade	Male
≤ Modal Grade	Female
≥ Modal Grade	Male
≥ Modal Grade	Female

Table 8-5

Preliminary Structure of Type of Math
Courses Taken by Self by Self-Reported Math Ability

Response to the following statements:

"Feel I am good at Math"	Type of Math Course Taken
Strongly Disagree	{Not used}*
Disagree	{Not used}*
Undecided	Algebra
Undecided	Not algebra
Agree	Algebra
Agree	Not algebra
Strongly agree	{Not used}*

* "Type of math course taken" was not used in conjunction with this response to "Feel I am good at Math" question

8.5.1 Procedures for Raking

The raking was accomplished by series of adjustments to the sets A, B, and C, and additional final adjustment to the set A. Let

$W^{(0)}_{smijk}$ = base weight of the m-th student in either monitored (s=0) or unmonitored half-sample (s=1), whose responses fell in the i-th class of A, j-th class of B, and k-th class of C.

$$N_{i..} = \sum_{s=0}^{1} \sum_{m} \sum_{j \epsilon B} \sum_{k \epsilon C} W^{(0)}_{smijk}$$

$$N_{.j.} = \sum_{s=0}^{1} \sum_{m} \sum_{i \epsilon A} \sum_{k \epsilon C} W^{(0)}_{smijk}$$

$$N_{..k} = \sum_{s=0}^{1} \sum_{m} \sum_{i \epsilon A} \sum_{j \epsilon B} W^{(0)}_{smijk}$$

The values of these cell totals (the N's) were printed and reviewed before we decided on the final number and configuration of cells in A, B, and C.

Then the weight total for, say, monitored schools (s=0) was:

$$\tilde{N}^{(0)}_{0ijk} = \sum_{m} W^{(0)}_{0mijk}$$

Each adjustment in the post-stratification scheme proceeded as follows:

1. $$\tilde{N}^{(1)}_{0ijk} = \tilde{N}^{(0)}_{0ijk} \frac{N_{i..}}{\sum_{j \epsilon B} \sum_{k \epsilon C} \tilde{N}^{(0)}_{0ijk}}$$

2. $$\tilde{N}^{(2)}_{0ijk} = \tilde{N}^{(1)}_{0ijk} \frac{N_{.j.}}{\sum_{i \epsilon A} \sum_{k \epsilon C} \tilde{N}^{(1)}_{0ijk}}$$

3. $$\tilde{N}^{(3)}_{0ijk} = \tilde{N}^{(2)}_{0ijk} \frac{N_{.k}}{\sum_{i\epsilon A} \sum_{j\epsilon B} \tilde{N}^{(2)}_{0ijk}}$$

4. $$\tilde{N}_{0ijk} = \tilde{N}^{(3)}_{0ijk} \frac{N_{.k}}{\sum_{i\epsilon A} \sum_{j\epsilon B} \tilde{N}^{(3)}_{0ijk}}$$

At this point we printed out the ratio of $\tilde{N}_{sijk}$ *to* $\tilde{N}^{(0)}_{sijk}$ for each ijk, separately for monitored and unmonitored schools. The highest acceptable ratio was 2. If unacceptably high ratios had been encountered, it would have been necessary to go to the sets A, B, and C and do more of collapsing of adjacent cells before repeating the raking procedure. In fact, this did not occur in any case. At the same time we printed the ratios of the final to initial marginal totals. We also checked the degree of convergence by looking at

$$\frac{\max |\mathit{Full\ Sample\ Marginal\ -\ Raked\ Half\ Sample\ Marginal}|}{\mathit{Full\ Sample\ Marginal}}$$

and ensuring that this quantity was trivially small.

The raking procedures were repeated on the replicate weights (see Section 8.6) using the same cell structure as for the full sample weights. The marginal totals used in the raking were those from the full sample (monitored and unmonitored schools together). At the end we printed the ratio of the final to initial $\tilde{N}'s$ for each ijk, separately for each replicate. If any of these ratios had been too large, we would have collapsed the cells used for raking and repeated the process. However, this was not necessary.

The final comparison sample weight for each assessed student was multiplied by the ratio adjustment $\frac{\tilde{N}_{sijk}}{N^{(0)}_{sijk}}$. All students in cell (s, i, j, k) received this adjustment to their weights. The replicate weights received a comparable adjustment, using the appropriate replicate estimate in the numerator of the ratio adjustment.

8.6 REPLICATE WEIGHT FORMATION

8.6.1 Variance Replicates

Replication estimates the variance of the full sample. This process involves repeatedly selecting portions of the sample to calculate the estimate of interest. The estimates that result are called replicate estimates. The variability among these calculated quantities is used to obtain the full sample variance. As described below, the process of forming these replicate estimates involves first dividing the sample elements among a set of replicate groups, then using the pattern of replicate groups in a systematic fashion to apply replicate weights to the file.

At the school level, two sets of replicate groups, each consisting of two or three schools, were formed for each state. The first set -- the Overall Replicates -- were used to create replicate estimates for estimating the variance for population estimates, regardless of whether assessment monitoring was used. The second set -- the Comparison Replicates -- were used to create replicates for estimating variances of comparisons between the monitored and unmonitored schools. The two sets of replicates differed primarily in the manner in which certainty schools were grouped into replicates. Replicates for the territories -- Guam and the Virgin Islands -- were constructed at the session level, within schools.

8.6.2 School Level Replicate Weights

Each school belonged to two replicate groups, one for overall replicates and one for monitor replicates and had a separate weight for each replicate. Since there were approximately 50 monitored and 50 unmonitored schools in each state, there were about 25 monitored and 25 unmonitored replicate groups. The allocation of schools and students to these replicate groups are described in the sections below. Thus, each school had about 25 monitored and 25 unmonitored replicate weights (one for each "dropped out" school), or 50 in total. The replicate weights were calculated by first deriving base replicate weights, via the formula:

$$\text{Base Replicate Weight (n)} = \begin{cases} \text{(K)(Base weight)} & \text{for the students retained in replicate group n} \\ 0 & \text{for the students that were "dropped" from replicate group n} \\ \text{Base weight} & \text{for all other students} \end{cases}$$

The value of K was either 2 if two schools formed the replicate group or 1.5 if three schools formed the replicate group.

The overall replicate weights were obtained by repeating the school and student nonresponse adjustments, but utilizing each of the base replicate weights in turn, from the overall replicates, and then applying these nonresponse adjustments to the appropriate corresponding base replicate weight. The comparison replicate weights were obtained by repeating the school and student nonresponse adjustments and the raking adjustment, but utilizing each of the base replicate weights in turn from the comparison replicates. The corresponding adjustments were then applied to each of the base replicate weights. Thus, the student and school nonresponse adjustments were repeated 56 times for the overall replicates and an additional 56 times for the comparison replicates. There were 56 replicate weights in each set, as discussed in Section 8.6.5. In each state, the iterative raking adjustments were repeated 56 times , once for each of the comparison replicates.

We did not replicate the excluded students' weighting procedures using the comparison replicates, since there is no reason to expect that the exclusion of students should be in any way related to monitor status. Schools identified excluded students before the schools had any knowledge as to their monitor status.

8.6.3 School Level Overall Replicate Groups

To form replicates, the noncertainty schools in each state were sorted in the following order: large schools or clusters in the order in which sampling was done, new schools added to the sample subsequent to the initial sampling, and then clusters of small schools. Monitored schools were sequentially paired with monitored schools, and unmonitored schools were paired with unmonitored schools. In some states three schools remained to be paired at the end of the sort configuration, resulting in two replicate groups being formed by the 3 schools. In North Dakota, Nebraska, Montana, and Oklahoma clusters of small schools were handled separately. That is, larger schools and new schools were never combined with clusters of small schools to form a replicate group.

Certainty schools were sorted and paired differently. In certainty schools with 60 or more selected students, either one or two replicate groups were formed. Certainty schools with sample sizes of 100 formed two replicates, and certainty schools with sample sizes of 60 formed one replicate. Certainty schools with 30 students selected were sorted by monitoring status, and within monitoring status by the largest certainty school then the smallest certainty school. Next came the second largest certainty school and the second smallest certainty school, and so on. If one school was left, its students were divided into two batches to form a single replicate. It is important to note that noncertainty schools were never paired with certainty schools.

A single replicate estimate was formed by dropping from the sample some randomly selected students from a single replicate group and appropriately re-weighting the remaining sample elements to give an approximately unbiased estimate (the replicate estimate). To determine which school should be dropped, the following rules were used:

1. The integers one and two were assigned at random to the members of the group (or one through three if the group contained three schools).

2. Schools with either the digit '1' or the digit '3' were dropped.

Although the formation of replicate groups was different for noncertainty and certainty schools, the process of creating replicates was the same.

8.6.4 School Level Comparison Replicate Groups

The pairs of monitored and unmonitored schools were identified (one monitored and one unmonitored school per pair). For states with many small schools, the monitor pairs were identified at the cluster level. The sort order for the formation of replicates was by monitor status and monitor pair, unless the state had many small schools (Type 3A/3B Cluster states). In this case, the sort order was by monitor status, school size, and monitor pair. This separated large schools (which were not clustered) and small schools (which were in clusters of two or more schools). The comparison replicate groups were formed by sequentially pairing monitored schools with monitored schools, and unmonitored schools with unmonitored schools. Thus, for the most part, replicate groups were formed using the algorithm described above, provided that the large and small schools were not combined to form replicate groups.

The only exceptions to this pairing method were Delaware and the District of Columbia. These two jurisdictions had too few replicate groups, so schools were formed into groups of three. This resulted in two replicate groups for each group of three schools, thus increasing the number of replicates for the state by fifty percent.

Having established the comparison replicate groups, the comparison replicate weights were formed in the same way as the overall replicate weights, using the algorithm described at the end of the previous section.

8.6.5 Number of Replicate Groups

Based on statistical and computer processing requirements, it was decided that 56 replicates should suffice for variance calculation. In a few states, there were initially more than 56 overall replicate groups using the procedures described above. However, to standardize the computer processing, it was useful to have exactly 56 replicate groups for each state. Thus, it was necessary to combine some replicate groups to reduce the number of replicates. When possible, we combined schools in the same cluster. In states with many small schools, we combined clusters of two or more small schools with a large school. Table 8-6 gives a list of such states and the final number of overall and comparison replicate groups for each state.

In some cases, particularly where no schools were selected with certainty, slightly fewer than 56 replicates were formed. The additional replicate weights, to give a uniform total of 56, were generated by repeating the full sample weight the requisite number of times. This procedure results in appropriate jackknife sampling errors, while giving uniformity across states in the number of replicate weights.

8.6.6 Comparison Replicates for Guam and the U.S. Virgin Islands

Since all students in the territories were assessed in each of the few schools in the population, each school conducted a sizeable number of assessment sessions (up to ten), half of which were monitored and half unmonitored. For the comparison replicates, groups of monitored and unmonitored sessions were formed in the same manner as described above, but at the session level rather than the school level. In the following example M stands for Monitored session and U stands for Unmonitored session.

TABLE 8-6: REPLICATE GROUPS

State	Final # Overall Rep. Groups	Final # Comparison Rep. Groups
Alabama	53	53
Arizona	55	55
Arkansas	54	54
California	54	54
Colorado	54	54
Connecticut	55	54
Delaware	52	23
District of Columbia	53	24
Florida	54	54
Georgia	55	55
Guam	28	33
Hawaii	56	29
Idaho	55	54
Illinois	54	54
Indiana	53	53
Iowa	55	56
Kentucky	56	56
Louisiana	54	54
Maryland	53	53
Michigan	53	53
Minnesota	56	54
Montana	56	56
Nebraska	55	56
New Hampshire	54	54
New Jersey	55	54
New Mexico	55	55
New York	53	53
North Carolina	54	54
North Dakota	56	56
Ohio	53	53
Oklahoma	56	56
Oregon	55	56
Pennsylvania	54	54
Rhode Island	56	26
Texas	54	54
Virginia	54	54
Virgin Islands	22	28
West Virginia	55	54
Wisconsin	56	55
Wyoming	56	36

Table 8-7 shows the assignment of sessions to replicate groups. Each replicate group consists of two or three sessions, depending upon the number of monitored and unmonitored sessions within the school.

Table 8-7

Assignment of Sessions to Replicate Groups

School	1	1	1	1	1	2	2	2	2
Session	1	2	3	4	5	1	2	3	4
Status	M	U	U	U	M	U	M	U	M
Replicate	1	2	2	2	1	3	4	3	4
Group(s)		5	5	5					

In forming replicate estimates, the method for deciding which session was to be dropped was identical. The overall replicates were formed in the same manner, except that groups with three schools were only assigned one replicate group instead of two. In deciding which session in each replicate group should be dropped, random numbers were assigned to each student, the file was then sorted by school, replicate group number, session, and then the random numbers. For one replicate estimate, every second student was dropped from the sample.

Chapter 9

THEORETICAL BACKGROUND AND PHILOSOPHY OF NAEP SCALING PROCEDURES

Eugene G. Johnson and Robert J. Mislevy

Educational Testing Service

9.1 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 series of subscales even when different students 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 particular. Details of the scaling procedures specific to the Trial State Assessment are presented in Chapter 10.

9.2 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 generated 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 a large number of items. The Trial State Assessment required 137 items. To reduce student burden, each assessed student was presented only a fraction of the full pool of items using multiple matrix sampling procedures.

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

It is useful to view the assessed items as random representatives of a conceptually infinite pool of items within the same domain and of the same type. In this random item concept, a set of items is taken to represent the domain of interest. An obvious measure of achievement within a domain of interest is the average percent correct across all presented items within that domain. The advantage of averaging is that it tends to cancel out the effects of peculiarities in items which can affect item difficulty in unpredictable ways. Furthermore, averaging makes it possible to compare more easily the general performances of subpopulations.

Despite their advantages, there are a number of significant problems with average percent correct 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 percent correct metric 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 percents correct on specific sets of items, it provides no simple way to report trends over time when the item pool changes. Finally, the average percent correct provides no estimate of the distribution of proficiency in the population when each student is administered only a fraction of the items. Average percent correct statistics describe the mean performance of students within subpopulations but provide no other information about the distributions of skills among students in the subpopulations.

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. 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 take all of the items within the pool. Using the 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 was carried out within the five mathematics content areas specified in the objectives because it was anticipated that different patterns of performance might exist for these essential subdivisions of the subject area. The five subscales corresponded to one of the following content areas: Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. By creating a separate subscale for each of these content areas, potential differences in subpopulation performance between the content areas are maintained. Analyses of the subscale level results from the 1990 Trial State Assessment and national mathematics assessment have shown that the subscales provide additional information that a single scale cannot -- for example gender differences in mathematics performance by subscale.

The creation of subscales to describe mathematics performance does not preclude the reporting of an overall mathematics composite as a single index of overall mathematics performance. A composite is computed as the weighted average of the subscale scores where the weights correspond to the relative importance given to each subscale as defined by the objectives. The composite scores provide a global measure of performance within the subject area while the constituent subscale scores allow the measurement of important interactions within educationally relevant subdivisions of the subject area.

9.3 SCALING METHODOLOGY

This section reviews the scaling models employed in the analyses of data from the Trial State Assessment and the 1990 national mathematics assessment, as well as the "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 and Sheehan (1987) and Beaton and Johnson (1990) for additional information on how the models are used in NAEP, and to Rubin (1987) for the theoretical underpinnings of the approach.

The 137 mathematics items administered in the Trial State Assessment were also administered to grade eight students in the national mathematics assessment. However, because the administration procedures differed, the Trial State Assessment data was scaled independently from the national data. The national data also included results for students in grade 4 and grade 12. Details of the scaling of the Trial State Assessment and the subsequent linking to the results from the national mathematics assessment are provided in Chapter 10.

9.3.1 The Scaling Model

The scaling model used by NAEP in the Trial State Assessment is the three-parameter logistic (3PL) model from item response theory (IRT; e.g., Lord, 1980). This is a "latent variable" model, defined separately for each of the five subscales, and quantifying respondents' tendencies to provide correct answers to the items contributing to a subscale as a function of a parameter that is not directly observed, called proficiency on the subscale.

The fundamental equation of the 3PL model is the probability that a person whose proficiency on subscale k is characterized by the *unobservable* variable θ_k will respond correctly to item j:

$$\begin{aligned} P(x_j=1\,|\,\theta_k,a_j,b_j,c_j) &= c_j + (1-c_j)/\{1+\exp[-1.7a_j(\theta_k - b_j)]\} \\ &\equiv P_j(\theta_k), \qquad (9.1) \end{aligned}$$

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 $0 \le c_j < 1$, is the lower asymptote parameter of item j, reflecting the chances of a correct response from students of very low proficiency; c parameters are estimated for multiple-choice items, but are fixed at zero for open-ended items.

A typical assumption of item response theory is the conditional independence of the probabilities of correct 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 $\underline{x} = (x_1,...,x_n)$ across a set of n items is simply the product of terms based on (9.1):

$$P(\underline{x}|\theta_k,\underline{a},\underline{b},\underline{c}) = \prod_j^n [P_j(\theta_k)]^{x_j}[1 - P_j(\theta_k)]^{1-x_j} \qquad (9.2)$$

It is also typically assumed that response probabilities are conditionally independent of background variables ($\underline{y}$), given θ_k, or

$$P(\underline{x}|\theta_k,\underline{a},\underline{b},\underline{c},y) = p(\underline{x}|\theta_k,\underline{a},\underline{b},\underline{c}). \qquad (9.3)$$

After $\underline{x}$ has been observed, equation 9.2 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 with a modified version of Mislevy and Bock's (1982) BILOG computer program, then treated as known in subsequent calculations. The parameters of the items constituting each of the five subscales were estimated independently of the parameters of the other subscales. Once items have been calibrated in this manner, a likelihood function for the subscale 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.

As stated previously, item parameter estimation was performed independently for the Trial State Assessment and for the national mathematics assessment. In both cases, the identical subscale definitions were used. The national mathematics data also included responses of students in grade four to 109 mathematics items and responses of students in grade 12 to 144 mathematics items; where 45 items were common between grades 4 and 8 and 63 items were common between grades 8 and 12. The subscales for national mathematics extends across the three grades.

Conditional independence is a mathematical assumption, not a necessary fact of nature. Although the IRT models are employed in NAEP only to summarize average performance, a number of checks are made to detect serious violations of conditional independence, and, when warranted, remedial efforts are made to mitigate its effects on inferences. These checks include the following:

1) Checks on relative item operating characteristics among distinct gender and ethnicity groups (i.e., differential item functioning, [DIF] (Holland and Thayer, 1988)). Some degree of relative differences are to be expected, of course, and modestly varying profiles among groups will exist beyond the differences conveyed by their differing θ distributions. The intent of the check at this stage is to detect and eliminate items that operate differentially for identifiable reasons that are unrelated to the skills intended to be measured in the subject area.

2) When a subscale extends over age groups as is the case for the national mathematics subscales, evidence is sought of different operating characteristics over ages. When such effects are found, an item in question is represented by different item parameters in different age groups.

Item-level factor analyses have diminished in importance as our perspective of the role of IRT in NAEP has evolved. The assumption that performance in a scaling area is driven by a single unidimensional variable is unarguably incorrect in detail. However, our use of the model is not theoretical, instead it is data analytic; interpretation of results is not trait-referenced, but domain-referenced. Scaling areas are determined *a priori* by considerations of content as collections of items for which overall performance is deemed to be of interest. The IRT summary is not expected to capture all meaningful variation in item response data, but to reflect distributions of overall proficiency -- to summarize the main patterns in item percents-correct in the populations and subpopulations of interest. Using a unidimensional IRT model when the true model is multidimensional captures these overall patterns even though it over- or under-estimates the covariances among responses to items in pairs. For inferences based on overall proficiency, violations of the model with respect to dimensionality are less serious than violations in the shapes of the marginal response curves -- hence our greater attention to routine checks of item-fit residuals for every item in every calibration run than to factor analytic results.

In all NAEP IRT analyses, missing responses at the end of each block a student was administered were considered "not-reached," and treated as if they had not been presented to the respondent. Missing responses 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.

The local independence assumption embodied in equation 9.2 implies that item response probabilities depend only on θ and the specified item parameters--not on the position of the item in the booklet, on the content of items around an item of interest, or on test-administration timing conditions. These effects are certainly present in any application. The practical question is whether the IRT probabilities obtained via (9.2) are "close enough" to be robust with respect to the context in which the data are to be collected and the inferences that are to be drawn.

The experience with adaptive testing has shown using the same item parameters regardless of when an item is administered does not materially bias estimates of the proficiencies of individual examinees. Our experience with the 1986 NAEP reading anomaly, has shown, however, that for measuring small changes over time, changes in item context and speededness conditions 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, and 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.

9.3.2 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 100 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 approximate 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 subscale 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 subscale 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 3PL model is a latent variable model, 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:

$$
\begin{aligned}
t^*(\underline{x},\underline{y}) &= E[t(\underline{\theta},\underline{y}) \mid \underline{x},\underline{y}] \\
&= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta} \mid \underline{x},\underline{y})\, d\underline{\theta} \ . \qquad (9.4)
\end{aligned}
$$

It is possible to approximate t^* using random draws from the conditional distributions, $p(\underline{\theta}|x_i,y_i)$, of the subscale proficiencies given the item responses x_i and background variables y_i 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 conditional distribution $p(\underline{\theta}|x_i,y_i)$. 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 (9.4); 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 9.3 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 9.4, 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.

9.3.3 Computing Plausible Values in IRT-based Scales

Plausible values for each respondent i are drawn from the conditional distribution $p(\theta|x_i,y_i)$. 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

$$\begin{aligned} p(\underline{\theta}|x_i,y_i) &\propto P(x_i|\underline{\theta},y_i)\, p(\underline{\theta}|y_i) \\ &= P(x_i|\underline{\theta})\, p(\underline{\theta}|y_i), \qquad (9.5) \end{aligned}$$

where, for vector-valued $\underline{\theta}$, $P(x_i|\underline{\theta})$ is the product over subscales of the *independent likelihoods* induced by responses to items within each subscale, and $p(\underline{\theta}|y_i)$ is the multivariate--and generally nonindependent -- *joint density* of proficiencies for the subscales, conditional on the observed value y_i of background responses.

In the analyses of the data from the Trial State Assessment and the data from the national mathematics assessment, a normal (Gaussian) form was assumed for $p(\underline{\theta}|y_i)$, with a common dispersion and with a mean given by a linear model based on the first 90 - 95 principal components of 170 selected main-effects and two-way interactions of the complete vector of background variables. The included background variables will be referred to as the *conditioning variables*, and will be denoted y^c. (The conditioning variables used in the Trial State Assessment analyses are listed in Appendix C). The following model was fit to the data within each state:

$$\theta = \Gamma' y^c + \varepsilon \,, \tag{9.6}$$

where ε is normally distributed with mean zero and dispersion Σ. The number of principal components of the conditioning variables used for each state was sufficient to account for 90% 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 subscale and Σ is the matrix *variance of residuals* between subscales. By fitting the model (9.6) separately within each state, interactions between each state and the conditioning variables are automatically included in the conditional joint density of subscale proficiencies. Like item parameter estimates, the estimates of the parameters of conditional distributions were treated as known true values in subsequent steps of the analyses.

Maximum likelihood estimates of Γ and Σ were obtained with Sheehan's (1985) MGROUP computer program, using a variant of the EM solution described in Mislevy (1985). The difference from the published algorithm lies in the numerical approximation that was employed. Note from (9.5) that $p(\underline{\theta}|x_i,y_i)$ is proportional to the product of two terms, the likelihood $P(x_i|\underline{\theta})$ and the conditional distribution $p(\underline{\theta}|y_i)$. The conditional distribution for person i has been assumed multivariate normal, with mean $\mu_i^c = \Gamma' y_i^c$ and covariance matrix Σ; if the likelihood is approximated by another normal distribution, with mean μ_i^L and covariance matrix Σ_i^L, then the posterior $p(\underline{\theta}|x_i,y_i)$ is also multivariate normal with covariance matrix

$$\Sigma_i^p = (\Sigma^{-1} + (\Sigma_i^L)^{-1})^{-1} \tag{9.7}$$

and mean vector

$$\bar{\underline{\theta}}_i = (\underline{\theta}_i^c \Sigma^{-1} + \underline{\theta}_i^L (\Sigma_i^L)^{-1})(\Sigma_i^p)^{-1}. \tag{9.8}$$

In the analyses of the Trial State Assessment, a normal approximation for $P(x_i|\underline{\theta})$ is accomplished in a given scale by the steps described below. (Recall that by the assumed conditional independence across scales, the joint conditional likelihood for multiple scales is the product of independent likelihoods for each of the scales.) These computations are carried out in the scale determined by BILOG (Mislevy & Bock, 1982) item parameter estimates, where the mean and standard deviation of the composite population formed by combining the three NAEP grade/ages has mean zero and standard deviation one. The steps were as follows.

1) Lay out a grid of Q equally spaced points from -5 to +5, a range that covers the region in each scale where all examinees are virtually certain to occur. The value of Q varies from 20 to 40, depending on the subscale being used; smaller values suffice for subscales with few items given to each respondent, while larger values are required for subscales with many items.

2) At each point X_q, compute the likelihood $L(x_i | \theta = X_q)$.

3) To improve the normal approximation in those cases in which likelihoods are not approximately symmetric in the range of interest -- as when all of a respondent's answers are correct -- multiply the values from Step 2 by the mild smoothing function

$$S(X_q) = \frac{\exp(X_q+5)}{[1+\exp(X_q+5)][1+\exp(X_q-5)]}. \qquad (9.9)$$

This is equivalent to augmenting each examinee's response vector with responses to two fictitious items, one extraordinarily easy item that everyone gets right and one extraordinarily difficult item that everyone gets wrong. This expedient improves the normal approximation for examinees with flat or degenerate likelihoods in the range where their conditional distributions lie, but has negligible effects for examinees with even modestly well-determined symmetric likelihoods.

4) Compute the mean and standard deviation of θ using the weights $S(X_q)L(x_i | \theta = X_q)$ obtained in Step 3.

At this stage the likelihood induced by a respondent's answers to the items in a given scale is approximated by a normal distribution. Since the mathematics assessment uses five subscales, independent normal distributions, one per subscale, are used to summarize information from responses to items from the several subscales.

This normalized-likelihood/normal posterior approximation was then employed in both the estimation of Γ and Σ and in the generation of plausible values. From the final estimates of Γ and Σ, a respondent's posterior distribution was obtained from the normal approximation using the four-step procedure outlined above. A plausible value was drawn from this multivariate normal distribution. Finally, weighted-average composites over subscales were also calculated after appropriate rescaling.

9.4 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 the uncertainty, namely the sampling of respondents. Item percents correct 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.

9.4.1 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 $(\underline{\theta},\underline{y}) \equiv (\theta_1,y_1,...,\theta_N,y_N)$], 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_i,y_i) rather than $(\underline{\theta}_i,y_i)$, we can approximate t by its expected value conditional on $(\underline{x},\underline{y})$, or

$$
\begin{aligned}
t^{*}(\underline{x},\underline{y}) &= E[t(\underline{\theta},\underline{y})\,|\,\underline{x},\underline{y}] \\
&= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta}\,|\,\underline{x},\underline{y})\, d\underline{\theta}\ . \qquad (9.10)
\end{aligned}
$$

It is possible to approximate t^* with random draws from the conditional distributions $\underline{p}(\underline{\theta}_i\,|\,x_i,y_i)$, which are obtained for all respondents by the method described in section 9.3.3. Let $\underline{\hat{\theta}}_m$ be the m^th 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 $m=1,...,M$.

2) Using the jackknife variance estimator defined in Chapter 8, compute the estimated sampling variance of $\hat{t}_m$, denoting the result U_m.

3) The final estimate of t is

$$t^{*} = \sum_{m=1}^{M} \frac{\hat{t}_m}{M}. \qquad (9.11)$$

4) Compute the average sampling variance over the M sets of plausible values, to approximate uncertainty due to sampling respondents:

$$U^{*} = \sum_{m=1}^{M} \frac{U_m}{M}. \qquad (9.12)$$

5) Compute the variance among the M estimates $\hat{t}_m$, to approximate uncertainty due to not observing θ values from respondents:

$$B_M = \sum_{m=1}^{M} \frac{(\hat{t}_m - t^{*})^2}{(M - 1)} \qquad (9.13)$$

6) The final estimate of the variance of t^{*} is the sum of two components:

$$V = U^{*} + (1 + M^{-1})\, B_{M^{*}} \qquad (9.14)$$

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.

9.4.2 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

$$\nu = \frac{1}{\dfrac{f_M^2}{M - 1} + \dfrac{(1 - f_M)^2}{d}} \qquad (9.15)$$

where f_M is the proportion of total variance due to not observing θ values:

$$f_M = (1+M^{-1})\, B_M / V_M \ . \qquad (9.16)$$

When B_M 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_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity

$$(T-t^*)\, V^{-1}\, (T-t^*)' \qquad (9.17)$$

is approximately F distributed, with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of f_M:

$$f_M = (1+M^{-1})\ \mathrm{Trace}\,(B_M V_M^{-1})/k. \qquad (9.18)$$

By the same reasoning as used for the normal approximation for scalar t, a chi-square distribution on k degrees of freedom often suffices.

9.4.3 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 (19901), 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 conditioning 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, indicate 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 170 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 (1991) shows that this puts an upper bound of 10 percent on the potential bias for all analyses involving the original conditioning variables.

9.5 SCALE ANCHORING[1]

Scale anchoring is a method for attaching meaning to a 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. In contrast, the NAEP scale anchoring is accomplished by describing what students at selected levels know and can do. This is the primary purpose of NAEP.

The anchoring process was performed on the national NAEP mathematics composite as follows. Composite plausible values for each student (in grades 4, 8, and 12 and/or for age 9, 13, and 17) who participated in the national mathematics assessment were created as a weighted average of the subscale plausible values, where each set of plausible values for a particular subscale was linearly adjusted to have a mean of 250.5 and a standard deviation of 50. The scale levels 200, 250, 300 and 350 on the 500 point scale were selected. These values (roughly standard deviation units apart) are far enough apart to be noticeably different but not so far apart as to be trivial.

The students are sorted by their plausible values, and students with a plausible value at or near each level (i.e. within 12.5 points) are grouped together. For the group at the lowest

[1] Appendix F contains a more detailed description of the scaling anchoring process.

scale score level, what they know and can do is defined by the items that at least 65% of the students answered correctly. At a higher score level, the question is: what is it that students at this level know and can do that students at the next lower level cannot. The answer is defined by the items that at least 65% of students at this level answered correctly, but a majority (at least 50%) at the next lower level answered incorrectly. Finally, the difference between the probabilities of success between the two levels must be at least 30 percentage points. The assessment items are, therefore, grouped by the levels between which they discriminate. It is important to note that the overall percentage of students who correctly answer an anchor item is not equal to the percentage scoring above that scale level.

Table 9-1 demonstrates the statistical anchoring process. Three items are displayed, identified by the labels "A", "B", and "C". Four anchoring levels are identified, corresponding to scale values of 200, 250, 300 and 350. In the table, Item "A" anchors at the 250 level since the probability of correct response for students with proficiencies around 250 is 80 percent while the probability of success for students at the next lower level (200) is 40 percent. Item "B" anchors at the 300 level since there is a steep rise in the probability of success between 250 and 300 and since the probabilities of success at the two levels satisfy the threshold values. Item "C" does not anchor at any of the four levels because the discrimination between adjacent levels is not sufficiently sharp. Of the 275 unique items in the 1990 national mathematics assessment, 143 (52 percent) satisfied the anchoring criteria with an additional 53 (19 percent) nearly satisfying the criteria.

Table 9-1

Three Example Items for Scale Anchoring

	Scale Values			
Item	200	250	300	350
A	40%*	80%	87%	92%
B	20%	23%	68%	84%
C	30%	56%	81%	87%

*percentages of students scoring at or near the scale value who responded correctly to the item

Following the determination of the anchor levels, a committee of mathematics experts, educators, and others was assembled to review the items and, using their knowledge of mathematics and student performance, to generalize from the items to more general constructs. To derive the descriptions of the four scale anchor points, the 19 panelists first worked in two independent groups and then as a whole. Although the two sets of descriptions did not differ substantively, the group felt that the cross-validation procedure was valuable. As a final step, the reconciled version was sent to all panelists for review.

Chapter 10

DATA ANALYSIS AND SCALING

John Mazzeo

Educational Testing Service

10.1 OVERVIEW

This chapter describes specific details of the analyses carried out in developing the Trial State Assessment content area scales and composite scale. The philosophical and theoretical underpinnings of the NAEP scaling procedures were described in the previous chapter.

There were five major steps in the analysis of the Trial State Assessment data:

- Conventional item and test analyses
- Item response theory (IRT) scaling
- Estimation of state and subgroup proficiency distributions based on the "plausible values" methodology[1]
- Linking of the Trial State Assessment content area scales to the corresponding scales from the 1990 national assessment
- Creation of the Trial State Assessment mathematics composite scale

Analysis details for each of the five steps are described in separate sections. 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.

10.2 DESCRIPTION OF ITEMS, ASSESSMENT BOOKLETS, AND ADMINISTRATION PROCEDURES

Each of the 137 mathematics items administered in the Trial State Assessment Program was categorized into one of five content areas: 1) Numbers and Operations (46 items), 2) Measurement (21 items), 3) Geometry (26 items), 4) Data Analysis, Statistics, and Probability (19 items), and, 5) Algebra and Functions (25 items). These 137 items, consisting of 102

[1] The word "state" is used in this chapter to refer to any of the 40 jurisdictions that participated in the Trial State Assessment Program, even though three of the jurisdictions -- the District of Columbia, Guam, and the Virgin Islands -- are not states.

multiple-choice and 35 open-ended items, were divided into seven mutually-exclusive blocks of items. The composition of each block of items, in terms of content and format, is given in Table 10-1. One or more open-ended items were included in six of the seven blocks, and one block of items consisted completely of open-ended items. Each of the open-ended items were scored by specially trained readers, as described in Chapter 6.

Included among the items were eight items that required a calculator for their solution (referred to as "calculator-active" items). The calculator-active items appeared in two of the seven blocks (three items in Block MH, five items in Block MI), and students assigned these blocks were provided a Texas Instruments TI-30 Challenger calculator for the 15-minute period(s) during which they worked on them. For each item in these calculator blocks, students were asked to indicate whether or not they used the calculator to answer each item. One of the seven blocks contained five items which required the use of a protractor/ruler for their solution. Students assigned this block were provided a protractor/ruler for the 15-minute period they worked on that block.

As described in Chapter 2, the seven blocks of items were used to create seven different assessment booklets according to a focused Balanced-Incomplete-Block (BIB) design. Each booklet contained three blocks of mathematics items, and each block of items appeared in exactly three booklets. To balance possible block position main effects, each block appeared once as the first block of mathematics items, once as the second block, and once as the third block. In addition, the BIB design required that each block of items be paired in a booklet with every other block of items exactly once.

Within each administration site, assessment booklets were spiraled together in a random sequence and distributed to students sequentially (e.g., in the order of the student's seating within the class). As a result of the BIB design and the "spiraling" of booklets, a considerable degree of balancing was achieved for 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 3/7 of the total sample size) within each state and across all states. In addition, within each state, and across all states, randomly equivalent samples of approximately equal size received each particular block of items as the first, second, or third block within a booklet.

As described in Chapter 5, a random half of the administration sessions within each state were observed by Westat trained Quality Control Monitors. Thus, within each state, and across all states, randomly equivalent samples of students received each block of items in a particular position within a booklet under monitored and unmonitored administration conditions. Equivalently, for both the monitored and unmonitored sessions within each state (and, therefore, across all states), randomly equivalent samples of students were administered each block of items in each of the three possible serial positions within a booklet.

Table 10-1
Format and Content Description of the Blocks of Exercises
Administered In the Trial State Assessment

		# of Exercises for each Item format		# of Exercises In each content area:				
Block	Total # of Exercises	# of Multiple-Choice Items	# of Constructed Response Items	# of Numbers & Operations Exercises	# of Measurement Exercises	# of Geometry Exercises	# of Data Analysis Statistics and Probability Exercises	# of Algebra & Functions Exercises
MC	23	19	4	9	4	3	4	3
MD	21	21	0	7	4	4	2	4
ME	16	0	16	3	1	6	3	3
MF[p]	21	16	5	7	5	3	3	3
MG	18	17	1	3	3	5	3	4
MH[c]	18	16	2	8	2	2	3	3
MI[c]	20	13	7	9	2	3	1	5
Total	137	102	35	46	21	26	19	25

[p] Students were provided a protractor to use in answering items in this block.

[c] Students were provided a calculator to use in answering items in this block.

10.3 CONVENTIONAL ITEM AND TEST ANALYSES

Table 10-2 contains summary statistics for each block of items. Block level statistics are provided both overall and by serial position of the block within booklet. All statistics were calculated using the final sampling weights provided by Westat. Table 10-2 shows the number of students assigned each block of items, the average proportion correct, the average biserial correlation, and the proportion of students attempting the last item in the block.

The average proportion correct for the block is the average, over items, of the proportion of students who correctly answered each item. In all 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 to which a student provided an answer) and missing responses prior to the last observed response. In the former case, the item is treated as having not been presented to the student. The latter type of missing response is treated as an intentional omission. In calculating the proportion correct for each item, students classified as having not been presented the item were excluded from the calculation of the statistic while students classified as intentionally omitting the item were treated as answering incorrectly.

The average biserial correlation is the average, over items, of the item-level biserial correlations (or, r-biserial). For each item-level r-biserial, total block number-correct score (including the item in question, and with students receiving zero points for all not presented items) was used as the criterion variable for the correlation, and students classified as having not been presented the item were omitted from the calculation of the statistic. The proportion of students attempting the last item is one minus the proportion of students that were classified as having not been presented the final item in the block. This proportion is often used as an index of the degree of speededness associated with the administration of a block of items.

As evident from Table 10-2, the difficulty and internal consistency of the blocks varied somewhat. Such variability was expected since the blocks were not created to be parallel in difficulty or content. Based on the proportion of students attempting the last item, the two blocks containing the calculator-active items appear to have been somewhat speeded for this group of examinees. The last item in Block MH was attempted by 70 percent of those administered that block; the last item of Block MI was attempted by 58 percent of those administered that block.

The data in Table 10-2 also indicate that there was little variability in the average proportion corrects or average biserial correlations for each block by serial position within the assessment booklet. This suggests that serial position within booklet had a small effect on the overall difficulty of the block, at least in terms of the proportion of attempted items which were answered correctly. However, one aspect of block level performance which did differ by serial position was the proportion of students attempting the last item in the block. For all seven blocks, the proportion attempting the last item was lowest when the block appeared in the first

Table 10-2

Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall.

	POS	MC	MD	ME	MF	MG	MH	MI
Unweighted sample size	1	14,403	14,384	14,336	14,380	14,389	14,416	14,399
	2	14,396	14,396	14,371	14,328	14,371	14,382	14,404
	3	14,371	14,383	14,378	14,382	14,360	14,310	14,348
	All	43,170	43,163	43,085	43,090	43,120	43,108	43,151
Average proportion correct	1	.68	.56	.53	.68	.41	.48	.51
	2	.67	.57	.54	.67	.41	.48	.50
	3	.67	.55	.53	.66	.41	.47	.50
	All	.67	.56	.53	.67	.41	.48	.50
Average r-biserial	1	.58	.53	.67	.66	.59	.58	.58
	2	.57	.53	.68	.65	.59	.58	.58
	3	.57	.51	.68	.66	.59	.58	.57
	All	.57	.52	.68	.66	.59	.58	.58
Proportion of students attempting last item	1	.91	.93	.89	.84	.96	.66	.52
	2	.91	.93	.92	.85	.97	.73	.63
	3	.91	.95	.94	.88	.97	.72	.58
	All	.91	.93	.92	.85	.97	.70	.58

[1] Includes Indivuals who reponded to at least 1 item in the block

position[2]. For all but the two calculator blocks, the proportion attempting the last item was highest when the block appeared in the last position.

An additional interesting result is that the proportion of examinees attempting the last item in Block MI (one of the calculator blocks) was largest when that block appeared as the second block of mathematics items within a booklet. The booklet in which this occurs contains Block MH (also a calculator block) in the first position.

As mentioned earlier, in an attempt to maintain rigorous standardized administration procedures across the states, a random half sample of all sessions within each state was observed by a Westat-trained Quality Control Monitor. Overall, there was little difference in the performance of students who attended monitored sessions and the performance of students who attended sessions that were unmonitored. The overall proportion correct (averaged over all seven blocks and over all 40 participating jurisdictions) for students from monitored sessions was .55 while the corresponding figure for students from unmonitored sessions was .54.

For each block of items, Table 10-3 provides the average proportion correct, average r-biserial, and the proportion of students attempting the last item for students who sessions were monitored and students whose sessions were not monitored. One notable feature in Table 10-3 is that for five of the seven blocks, the proportion of students attempting the last item in a block was higher for the students in unmonitored sessions than the corresponding proportion of students from monitored sessions. The largest difference between the two types of sessions occurred for the two calculator blocks (Blocks MH and MI). However, the higher proportion of students from unmonitored sessions attempting the last item in each block did not result in higher average proportion corrects for the items they attempted. In fact, the average proportion correct for items each group attempted differed only slightly, with students from monitored sessions performing, on average, about one percent higher than students in unmonitored sessions.

[2] The differences for some of the blocks are not evident in Table 12-2 since the results are rounded to the nearest integer.

Table 10-3

Block-level Descriptive Statistics for Monitored and Unmonitored Sessions

	MC	MD	ME	MF	MG	MH	MI
Unwtd sample size							
unmoniterd	22,006	21,978	21,957	21,937	21,895	21,954	21,905
monitored	21,166	21,185	21,128	21,153	21,225	21,154	21,246
Average proportion correct							
unmonitored	.67	.56	.53	.67	.41	.47	.50
monitored	.68	.56	.53	.68	.41	.48	.51
Average r-biserial							
unmonitored	.57	.52	.68	.66	.58	.58	.58
monitored	.57	.52	.68	.65	.59	.58	.58
Proportion of students attempting last item							
unmonitored	.99	.94	.92	.86	.97	.72	.60
monitored	.99	.93	.92	.84	.96	.68	.55

Figure 10-1 presents a stem-and-leaf display of the differences in average proportion correct (over all seven blocks) for students from monitored sessions and from unmonitored sessions for all 40 jurisdictions[3]. The median difference (monitored minus unmonitored) was .006. For fifteen jurisdictions, the difference was negative (i.e., the average proportion correct for students from unmonitored sessions was higher than that of students from monitored sessions), with the largest difference being -.02. For the remaining twenty five jurisdictions, the difference was at least zero (i.e., the average proportion correct for students from monitored sessions at least that of students from unmonitored sessions), with the largest difference being .03. For 34 of the 40 participants (88%), the absolute difference between average proportion corrects for students from monitored and unmonitored sessions was less than .02.

10.4 ITEM RESPONSE THEORY (IRT) SCALING

IRT-based content area scales were developed, using the 3-parameter logistic (3PL) model described in the previous chapter, by separately calibrating the sets of items in each of the five content areas. Item parameter estimates on a provisional scale were obtained using a modified version of the BILOG program (Mislevy & Bock, 1982). The BILOG item calibrations were based on the data from a systematic random sample of about 25% of the students who participated in the Trial State Assessment. This sample of students (650 students from each of the 40 participating jurisdictions) will be referred to as the "calibration sample".

Figures 10-2 through 10-6 contain stem-and-leaf displays of the item proportion corrects (over all 40 participating jurisdictions) for the collections of items comprising each of the five content area scales. The proportion corrects are based on students in the "calibration sample", and were calculated using the final sampling weights. On average, students found the set of items comprising the Measurement scale and Numbers and Operation scale to be easier than the other three scales.

10.4.1 BILOG Scaling

The Trial State Assessment analysis plans called for a single set of item parameters to be estimated for each item. This common set of item parameters was to be used for obtaining the scaled score results for all 40 states. Several factors contributed to the decision to use a single set of item parameters for all states. One factor was the desire to ensure equity for all participants and maintain an equal measure for establishing comparisons among participating jurisdictions.

In addition to equity considerations, there were compelling practical reasons to use a single common estimate of each item characteristic curve (ICC) rather than estimate separate ICC's for each state.[4] Since the sample size for an individual state was considerably smaller than that for the entire collection, and (because of the BIB design) only three-sevenths of the

[3] Westat produced a special set of sampling weights to be used in comparing performance from monitored and unmonitored sessions. Unless otherwise noted, all analyses involving comparisons between the monitored and unmonitored sessions were conducted using the weights provided for that purpose.

[4] An item characteristic curve relates the expected probability of success on an item to scale score level.

Figure 10-1

Stem and Leaf Display of State Differences Between Monitored and Unmonitored Sessions in Average Proportion-Correct Statistics (Monitored minus Unmonitored)

```
-0 : 2
-0 : 11111111111000
 0 : 00000111111111111111
 0 : 22223
```

N = 40   Median = 0.0055
Quartiles = -0.005, 0.0125

Decimal point is 1 place to the left of the colon

Figure 10-2

Stem and Leaf Display of Item Proportion-Correct Statistics for the Trial State Assessment Calibration Sample for the Numbers and Operations Scale

```
1 : 488
2 : 37
3 : 5677
4 : 003358899
5 : 23357
6 : 112679
7 : 2455889
8 : 0123348
9 : 226
```

Number of items = 46   Mean = 0.59   Median = 0.59
Quartiles = 0.43, 0.78

Decimal point is 1 place to the left of the colon

Figure 10-3

Stem and Leaf Display of Item Proportion-Correct Statistics for the Trial State Assessment Calibration Sample for the Measurement Scale

```
1 : 8
2 : 01
3 : 2
4 : 59
5 : 6889
6 : 045
7 : 1
8 : 03458
9 : 12
```

Number of items = 21 Mean = 0.61 Median = 0.60
Quartiles = 0.49, 0.83

Decimal point is 1 place to the left of the colon

Figure 10-4

Stem and Leaf Display of Item Proportion-Correct Statistics the Trial State Assessment Calibration Sample for the Geometry Scale

```
2 : 05679
3 : 669
4 : 0238
5 : 459
6 : 0223689
7 : 049
8 : 1
```

Number of items = 26 Mean=.51 Median = 0.545
Quartiles = 0.36, 0.66

Decimal point is 1 place to the left of the colon

Figure 10-5

Stem and Leaf Display of Item Proportion-Correct Statistics for the Trial State Assessment Calibration Sample for the Data Analysis, Probability, Statistics Scale

```
1 : 1279
2 :
3 : 34
4 : 168
5 : 3
6 : 05
7 : 2256
8 : 35
9 : 0
```

Number of items = 19 Mean=0.52 Median = 0.53
Quartiles = 0.33, 0.75

Decimal point is 1 place to the left of the colon

Figure 10-6

Stem and Leaf Display of Item Proportion-Correct Statistics for the Trial State Assessment Calibration Sample for the Algebra and Functions Scale

```
1 : 568
2 : 5
3 : 23346
4 : 1458
5 : 014
6 : 67
7 : 4899
8 :
9 : 155
```

Number of items = 25 Mean=0.52 Median = 0.48
Quartiles = 0.33, 0.74

Decimal point is 1 place to the left of the colon

sampled students attempted any particular item, state-specific item parameters could not be as precisely estimated as those estimated using data from all participants.

The decision to use a single common ICC estimate for all 40 participants added complexities to the IRT scaling process. Because of differences in curricular emphases across the states, it was reasonable to expect that ICCs might differ across the states for at least some items. Whatever procedure was to be used for obtaining the estimates of the common ICCs needed to be as fair as possible for all of the participants. The procedure to be used needed to avoid building in biases that would result in parameter estimates that were more appropriate for certain types of states and less appropriate for other types of states.

Work by Yamamoto and Muraki (1990) indicated that it was desirable to use sampling weights in estimating the Trial State Assessment item parameters. As described in previous chapters, final sampling weights (adjusted for both school-level and student-level nonresponse) were developed by Westat and used in the computer programs designed to derive the IRT estimates. The sum of these weights for each of the states was proportional to the number of eighth-grade public-school students in that state. Thus, direct use of these weights would have resulted in item parameter estimates that best fit the data from the largest states. Such estimates could be considered "unfair" for those states with smaller sample sizes and with ICCs which differed from those of the largest states.

To obtain a result that used sampling weights (and, therefore, correctly reflected the demographic composition within each state) and was equitable for each state, the Westat weights were rescaled prior to calibration so that the sum of the weights for each state was equal. Using these rescaled weights during item calibration, if ICC's differed by state, the fit of the common ICC to the data from an individual state would not depend on the state's size.

Further restrictions on the rescaled weights were required, however, because of the presence of data from both monitored and unmonitored sessions. As described earlier, a random half sample of the students were assessed in sessions attended by a Westat-trained Quality Control Monitor. Although the items administered in the monitored and unmonitored sessions were identical, these types of assessment sessions represented two slightly different measurement situations. It was possible that for a given item, the ICC under monitored administrations conditions and the ICC under unmonitored administration conditions might not be identical.

Consideration was given to performing separate BILOG calibrations for the two types of administrations and equating the scales of the two calibration runs. However, several factors argued against such an approach. First, based on the results reported in the previous section, there were not substantial differences in the item-level performance of students from the two types of sessions in terms of the proportion of attempted items which were correctly answered and the average biserial correlations. Figures 10-7 through 10-11 contain plots of item-level proportion-correct statistics based on calibration data from the unmonitored sessions (horizontal) and from the monitored sessions (vertical). As noted in the previous section, the average proportion correct for all items was slightly lower for students from the unmonitored sessions. However, the relative difficulty of items within a scale was nearly identical across the two types

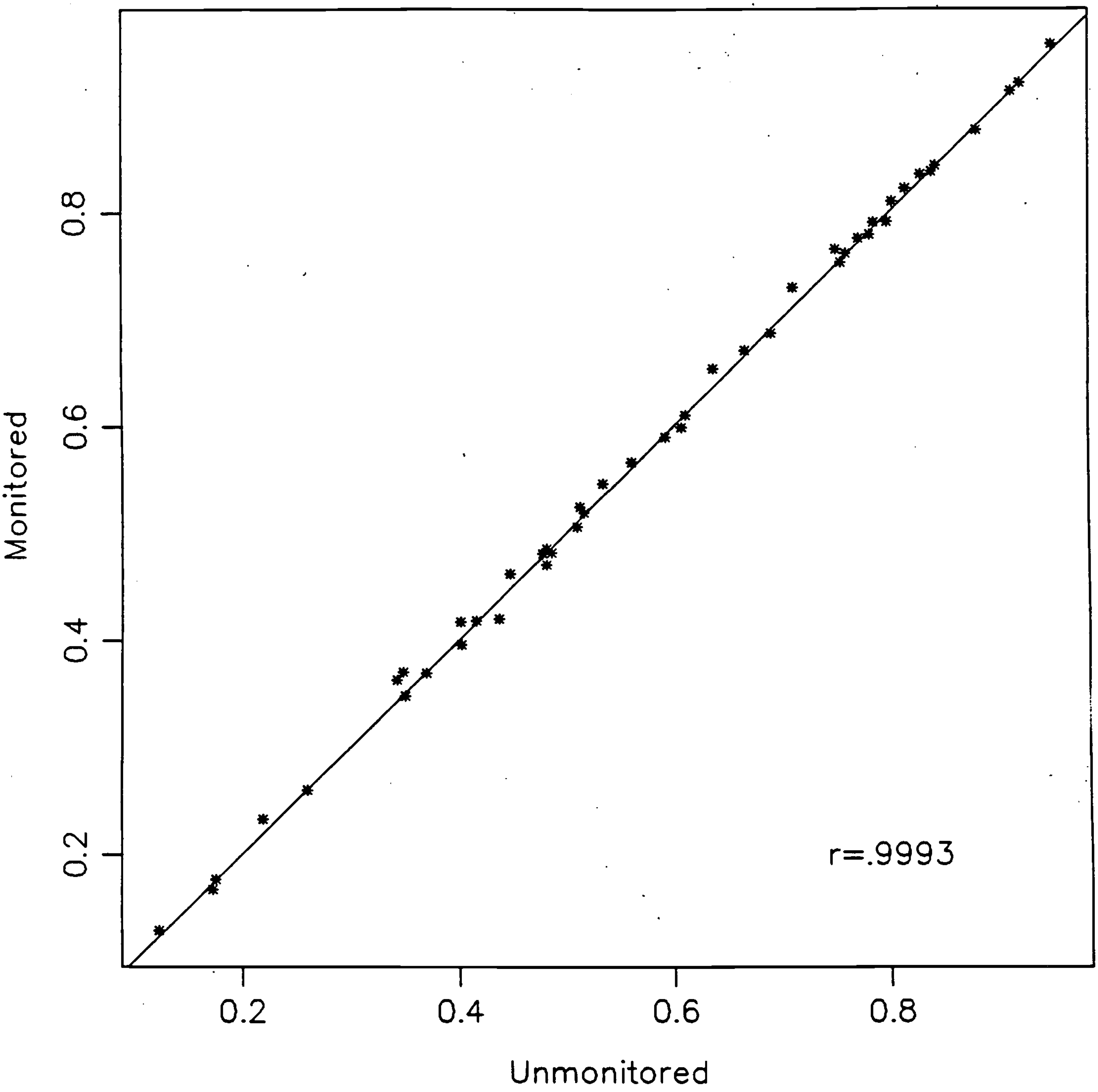


Figure 10-7

Proportion-correct Statistics for
Numbers and Operations Items
for Monitored and Unmonitored Sessions

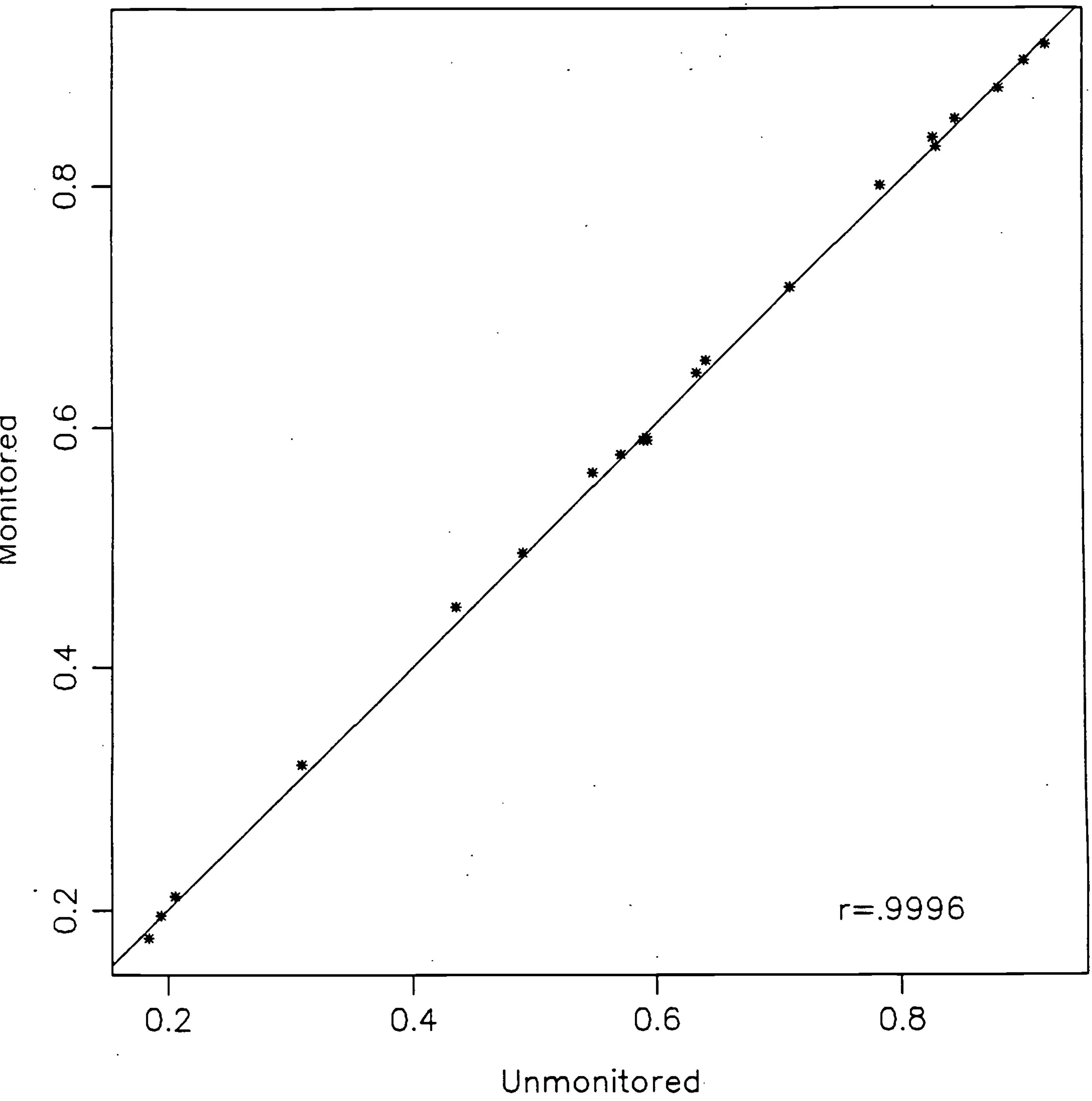


Figure 10-8

Proportion-correct Statistics for Measurement Items for Monitored and Unmonitored Sessions

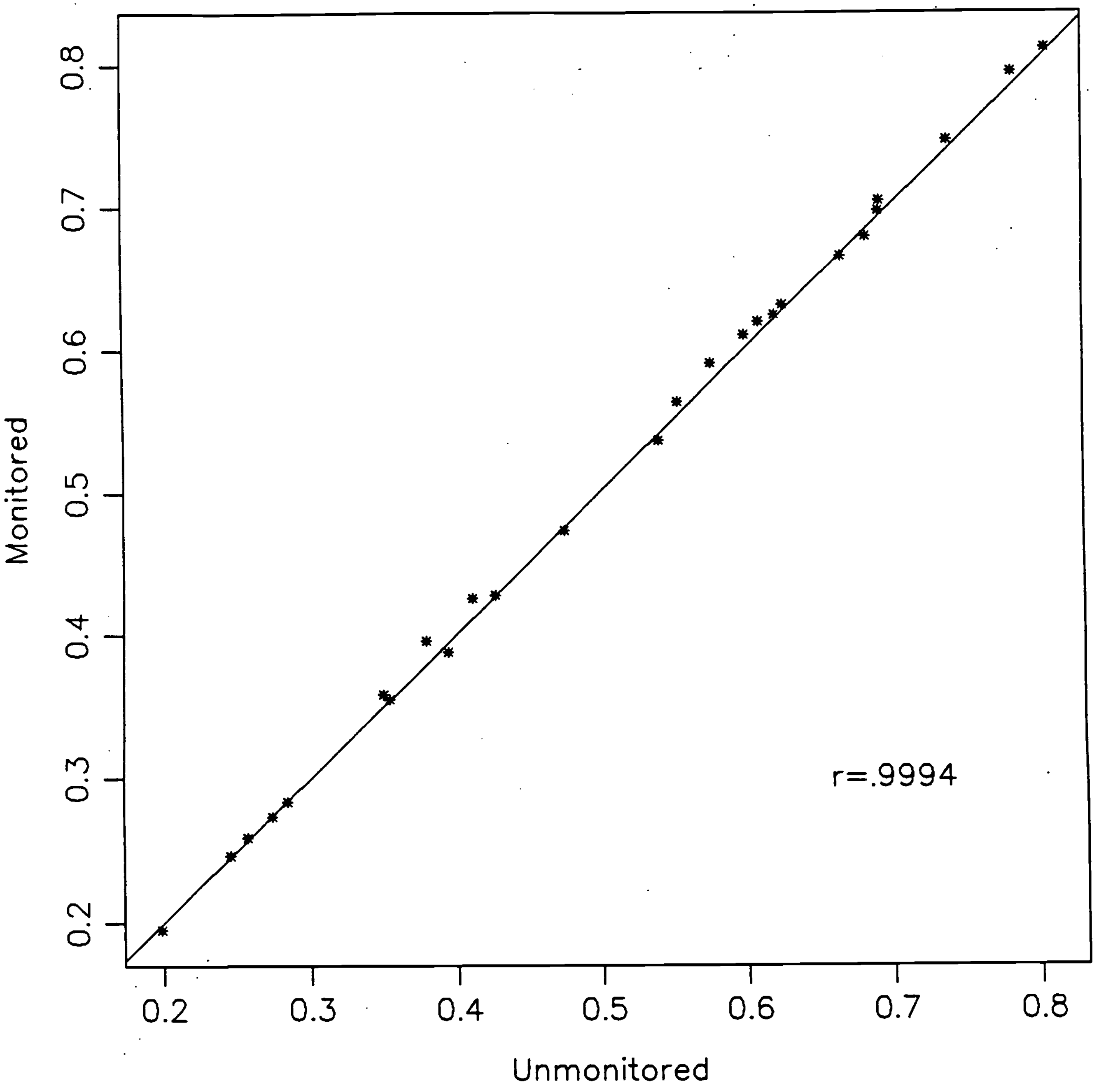


Figure 10-9

Proportion–correct Statistics for
Geometry Items
for Monitored and Unmonitored Sessions

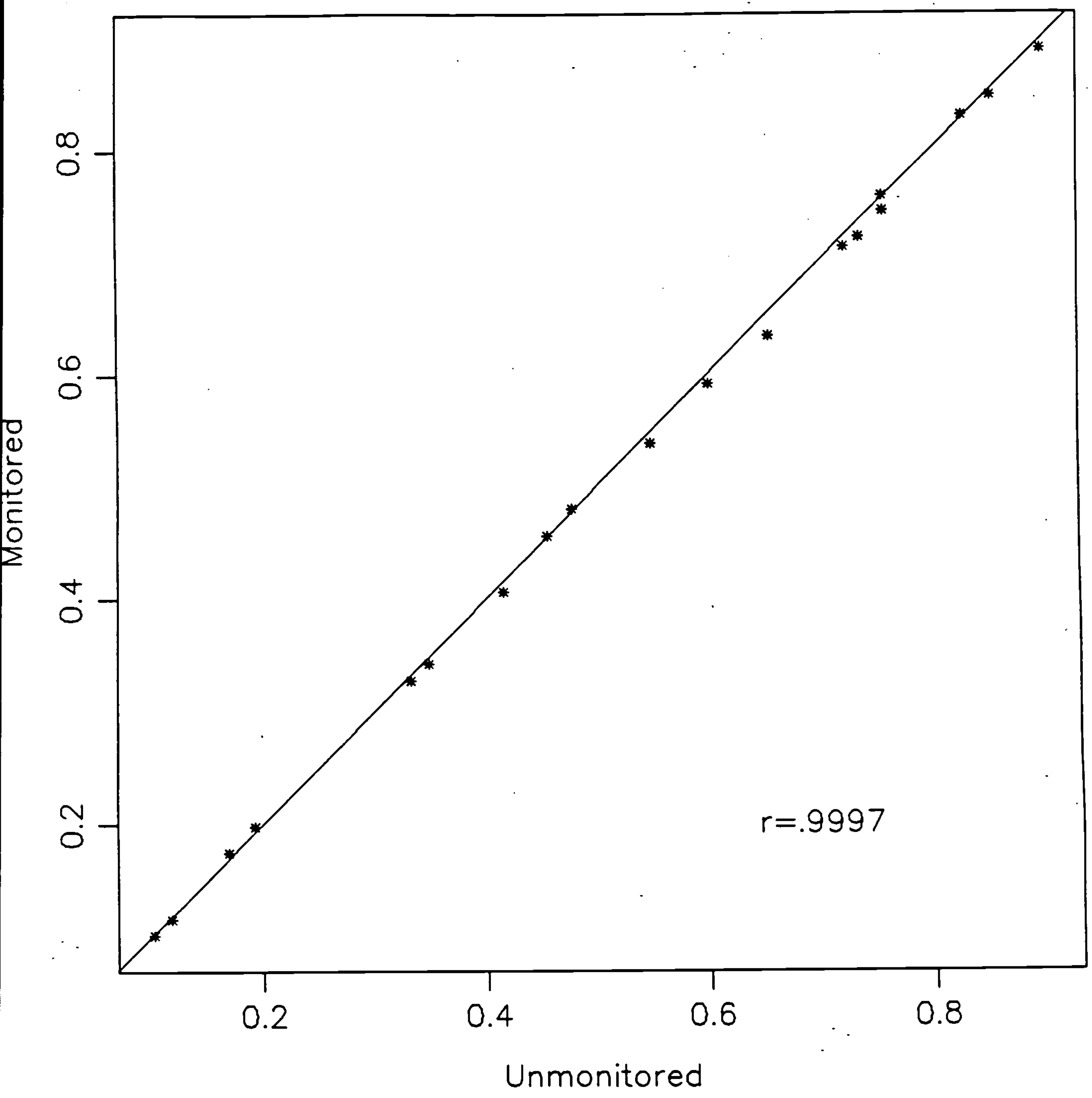


Figure 10-10

Proportion-correct Statistics for
Data Analysis, Probability, and Statistics Items
For Monitored and Unmonitored Sessions

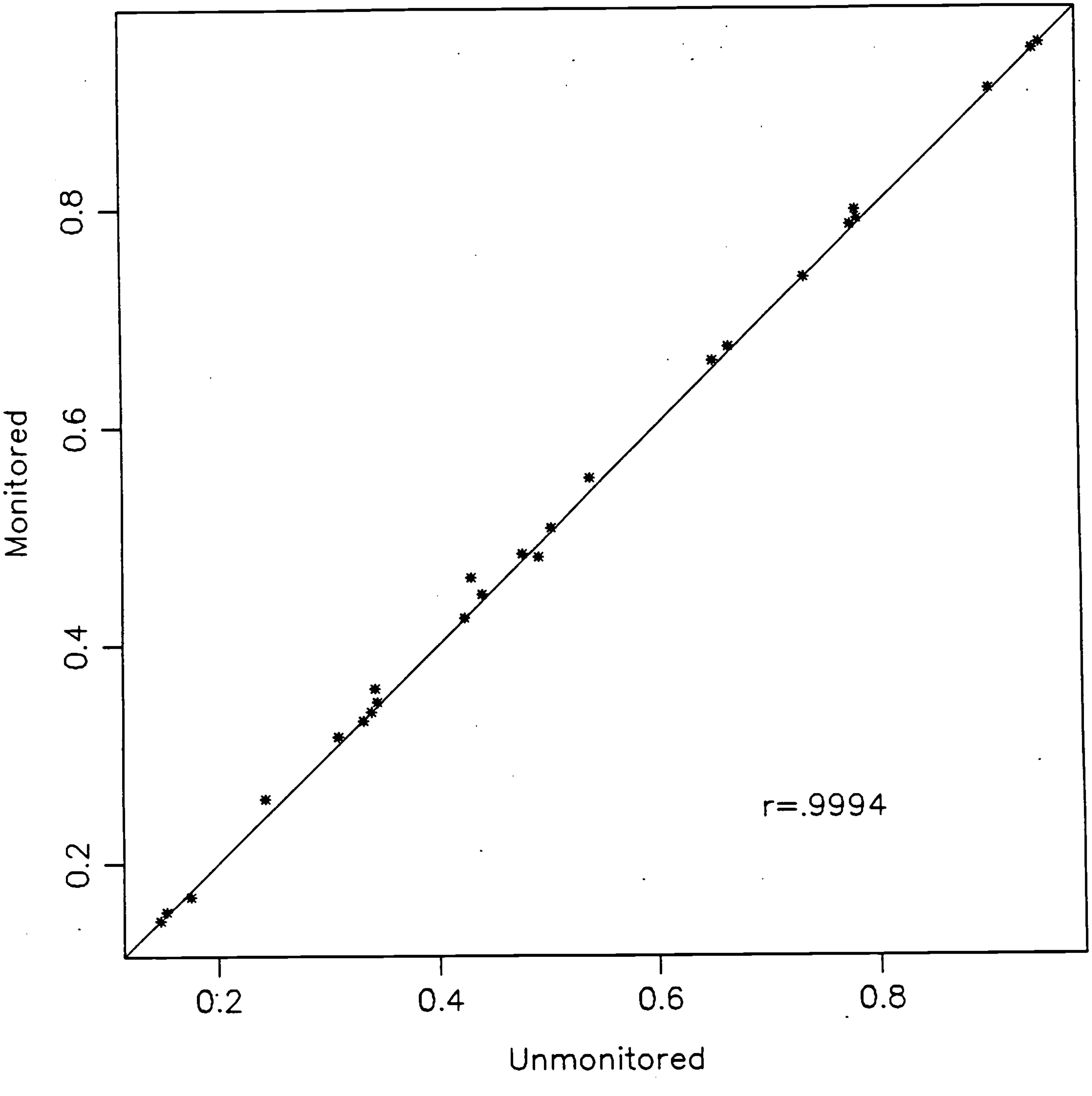


Figure 10-11

Proportion-correct Statistics for
Algebra and Functions Items
for Monitored and Unmonitored Sessions

of sessions. While similarity of item difficulties does not strictly imply that ICCs for the two types of sessions are similar, it is consistent with such a state of affairs.

A second reason for avoiding separate ICC estimates for the two types of sessions was a practical one. If separate ICC estimates were obtained, the scales for the two sets of item parameters would need to be equated. This additional equating would add an undesirable additional level of complexity to an already complex statistical and data processing system. Last, and perhaps most importantly, since random half samples within each state were administered the assessment under monitored or unmonitored conditions, it was assumed that the effects of any differences in item parameters by type of session would have a relatively small impact on comparisons across states, and across major subgroups within a state. Based on these considerations, a decision was made to use results from the monitored and unmonitored sessions for the calibration and that data from each would be equally represented.

To obtain a single set of item parameters in which 1)sampling weights were used to reflect the demographic composition within each state, 2)each state's data contributed equally to the estimation process, and 3)data from monitored and unmonitored sessions contributed equally, the final sampling weights were rescaled only for item parameter estimation.

The sampling for item calibration and the rescaling of weights included the following:

- Samples of 650 records were drawn for each state. 325 records were drawn from the monitored sessions and 325 from the unmonitored sessions using systematic sampling. This resulted in a total sample of 26,000 records.

- For each state, the sum of the Westat sampling weights for the set of monitored and unmonitored records selected for the sample was obtained (these sums are denoted as WM_s and WU_s, respectively).

- For each state, the Westat weights for the individuals in the sample (denoted as w_{si}) were rescaled so the sum of the weights for the monitored and unmonitored sessions would each be equal to 325. Thus, for the monitored students in the sample,

$$w^{*}_{si} = w_{si} (325/WM_s),$$

and for the unmonitored students,

$$w^{*}_{si} = w_{si} (325/WU_s),$$

where w^{*}_{si} denotes the rescaled weight for individual i from state s.

Figures 10-12 through 10-16 contain scatterplots of item-level proportion corrects for the sets of items comprising up each of the five scales (one for each content area). The proportion corrects plotted on the horizontal axis were determined using the calibration sample and the rescaled calibration weights while those plotted on the vertical axis were determined on the same sample of students, but using the final sampling weights. As apparent from these figures,

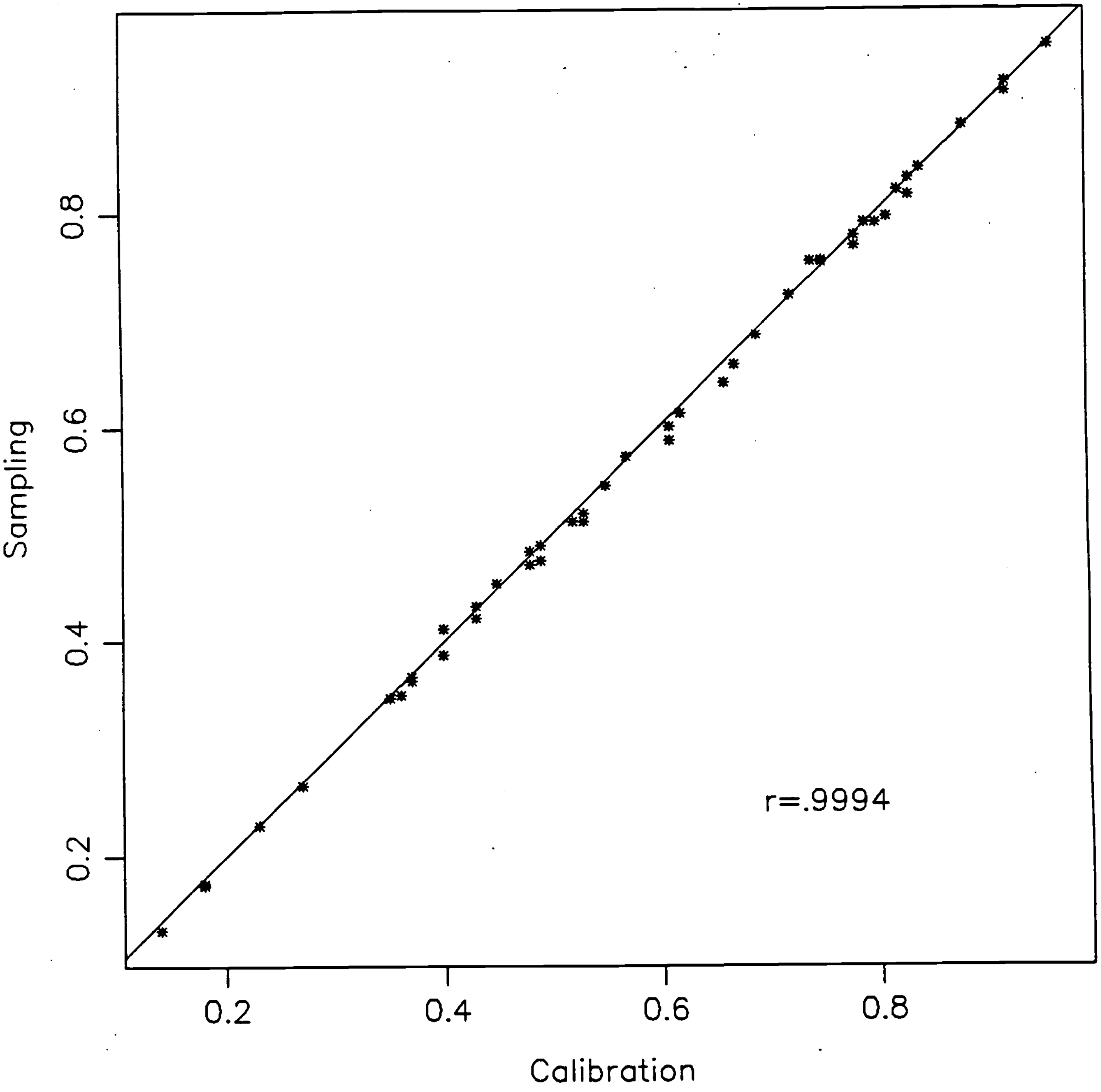


Figure 10-12

Proportion-correct Statistics for
Numbers and Operations Items
Using Calibration and Sampling Weights

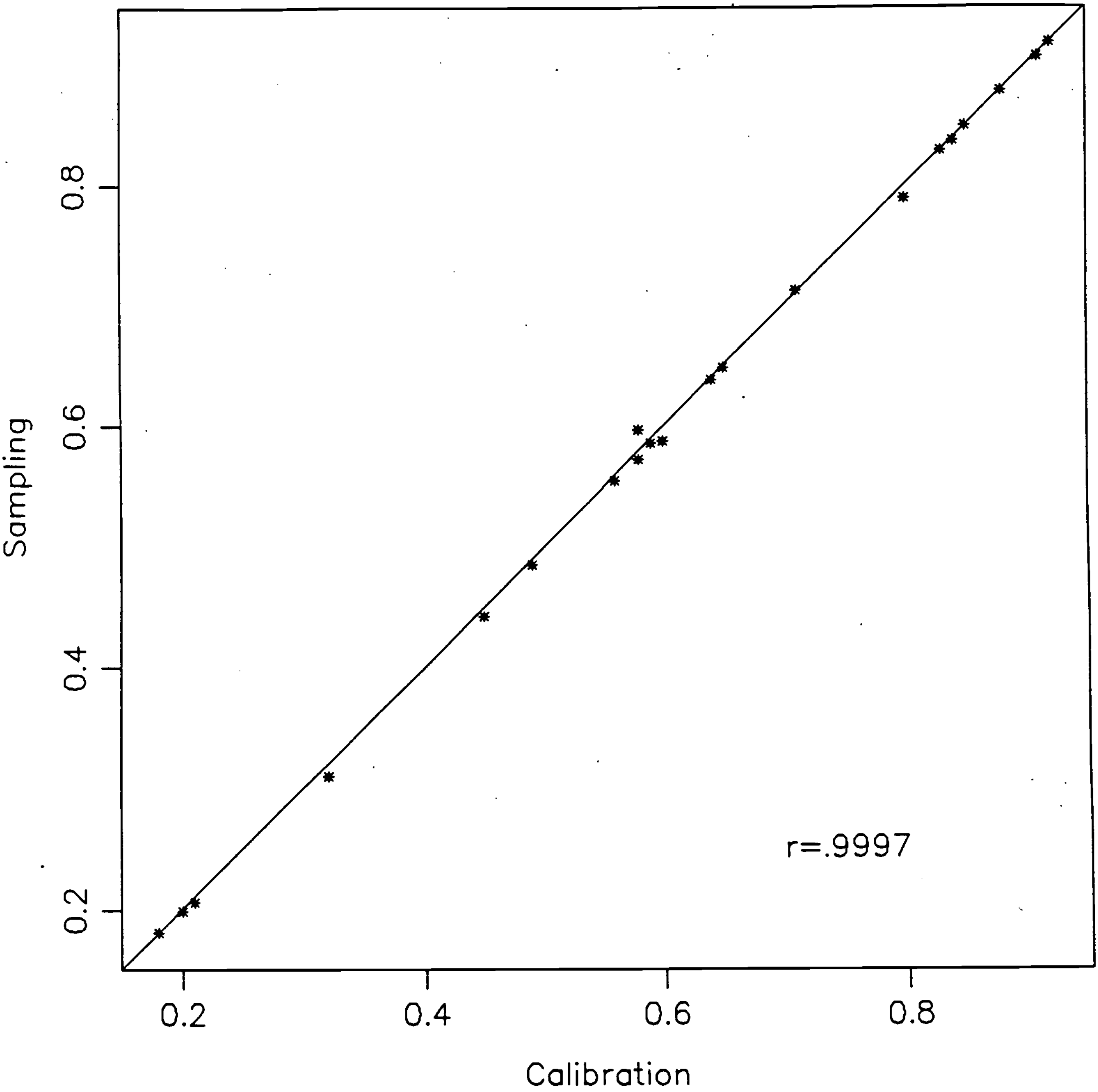


Figure 10-13

Proportion—correct Statistics for
Measurement Items
Using Calibration and Sampling Weights

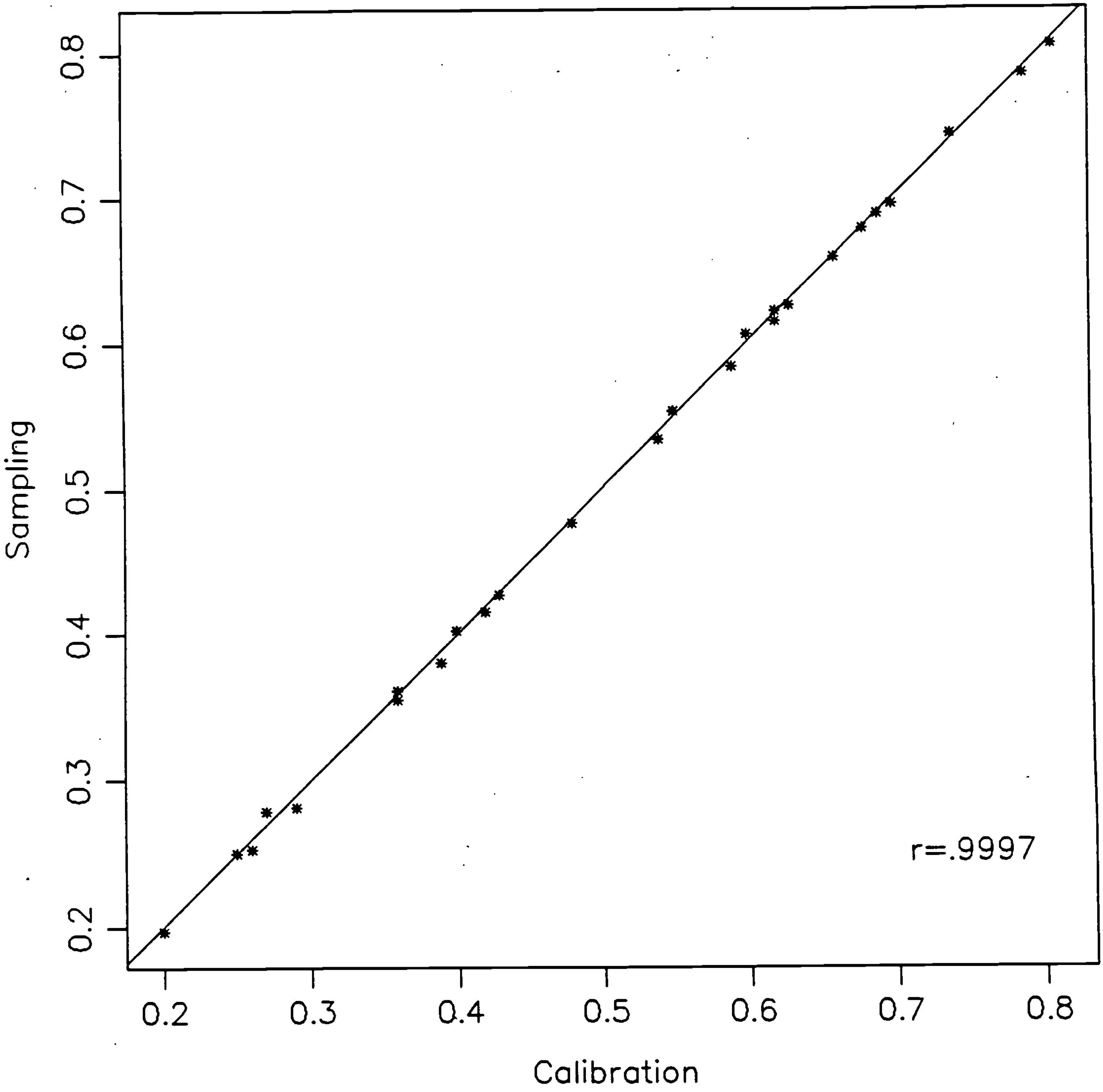


Figure 10-14

Proportion–correct Statistics for
Geometry Items
Using Calibration and Sampling Weights

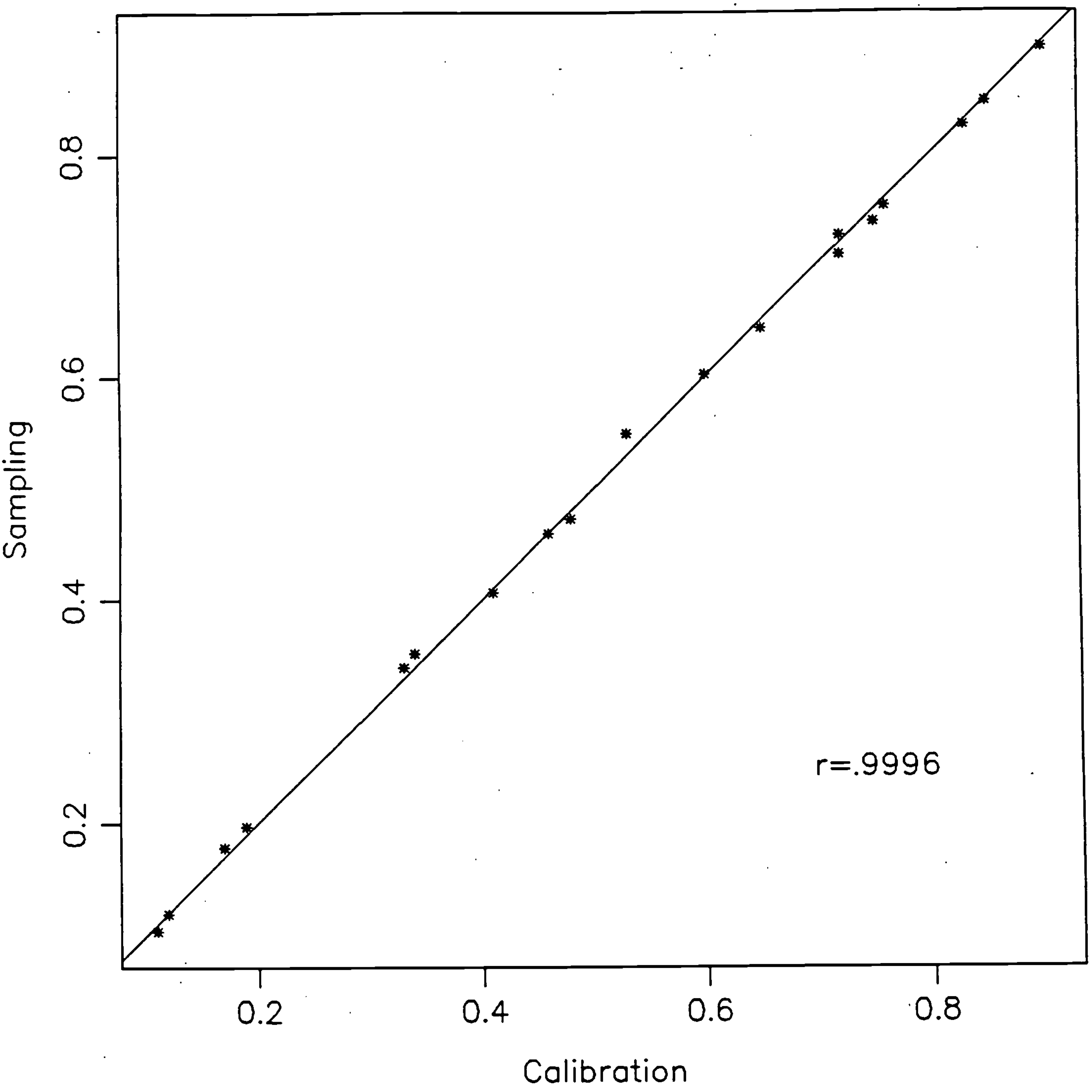


Figure 10-15

Proportion–correct Statistics for
Data Analysis and Statistics Items
Using Calibration and Sampling Weights

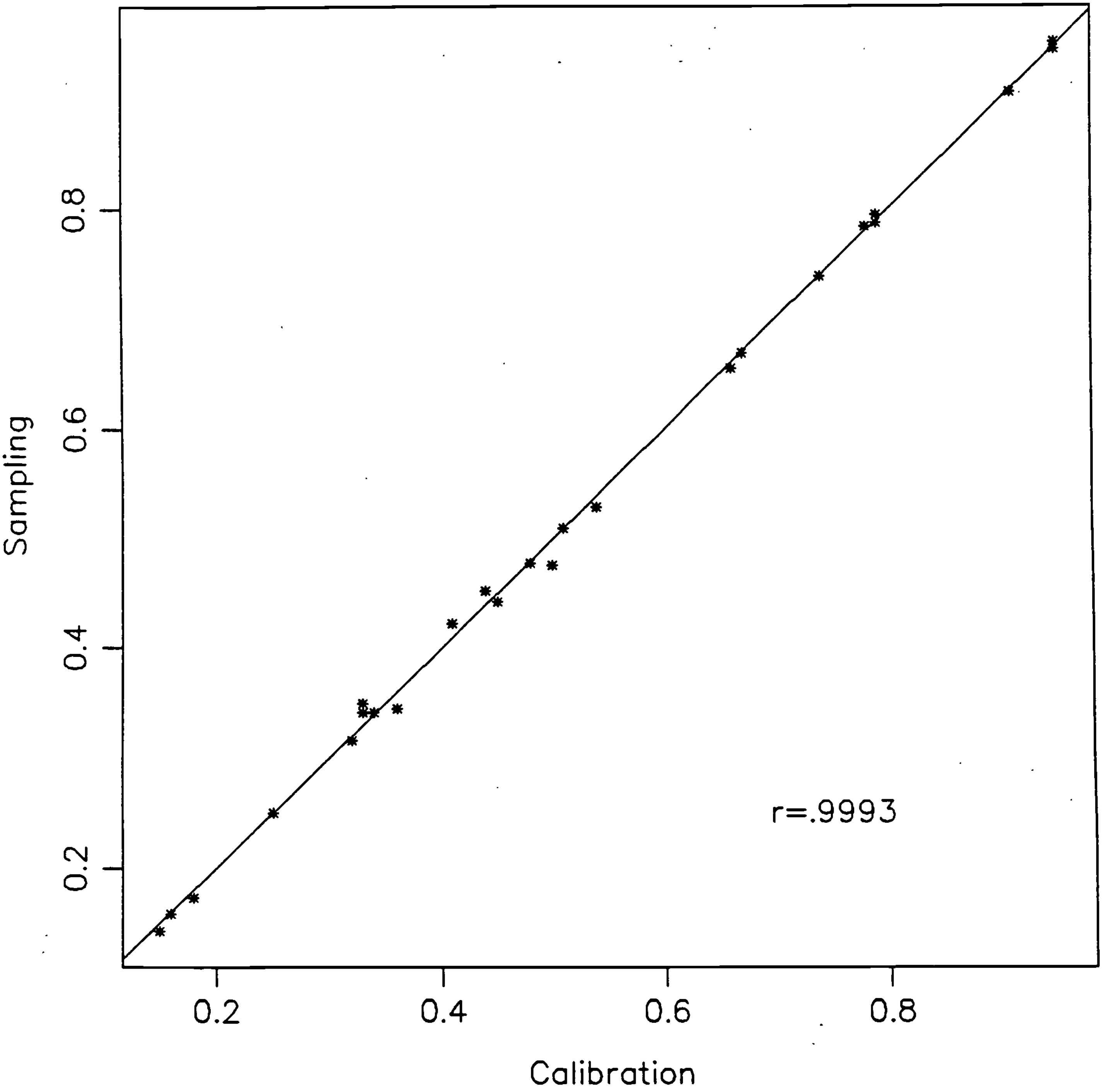


Figure 10-16

Proportion—correct Statistics for
Algebra and Functions Items
Using Calibration and Sampling Weights

using the rescaled sampling weights had almost no effect on the overall difficulty of the set of items in each scale or the relative difficulty of items within a scale. As mentioned earlier, the similarities in item proportion correct statistics does not imply, but is consistent with, equal ICCs for the two types of weighting schemes.

IRT calibrations were carried out separately for each scale using a version of the BILOG program which has been modified for use in NAEP. Prior distributions were imposed on item parameters with the following starting values: thresholds (normal[0,2]); slopes (log-normal[0,.5]); and, asymptotes (2-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. Items presented to, but not reached by, students were treated as "not-presented" items. Intentional omissions were treated as fractionally correct with probability equal to the reciprocal of the number of response options for each item.[5]

Item parameter estimation proceeded in two phases. First, the subject ability distribution was assumed fixed [normal(0,1)] and a stable solution was obtained. The parameter estimates from this 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 restandardized to have a mean of zero and standard deviation of one and, correspondingly, parameter estimates for that cycle were also linearly restandardized.

Model fit was evaluated by examining BILOG likelihood ratio chi-square statistics[6] and by examining plots of nonmodel-based estimates of the expected conditional (on θ) proportion correct versus the proportion correct predicted by the estimated ICC at each of set of θ levels (see, Mislevy & Sheehan, 1987, p. 302). In general, the fit of the model was quite good. During the estimation process, difficulties obtaining a stable set of parameter estimates were encountered for only two of the 137 items. One of these items was removed from the measurement scale since preliminary graphical analyses suggested a poor fit to the 3PL model. This item also was removed from the Measurement scale for the national assessment. For the other item (algebra and functions scale), the 3PL model appeared to fit well, however, difficulty was encountered obtaining a converged slope estimate. A decision was made to retain the item and fix the slope at the value obtained after ten BILOG estimation cycles. The IRT parameters for the items included in the Trial State Assessment are listed in Appendix D.

10.5 ESTIMATION OF STATE AND SUBGROUP PROFICIENCY DISTRIBUTIONS

The proficiency distributions (for the total population in each state, and for important subgroups within each state) were estimated by using the multivariate plausible values methodology described in the previous chapter (see also Mislevy, 1988). The background variables included in the model (denoted $\underline{y}$ in the previous chapter) were principal component

[5] The probability was set at zero for open-ended items.

[6] These sampling distributions of these statistics are probably not strictly χ^2 with the indicated degrees of freedom. Therefore, they were used as descriptive indices of relative model fit rather than in a statistically rigorous fashion.

scores derived from the correlation matrix of selected main-effects and two-way interactions associated with a wide range of student, teacher, school, and community variables. A set of five multivariate plausible values was drawn for each individual who participated in the Trial State Assessment.

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), student attitudes toward mathematics, student behaviors both in and out of school (e.g., amount of TV watched daily, amount of mathematics homework each day), the type of mathematics class being taken (e.g., algebra, or general eighth-grade mathematics), the amount of emphasis on various topics included in the assessment provided by the students' teachers, as well as 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. Overall, relationships between proficiency and more than 50 variables, taken directly or derived from the student, teacher, and school questionnaires, or provided by Westat, were estimated and reported.

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 all variables to be reported on as independent variables in the conditioning model. When expressed in terms of contrast-coded main effects and interactions, the variables cited above resulted in 167 variables to be included in the conditioning model. A listing of the complete set of variables included in the conditioning model is provided in Appendix C.

The conditioning model, including all the contrasts listed in Appendix C, included up to 167 contrasts[7]. Many of these contrasts were highly correlated with other contrasts in the model; other contrasts involved relatively small numbers of individuals. Under such conditions, it can be difficult to obtain converged estimates of Γ and Σ (described in the previous chapter) based on the iterative numerical procedures used in MGROUP, (the computer program developed by Sheehan & Mislevy (1985), which is used by NAEP to estimate conditioning models and generate plausible values). To minimize such potential convergence problems, the original background variable contrasts were standardized and transformed into a set of linearly independent variables by extracting principal components from the correlation matrix of the original contrast variables. The principal components, rather than the original variables, were used as the independent variables in the conditioning model.

Principal components are a set of uncorrelated linear combinations of the original standardized variables (Harris, 1975). They retain information about variability and intercorrelation among the original variables. Previous analyses of the NAEP 1988 Reading data suggested that conditioning using principal components virtually eliminated biases in analyses involving the original effects from which the components were derived (Mislevy, 1988). In addition, because principal components are uncorrelated, the MGROUP estimation problems

[7] In some states, one or more contrasts were not possible since all individuals were at the same level of that contrast.

which might have resulted from the high degree of multicollinearity among the original variables were avoided.

The same variables and codings described in Appendix C were included in the conditioning model for all 40 Trial State Assessment participants. In addition, a single common set of IRT item parameters were used. However, principal components were extracted separately and separate conditioning models were estimated for each of the 40 Trial State Assessment participants.

Estimating separate conditioning models for each state was more complex than the simpler alternative of estimating a single model for all 40 states. However, there were significant potential problems associated with the simpler approach to warrant the more complicated approach. The need for separate conditioning models for each state can be understood by examining the potential problems associated with estimating a single common model. The problems can be clearly illustrated in the context of using the original background variables (rather than the principal component scores that were actually used as conditioning variables).[8]

Under the assumptions of the model, estimating a single conditioning model for all 40 states would produce consistent estimates of the means for subgroups for which contrasts were explicitly included in the model. For example, since a Race/Ethnicity contrast was included for Asian American students, a consistent estimate of the mean proficiency of the total group of Asian American students represented by those students who participated in the Trial State Assessment, could be obtained from the single conditioning model.

Trial State Assessment results were reported separately for each state and for subgroups within the states. Given this reporting structure, the single model approach is problematic because it will produce consistent estimates of the mean proficiency of subgroups within each state only if the magnitude of the effect associated with a particular contrast is identical across all 40 states (i.e., the single model approach is tantamount to assuming there are no state-by-contrast interactions). Using the example in the previous paragraph, the single model approach would provide consistent estimates of the mean for Asian American students within a particular state if the difference between the predicted mean for Asian American students and the predicted mean for all other students is identical across all 40 states. If that is not true, the types of biases described in the previous chapter will affect the state-specific estimates of subgroup means.

There is little prior research or information to suggest that the nature and relative magnitude of relationships between proficiencies and the conditioning variables are consistent and similar across states. As an example, it is more reasonable to assume that Asian American students (or students from any racial/ethnic group) from different states might, on average, differ with respect to a large number of variables such as their economic situations, the length of time spent in this country, their facility with English, and various other factors. Thus, to ensure

[8] The same problems *exist* when principal components are employed in the model rather than the original conditioning contrasts. However, the reasons for the problems are more easily *explained* in the context of using the original conditioning contrasts.

consistent estimates of proficiency distributions of subgroups within each state, the conditioning model needs to include state-by-contrast interaction effects for all the contrasts in the model. Alternatively, separate conditioning models can be estimated for each state using only that state's data (i.e., separate state-specific estimates of the conditioning effects [the Γ matrix defined in the previous chapter] and the residual variance matrix [the Σ matrix described in the previous chapter] are obtained).

As mentioned above, in addition to estimating separate conditioning models for each state, principal components were extracted separately for each state (for reasons similar to those given above). In theory, the number of principal components that could be extracted is equal to the total number of the original contrast variables minus the number of these variables that are exactly collinear with other variables (or collections of variables) in the model. Analyses by Kaplan and Nelson (see Mislevy, 1990) on the 1988 NAEP Reading data suggested that a relatively small number of principal components will capture almost all of the variance and most of the complex intercorrelations among the set of original variables and will reduce most of the potential bias for primary and secondary analyses. For the Trial State Assessment analysis, the number of principal components included for each state was that number required to account for approximately 90 percent of the variance in the original contrast variables.

Table 10-4 contains a listing for each of the 40 states of the number of principal components included in and the proportion of variance accounted for by conditioning model. It is important to note that the proportion of variance accounted for by the conditioning model differs across scales within a state as well as across states within a scale. Such variability is not unexpected for at least two reasons. First, there is no reason to expect that the strength of the relationship between proficiency and demographics to be identical across all states. In fact, one of the reasons for fitting separate conditioning models is that the strength and nature of this relationship may differ across states. Second, the homogeneity of the demographic profile also differs across states. As with any correlational analysis, the restriction of the range in the predictor variables will attenuate the relationship.

As discussed in the previous chapter, NAEP scales are viewed as summaries of consistencies and regularities 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 which compared state-level and sub-group 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 these analyses. One set of such analyses are presented in Figures 10-17 through 10.21. The figures contain scatterplots of the state proportion-correct means versus the state scale score mean (expressed on the provisional BILOG scale), for each of the five mathematics content areas. As evident from the figures, there is an extremely strong relationship between the estimates of state-level performance in the scale-score and proportion-correct metrics for all five content areas.

Table 10-4

Number of Principal Components Included in, and the Proportion of Variance Accounted for by the Conditioning Model for Each of the 40 Jurisdictions in the Trial State Assessment

Jurisdiction	# of Principal Components	Proportion of Variance Accounted For:				
		Scale 1	Scale 2	Scale 3	Scale 4	Scale 5
Alabama	90	0.620	0.584	0.545	0.682	0.623
Arizona	90	0.610	0.511	0.518	0.611	0.606
Arkansas	90	0.611	0.577	0.588	0.660	0.643
California	90	0.653	0.581	0.553	0.678	0.682
Colorado	89	0.591	0.518	0.525	0.585	0.624
Connecticut	89	0.655	0.610	0.630	0.716	0.691
Delaware	86	0.649	0.642	0.576	0.703	0.695
District of Columbia	87	0.612	0.514	0.572	0.719	0.651
Florida	91	0.667	0.565	0.616	0.671	0.714
Georgia	89	0.632	0.586	0.616	0.691	0.661
Guam	82	0.681	0.559	0.580	0.714	0.725
Hawaii	88	0.672	0.634	0.660	0.743	0.701
Idaho	89	0.578	0.480	0.454	0.681	0.636
Illinois	88	0.628	0.580	0.574	0.713	0.652
Indiana	91	0.609	0.556	0.540	0.665	0.654
Iowa	89	0.541	0.468	0.475	0.563	0.543
Kentucky	90	0.609	0.543	0.533	0.667	0.624
Louisiana	90	0.606	0.551	0.551	0.651	0.604
Maryland	89	0.687	0.690	0.672	0.725	0.711
Michigan	90	0.635	0.584	0.510	0.661	0.649
Minnesota	90	0.574	0.504	0.481	0.602	0.569
Montana	87	0.556	0.477	0.475	0.565	0.527
Nebraska	88	0.591	0.546	0.565	0.563	0.573
New Hampshire	89	0.602	0.524	0.514	0.686	0.633
New Jersey	89	0.657	0.656	0.612	0.701	0.716
New Mexico	90	0.638	0.574	0.506	0.685	0.660
New York	91	0.661	0.606	0.606	0.712	0.691
North Carolina	90	0.665	0.605	0.611	0.692	0.704
North Dakota	86	0.590	0.502	0.511	0.623	0.664
Ohio	90	0.624	0.529	0.547	0.653	0.640
Oklahoma	90	0.570	0.549	0.484	0.571	0.617
Oregon	91	0.587	0.483	0.513	0.630	0.619
Pennsylvania	90	0.691	0.603	0.630	0.734	0.712
Rhode Island	90	0.658	0.627	0.654	0.746	0.685
Texas	90	0.619	0.549	0.554	0.666	0.645
Virgin Islands	82	0.577	0.484	0.535	0.734	0.600
Virginia	89	0.708	0.639	0.666	0.706	0.705
West Virginia	91	0.581	0.496	0.489	0.618	0.627
Wisconsin	89	0.582	0.532	0.522	0.594	0.608
Wyoming	88	0.545	0.493	0.460	0.590	0.562

Note: Scale 1 = Numbers & Operations, Scale 2 = Measurement, Scale 3 = Geometry, Scale 4 = Data Analysis, Probability, & Statistics, Scale 5 = Algebra & Functions.

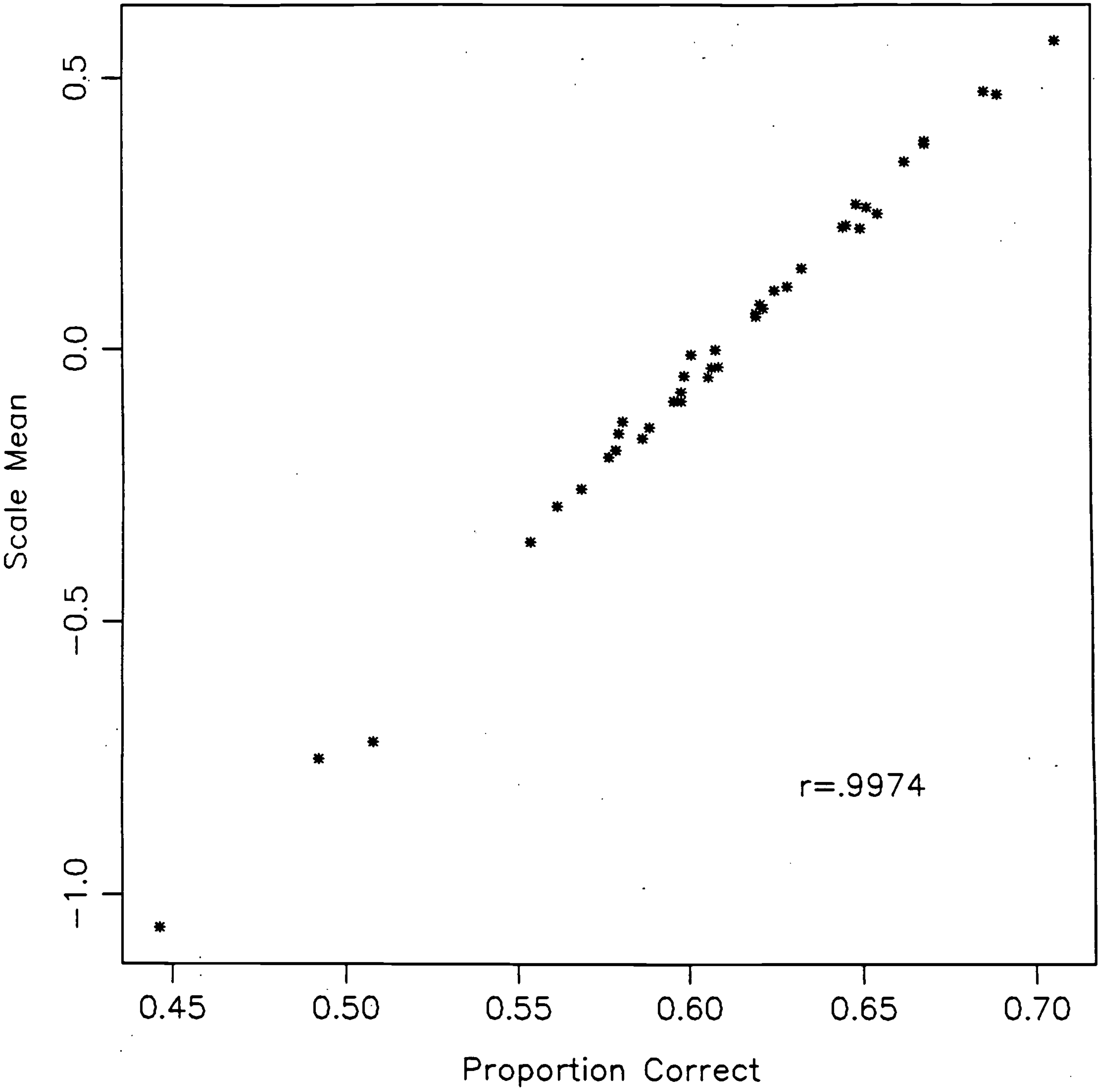


Figure 10-17

State Proportion-correct Mean Versus
State Provisional Scale Mean
for Numbers and Operations Items

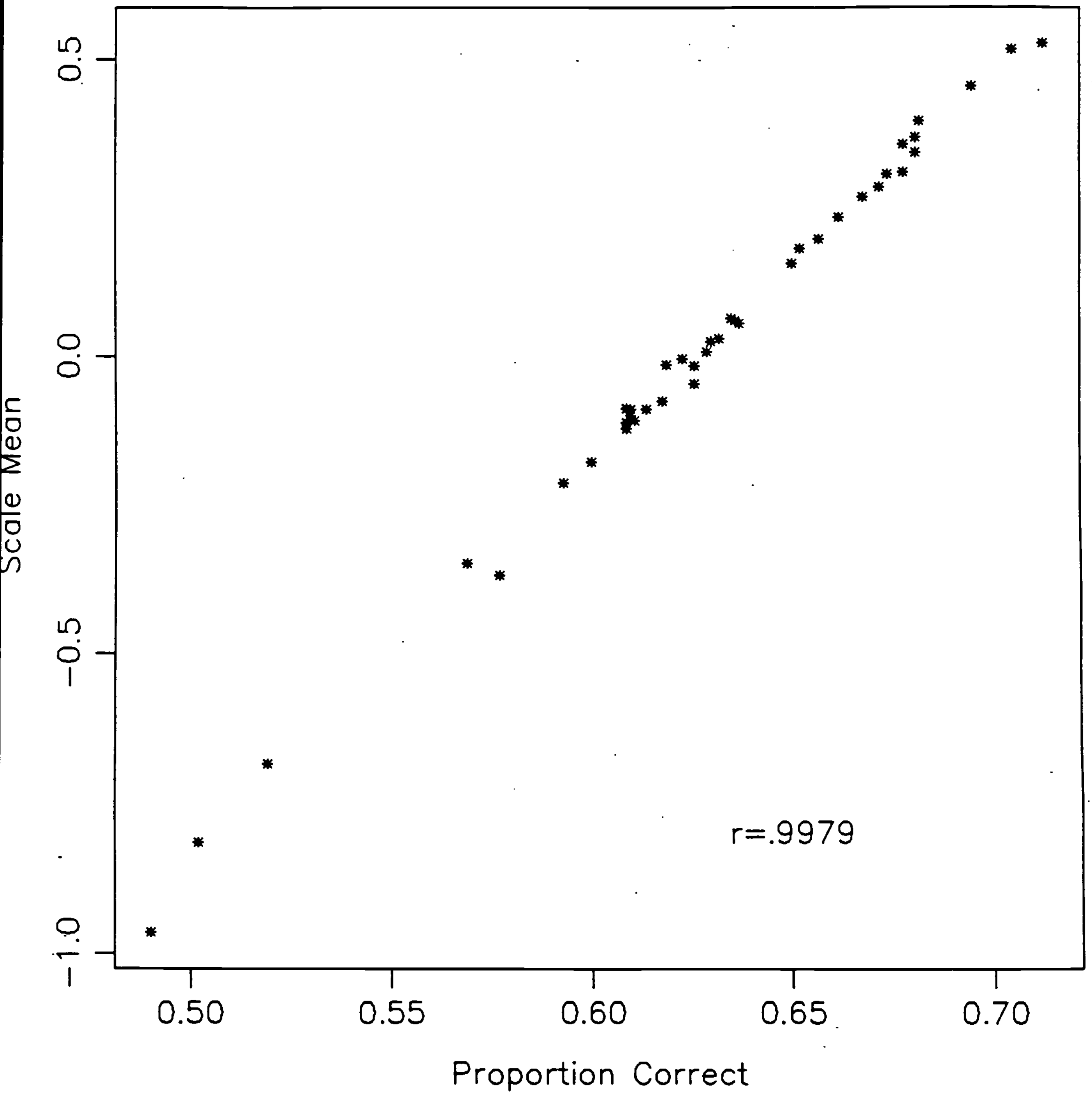


Figure 10-18

State Proportion–correct Mean Versus
State Provisional–Scale Mean
for Measurement Items

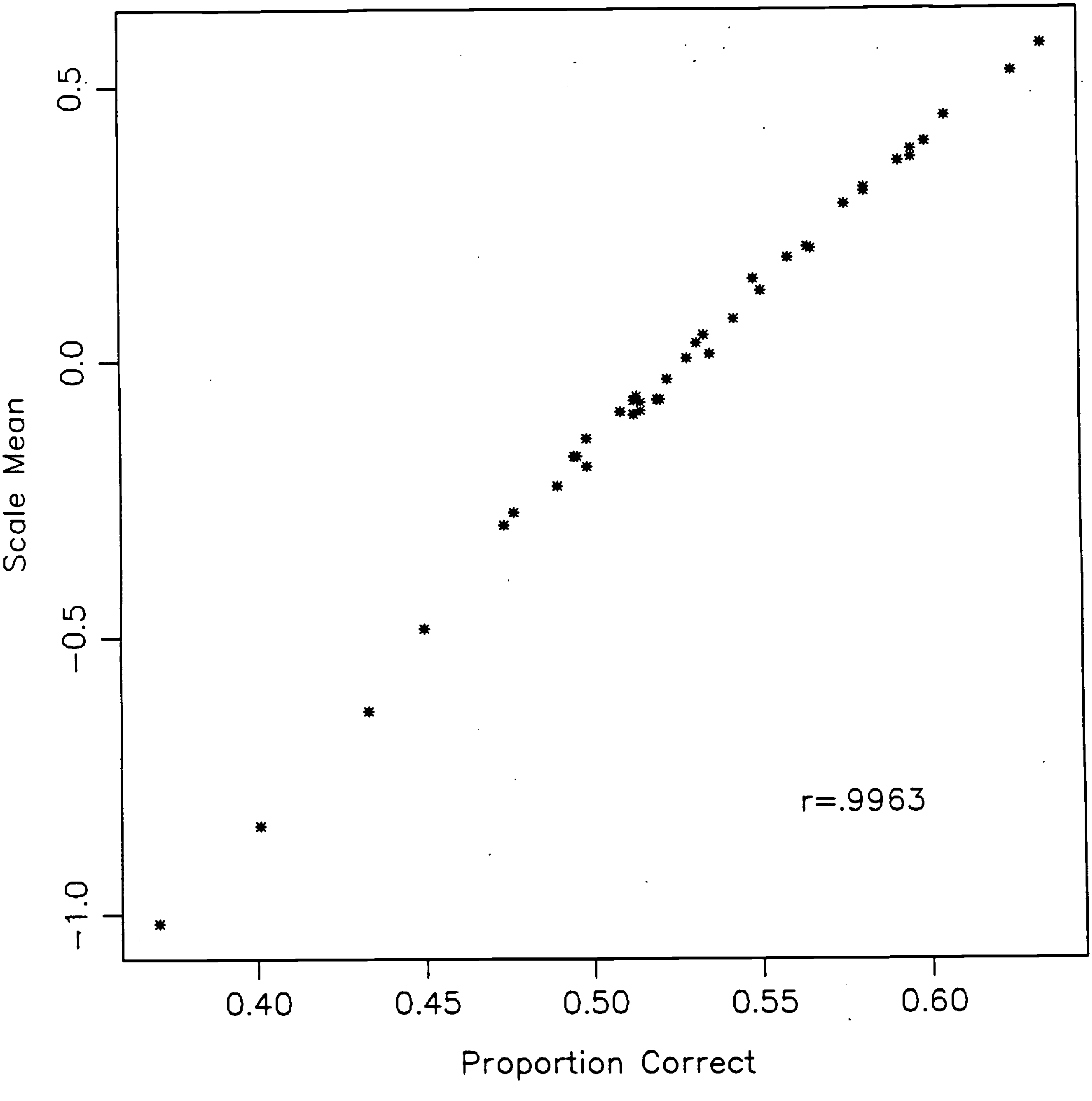


Figure 10-19

State Proportion-correct Mean Versus
State Provisional Scale Mean
for Geometry Items

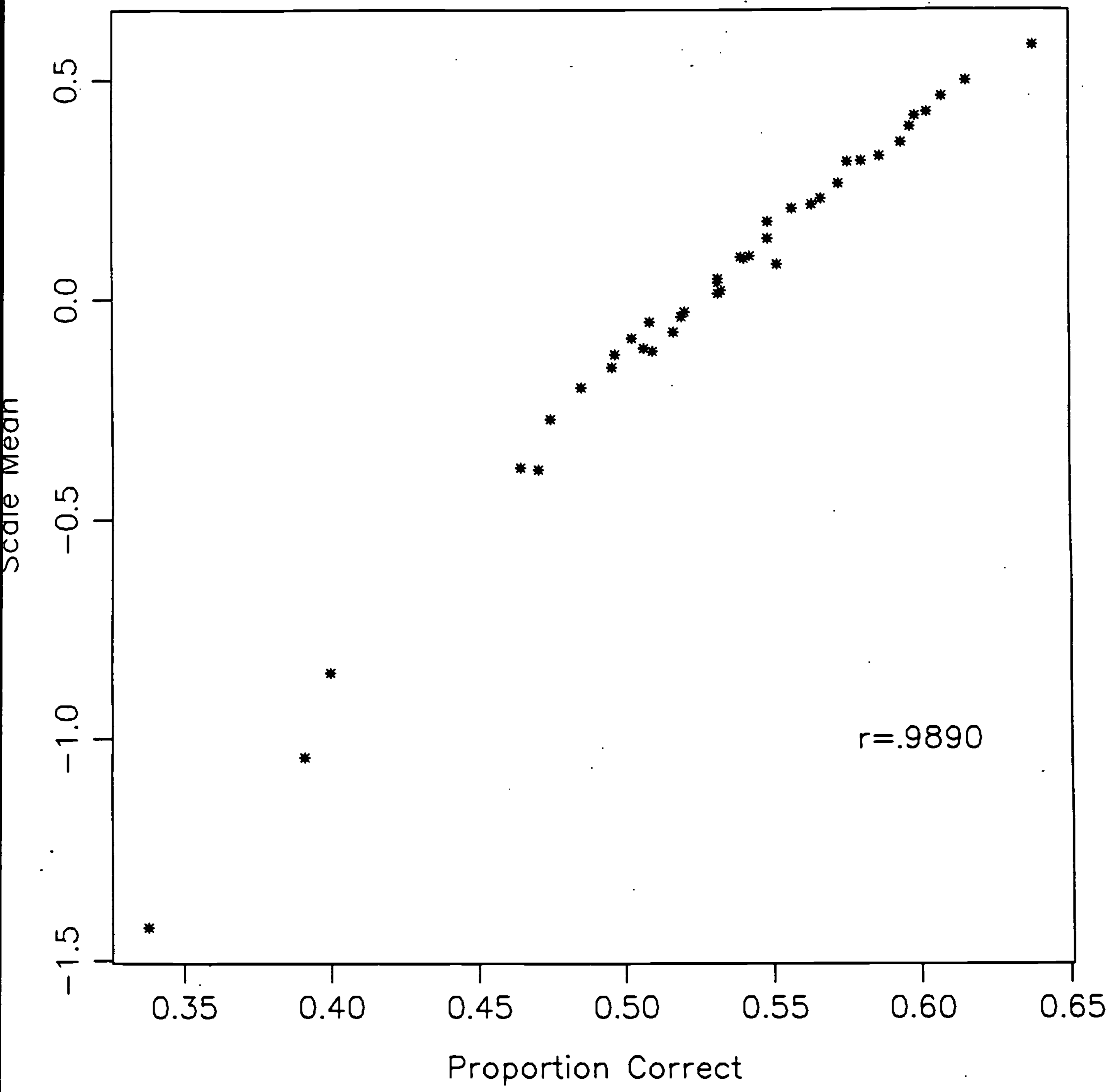


Figure 10-20

State Proportion–correct Mean Versus
State Provisional–Scale Mean
for Data Analysis & Statistics Items

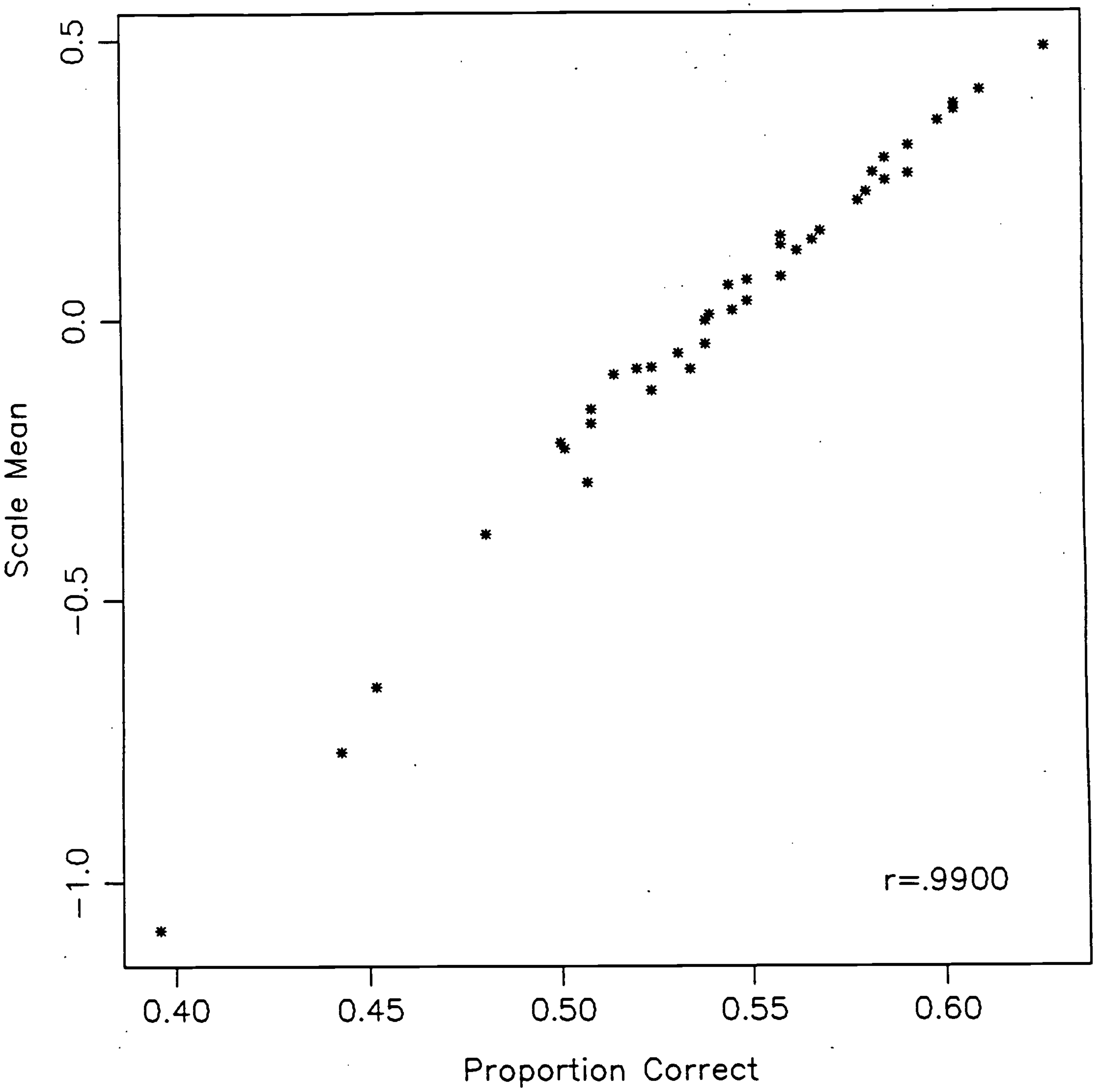


Figure 10-21

State Proportion–correct Mean Versus
State Provisional–Scale Mean
for Algebra & Functions Items

10.6 LINKING STATE AND NATIONAL SCALES

One of the purposes of the Trial State Assessment Program was to allow each participating jurisdiction to compare its results with the nation as a whole and with the region of the country in which that jurisdiction is located.[9] To permit such comparisons, a nationally representative sample of public-school students in the eighth-grade was tested as part of the national assessment using the same assessment booklets as in the Trial State Assessment. In addition, a subsample of the national assessment was tested at about the same time of the year (January to March 1990) as were students participating in the Trial State Assessment (February 5 to March 2, 1990).

For valid comparisons to be made between each of the Trial State Assessment participants and the relevant national subsample, results from the two assessments had to be expressed in terms of a similar system of scale units. As described above, the provisional BILOG scales for the Trial State Assessment (and subsequent estimation of proficiency distributions using plausible values) were computed independently from the scaling used for the national assessment[10].

There were several reasons for the decision to scale the Trial State Assessment and national mathematics assessments separately. First, there was one substantial difference in administration procedures between the two assessments -- Westat staff collected the data for the national assessment, while data collection activities for the Trial State Assessment (such as ensuring school and student participation, assessing students according to standardized procedures, and maintaining test security) were the responsibility of each of the participating states. Second, because of the political sensitivity of the Trial State Assessment results, the stakes of that assessment are somewhat higher than they are for the national assessment. Because of the higher stakes, motivational differences might exist between the sample of students participating in the Trial State Assessment and those in the national assessment. These motivational differences might translate into differential performance.

The systems of scale units that result from separate IRT scalings are not typically comparable, even if the same set of items are used in both scalings. The units and origin of the provisional scales for both the national and Trial State Assessment assessments were set by standardizing the ability distributions for their respective calibration samples to have a mean of zero and standard deviation of one. One major difference between the Trial State Assessment and national calibration samples that is particularly relevant to the issue of aligning the scales from the two assessments was the availability in the national assessment of data from other age groups. These additional data are of use in both calibrating items as well as in locating the proficiency distribution of the national grade eight sample compared to the grade 4 and grade 12 samples. Because all three age groups are used in the national item calibration, the origin and scale unit for the national results are based on an aggregate distribution which is the sum of

[9] There are no regions designated for the territories.

[10] Care was taken to ensure that the five scales were produced for both the national and Trial State Assessment, and that all the items included in the Trial State Assessment were also included in the national assessment. Because the national assessment spans three age/grades, additional items are used in developing the national scales which were not part of the Trial State Assessment.

the three age ability distributions. In contrast, the unit and origin of the scales for the Trial State Assessment are based on an ability distribution for a single grade (grade eight). Clearly, without a transformation, the two metrics are not comparable and special procedures must be conducted to ensure a similar set of scale units.

In the context of standard fixed-length test forms (e.g., X and Y), a commonly used definition of equating is that scores on test X and test Y are equated for some population of examinees if the score distributions of test X and test Y are identical for the population in question (Braun & Holland, 1982). One method to equate tests is to obtain two large random samples of the population and administer test X to one, test Y to the other. A transformation is then derived for one of the tests (e.g., X), such that the distribution of the transformed X scores for the sample taking that test is identical to the distribution of Y scores for the sample taking test Y. A linear transformation is used when the distributions of scores on test X and test Y have the same shape and differ only with respect to their means and standard deviations. When the distributions differ somewhat in shape, a linear transformation is still often chosen for simplicity purposes. However, in such cases, equivalent scores from the two tests are comparable only to the extent of indicating the same relative distance from the mean of each test's score distribution.

A procedure analogous to linearly equating test forms was used to link the Trial State Assessment and national scales. The Trial State Assessment and national scales were made comparable in the sense that estimated proficiency distributions from two samples (the Trial State Assessment and a special sample of the national assessment [called the State Aggregate Comparison Sample and described below]) from the same population (eighth-grade students in public schools in the 37 states and the District of Columbia) were constrained to have the same mean and standard deviation.[11]

The State Aggregate Comparison (SAC) sample was a subsample of 2,467 students from the winter subsample of the national assessment. The SAC subsample consists of all eighth-grade students in public schools in the 37 participating states and the District of Columbia who were assessed as part of the winter administration of the national mathematics assessment. With appropriate weighting (provided by Westat), the SAC is a representative sample of the population of all grade-eligible public-school students within the 37 states and the District of Columbia participating in the Trial State Assessment and was assessed at a reasonably similar point in time as the Trial State Assessment.

The following steps were followed to linearly link the scales of the two assessments:

1) For each scale, an estimate of the proficiency distribution of the total Trial State Assessment sample (minus the students from Guam and the Virgin Islands) was obtained using the full set of plausible values generated by the MGROUP program. Recall that these plausible values are expressed on the provisional Trial State Assessment scale and were generated using the common state item parameters, but separate state-specific conditioning coefficients. The weights used were the final

[11] Data from the two territories (Guam and the Virgin Islands) were excluded for the purposes of establishing the link to the national scale.

sampling weights. Thus, the resulting estimate pertains to the distribution of proficiency in the aggregated group of eighth grade public school students in the 37 states and the District of Columbia.

The arithmetic mean of the five sets of plausible values was taken as the estimated mean of the Trial State Assessment distribution, and the geometric mean of the standard deviations of the five sets of plausible values was taken as the estimated standard deviation of the distributions for each scale.

2) For each scale, an estimate of the proficiency distribution of the total SAC subsample of the national eight-grade winter half-sample was obtained using the full set of plausible values for this group. These plausible values were expressed in terms of the scale that was intended to be used for reporting the results for the national mathematics assessment and were generated using the national assessment item parameters and a common set of eighth-grade specific conditioning coefficients. The weights used were specially provided by Westat to allow for the estimation of proficiency for the same population of students as for state data (i.e., the aggregated group of eighth-grade public- school students in 37 states and the District of Columbia).

   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..

3) For each content area scale, a set of linear transformation coefficients to link the state scale to the corresponding national scale were obtained. The linking was of the form,

$$Y^{*} = \alpha + \beta X_{TSA}$$

   where,

   X_{TSA} = a scale level in terms of the system of units of the provisional BILOG scale

   Y^{*} = scale level in terms of the system of units comparable to those used for reporting the national mathematics results

   β = (SD_{SAC}/SD_{TSA}),

   α = $(M_{SAC} - \beta M_{TSA})$

   SD_{SAC} = the estimated standard deviation of the SAC sample proficiency distribution

   SD_{TSA} = the estimated standard deviation of the Trial State Assessment equating sample proficiency distribution (with Guam and Virgin Islands removed)

M_{SAC} = the estimated mean of the SAC sample proficiency distribution

M_{TSA} = the estimated mean of the Trial State Assessment equating sample proficiency distribution (with Guam and Virgin Islands removed)

The final conversion parameters for transforming plausible values from the provisional BILOG scales to the final Trial State Assessment reporting scales are given in Table 10-5. All Trial State Assessment results are reported in terms of the Y^* metric.

Figures 10-22 through 10-26 provide plots of the estimated proficiency distributions for the aggregate of the Trial State Assessment data (minus Guam and the Virgin Islands) and the SAC sample. These Trial State Assessment results are expressed on the final Trial State Assessment scale, and SAC sample results are expressed on the national mathematics scale. The estimated distributions for each scale, and for the composite, were obtained by combining the five sets of plausible values and obtaining histograms of these collections. The histograms were then smoothed using a cubic spline routine given in Reinsch (1967) and available in the Statistical Analysis System (SAS). For all five content area scales, the shapes of the Trial State Assessment and SAC samples were fairly similar.

Table 10-5

Scaling parameters used to transform Trial State Assessment Results From Provisional BILOG Scales to Final Reporting Scale

Scale	α	β
Numbers & Operations	265.2791	36.1399
Measurement	256.6941	43.9237
Geometry	258.9029	35.8072
Data Anal., Prob., & Stat.	259.6359	44.8398
Algebra & Functions	259.7096	38.1248

Figure 10-22

Smoothed Histograms of Numbers and Operations Plausible Values for the Trial State Assessment and State Aggregate Equating Samples

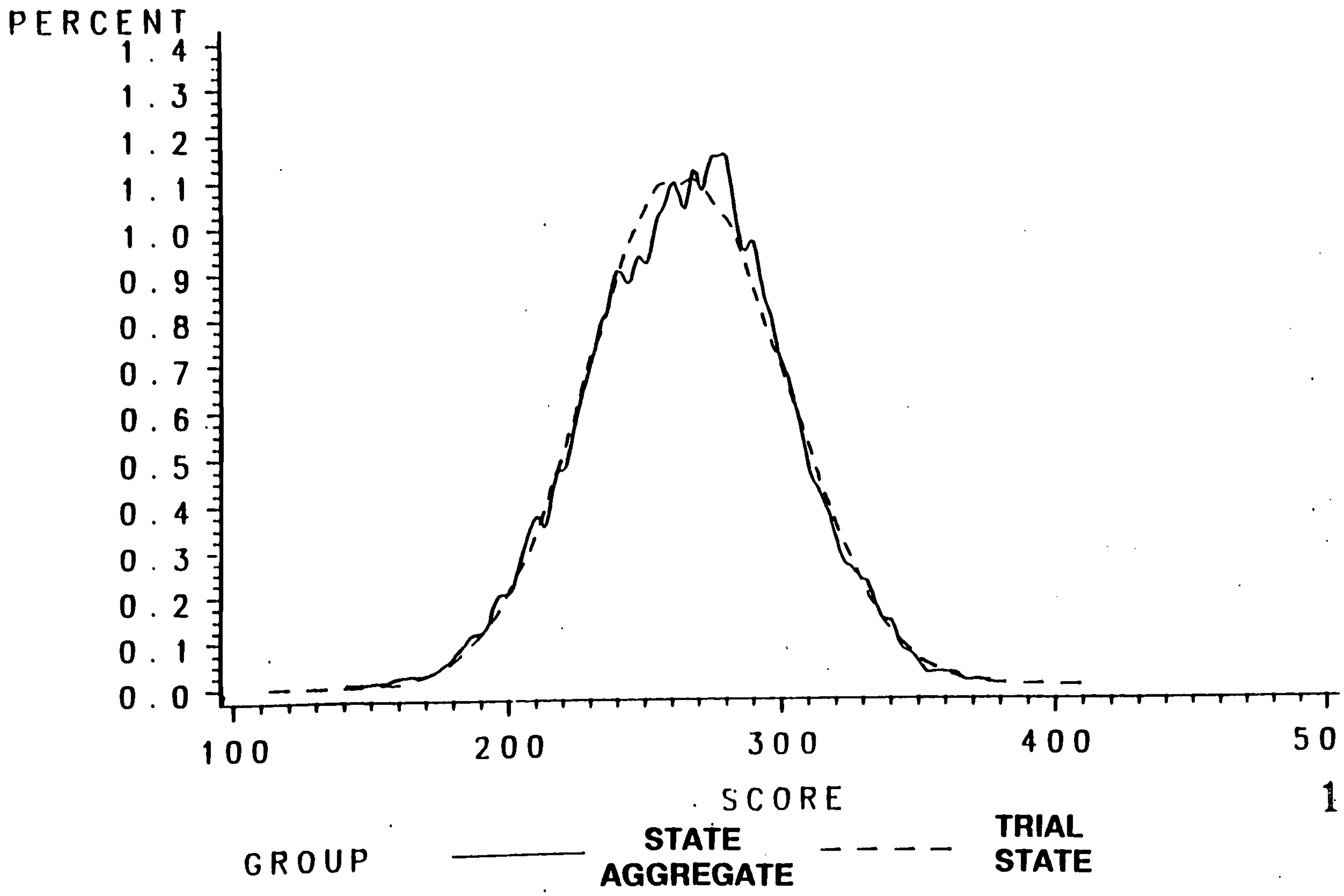


Figure 10-23

Smoothed Histograms of Measurement Plausible Values for the Trial State Assessment and State Aggregate Equating Samples

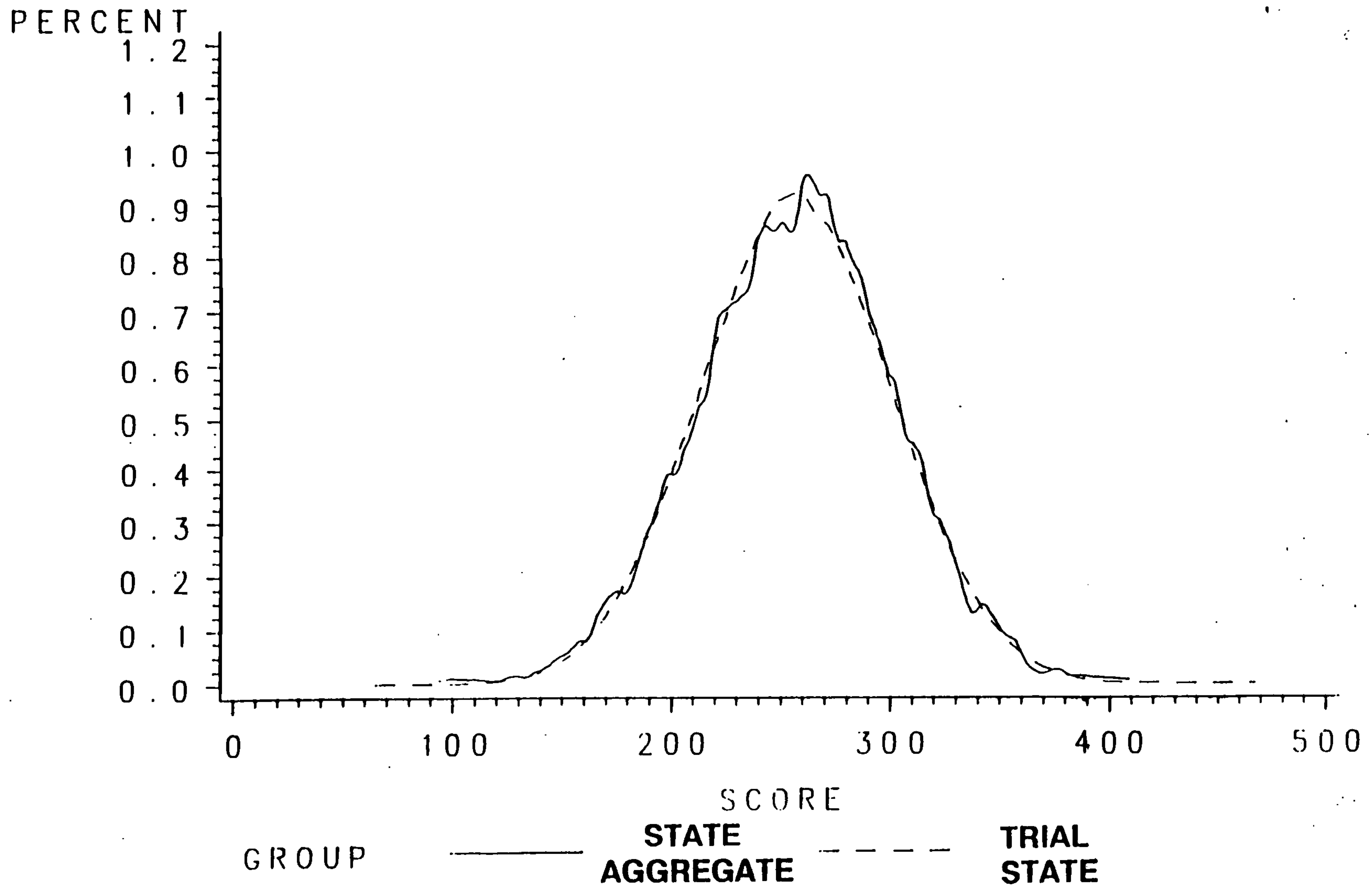


Figure 10-24

Smoothed Histograms of Geometry Plausible Values for the Trial State Assessment and State Aggregate Equating Samples

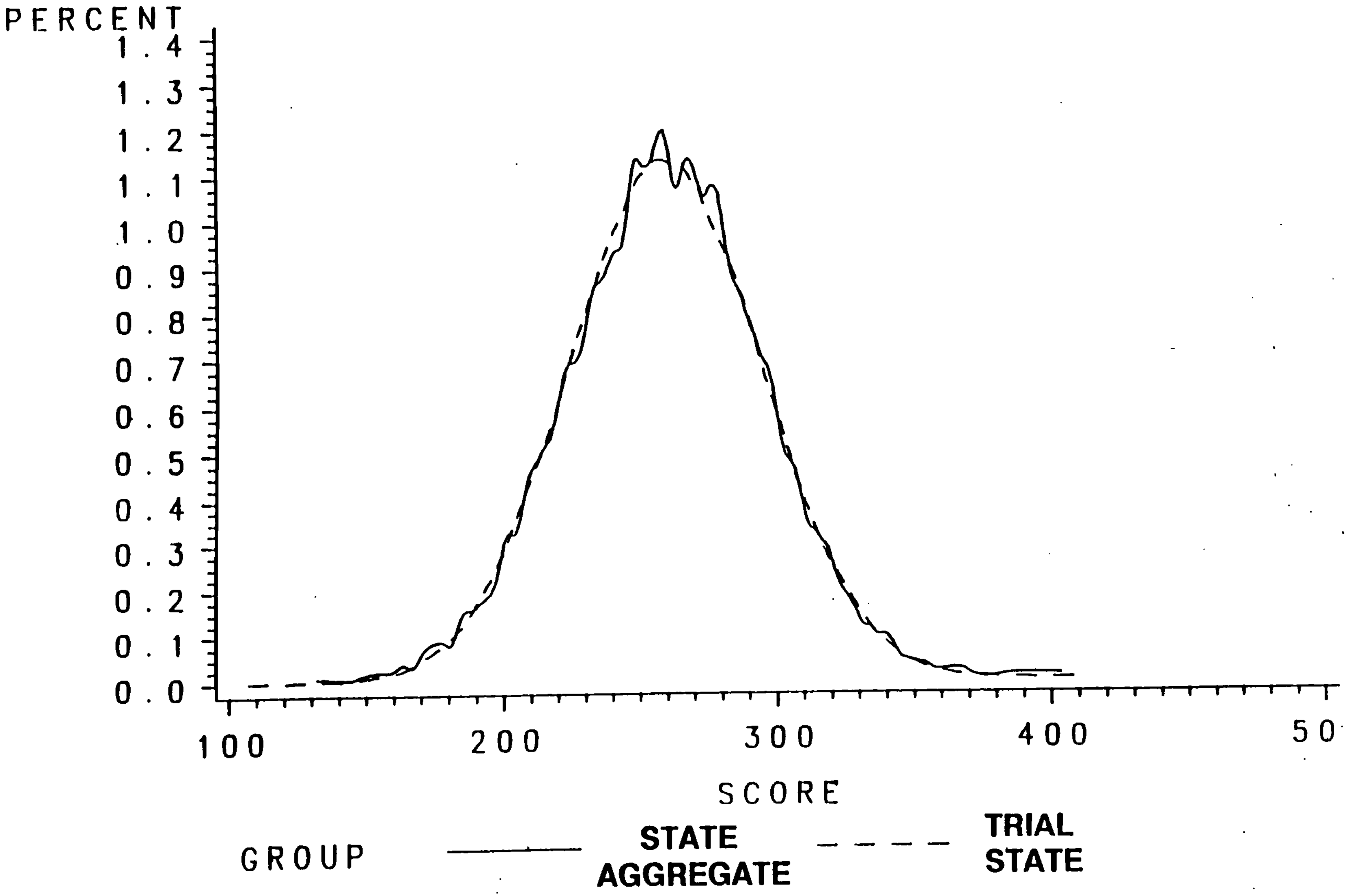


Figure 10-25

Smoothed Histograms of Data Analysis, Probability, and Statistics Plausible Values for the Trial State Assessment and State Aggregate Equating Samples

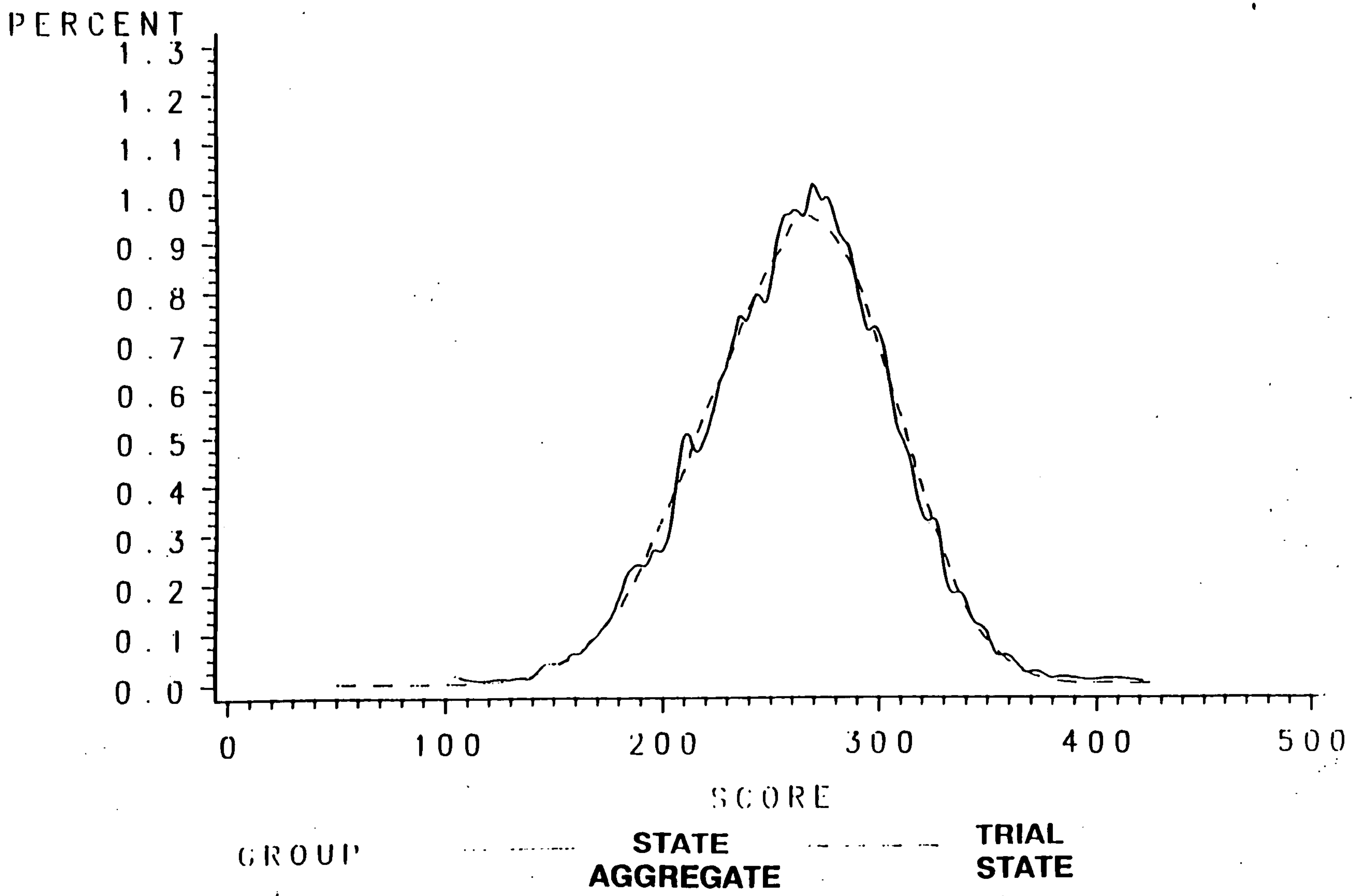
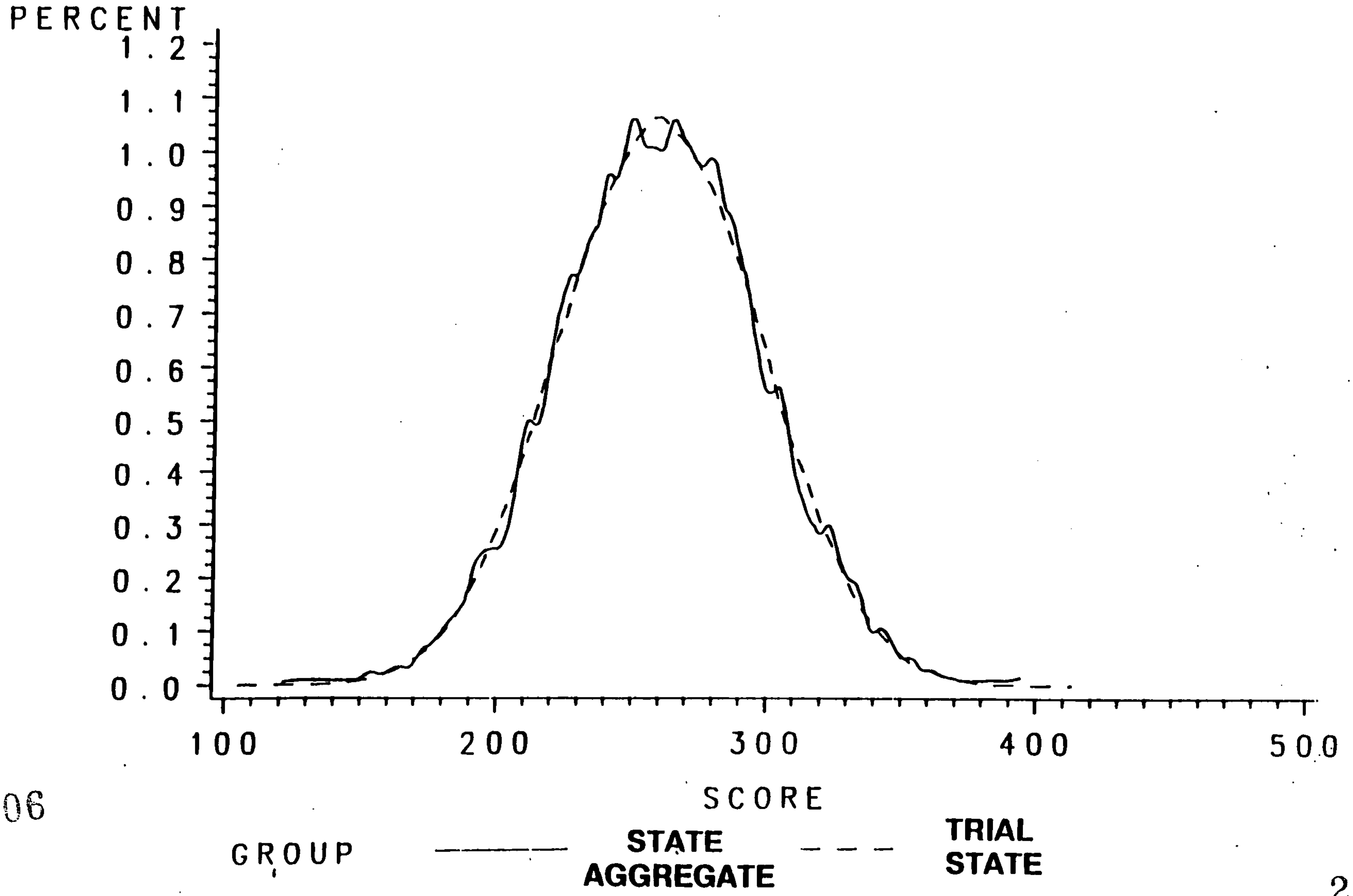


Figure 10-26

Smoothed Histograms of Algebra and Functions Plausible Values for the Trial State Assessment and State Aggregate Equating Samples



10.7 PRODUCING A MATHEMATICS COMPOSITE SCALE

For the national assessment, a grade eight composite scale was created as an overall measure of mathematics proficiency for students at that grade. The composite was a weighted average of plausible values on the five content area scales (Numbers and Operations; Measurement; Geometry; Data Analysis, Probability, and Statistics; and Algebra and Functions). The weights for the national scale were proportional to the relative importance assigned to each content area in the assessment specifications developed by the Mathematics Objectives Panel. The weights for each content area were similar to the actual proportion of items from that content area in the entire eighth-grade item pool.

A Trial State Assessment composite scale was developed using weights identical to those used to produce the grade eight composite for the 1990 National Mathematics assessment. The weights were as follows:

Table 10-6
Weights for Composite Scale

Content Area Scale	Weight for Composite	Proportion of item pool
Numbers and Operations	.30	.34
Measurement	.15	.15
Geometry	.20	.19
Data Analysis, Probability, and Statistics	.15	.14
Algebra and Functions	.20	.18

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 scales.)

As mentioned earlier, NAEP scale scores are convenient summaries of aggregate performance over sets of items. As such, they should agree with other reasonable aggregate measures of performance. Therefore, as one check on the plausibility of the final scaling results, each state's estimated composite scale mean was compared to the average proportion correct for that state over the entire eighth-grade item pool. The plot is given in Figure 10-27. The correlation between the transformed average proportion correct and composite scale means is .996.

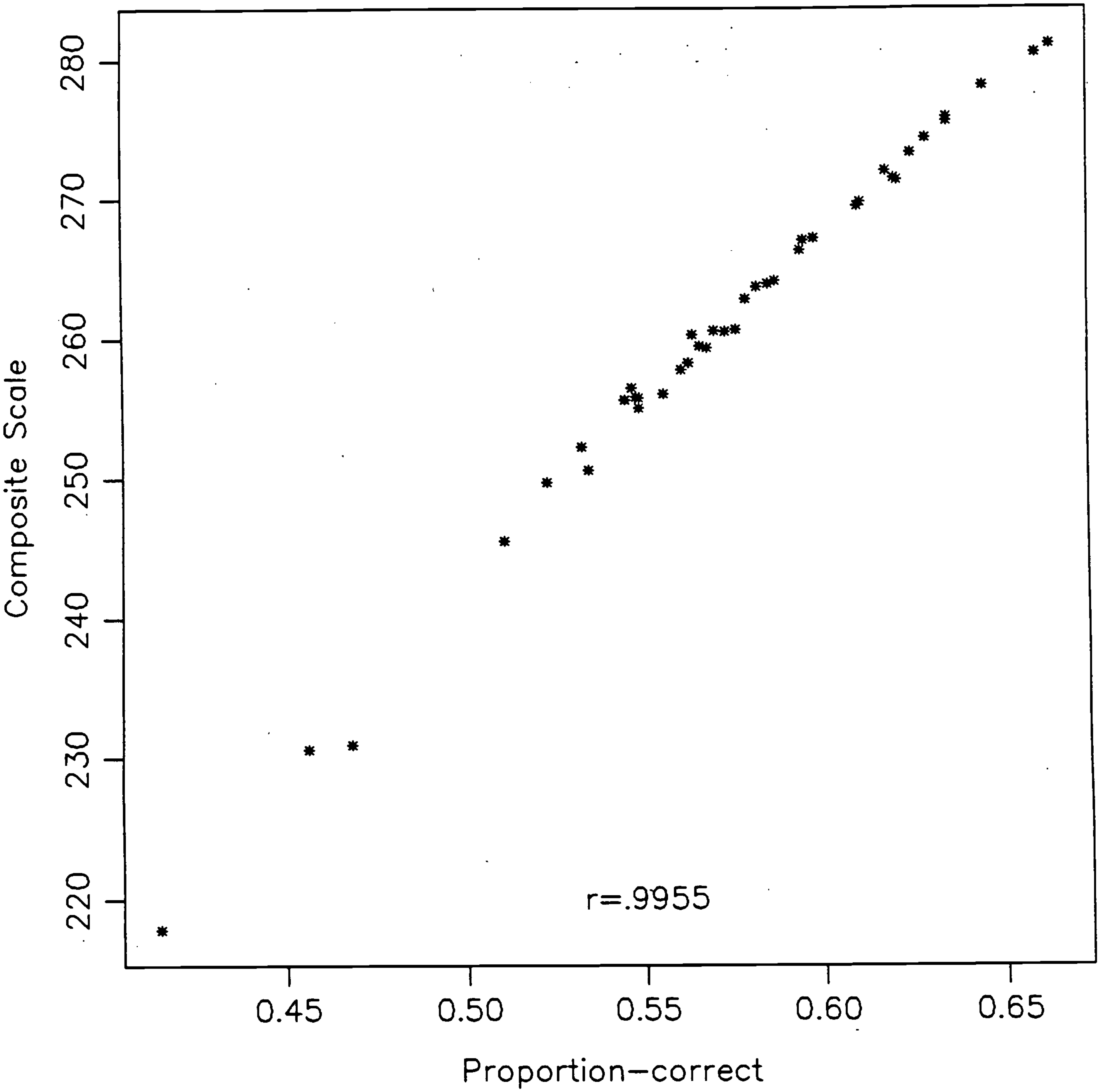


Figure 10-27

State Proportion—correct Mean on the
Entire Item Pool Versus State Composite Scale Mean

Chapter 11

CONVENTIONS USED IN REPORTING THE RESULTS

John Mazzeo

Educational Testing Service

11.1 OVERVIEW

Results for the Trial State Assessment were disseminated in a variety of different reports. Each participating state and jurisdiction received copies of two types of summary reports. The first was a State Report that provided results, with accompanying text and graphics, for that state or territory, as well as national and regional comparisons[1]. The State Reports were produced by a computerized report generation system developed by ETS report writers, statisticians, data analysts, graphic designers, and editors. The reports contained state-level estimates of mean proficiencies (and the proportions of students above selected scale points) for the state as a whole, and for categories of key reporting variables such as gender, race/ethnicity, level of parental education and type of community. In addition, results were reported for a variety of other subpopulations based on variables taken directly from, or derived from, the Student, Teacher, and School Questionnaires. Results were also reported by a number of school and community demographic variables provided by Westat[2].

Each participating jurisdiction also received a Composite Report, in addition to the State Report. The first part of the Composite Report provided the results of the 1990 national mathematics assessment. The second part presented summary information for all 40 participating jurisdictions, along with comparison results for each of the four regions of the country and for the nation. The same variables used to report each jurisdiction's results in the State Report were included in the Composite Report. However, additional variables were also reported on in the Composite Report.

Data about school and student participation rates were reported for each jurisdiction to provide information about the generalizability of the results. School participation rates were 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 were reported, including the overall rate, the percentage of students excluded from the assessment and the exclusion rates for Limited English Proficiency (LEP) students and for students with Individualized Education Plans (IEP).

[1] The national and regional results included in the State Reports and the second section of the Composite Report are based on data from the winter half sample of the 1990 national mathematics assessment and includes only eighth-grade students enrolled in public schools.

[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.

Results were also reported to all participants in a five-section Almanac. Three of the sections of the Almanac (referred to as Proficiency sections) presented analyses based on responses to each of the questionnaires (Student, Mathematics Teacher, and School) that was administered as part of the Trial State Assessment. For most background questions contained in these questionnaires, the proportion of students responding to each option and the mathematics composite proficiency mean for these students were reported with their jackknifed standard errors.[3] The Student Proficiency section of the Almanac also contained the percentage of students at or above the mathematics anchor points. Results were provided for the total group of students in each participating jurisdiction, as well as for groups defined by several traditional NAEP reporting variables (Gender, Race/Ethnicity, Type of Community, and Level of Parent Education).

The fourth section of the Almanac (referred to as a Subscale section) reported proficiency means (and associated standard errors) for the five mathematics content-area scales. Results in this section were also reported for the total group in each state, as well as for select subgroups of interest. The final section of the Almanac ("P-value" section) provided the total-group proportion of correct responses to each item included in the assessment.

The production of the State and Composite Reports required many decisions regarding issues, such as which of the categories of the key reporting variables had sufficient data to permit the reliable reporting of subgroup results, and which, if any, estimates were sufficiently unreliable that they needed to be flagged. Further, the State Report contained computer-generated text that described the results for a particular State and compared total and subgroup performance within the State 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 statistically sound analyses.

The purpose of this chapter is to describe and document the procedures used in generating the State and Composite Reports. Section 11.2 discusses the issues pertaining to the generation and proper interpretation of both the State and Composite Reports. Section 11.3 describes the logical and statistical rules developed to guide the production of the computer-generated State report.

The production of the State and Composite Reports involved the creation of several composite variables derived from responses to questions in the student, teacher, or school questionnaires. A description of these and other NAEP derived variables is included in Appendix E.

[3] Some of the items in the Mathematics Teacher and School Questionnaires were open-ended. Results were reported in a slightly different format for those open-ended items.

11.2 MINIMUM SAMPLE SIZES FOR REPORTING SUBGROUP RESULTS

In both reports, estimates of quantities such as the composite and content area proficiency means, percentage of students above the four anchor points, and the percentages of students indicating particular levels of background variables (as measured in the Student, Teacher, and School Questionnaires) were reported for the total population of eighth-grade students in each jurisdiction, as well as for certain key subgroups. The subgroups were defined by four standard NAEP reporting variables-- race/ethnicity, type of community, parents' education, and gender. NAEP maintains separate results for five racial/ethnic subgroups (White, Black, Hispanic, Asian American/Pacific Islander, and American Indian), four types of communities (Advantaged Urban, Disadvantaged Urban, Extreme Rural, and Other Non-extreme Communities), and four levels of parental eduction (high school nongraduate, high school graduate, some college, college graduate). However, in some jurisdictions, and for some regions of the country, sample sizes were not large enough to permit accurate estimation of proficiency and/or background variable results for one or more of the five racial/ethnic subgroups or of the four types of communities[4]. Results were provided for only those subgroups with sufficiently large sample sizes.

For results to be reported for a racial/ethnic subgroup or type of community, a minimum sample size of 62 individuals was required. This number was arrived at by determining the sample size, under simple random sampling, required to detect an effect size of .2 with a probability of .8 or greater. The effect size of .2 pertains to the "true" difference in mean proficiency between the subgroup in question and the total eighth-grade public school population in the state, divided by the standard deviation of proficiency in the total population. An effect size of .2 was chosen following Cohen (1977) who classifies effect sizes less than .2 as "small".

Under simple random sampling, if the true difference between subgroup and total group means is .2 total-group standard deviation units, a sample size of 31 would be required to detect such a difference with a probability of .8. However, as described in Chapter 3 of this Technical Report, the sampling for the Trial State Assessment Program was based on a multi-stage sampling procedure, not on simple random sampling. As a result, the standard errors of statistics are larger than they would be for a simple random sample of equivalent size. The ratio of the standard errors that take the sample design into account to standard errors assuming simple random sampling is called the design effect. To take into account the sampling design used for the Trial State Assessment, an average design effect of 2 was assumed based on experience from previous NAEP assessments. Given the assumed average design effect, the required sample size based on simple random sampling was doubled, yielding the number 62.

Both the State and Composite Reports included large numbers of tables that provided estimates of the proportion of the students responding to each category of a background 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 a background variable was less than

[4] In all 40 jurisdictions, sufficient data were obtained for all levels of the parental education variable and for both males and females to permit reporting of results for these subgroups.

62. For those instances, the minimum sample size restriction of 62 was applied to these subgroups, and the resulting estimated mean proficiency was not reported.

11.3 ESTIMATES OF STANDARD ERRORS WITH LARGE MEAN SQUARED ERRORS

Standard errors of mean proficiencies 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. In certain cases, typically when the number of students upon which the standard error is based is small or when this group of students all come from a small number of schools that participated in the assessment, the mean squared error[5] associated with the estimated standard errors may be quite large. Throughout the State and Composite Reports, estimated standard errors 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 CV 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 .2, the mean squared error of the estimated standard errors of means and proportions based on samples of this size may be quite large. Therefore, the standard errors of means and proportions for all subgroups for which the CV of the population size exceeded .2 are followed by a "!" in the tables of both the State and Composite Reports. These standard errors, and any confidence intervals or significance tests involving these standard errors, should be interpreted cautiously.

[5] The mean squared error of the estimated standard error is defined as $\mathcal{E}[\hat{S} - \sigma]^2$, where $\hat{S}$ is the estimated standard error, σ is the "true" standard error, and $\mathcal{E}$ is the expectation operator.

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

Responses to the Student, Teacher, and School Questionnaires played a prominent role in both reports. Although the return rate on all three questionnaires was high, there were missing data from each questionnaire.

For all three questionnaires, 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 was available. In the terminology of Little and Rubin (1987), the analyses pertaining to a particular background variable presented in the State and Composite Reports 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 is likely the case, the assumptions are not correct. There was sufficiently little missing data (usually, less than two percent) for most of the variables obtained from the Student and School questionnaires to presume that the amount of potential nonresponse bias was tolerably small. However, for particular background questions from the Student and School Questionnaires, the level of nonresponse in certain jurisdictions was somewhat higher than the level of nonresponse observed on average. Background questions for which more than 10 percent of the returned questionnaires were missing are identified in Background Almanacs produced for each jurisdiction. Results for analyses involving these questions should be interpreted with caution.

To analyze the data from the mathematics teacher questionnaires with respect to the students' data, each teacher's questionnaire had to be matched to all of the students who were taught mathematics by that teacher. Table 11-1 provides the percentages of students that were matched to Teacher Questionnaires for each of the 40 jurisdictions that participated in the Trial State Assessment. Two separate match rates are given. The first is the percentage of students that could be matched to both the first and second parts of the Teacher Questionnaire. The second match rate is the percentage of students that could be matched to only the first part of the Teacher Questionnaire. 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 Background Almanacs produced for each jurisdiction.

TABLE 11-1

Teacher Questionnaire Match Rates (Percents) By State

State/Territory	No[1] Match	Partial[2] Match	Complete[3] Match
Alabama	1.9	3.8	94.3
Arizona	8.2	7.4	84.4
Arkansas	1.7	6.1	92.2
Califorina	6.2	8.4	85.4
Colorado	7.3	7.7	85.0
Connecticut	2.8	9.0	88.2
Delaware	5.6	9.8	84.5
District of Columbia	2.6	3.0	94.4
Florida	5.8	6.0	88.2
Georgia	6.2	5.1	88.7
Guam	0.4	1.9	97.7
Hawaii	4.4	4.1	91.6
Idaho	5.8	7.3	87.0
Illinois	2.4	12.7	84.9
Indiana	6.2	7.1	86.7
Iowa	5.4	6.6	88.0
Kentucky	1.8	4.6	93.6
Louisiana	4.1	5.5	90.4
Maryland	2.7	7.6	89.7
Michigan	4.1	4.8	91.1
Minnesota	6.0	6.0	88.0
Montana	2.1	4.4	93.5
Nebraska	4.3	8.2	87.5
New Hampshire	6.1	7.6	86.3
New Jersey	3.1	6.3	90.6
New Mexico	3.6	6.2	90.2
New York	9.2	5.1	85.7
North Carolina	2.0	6.6	91.4
North Dakota	2.9	4.1	93.1
Ohio	4.8	12.1	83.1
Oklahoma	5.3	3.5	91.3
Oregon	8.0	8.4	83.6
Pennsylvania	2.6	10.2	87.3
Rhode Island	4.7	9.6	85.6
Texas	8.7	6.7	84.6
Virginia	3.0	4.4	92.6
Virgin Islands	8.0	4.1	87.9
West Virginia	2.9	6.3	90.8
Wisconsin	5.2	6.7	88.1
Wyoming	6.7	9.1	84.2

[1] No Match: Students with no teacher code or no teacher questionnaire returned.

[2] Partial Match: Students who match on teacher background, but do not match classroom period.

[3] Complete Match: Students who match on teacher background and classroom period. Also, includes cases when only 1 classroom period exists on questionnaire.

11.5 STATISTICAL RULES USED FOR PRODUCING THE STATE REPORTS

As described earlier, the State reports contained state-level estimates of mean proficiencies (and the proportions of students above selected scale points) for the state as a whole and for categories of key reporting variables such as gender, race/ethnicity, level of parental education, and type of community. In addition to these key reporting variables, results were reported for a variety of other subpopulations based on variables taken directly from, or derived from, the Student, Teacher, and School Questionnaires, as well as from school and community demographic variables provided by Westat. Similar estimates of means and proportions by subgroup were provided for the nation and, where sample sizes permitted, for the region to which each state belongs[6]. The State Reports were produced entirely by computer. 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.

The State Reports are mainly descriptive reports. Estimates of the mean proficiency, the percentages of students above scale anchor points, and the percentages of students responding in particular ways to background questions were provided for the total population of eighth-grade public-school students in the jurisdiction, as well as for demographic subgroups typically of interest to educators and policy makers. Similar results were reported for the nation and region. The principal way in which the results were presented was through a sequence of figures and tables containing estimated means and proportions, along with their standard errors.

Computer-generated interpretative text was also provided in addition to the graphical and tabular presentation of results. In some cases, the computer-generated interpretative text was primarily descriptive in nature and reported the total group and subgroup proficiency means and proportions of interest. However, some of the interpretative text was intended to focus the reader on interesting and potentially important group differences in mathematics proficiency differences or on the percentages of students responding in particular ways to the background questions. For example, one question of considerable interest to each jurisdiction was whether, on average, its students performed higher than, lower than, or about the same as students in the nation. Another question of interest was whether students from disadvantaged urban areas were less likely to be enrolled in an eighth-grade algebra course than were students from advantaged urban areas. Other interpretive text was intended to focus on potentially interesting patterns of achievement across the five mathematics content areas or on the pattern of response to a particular background question in the state. For example, do more students report spending 30 minutes or 15 minutes on homework each day?

Rules for the production of the computer-generated text for questions involving the comparison of results for subgroups and interpretations of patterns of results were developed for the State Report. The 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 which were within the limitations imposed by the availability of computational

[6] Because United States Territories are not classified into NAEP regions, no regional comparisons were provided for Guam and the Virgin Islands.

resources and the time frame for the production of the report. The following sections describe these procedures and rules.

11.5.1 Comparing Means and Proportions for Mutually Exclusive Groups of Students

Many group comparisons that were commented upon in the State Reports involved contrasting the mean proficiencies, proportions of students above anchor points, or proportions of students responding to a background question in a particular way for mutually exclusive sets of students. One common example of such a comparison is the contrast between the mean composite proficiency in a particular state to the mean composite proficiency in the nation. Other examples include comparisons of the average proficiency (or proportions of students above anchor points, or the proportions of students responding in a particular way to a background question) for: 1) males and females within a jurisdiction, 2) White students and Hispanic students within a jurisdiction, 3) students from advantaged urban schools and disadvantaged urban schools, and, 4) students who reported watching six or more hours of television each night and students who report watching less than one hour each night.

Throughout the State Report, computer-generated text indicated that means or proportions from two groups were different only when the differences in the point estimates for the groups being compared were statistically significant at an approximate α level of .05. An approximate procedure was used for determining statistical significance which 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:

$$|t_i - t_j| \geq Z_{\frac{.025}{c}} \sqrt{S\hat{E}^2(t_i) + S\hat{E}^2(t_j)}$$

where Z_α is the (1 - α) percentile of the standard normal distribution, and c is the number of contrasts being tested. In most cases, group comparisons were treated as individual units. Therefore, the value of c was taken as one, and the test statistic was equivalent to a standard two-tailed z-test for the difference between group means or proportions from independent samples with the α level set at .05.

Frequently in the State Reports, a group of comparisons were made and presented as a single set. For these sets of contrasts, a Bonferroni procedure was used for determining the value of Z_α, where c was the number of contrasts in the set.

The procedure described above was used for testing differences of both means *and* proportions. The normal 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 cautiously if at least one of the groups being compared is small in size and/or somewhat extreme proportions are being compared. When the proportions for groups being compared exceeded .9 or were below .1, the statistical significance of differences between groups was not tested, and the accompanying computer-generated text was only descriptive.

The procedures described above 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 state may not be strictly independent. Therefore, the significance tests employed are, in many cases, only approximate. As described in the next section, another procedure that does not assume independence could have been conducted. However, the procedure is computationally burdensome and resources precluded its application for all the comparisons in the State Reports. It was the judgement 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.

When single comparisons were being 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 effect sizes was used. The effect size for comparing means from two groups was defined as:

$$\textit{effect size} = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{\sqrt{\frac{S_i^2 + S_j^2}{2}}}$$

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 5 plausible values for the students in subgroup i and were calculated using the Westat sampling weights.

Following Cohen (1977), the effect size for comparing proportion was defined as:

$|f_i - f_j|$, where $f_i = 2 \arcsin\sqrt{p_i}$, 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 effect size exceeded .1. However, when a significant difference was found but the effect size was less than .1, the qualifier *somewhat* was used. For example, if the mean proficiency for females was significantly higher than that of males but the effect size of the difference was less than .1, females were described as performing *somewhat higher* than males.

11.5.2 Determining the Highest and Lowest Scoring Groups from a Set of Ranked Groups

Three analyses in the State Report consisted of determining which of a set of several groups had the highest and/or lowest proficiency among the set. For example, an analysis compared the average proficiency of students who reported watching various amounts of television each day. There were five levels of television watching (one hour or less, two hours, three hours, four to five hours, six or more hours). Based on their answers to this question in the Student Background Questionnaire, students were classified into one of the five levels of television watching, and the mean composite proficiency was obtained for students at each level. The analysis focussed 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 *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.

11.5.3 Comparing Dependent Proportions

Several 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, the NAEP staff determined that the dependencies in the data could not be ignored.

Unlike the case for the analyses of the type described in Section 11.5.1, 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, 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 the case for the procedures described in the previous section). Such underestimation can result in too many "nonsignificant" differences being identified as significant.

To avoid such differences being identified as significant, the procedures of Section 11.5.1 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 obtaining a separate estimate of the difference in question for each jackknife replicate, using the first plausible value only, and 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 the previous section only with respect to estimating the standard error of the difference. All other aspects of the procedures were identical to those described in the previous section.

11.5.4 Descriptions of the Magnitude of Percentage

Percentages reported in the text of the State Reports were sometimes described in qualitative fashion. For example, the number of students being taught by teachers with masters degrees in mathematics 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. A list of the rules used to select the descriptive phrases in the report are provided in Table 11-2.

Table 11-2

Rules for Selection Descriptions of Percentages

Percentage	Description of Text in Report
$p = 0$	None
$0 < p < 10$	Relatively few
$10 < p < 20$	Some
$20 < p < 30$	About one-quarter
$30 < p < 44$	Less than half
$44 < p < 55$	About half
$55 < p < 69$	More than half
$69 < p < 79$	About three-quarters
$79 < p < 89$	Many
$89 < p < 99$	Almost all
$p=100$	All

GLOSSARY OF TERMS

GLOSSARY OF TERMS

anchoring. The process of characterizing score levels in terms of predicted observable behavior.

assessment session. The period of time during which a NAEP booklet is administered to one or more individuals.

background questionnaires. The instruments used to collect information about students' demographics and educational experiences.

bias. In statistics, the difference between the expected value of an estimator and the population parameter being estimated. If the average value of the estimator over all possible samples (the estimator's expected value) equals the parameter being estimated, the estimator is said to be **unbiased**; otherwise, the estimator is **biased**.

BIB (Balanced Incomplete Block) spiraling. A complex variant of multiple matrix sampling, in which items are administered in such a way that each pair of items is administered to a nationally representative sample of respondents.

BILOG. A computer program for estimating item parameters.

block. A group of assessment items created by dividing the item pool for an age/grade into subsets. Used in the implementation of the BIB spiral sample design.

booklet. The assessment instrument created by combining blocks of assessment items.

calibrate. To estimate the parameters of a set of items from responses of a sample of examinees.

clustering. The process of forming sampling units as groups of other units.

coefficient of variation. The ratio of the standard deviation of an estimate to the value of the estimate.

common block. A group of background items included in the beginning of every assessment booklet.

conditional probability. Probability of an event, given the occurrence of another event.

conditioning variables. Demographic and other background variables characterizing a respondent. Used in construction of plausible values.

degrees of freedom. [of a variance estimator] The number of independent pieces of information used to generate a variance estimate.

derived variables. Subgroup data that were not obtained directly from assessment responses, but through procedures of interpretation, classification, or calculation.

design effects. The ratio of the variance for the sample design to the variance for a simple random sample of the same size.

distractor. An incorrect response choice included in a multiple-choice item.

excluded student questionnaire. An instrument completed for every student who was sampled but excluded from the assessment.

excluded students. Sampled students determined by the school to be unable to participate because they have limited English proficiency, are mildly mentally retarded (educable), or are functionally disabled.

expected value. The average of the sample estimates given by an estimator over all possible samples. If the estimator is unbiased, then its expected value will equal the population value being estimated.

field test. A pretest of items to obtain information regarding clarity, difficulty levels, timing, feasibility, and special administrative situations; performed before revising and selecting items to be used in the assessment.

focused-BIB spiraling. A variation of BIB spiraling in which items are administered in such a way that each pair of items *within a subject area* is administered to a nationally representative sample of respondents.

foils. The correct and incorrect response choices included in a multiple-choice item.

group effect. The difference between the mean for a group and the mean for the nation.

imputation. Prediction of a missing value according to some procedure, using a mathematical model in combination with available information. See **plausible values.**

imputed race/ethnicity. The race or ethnicity of an assessed student, as derived from his or her responses to particular common background items. A NAEP **reporting subgroup.**

item response theory (IRT). Test analysis procedures that assume a mathematical model for the probability that a given examinee will respond correctly to a given exercise.

jackknife. A procedure to estimate standard errors of percentages and other statistics. Particularly suited to complex sample designs.

major strata. Used to stratify the primary sampling frame within each region. Involves stratification by size of community and degree of ruralization (SDOC).

master catalog. Computer processing control information, IRT parameters, foil codes, and labels in a computer-readable format.

Metropolitan statistical area (MSA). An area defined by the federal government for the purposes of presenting general-purpose statistics for metropolitan areas. Typically, an MSA contains a city with a population of at least 50,000 plus adjacent areas.

multistage sample design. Indicates more than one stage of sampling. An example of three-stage sampling: 1) sample of counties (primary sampling units or PSUs); 2) sample of schools within each sample county; 3) sample of students within each sample school.

multiple matrix sampling. Sampling plan in which different samples of respondents take different samples of items.

NAEP scales. The anchored scales common across age/grade levels and assessment years used to report NAEP results.

nonresponse. The failure to obtain responses or measurements for all sample elements.

nonsampling error. A general term applying to all sources of error except sampling error. Includes errors from defects in the sampling frame, response or measurement error, and mistakes in processing the data.

objective. A desirable education goal agreed upon by scholars in the field, educators, and concerned laypeople, and established through the consensus approach.

observed race/ethnicity. Race or ethnicity of an assessed student as perceived by the exercise administrator.

open-ended response item. A nonmultiple-choice item that requires some type of written or oral response.

oversampling. Deliberately sampling a portion of the population at a higher rate than the remainder of the population.

parental education. The level of education of the mother and father of an assessed student as derived from the student's response to two assessment items. A NAEP **reporting subgroup**.

percent correct. The percent of a target population that would answering a particular exercise correctly.

plausible values. Proficiency values drawn at random from a conditional distribution of a NAEP respondent, given his or her response to cognitive exercises and a specified subset of background variables (conditioning variables). The selection of a plausible value is a form of **imputation**.

poststratification. Classification and weighting to correspond to external values of selected sampling units by a set of strata definitions after the sample has been selected.

Principal's Questionnaire. A questionnaire sent to every sampled school that agreed to participate in the Trial State Assessment. It requested aggregate information on enrollment by grade, race, and ethnicity of the student population, community size, and the distribution of employment status of parents of attending students.

primary sampling unit (PSU). The basic geographic sampling unit for NAEP. Either a single county or a set of contiguous counties.

principal questionnaire. A data collection form given to school principals before assessments. The principals respond to questions concerning enrollment, size and occupational composition of the community, etc.

probability sample. A sample in which every element of the population has a known, nonzero probability of being selected.

pseudoreplicate. The value of a statistic based on an altered sample. Used by the **jackknife** variance estimator.

QED. Quality Education Data, Inc. A supplier of lists of schools, school districts, and other school data.

random variable. A variable that takes on any value of a specified set with a particular probability.

region. One of four geographic areas used in gathering and reporting data: Northeast, Southeast, Central, and West (as defined by the Office of Business Economics, U.S. Department of Commerce). A NAEP **reporting subgroup.**

reporting subgroup. Groups within the national population for which NAEP data are reported: for example, gender, race/ethnicity, grade, age, level of parental education, region, and size and type of community.

respondent. A person who is eligible for NAEP, is in the sample, and responds by completing one or more items in an assessment booklet.

response options. In a multiple-choice question, alternatives that can be selected by a respondent.

sample. A portion of a population, or a subset from a set of units, selected by some probability mechanism for the purpose of investigating the properties of the population. NAEP does not assess an entire population but rather selects a representative sample from the group to answer assessment items.

sampling error. The error in survey estimates that occurs because only a sample of the population is observed. Measured by sampling **standard error.**

sampling frame. The list of sampling units from which the sample is selected.

sampling weight. A multiplicative factor equal to the reciprocal of the probability of a respondent being selected for assessment with adjustment for nonresponse and perhaps also for poststratification. The sum of the weights provides an estimate of the number of persons in the population represented by a respondent in the sample.

school characteristics and policy questionnaire. A questionnaire completed for each school by the principal or other official; used to gather information concerning school administration, staffing patterns, curriculum, and student services.

selection probability. The chance that a particular sampling unit has of being selected in the sample.

session. A group of students reporting for the administration of an assessment. Most schools conducted only one session, but some large schools conducted as many as 10 or more.

simple random sample. Process for selecting n sampling units from a population of N sampling units so that each sampling unit has an equal chance of being in the sample and every combination of n sampling units has the same chance of being in the sample chosen.

standard error. A measure of sampling variability and measurement error for a statistic. Because of NAEP's complex sample design, sampling standard errors are estimated by **jackknifing** the samples from first-stage sample estimates. Standard errors may also include a component due to the error of measurement of individual scores estimated using plausible values.

stratification. The division of a population into parts, called strata.

stratified sample. A sample selected from a population that has been stratified, with a sample selected independently in each stratum. The strata are defined for the purpose of reducing sampling error.

student ID number. A unique identification number assigned to each respondent to preserve his or her anonymity. NAEP does not record the names of any respondents.

subject area. One of the areas assessed by National Assessment; for example, art, civics, computer competence, geography, literature, mathematics, music, reading, science, U.S. history, or writing.

systematic sample (systematic random sample). A sample selected by a systematic method; for example, when units are selected from a list at equally spaced intervals.

teacher questionnaire. A questionnaire completed by selected teachers of sample students; used to gather information concerning years of teaching experience, frequency of assignments, teaching materials used, and availability and use of computers.

Trial State Assessment Program. The NAEP program, authorized by Congress in 1988, which was established to provide for a program of voluntary state-by-state assessments on a trial basis.

trimming A process by which extreme weights are reduced (trimmed) to diminish the effect of extreme values on estimates and estimated variances.

type 1 Cluster. Individual schools in states where all eighth-grade schools were included in the sample.

type 2 Cluster. A school or group of schools in states where small schools were grouped geographically with large ones.

type 3A Cluster. A large school in states where large and small schools were stratified and sampled separately.

type 3B Cluster. A group of small schools in states where large and small schools were stratified and sampled separately.

type of community (TOC). One of the NAEP **reporting subgroups**, dividing the communities in the nation into four groups on the basis of the proportion of the students living in each of three sizes of communities and on the percentage of parents in each of five occupational categories.

variance. The average of the squared deviations of a random variable from the expected value of the variable. The variance of an estimate is the squared standard error of the estimate.

REFERENCES CITED IN TEXT

REFERENCES CITED IN TEXT

Beaton, A.E. & Johnson, E.G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15,* 9-38.

Braun, H.I. & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & H.I. Braun (Eds.), *Test equating.* New York, NY: Academic Press.

Cochran, W.G. (1977). *Sampling techniques*. New York, NY: John Wiley & Sons.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York, NY: Academic Press.

Dossey, J. A., Mullis, I.V.S., Lindquist, M.M., and Chambers, D.L. (1988). *The mathematics report card: are we measuring up?* Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Educational Testing Service (1987). *ETS standards for quality and fairness.* Princeton, NJ: Educational Testing Service.

Harris, R.J. (1975). *A primer of multivariate statistics*. New York, NJ: Academic Press.

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.

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. *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.

Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359-381.

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993-997.

Mislevy, R.J. (1988). *Randomization-based inferences about latent variables from complex samples.* (ETS Research Report RR-88-54-ONR) Princeton, NJ: Educational Testing Service.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika.*

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. & Bock, R.D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software.

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-TA-20) Princeton, NJ: National Assessment of Educational Progress, 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.

National Assessment of Educational Progress. (1988). *Mathematics objectives: 1990 assessment.* Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1987). *Mathematics objectives: 1985-86 assessment.* Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1990). *1990 policy information framework.* Princeton, NJ: Educational Testing Service.

National Council of Teachers of Mathematics. (1987). *Curriculum and evaluation standards for school mathematics.* Reston, VA.

Raisen, S. and Jones, L., (Eds.), (1985). *Indicators of precollege education in science and mathematics: A Preliminary Review.* Washington, DC: National Academy Press.

Reinsh, C.H. (1967). Smoothing by spline functions. *Numerische Mathematik, 10,* 177-183.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Sheehan, K.M. (1985). *M-GROUP: Estimation of group effects in multivariate models* [Computer program] Princeton, NJ: Educational Testing Service.

Tobias, S. (1987). *Succeed with math: Every student's guide to conquering mathematics anxiety.* New York: The College Entrance Examination Board.

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., & Muraki, E.J. (1990). *Effect of non-normal ability distributions on item parameter estimation.* Unpublished manuscript.

APPENDIX A

PARTICIPANTS IN THE OBJECTIVE AND ITEM DEVELOPMENT PROCESS

APPENDIX A

PARTICIPANTS IN THE OBJECTIVES AND ITEM DEVELOPMENT PROCESS

The National Assessment of Educational Progress extends its deep appreciation to all those individuals who participated in the development of the framework, objective, and items for the Trial State Assessment Program.

NATIONAL ASSESSMENT PLANNING PROJECT

Steering Committee

Robert Astrup	National Education Association
Lillian Barna	Council of the Great City Schools
Richard A. Boyd	Council of Chief State School Officers
Glenn Bracht	Council for American Private Education and National Association of Independent Schools
William M. Ciliate	National School Boards Association
Antonia Cortese	American Federation of Teachers
Mary Brian Costello	National Community on Catholic Education Association
Wilhelmina Delco	National Council of State Legislators
Nancy DiLaura	National Governors' Association
Thomas Fisher	Association of State Assessment Programs
Alice Houston	Association for Supervision and Curriculum Development
C. June Knight	National Association of Elementary School Principals
Stephen Lee	National Association of Secondary School Principals
Paul LeMahieu	National Association of Test Directors
Glen Ligon	Directors of Research and Evaluation
Barbara Roberts Mason	National Association of State Boards of Education
James E. Morrell	American Association of School Administrators Austin Independent School District, Texas

Mathematics Objectives Committee

Joan Burks	Damascus High School, Damascus, Maryland
Phillip Curtis	University of California at Los Angeles, Los Angeles, California
Walter Denham	California Department of Education, Sacramento, California
Thomas Fisher	Florida Department of Education, Tallahassee, Florida
Ann Kahn	The National Parent-Teacher Association, Fairfax, Virginia
Mary M. Lindquist	Columbus College, Columbus, Georgia
Susan Purser	Whitten Junior High School, Jackson, Mississippi
Dorothy Strong	Chicago Public Schools, Chicago, Illinois
Thomas W. Tucker	Colgate University, Hamilton, New York
Charles Watson	Arkansas Department of Education, Little Rock, Arkansas
O. R. Wells, Jr.	Rice University, Houston, Texas

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Item Development Panel

Bruce Brombacher	Jones Junior High School, Westerville, Ohio
Iris Carl	Houston Independent School District, Houston, Texas
John Dossey	Illinois State University, Normal, Illinois
Linda Foreman	Portland State University, Portland, Oregon
Audrey Jackson	Parkway School District, Chesterfield, Missouri
Jeremy Kilpatrick	University of Georgia, Athens, Georgia
Mary Lindquist	Columbus College, Columbus, Georgia
Thomas Tucker	Colgate University, Hamilton, New York

Test Development Consultants

James Braswell	College Board Programs, Educational Testing Service
Jeanne Elbich	College Board Programs, Educational Testing Service
Jeffrey Haberstroh	College Board Programs, Educational Testing Service
Chancey Jones	College Board Programs, Educational Testing Service
Jane Kupin	College Board Programs, Educational Testing Service
Marlene Supernavage	College Board Programs, Educational Testing Service
Beverly Whittington	College Board Programs, Educational Testing Service

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Scale Anchoring Panel

Laurie Boswell	Profile High School, Bethlehem, New Hampshire
Bruce Brombacher	Jones Junior High School, Westerville, Ohio
Catherine Brown	Virginia Polytechnic Institute, Blacksburg, Virginia
Joe Crosswhite	Springfield, Missouri
John Dossey	Illinois State University, Normal, Illinois
Henry Kepner, Jr.	University of Wisconsin at Milwaukee, Milwaukee, Wisconsin
Linda Kolnowski	Detroit Public Schools, Detroit, Michigan
Gordon Lewis	Washington DC Public Schools, Washington, D.C.
Mary M. Lindquist	Columbus College, Columbus, Georgia
Donna Long	Indiana Dept. of Education, Indianapolis, Indiana
Vena Long	Missouri Department of Education, Jefferson City, Missouri
William Masalski	University of Massachusetts, Amherst, Massachusetts
Wendell Meeks	Illinois State Board of Education, Springfield, Illinois
Andy Reeves	Florida Department of Education, Tallahassee, Florida
Diane Thiessen	University of Northern Iowa, Cedar Falls, Iowa
Alba Thompson	CRMSE, San Diego, CA 92120
Shiela Vice	Kentucky Department of Education, Frankfort, Kentucky
Charles Watson	Arkansas Department of Education, Little Rock, Arkansas
Vernon Williams	H.W. Longfellow Intermediate School, Falls Church, Virginia

APPENDIX B

SUMMARY OF PARTICIPATION RATES

Introduction

State representatives, the National Assessment Governing Board (NAGB), and several committees of external advisors to NAEP and NCES have engaged in numerous discussions about the procedures for reporting the NAEP Trial State Assessment results. As part of these discussions, it was recognized that sample participation rates across the states would need to be uniformly high to permit fair and valid comparisons. Therefore, NCES decided to establish guidelines for levels of school and student participation in the state assessments. In the event that any state had participation levels that did not meet the guidelines, a notation would be made in the state reports.

Virtually every state met or exceeded the four guidelines that were established. However, an explanation of the guidelines is provided in the first section of this appendix for use in interpreting the attached information on the 1990 data collection experience and for reference in preparing for the 1992 state assessments. The guidelines are based on the standards for sample surveys that are set forth in the U.S. Department of Education's Standards and Policies (1987). In brief, they cover levels of school and student participation, both overall and for particular population subgroups.

The procedures used to derive the weighted school and student participation rates are explained in the second and third sections of this appendix. Consistent with the NCES standards, weighted data are to be used to calculate all participation rates for sample surveys, and weighted rates will be provided in the final reports. However, the unweighted participation rates, based on the counts of schools and students, also are provided in the enclosed information summarizing the field data collection experience in your state.

The last section of the Appendix consists of an update of the earlier draft of the combined counts across all states that participated in the 1990 Trial State Assessment. The preliminary version of this information was initially distributed at the April 30, 1990 NETWORK meeting. Because the aggregate across all states is not representative of any meaningful sample, the weighted participation rates across states have not been analyzed. However, the overall counts do provide some context for interpreting the summary of activities in each individual state.

Finally, it should be explained that in several states, a few materials were lost by the U.S. Postal Service or did not arrive back from the field. While these situations had minimal effect on the participation rates, if they occurred there is a notation in the state summary. Documentation also is provided in the state summary about any other unusual circumstances or occurrences that may have affected the participation rates in some small way.

Notations for Use in Reporting Trial State Assessment School and Student Participation Rates

The following notations concerning school and student participation rates in the NAEP state assessments were established to address four significant ways in which nonresponse bias could be introduced into the state sample estimates. Nonresponse bias can occur if data are not obtained from portions of the state population to the extent that overall sample representativeness could be affected.

1. **Both the weighted participation rate for the initial sample of schools was below 85 percent AND the weighted school participation rate after substitution was below 90 percent.**

All states participating in the 1990 Trial State Assessment had school participation rates that exceeded either the first or second part of the guideline (or both parts). Thus, this note will not be used. However, as mentioned previously, an explanation is provided below for reference purposes.

For states that did not use substitute schools, the 1990 Trial State Assessment results will be based on participating schools from the original sample. In these situations, the NCES standards specify weighted school participation rates of 85 percent or better to guard against potential bias due to school nonresponse. Thus, the first part of this guideline, which refers 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 state participating in the 1990 Trial State Assessment Program, NAEP provided substitutes for nonparticipating schools. When possible, a substitute school was provided for each initially selected school that declined participation before November 10, 1989. Thus, for states that did use substitute schools, the 1990 Trial State Assessment results will be based on all participating schools from both the original sample and the list of substitutes.

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 guideline was set at 90 percent.

2. **The nonparticipating schools included a group of schools with similar characteristics, who together accounted for more than 5 percent of the state's total eighth-grade population in public schools. The types of schools from which a state needed minimum levels of student representation were determined by urbanicity, minority enrollment, and median family income.**

The NCES standards also specify that attention should be given to the representativeness of the sample coverage. Thus, if some important segment of the state's population is not adequately represented, it is of concern, regardless of the overall participation rate. Virtually all states met this guideline.

This notation addresses 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 strata have been formed within each state, and the schools within each stratum are similar with respect to minority enrollment, urbanicity, and/or median household income, as appropriate for each state.

If more than 5 percent (weighted) of the sampled schools are nonparticipants from a single adjustment stratum, then the potential for nonresponse bias may be too great. (The weight of a sampled school estimates the number of students in the population who are represented by that school.) This guideline is based on the NCES standard for stratum-specific school nonresponse rates.

3. **The weighted student response rate within participating schools was below 85 percent.**

This guideline follows 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 makeup session. (The method used to calculate the weighted student participation rate is defined in the next section.) If the rate falls below 85 percent, then the potential for bias due to student nonresponse may be too great. Again, however, virtually all states who participated in the 1990 Trial State Assessment met this guideline.

4. **The nonresponding students within participating schools included a group of students with similar characteristics, who together accounted for more than 5 percent of the state's assessable public-school population. Student groups from which a state needed minimum levels of participation were determined by age of respondent and type of assessment session (unmonitored or monitored), as well as school urbanicity, minority enrollment, and median family income.**

All states met this guideline. However, this notation would 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 strata have been formed using the school-level nonresponse adjustment strata, together with the student's age and the nature of the assessment session (unmonitored or monitored). If more than 5 percent (weighted) of the invited students who do not participate in the assessment are from a single stratum, then the potential for nonresponse bias may be too great. This guideline is 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 state give the percentages of eighth-grade students in public schools who are represented by the schools participating in the assessment, prior to statistical adjustments for school nonresponse. Two weighted school participation rates are computed for each state. The first rate is based only on participating schools that were initially selected for the assessment, while the second rate includes schools selected as substitutes for nonparticipating schools. The numerator in the before-substitution rate is the sum of the number of students represented by each initially selected school that participated in the assessment. The numerator in the after-substitution rate is the sum of the number of students represented by each of the initially selected participating schools and each of the participating substitute schools. The denominator of both rates is the sum of the number of students represented by each of the initially selected schools (both participating and nonparticipating)--an estimate of the total number of eighth-grade students in the state's public schools.

In general, different schools in the sample can represent different numbers of students in the state population. The number of students represented by an initially selected school (the school weight) is the eighth-grade enrollment of the school divided by the probability that the school was included in the sample. The number of students represented by a substitute school is the number of students represented by the replaced nonparticipating school. A school with a selection probability of less than 1.0 represents more students in the population than its enrollment, while a school with a selection probability of 1.0 represents only the students attending that school. Thus, a selected school with an eighth-grade enrollment of 150 and a selection probability of 0.2 represents 750 students from that state, while a school with an enrollment of 1,000 and a selection probability of 1.0 represents only the 1,000 students attending that school itself.

Because each school represents different numbers of students in the population, the weighted school participation rates differ somewhat from the simple unweighted rates. (The unweighted rates are calculated from the initial raw counts by dividing the number of participating schools by the number of schools in the sample.) The difference between the weighted and the unweighted rates is potentially largest in smaller states where all schools with eighth-grade students were included in the sample. In those states, 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 a greater impact than that of smaller schools on reducing weighted school participation rates in larger states where less 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 states where all schools were selected.

In general, the larger the state, the less the difference is between the weighted and unweighted school participation rates. However, even in the smaller states, the differences tend to be small--typically within one percentage point. Furthermore, in the 1990 Trial State Assessment, whenever the difference exceeded one percentage point, it was always because the weighted participation rate exceeded the unweighted rate.

Weighted Student Participation Rate

The weighted student participation rate shows the percentage of the eligible student population within the state that is represented by the students who participated in the assessment (in either an initial session or a make-up session), after accounting for school nonparticipation. The eligible student population within a state consists of all public-school students who were in the eighth grade and who, if selected, would not be 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. 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. In other words, the denominator is an estimate of the total number of assessable students in the state.

The number of students represented by a single selected student (the student weight) is 1.0 divided by the probability that the student was selected for assessment, with adjustments to account for nonparticipation of schools. In general, each sampled student within a state represents approximately the same number of students from that state's population. Consequently, there is little difference between the weighted student participation rate and the unweighted student participation rate.

Additional Weighted Percentages Included in the Summaries

Weighted Percentage of Excluded Students

The weighted percentage of excluded students estimates the percentage of the eighth-grade population in the state's public 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 state's schools.

Weighted Percentage of Individualized Education Plan (IEP) Students

The weighted percentage of IEP students estimates the percentage of the eighth-grade population in the state's public schools that is 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 state's schools.

Weighted Percentage of Excluded IEP Students

The weighted percentage of IEP students who were excluded estimates the percentage of students in the state that is represented by those IEP students who were 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 students who was sampled and had not withdrawn from the state's schools.

Weighted Percentage of Limited English Proficiency (LEP) Students

The weighted percentage of LEP students estimates the percentage of the eighth-grade population in the states public schools that is 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 state's schools.

Weighted Percentage of Excluded LEP Students

The weighted percentage of LEP students who were excluded estimates the percentage of students in the state this is represented by those LEP students who were 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 students who was sampled and had not withdrawn from the state's schools.

TABLE B.1 | **School Participation Rates**

GRADE 8 PUBLIC SCHOOLS	Weighted Percentage School Participation Before Substitution	Weighted Percentage School Participation After Substitution	Number Schools in Original Sample	Number Schools Not Eligible
NATION	88	92	145	13
Northeast	72	90	25	3
Southeast	94	94	40	1
Central	94	94	31	4
West	88	90	49	5
STATES				
Alabama	86	97	106	5
Arizona[3]	97	97	110	7
Arkansas	100	100	107	0
California	94	94	106	2
Colorado	100	100	107	2
Connecticut	100	100	108	5
Delaware[2]	100	100	30	0
District of Columbia[2]	100	100	36	0
Florida[2]	98	98	108	6
Georgia	100	100	109	3
Hawaii[2]	100	100	57	4
Idaho	97	97	108	2
Illinois	78	96	107	2
Indiana[3]	89	94	105	1
Iowa[1]	91	91	108	7
Kentucky	100	100	112	8
Louisiana	100	100	108	9
Maryland	100	100	107	2
Michigan	90	97	105	4
Minnesota	90	93	108	3
Montana	90	90	124	8
Nebraska	87	94	121	8
New Hampshire	91	97	107	3
New Jersey	97	98	112	3
New Mexico	100	100	108	2
New York[2]	86	86	105	0
North Carolina	100	100	111	5
North Dakota	96	100	111	5
Ohio	96	98	105	2
Oklahoma	78	99	112	0
Oregon	100	100	109	3
Pennsylvania	90	93	106	4
Rhode Island[2]	94	97	52	0
Texas[4]	88	97	107	4
Virginia	99	99	106	1
West Virginia	100	100	107	6
Wisconsin[2]	99	99	109	3
Wyoming[2]	100	100	69	0
TERRITORIES				
Guam[2]	100	100	7	1
Virgin Islands[2]	100	100	6	0

[1] The nonparticipating schools included a group of schools with similar characteristics, who together accounted for more than 5 percent of the state's eighth-grade population in public schools. The types of schools from which a state needed minimum levels of student representation were determined by urbanicity, minority enrollment, and median family income. See Appendix for explanations of the notations and guidelines about sample representativeness and for the derivation of weighted participation. [2] The Trial State Assessment was based on all eligible schools. There was no sampling of schools.

TABLE B.1 | **School Participation Rates (continued)**

GRADE 8 PUBLIC SCHOOLS	Number Schools in Original Sample that Participated	Number Substitute Schools Provided	Number Substitute Schools that Participated	Total Number Schools that Participated
NATION	117	15	3	120
Northeast	17	5	2	19
Southeast	35	4	0	35
Central	26	1	0	26
West	39	5	1	40
STATES				
Alabama	87	13	11	98
Arizona	102	0	0	102
Arkansas	107	0	0	107
California	98	0	0	98
Colorado	105	0	0	105
Connecticut	103	0	0	103
Delaware	30	0	0	30
District of Columbia	36	0	0	36
Florida	101	0	0	101
Georgia	106	0	0	106
Hawaii	53	0	0	52
Idaho	101	4	0	101
Illinois	82	21	19	101
Indiana	92	9	6	98
Iowa	92	9	0	92
Kentucky	104	0	0	104
Louisiana	99	0	0	99
Maryland	105	0	0	105
Michigan	90	9	8	98
Minnesota	94	5	3	97
Montana	100	4	0	100
Nebraska	94	10	9	103
New Hampshire	94	4	4	98
New Jersey	106	2	1	107
New Mexico	106	0	0	106
New York	91	0	0	91
North Carolina	106	0	0	106
North Dakota	98	8	8	106
Ohio	99	4	2	101
Oklahoma	85	26	23	108
Oregon	106	0	0	106
Pennsylvania	92	4	3	95
Rhode Island	49	2	2	51
Texas	92	10	9	101
Virginia	104	0	0	104
West Virginia	101	0	0	101
Wisconsin	106	0	0	106
Wyoming	69	0	0	69
TERRITORIES				
Guam	6	0	0	6
Virgin Islands	6	0	0	6

[3]For one school, an assessment was conducted, but the materials were destroyed in shipping via the U.S. Postal Service. The school was included in the counts of participating schools, both before and after substitution. However, in the weighted results, the school was treated in the same manner as a nonparticipating school because no student responses were available for analysis and reporting. In Arizona, materials for two schools were destroyed in shipping. [4]One school in the original sample initially declined and then decided to participate after a substitute for that school had been provided. Although the substitute school also participated, the state's estimates will be based on the sampling including the original school and not the substitute school.

TABLE B.2 | Student Participation Rates

GRADE 8 PUBLIC SCHOOLS	Weighted Percentage Student Participation After Make-ups	Number Students Original Sample	Number Students Supplemental Sample	Number Students Withdrawn
NATION[1]	0	0	0	0
Northeast	.0	0	0	0
Southeast	0	0	0	0
Central	0	0	0	0
West	0	0	0	0
STATES				
Alabama	95	2,908	99	186
Arizona	93	2,945	161	206
Arkansas	95	3,104	127	183
California	93	2,933	63	135
Colorado	94	3,074	103	192
Connecticut	95	3,085	58	115
Delaware	93	2,455	83	163
District of Columbia	88	2,758	72	237
Florida	92	3,005	148	209
Georgia	94	3,175	126	254
Hawaii	93	2,933	82	120
Idaho	96	2,941	90	123
Illinois	96	3,021	66	103
Indiana	95	2,910	81	143
Iowa	96	2,714	40	73
Kentucky	95	3,068	88	179
Louisiana	94	2,949	108	204
Maryland	94	3,151	82	115
Michigan	95	2,941	64	140
Minnesota	95	2,857	50	105
Montana	96	2,684	70	99
Nebraska	95	2,766	58	93
New Hampshire	95	2,870	52	80
New Jersey	94	3,149	83	113
New Mexico	94	3,091	122	236
New York	93	2,704	56	98
North Carolina	95	3,160	97	142
North Dakota	96	2,672	55	58
Ohio	95	3,030	90	138
Oklahoma[2]	80	3,007	107	194
Oregon	93	3,073	110	188
Pennsylvania[3]	94	2,849	51	77
Rhode Island	93	3,152	91	178
Texas	96	2,909	140	196
Virginia	94	3,120	85	195
West Virginia	94	3,008	77	152
Wisconsin	94	3,101	52	92
Wyoming	96	2,973	83	126
TERRITORIES				
Guam	93	1,810	82	58
Virgin Islands	93	1,490	1	16

[1] Because the national sampling is conducted for all subject areas assessed (reading and science in addition to mathematics in 1990) up to the stage of assignment different subject area booklets in assessment sessions, comparable information for just mathematics student sampling is not available for the nation and regions. [2] The weighted student response rate within participating schools was below 85 percent. Oklahoma, however, was the only state that required signed parental permission forms on a statewide basis. See Appendix for explanations of the notations and the guidelines about sample representativeness and for the derivation of weighted participation rates.

TABLE B.2 | **Student Participation (continued)**

GRADE 8 PUBLIC SCHOOLS	Number Students Excluded	Number Students to be Assessed	Number Students Assessed Initial Sessions	Number Students Assessed Make-ups	Total Number Students Assessed
NATION	741	0	0	0	9,922
Northeast	96	0	0	0	1,633
Southeast	119	0	0	0	2,752
Central	219	0	0	0	2,039
West	307	0	0	0	3,498
STATES					
Alabama	182	2,859	2,511	20	2,531
Arizona	158	2,742	2,480	78	2,558
Arkansas	244	2,804	2,840	29	2,869
California	242	2,619	2,353	71	2,424
Colorado	142	2,843	2,632	43	2,675
Connecticut	213	2,815	2,646	26	2,672
Delaware	122	2,253	2,052	58	2,110
District of Columbia	156	2,437	2,017	118	2,135
Florida	200	2,744	2,475	59	2,534
Georgia	117	2,930	2,736	30	2,766
Hawaii	151	2,744	2,452	99	2,551
Idaho	78	2,830	2,707	9	2,716
Illinois	171	2,813	2,637	46	2,683
Indiana	144	2,704	2,534	35	2,569
Iowa	104	2,577	2,462	12	2,474
Kentucky	158	2,819	2,660	20	2,680
Louisiana	130	2,723	2,544	28	2,572
Maryland	152	2,966	2,732	62	2,794
Michigan	129	2,736	2,524	63	2,587
Minnesota	87	2,715	2,537	47	2,584
Montana	69	2,588	2,459	27	2,486
Nebraska	84	2,647	2,497	22	2,519
New Hampshire	132	2,710	2,548	20	2,568
New Jersey	234	2,865	2,675	35	2,710
New Mexico	185	2,792	2,600	43	2,643
New York	171	2,491	2,242	60	2,302
North Carolina	107	3,008	2,791	52	2,843
North Dakota	91	2,578	2,483	2	2,485
Ohio	174	2,808	2,642	31	2,673
Oklahoma	184	2,758	2,208	14	2,222
Oregon	92	2,903	2,634	74	2,708
Pennsylvania	148	2,675	2,506	22	2,528
Rhode Island	208	2,857	2,633	42	2,675
Texas	196	2,657	2,525	17	2,542
Virginia	174	2,836	2,633	28	2,661
West Virginia	172	2,761	2,532	68	2,600
Wisconsin	145	2,916	2,705	45	2,750
Wyoming	106	2,824	2,662	39	2,701
TERRITORIES					
Guam	75	1,739	1,573	44	1,617
Virgin Islands	48	1,427	1,299	27	1,326

[3] For six students, the assessment was conducted, but the materials were destroyed in shipping via the U.S. Postal Service. Therefore, these students were treated in the same manner as absent students because no student responses were available for analysis and reporting.

TABLE B.3 | **Weighted Percentages of Students Excluded (IEP and LEP) from Original Sample**

GRADE 8 PUBLIC SCHOOLS	Total Percentage Students Identified IEP and LEP	Total Percentage Students Excluded	Percentage Students Identified IEP	Percentage Students Excluded IEP	Percentage Students Identified LEP	Percentage Students Excluded LEP
NATION	0	0	0	0	0	0
Northeast	0	0	0	0	0	0
Southeast	0	0	0	0	0	0
Central	0	0	0	0	0	0
West	0	0	0	0	0	0
STATES						
Alabama	10	6	10	6	0	0
Arizona	13	5	7	4	6	2
Arkansas	12	8	11	8	0	0
California	16	8	7	4	9	5
Colorado	10	5	9	4	1	1
Connecticut	12	7	10	6	2	1
Delaware	10	5	9	4	1	1
District of Columbia	7	6	5	5	1	1
Florida	12	7	9	5	3	2
Georgia	7	4	7	4	0	0
Hawaii	10	5	7	4	3	1
Idaho	7	3	6	2	1	0
Illinois	10	6	9	5	1	1
Indiana	8	5	7	5	0	0
Iowa	10	4	10	4	0	0
Kentucky	8	5	8	5	0	0
Louisiana	7	5	6	4	0	0
Maryland	11	5	10	4	1	1
Michigan	9	5	8	4	1	0
Minnesota	9	3	8	3	1	0
Montana	7	2	7	2	0	0
Nebraska	9	3	8	3	0	0
New Hampshire	12	5	12	5	0	0
New Jersey	13	8	10	6	2	2
New Mexico	10	7	9	6	2	1
New York	12	7	9	5	4	2
North Carolina	9	3	9	3	0	0
North Dakota	8	3	8	3	1	0
Ohio	8	6	8	6	0	0
Oklahoma	9	6	8	5	1	0
Oregon	9	3	8	3	1	0
Pennsylvania	11	8	10	5	1	0
Rhode Island	15	7	12	5	4	2
Texas	14	7	8	5	5	2
Virginia	10	6	8	5	2	1
West Virginia	10	6	10	6	0	0
Wisconsin	8	5	8	4	1	0
Wyoming	9	4	8	4	1	0
TERRITORIES						
Guam	7	4	5	4	2	1
Virgin Islands	4	3	4	3	0	0

IEP = Individual Education Plan and LEP = 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. Weighted percentages for the nation and region are based on students sampled for all subject areas assessed in 1990 (reading, science, and mathematics). However, based on the national sampling design the rates shown also are the best estimates for the mathematics assessment.

TABLE B.4 | Weighted Percentages of Absent, IEP, and LEP Students Based on Those Invited to Participate in the Assessment

GRADE 8 PUBLIC SCHOOLS	Weighted Percentage Student Participation After Make-Ups	Weighted Percentage Absent	Weighted Percentage Assessed IEP	Weighted Percentage Absent IEP	Weighted Percentage Assessed LEP	Weighted Percentage Absent LEP
NATION	0	0	0	0	0	
Northeast	0	0	0	0	0	0
Southeast	0	0	0	0	0	0
Central	0	0	0	0	0	0
West	0	0	0	0	0	0
STATES						
Alabama	95	5	92	8	100	0
Arizona	93	7	90	10	89	11
Arkansas	95	5	91	9	100	0
California	93	7	97	3	94	6
Colorado	94	6	92	8	100	0
Connecticut	95	5	93	7	100	0
Delaware	93	7	94	6	80	20
District of Columbia	88	12	92	8	0	0
Florida	92	8	88	12	79	21
Georgia	94	6	97	3	93	7
Hawaii	93	7	85	15	100	0
Idaho	96	4	97	3	100	0
Illinois	95	5	92	8	100	0
Indiana	95	5	93	7	100	0
Iowa	96	4	97	3	100	0
Kentucky	95	5	94	6	100	0
Louisiana	94	6	96	4	100	0
Maryland	94	6	88	12	100	0
Michigan	95	5	94	6	100	0
Minnesota	95	5	96	4	100	0
Montana	96	4	90	10	100	0
Nebraska	95	5	95	5	100	0
New Hampshire	95	5	95	5	100	0
New Jersey	94	6	88	12	94	6
New Mexico	94	6	95	5	95	5
New York	93	7	94	6	100	0
North Carolina	95	5	93	7	100	0
North Dakota	96	4	95	5	100	0
Ohio	95	5	97	3	100	0
Oklahoma	80	20	76	24	100	0
Oregon	93	7	91	9	100	0
Pennsylvania	94	6	95	5	100	0
Rhode Island	93	7	92	8	91	9
Texas	96	4	97	3	94	6
Virginia	94	6	91	9	90	10
West Virginia	94	6	94	6	100	0
Wisconsin	94	6	93	7	93	7
Wyoming	96	4	93	7	100	0
TERRITORIES						
Guam	93	7	75	25	100	0
Virgin Islands	93	7	73	27	100	0

TABLE B.5 | Questionnaire Response Rates

GRADE 8 PUBLIC SCHOOLS	Weighted Percentage of Students Matched to Mathematics Teacher Questionnaires	Percentage of Mathematics Teacher Questionnaires Returned	Weighted Percentage of Students Matched to School Characteristics / Policy Questionnaires	Percentage of School Characteristics / Policies Questionnaires Returned	Percentage of Excluded Student Questionnaires Returned
NATION	76	0	86	84	0
Northeast	65	0	94	88	0
Southeast	78	0	91	87	0
Central	79	0	70	75	0
West	77	0	88	88	0
STATES					
Alabama	94	0	100	100	100
Arizona	85	0	99	99	98
Arkansas	92	0	98	98	100
California	86	0	98	97	95
Colorado	85	0	99	99	100
Connecticut	89	0	99	99	96
Delaware	85	0	96	97	98
District of Columbia	94	0	98	97	99
Florida	88	0	98	97	97
Georgia	87	0	96	98	100
Hawaii	91	0	99	98	99
Idaho	87	0	98	99	97
Illinois	85	0	95	96	98
Indiana	87	0	98	98	93
Iowa	89	0	99	99	99
Kentucky	93	0	100	100	100
Louisiana	90	0	99	99	100
Maryland	89	0	99	99	99
Michigan	91	0	100	100	99
Minnesota	88	0	99	99	97
Montana	94	0	100	100	100
Nebraska	89	0	99	99	99
New Hampshire	88	0	100	100	99
New Jersey	91	0	99	99	97
New Mexico	90	0	97	97	93
New York	85	0	98	99	98
North Carolina	91	0	98	98	97
North Dakota	94	0	85	97	99
Ohio	83	0	100	100	98
Oklahoma	91	0	99	99	99
Oregon	84	0	94	94	100
Pennsylvania	87	0	96	98	97
Rhode Island	87	0	100	100	100
Texas	84	0	99	99	99
Virginia	93	0	98	97	99
West Virginia	91	0	99	99	100
Wisconsin	87	0	99	99	98
Wyoming	84	0	100	99	99
TERRITORIES					
Guam	96	0	100	100	100
Virgin Islands	88	0	100	100	100

The Mathematics Teacher Questionnaire was in two parts -- the first requesting background information about the teacher and the second asking about instruction in particular classes. The percentage of students matched to questionnaires is provided for Part II. If they differed, the match rates for Part I were higher. For the nation and regions, the percentage of excluded student questionnaires returned is based on students sampled for all subjects assessed in 1990. However, based on the sampling design this rate also is the best estimate of the comparable rate for the mathematics assessment.

APPENDIX C

LIST OF CONDITIONING VARIABLES AND ASSOCIATED CONTRAST CODINGS

APPENDIX C

LIST OF CONDITIONING VARIABLES AND ASSOCIATED CONTRAST CODINGS

This appendix contains information about the conditioning variables used in the construction of plausible values for the 1990 Trial State Assessment Program. Two kinds of conditioning variables were used, continuous or quasi-continuous variables (such as average school mathematics score or median school income), and categorical variables (such as student responses to demographic background questions). The continuous variables, and the range of possible values for these variables, are given in the tables that follow. The categorical variables were incorporated into the conditioning process by first recoding them into a series of binary, (and sometimes linear and quadratic) contrasts. The possible response categories and contrast coding schemes for each categorical variable are also provided in the tables that follow.

It should be noted that, as described in Chapter 10, the linear conditioning model that was employed did not use directly the listed conditioning variables. First, principal components were derived from the variables listed in this Appendix. These principal components were then used as the predictor variables in the linear conditioning models used to construct plausible values.

CONTINUOUS VARIABLES

Variable	Source	Range of Values
Teacher Emphasis on Numbers & Operations Topics	Teacher Questionnaire, Part II. Derived from items 16,17,18, & 20 of T031500	0 to 12
Percent enrolled in School Lunch Program	Westat, Principal's Questionnaire	0 to 100
School Median Income (in thousands of dollars)	Westat, Principal's Questionnaire	0 to 100,000
School Average Mathematics Proficiency	ETS, school mean of logit-percent correct over all items taken	-4 to 4

CONTRAST-CODED CATEGORICAL VARIABLES

ETS and Westat Derived Variables

Variable	Response Categories	Contrast Coding	Number of Contrasts
Overall	---	1	1
Gender	1 Male	1	1
	2 Female	0	
Race/Ethnicity	1 White	000	3
	2 Black	100	
	3 Hispanic	010	
	4 Asian American	001	
	5 American Indian	000	
	6 Unclassified	000	
	BLK Missing	000	

Multi-column entries without overbars indicate multiple contrasts. A multi-column entry with a bar over it indicates a single contrast.

Variable	Response Categories	Contrast Coding	Number of Contrasts
Type of Community	1 Rural	01	2
	2 Disadvantage Urban	00	
	3 Advantaged Urban	10	
	4 Other	01	
	BLK Missing	01	
Parental Education	1 < High School	0000	4
	2 High School Grad.	1000	
	3 Some College	0100	
	4 College Grad.	0010	
	BLK Missing & I don't know	0001	
# of Reading Items in the Home (Items asked about: 1 - Newspapers, 2 - > 25 books, 3 - encyclopedia, 4 - magazines)	1 0 to 2 out of 4	00	2
	2 3 out of 4	10	
	3 4 out of 4	01	
	BLK Missing	00	
Percent White in School	1 0-49 , White Minority	10	2
	2 50-79, Integrated	01	
	3 80-100, Predom. White	00	
	BLK Missing	00	
Age	1 < Modal Age	100	3
	2 Modal Age	010	
	3 > Modal Age	001	
	BLK Missing	000	
Actual Monitored Status	1 Unmonitored Session	0	1
	2 Monitored Session	1	

Multi-column entries without overbars indicate multiple contrasts. A multi-column entry with a bar over it indicates a single contrast.

Variables from Student Background Questions

Variable	Response Categories	Contrast Coding	Number of Contrasts
TV Watching	1 None	0 $\overline{00}$	2
(B001801)[1]	2 1 hour or less	1 01	
	3 2 hours	2 04	
	4 3 hours	3 09	
	5 4 hours	4 16	
	6 5 hours	5 25	
	7 6 or more hours	6 36	
	BLK Missing	3 09	
Daily Homework	1 Don't have any	10 0 $\overline{00}$	4
in All Subjects	2 Don't do any	01 0 00	
(B003901)	3 1/2 hour	01 1 01	
	4 1 hour	01 2 04	
	5 2 hours	01 3 09	
	6 > 2 hours	01 4 16	
Speak Other Language	1 Never	0	1
at Home	2 Sometimes	1	
(B003201)	3 Always	1	
	BLK Missing	0	
Help with homework	1 almost every day	1	1
(B006701)	2 1-2 a week	1	
	3 1-2 a month	0	
	4 never or hardly ever	0	
	5 don't have hw	0	
	BLK Missing	0	
Single/Multiple	1 Yes to Father and	1	1
Parent at Home	and Mother at Home		
(B005601 & B005701)	2 Any other responses	0	
	BLK Missing	0	
Mother at Home	1 Yes	1	1
(B005601)	2 No	0	
	BLK Missing	0	

[1]NAEP Background Question ID # is given in parentheses

Multi-column entries without overbars indicate multiple contrasts. A multi-column entry with a bar over it indicates a single contrast.

Variable	Response Categories	Contrast Coding	Number of Contrasts
Pages a Day Read for School & Homework (B001101)	1 > 20	11	2
	2 16-20	11	
	3 11-15	11	
	4 6-10	10	
	5 5 or fewer	00	
	BLK Missing	00	
Expect to Graduate from High School (S003401)	1 Yes	1	1
	2 No	0	
	3 I don't know	0	
	BLK Missing	0	
School Days Missed Last Month (S004001)	1 None	1	1
	2 1 or 2 Days	1	
	3 3 or 4 Days	0	
	4 5 to 10 Days	0	
	5 > 10 Days	0	
	BLK Missing	0	
School Rules for Behavior are Strict (B007001)	1 Strongly Agree	11	2
	2 Agree	12	
	3 Disagree	13	
	4 Strongly Disagree	14	
	BLK Missing	00	
I Don't Feel Safe at School (B007002)	1 Strongly Agree	11	2
	2 Agree	12	
	3 Disagree	13	
	4 Strongly Disagree	14	
	BLK Missing	00	
Students Often Disrupt Class (B007003)	1 Strongly Agree	11	2
	2 Agree	12	
	3 Disagree	13	
	4 Strongly Disagree	14	
	BLK Missing	00	
Do textbook problems (M810101)	1 almost every day	1000	4
	2 several times a week	0100	
	3 once a week	0010	
	4 less than once a week	0001	
	5 never	0000	
	BLK Missing	0000	

Multi-column entries without overbars indicate multiple contrasts. A multi-column entry with a bar over it indicates a single contrast.

Variable	Response Categories	Contrast Coding	Number of Contrasts
Do worksheet problems	1 almost every day	1000	4
(M810102)	2 several times a week	0100	
	3 once a week	0010	
	4 less than once a week	0001	
	5 never	0000	
	BLK Missing	1000	
Work in small groups	1 almost every day	1000	4
(M810103)	2 several times a week	0100	
	3 once a week	0010	
	4 less than once a week	0001	
	5 never	0000	
	BLK Missing	1000	
Work with objects	1 almost every day	1000	4
(M810104)	2 several times a week	0100	
	3 once a week	0010	
	4 less than once a week	0001	
	5 never	0000	
	BLK Missing	1000	
Uses Calculator	1 almost every day	1000	4
(M810105)	2 several times a week	0100	
	3 once a week	0010	
	4 less than once a week	0001	
	5 never	0000	
	BLK Missing	0000	
Uses Computer	1 almost every day	1000	4
(M810106)	2 several times a week	0100	
	3 once a week	0010	
	4 less than once a week	0001	
	5 never	0000	
	BLK Missing	1000	
Takes Tests	1 almost every day	1000	4
(M810107)	2 several times a week	0100	
	3 once a week	0010	
	4 less than once a week	0001	
	5 never	0000	
	BLK Missing	1000	

Multi-column entries without overbars indicate multiple contrasts. A multi-column entry with a bar over it indicates a single contrast.

Variable	Response Categories	Contrast Coding	Number of Contrasts	
Writes Reports (M810108)	1 almost every day	1000	4	
	2 several times a week	0100		
	3 once a week	0010		
	4 less than once a week	0001		
	5 never	0000		
	BLK Missing	1000		
Teacher explains calculator use (M810108)	1 Yes	10	2	
	2 No	01		
	BLK Missing	00		
Uses calculator on class problems (M810301)	1 Almost Always	10	2	
	2 Sometimes	01		
	3 Never	00		
	BLK Missing	10		
Uses calculator on home problems (M810302)	1 Almost always	10	2	
	2 Sometimes	01		
	3 Never	00		
	BLK Missing	10		
Uses calculator on tests (M810303)	1 Almost always	10	2	
	2 Sometimes	01		
	3 Never	00		
	BLK Missing	10		
What kind of class are you taking (M810501)	1 Not taking mathematics		0000	4
	2 8th grade mathematics		1000	
	3 pre-algebra	0100		
	4 algebra	0010		
	5 other	0001		
	BLK Missing	0000		
Amount of Mathematics Done Daily (M810601)	1 None	1000	4	
	2 15 Minutes	0100		
	3 30 Minutes	0100		
	4 45 Minutes	0010		
	5 60 Minutes	0001		
	6 > 1 hour	0001		
	7 Not taking mathematics		0000	
	BLK Missing	0000		

Multi-column entries without overbars indicate multiple contrasts. A multi-column entry with a bar over it indicates a single contrast.

Variable	Response Categories	Contrast Coding	Number of Contrasts
Likes mathematics (M810701)	1 strongly agree	0001	4
	2 agree	0010	
	3 undecided	0100	
	4 disagree	1000	
	5 strongly disagree	0000	
	BLK Missing	0000	
People use mathematics (M81072)	1 strongly agree	0001	4
	2 agree	0010	
	3 undecided	0100	
	4 disagree	1000	
	5 strongly disagree	0000	
	BLK Missing	0000	
Good at math (M810703)	1 strongly agree	0001	4
	2 agree	0010	
	3 undecided	0100	
	4 disagree	1000	
	5 strongly disagree	0000	
	BLK Missing	0000	
Mathematics is for boys (M810704)	1 strongly agree	100	3
	2 agree	100	
	3 undecided	010	
	4 disagree	010	
	5 strongly disagree	001	
	BLK Missing	000	
Useful for everyday problems (M810705)	1 strongly agree	0001	4
	2 agree	0010	
	3 undecided	0100	
	4 disagree	1000	
	5. strongly disagree	0000	
	BLK Missing	0000	

Multi-column entries without overbars indicate multiple contrasts. A multi-column entry with a bar over it indicates a single contrast.

Variables from Teacher Background Questions

Variable	Response Categories	Contrast Coding	Number of Contrasts
Mathematics Courses Taken (T030400)			
Geometry	1 None	100	3
	2 One	010	
	3 Two	001	
	4 Three	001	
	BLK Missing	100	
Abstract Algebra	1 None	100	3
	2 One	010	
	3 Two	001	
	4 Three	001	
	BLK Missing	100	
Calculus	1 None	100	3
	2 One	010	
	3 Two	001	
	4 Three	001	
	BLK Missing	100	
Amount of resources (T030801)	1 all I need	100	3
	2 most of what I need	010	
	3 some of what I need	001	
	4 no resources	001	
	BLK Missing	001	
Ability of Class (T031001)	1 primarily high	100	3
	2 primarily average	010	
	3 primarily low	001	
	4 mixed	010	
	BLK Missing	010	
	Unmatched	000	
Amount of Mathematics HW done (T031201)	1 None	11	2
	2 15 minutes	11	
	3 30 minutes	21	
	4 45 minutes	31	
	5 60 minutes	31	
	6 > 1 hour	21	
	BLK Missing	21	
	Unmatched	00	

Variable	Response Categories	Contrast Coding	Number of Contrasts
How Often to Students in Class do the following (T031400)			
Textbook problems	1 almost every day	100	3
	2 several times a week	010	
	3 once a week	001	
	4 less than once a week	001	
	5 never	001	
	BLK Missing	001	
Worksheets	1 almost every day	100	3
	2 several times a week	010	
	3 once a week	001	
	4 less than once a week	001	
	5 never	001	
	BLK Missing	001	
Use calculators	1 Yes	11	2
	2 No	01	
	BLK Missing	01	
Can use calculators on tests (T031701)	1 Yes	11	2
	2 No	01	
	BLK Missing	01	
Emphasis Placed on the following: (T031500)			
Measurement	1 None	01	2
	2 Little	11	
	3 Moderate	21	
	4 Heavy	31	
	BLK Missing	21	
	UM Unmatched	00	
Algebra	1 None	01	2
	2 Little	11	
	3 Moderate	21	
	4 Heavy	31	
	BLK Missing	21	
	UM Unmatched	00	

Variables Related to Nonresponse

Response Variable	Contrast Categories	Coding	Number of Contrasts
Teacher Questionnaire Match Status	1 Both Sections	00	2
	2 Section 1 Only	10	
	3 Neither Section	02	
Teacher Quest. Match Status by Race/Ethnicity	1 White/Both Sections	8 0 0 0 0 0 0 0	8
	2 White/Section 1 Only	-4 4 0 0 0 0 0 0	
	3 White/Neither	-4-4 0 0 0 0 0 0	
	4 Black/Both Section	-2 0 6 0 0 0 0 0	
	5 Black/Section 1 Only	-1-1-3 3 0 0 0 0	
	6 Black/Neither Section	1 1-3-3 0 0 0 0	
	7 Hispanic/Both Sections	-2 0-2 0 4 0 0 0	
	8 Hispanic/Section 1 Only	1-1 1-1-2 2 0 0	
	9 Hispanic/Neither Section	1 1 1 1-2-2 0 0	
	10 Asian/Both Sections	-2 0-2 0-2 0 2 0	
	11 Asian/ Section 1 Only	1-1 1-1 1-1-1 1	
	12 Asian/ Neither Section	1 1 1 1 1 1-1-1	
	13 Other/Both Sections	-2 0-2 0-2 0-2 0	
	14 Other/Section 1 Only	1-1 1-1 1-1 1-1	
	15 Other/Neither Section	1 1 1 1 1 1 1 1	
Teacher Quest. Match Status by Parental Education	1 Coll Grad/Both Sections	8 0 0 0 0 0 0 0	8
	2 Coll Grad/Section 1 Only	-4 4 0 0 0 0 0 0	
	3 Coll Grad/Neither Section	-4-4 0 0 0 0 0 0	
	4 Some College/Both Sections	-2 0 6 0 0 0 0 0	
	5 Some College/Section 1 Only	1-1-3 3 0 0 0 0	
	6 Some College/Neither Section	1 1-3-3 0 0 0 0	
	7 HS Grad/Both Sections	-2 0-2 0 4 0 0 0	
	8 HS Grad/Section 1 Only	1-1 1-1-2 2 0 0	
	9 HS Grad/Neither Section	1 1 1 1-2-2 0 0	
	10 Not HS Grad/Both Sections	-2 0-2 0-2 0 2 0	
	11 Not HS Grad/Section 1 Only	1-1 1-1 1-1-1 1	
	12 Not HS Grad/Neither Section	1 1 1 1 1 1-1-1	
	13 Don't Know/Both Sections	-2 0-2 0-2 0-2 0	
	14 Don't Know/Section 1 Only	1-1 1-1 1-1 1-1	
	15 Don't Know/Neither Section	1 1 1 1 1 1 1 1	

Variable	Response Categories	Contrast Coding	Number of Contrasts
Teacher Quest. Match Status by Gender	1 Male/Both Sections	2 0	2
	2 Male/Section 1 Only	-1 1	
	3 Male/Neither Section	-1-1	
	4 Female/Both Sections	-2 0	
	5 Female/Section 1 Only	1-1	
	6 Female/Neither Section	1 1	
Percent Enrolled in School Lunch Program	1 Variable was missing	1	1
	2 Variable available	0	
School Median Income	1 Variable was missing	1	1
	2 Variable available	0	
School Average Math Proficiency	1 Variable was missing	1	1
	2 Variable available	0	

APPENDIX D

IRT PARAMETERS FOR MATHEMATICS ITEMS

IRT PARAMETERS FOR MATHEMATICS ITEMS

NUMBERS AND OPERATIONS

NAEP ID	A	S.E.	B	S.E.	C	S.E.	BLOCK	ITEM
M011131	0.643	(0.022)	-1.477	(0.098)	0.155	(0.042)	M8	13
M012431	0.828	(0.024)	-0.396	(0.040)	0.080	(0.019)	M8	3
M012531	0.661	(0.025)	0.655	(0.033)	0.066	(0.013)	M8	4
M012931	0.919	(0.050)	1.213	(0.026)	0.212	(0.009)	M8	8
M013431	0.956	(0.037)	0.191	(0.032)	0.131	(0.014)	M8	15
M013531	0.638	(0.044)	1.796	(0.045)	0.085	(0.010)	M8	16
M013631	1.344	(0.052)	0.937	(0.015)	0.058	(0.005)	M8	17
M015501	0.969	(0.033)	0.224	(0.026)	0.082	(0.012)	M7	2
M015901	0.685	(0.047)	1.246	(0.039)	0.219	(0.014)	M7	6
M016501	1.075	(0.061)	1.695	(0.030)	0.079	(0.005)	M7	12
M017401	0.258	(0.016)	-5.220	(0.387)	0.198	(0.057)	M4	1
M017701	0.844	(0.025)	-1.050	(0.057)	0.125	(0.029)	M4	4
M017901	1.147	(0.035)	-0.892	(0.038)	0.105	(0.023)	M4	6
M018201	0.601	(0.018)	-0.756	(0.064)	0.090	(0.025)	M4	9
M018401	1.202	(0.050)	-0.743	(0.050)	0.322	(0.024)	M4	11
M018501	1.620	(0.067)	0.541	(0.017)	0.237	(0.007)	M4	12
M018601	0.598	(0.036)	1.201	(0.040)	0.135	(0.015)	M4	13
M020001	0.667	(0.013)	-0.214	(0.014)	0.000	(0.000)	M5	4
M020101	1.304	(0.025)	-0.329	(0.009)	0.000	(0.000)	M5	5
M020501	0.847	(0.016)	-0.390	(0.012)	0.000	(0.000)	M5	9
M021901	0.868	(0.025)	-1.387	(0.063)	0.135	(0.035)	M6	1
M022001	1.025	(0.030)	-0.802	(0.043)	0.135	(0.024)	M6	2
M022301	0.626	(0.021)	-2.456	(0.113)	0.176	(0.051)	M6	5
M022701	0.859	(0.029)	-0.813	(0.060)	0.170	(0.029)	M6	9
M022901	1.161	(0.046)	-0.292	(0.039)	0.312	(0.017)	M6	12
M023001	1.105	(0.039)	-0.230	(0.034)	0.225	(0.016)	M6	13
M023801	1.261	(0.043)	0.310	(0.019)	0.101	(0.009)	M6	21
M027031	0.402	(0.023)	-4.564	(0.263)	0.193	(0.056)	M9	1
M027331	0.778	(0.014)	0.600	(0.014)	0.000	(0.000)	M9	4
M027831	1.013	(0.017)	0.059	(0.010)	0.000	(0.000)	M9	9
M028031	0.950	(0.040)	0.622	(0.027)	0.181	(0.011)	M9	11
M028131	0.541	(0.012)	0.832	(0.021)	0.000	(0.000)	M9	12
M028231	0.687	(0.025)	0.486	(0.035)	0.060	(0.014)	M9	13
M028631	1.276	(0.033)	1.499	(0.018)	0.000	(0.000)	M9	17
M028731	1.729	(0.103)	1.541	(0.020)	0.082	(0.003)	M9	18
M028931	0.629	(0.058)	1.430	(0.058)	0.258	(0.018)	M9	20
N202831	0.627	(0.023)	-1.998	(0.125)	0.198	(0.054)	M8	12
N258801	1.167	(0.068)	0.668	(0.031)	0.411	(0.010)	M3	11
N260101	1.075	(0.041)	-0.245	(0.039)	0.228	(0.018)	M3	18
N274801	0.685	(0.041)	-0.284	(0.105)	0.417	(0.029)	M3	10
N275301	0.280	(0.014)	-3.068	(0.263)	0.172	(0.053)	M3	14
N276803	0.223	(0.011)	-3.735	(0.176)	0.000	(0.000)	M3	1

IRT PARAMETERS FOR MATHEMATICS ITEMS

NUMBERS AND OPERATIONS

NAEP ID	A	S.E.	B	S.E.	C	S.E.	BLOCK	ITEM
N277602	0.418	(0.012)	-2.415	(0.065)	0.000	(0.000)	M3	2
N286201	0.806	(0.027)	-1.157	(0.074)	0.151	(0.037)	M3	6
N286301	1.112	(0.042)	0.178	(0.029)	0.180	(0.013)	M3	21
N286602	0.641	(0.012)	-0.141	(0.014)	0.000	(0.000)	M3	13

IRT PARAMETERS FOR MATHEMATICS ITEMS

MEASUREMENT

NAEP ID	A	S.E.	B	S.E.	C	S.E.	BLOCK	ITEM
M012331	0.717	(0.035)	-1.427	(0.116)	0.200	(0.051)	M8	2
M013331	0.878	(0.048)	-1.356	(0.105)	0.211	(0.052)	M8	14
M015401	0.710	(0.051)	0.043	(0.085)	0.190	(0.032)	M7	1
M015701	0.837	(0.039)	-2.000	(0.111)	0.227	(0.058)	M7	4
M016201	0.887	(0.077)	0.787	(0.041)	0.211	(0.017)	M7	9
M017501	0.431	(0.025)	-2.430	(0.232)	0.288	(0.063)	M4	2
M018101	0.804	(0.062)	-0.073	(0.090)	0.269	(0.033)	M4	8
M019101	1.482	(0.241)	2.032	(0.072)	0.175	(0.006)	M4	18
M019201	1.450	(0.205)	1.894	(0.061)	0.147	(0.006)	M4	19
M020301	1.000	(0.030)	-0.354	(0.014)	0.000	(0.000)	M5	7
M022601	1.129	(0.108)	0.780	(0.037)	0.381	(0.013)	M6	8
M022801	1.751	(0.057)	-0.608	(0.012)	0.000	(0.000)	M6	10
M022802	1.604	(0.048)	-0.929	(0.015)	0.000	(0.000)	M6	11
M023401	0.860	(0.075)	0.295	(0.069)	0.364	(0.024)	M6	17
M023701	0.519	(0.017)	1.038	(0.033)	0.000	(0.000)	M6	20
M027631	1.067	(0.086)	0.094	(0.053)	0.209	(0.024)	M9	7
M028831	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	M9	19
N252101	0.654	(0.062)	0.275	(0.107)	0.268	(0.035)	M3	17
N265201	0.755	(0.044)	-1.872	(0.158)	0.339	(0.066)	M3	9
N265901	0.742	(0.069)	0.651	(0.066)	0.250	(0.024)	M3	16
N267201	0.796	(0.061)	-1.009	(0.162)	0.401	(0.056)	M3	3

IRT PARAMETERS FOR MATHEMATICS ITEMS

GEOMETRY

NAEP ID	A	S.E.	B	S.E.	C	S.E.	BLOCK	ITEM
M012731	0.646	(0.058)	1.325	(0.053)	0.174	(0.019)	M8	6
M012831	1.185	(0.071)	0.649	(0.026)	0.119	(0.012)	M8	7
M015601	0.358	(0.030)	-0.078	(0.251)	0.234	(0.054)	M7	3
M016301	0.608	(0.028)	-0.289	(0.085)	0.123	(0.032)	M7	10
M016401	1.580	(0.118)	1.234	(0.022)	0.167	(0.006)	M7	11
M016601	0.833	(0.049)	1.375	(0.031)	0.080	(0.010)	M7	13
M016701	1.236	(0.093)	1.718	(0.034)	0.119	(0.005)	M7	14
M017601	0.459	(0.022)	-1.744	(0.169)	0.184	(0.054)	M4	3
M018001	0.755	(0.049)	0.044	(0.084)	0.218	(0.031)	M4	7
M019001	0.733	(0.050)	0.776	(0.050)	0.150	(0.020)	M4	17
M019601	0.720	(0.063)	1.650	(0.047)	0.128	(0.013)	M4	21
M019801	0.982	(0.023)	-0.578	(0.014)	0.000	(0.000)	M5	2
M019901	0.675	(0.018)	-1.438	(0.032)	0.000	(0.000)	M5	3
M020901	0.563	(0.016)	1.314	(0.033)	0.000	(0.000)	M5	11
M021001	0.862	(0.019)	0.277	(0.013)	0.000	(0.000)	M5	12
M021301	1.194	(0.027)	0.125	(0.011)	0.000	(0.000)	M5	15
M021302	1.165	(0.026)	-0.079	(0.011)	0.000	(0.000)	M5	16
M022201	0.539	(0.015)	-0.645	(0.023)	0.000	(0.000)	M6	4
M022501	0.800	(0.020)	-0.368	(0.015)	0.000	(0.000)	M6	7
M023101	1.087	(0.053)	0.029	(0.038)	0.118	(0.019)	M6	14
M027231	0.704	(0.057)	-0.303	(0.144)	0.401	(0.042)	M9	3
M027431	0.669	(0.033)	-0.627	(0.102)	0.167	(0.040)	M9	5
M028331	1.595	(0.228)	1.602	(0.042)	0.351	(0.007)	M9	14
N253701	0.525	(0.042)	-0.309	(0.194)	0.309	(0.052)	M3	12
N254602	1.322	(0.100)	1.029	(0.024)	0.196	(0.009)	M3	22
N269901	0.816	(0.061)	-0.152	(0.104)	0.337	(0.036)	M3	15

IRT PARAMETERS FOR MATHEMATICS ITEMS

DATA ANALYSIS, STATISTICS, AND PROBABILITY

NAEP ID	A	S.E.	B	S.E.	C	S.E.	BLOCK	ITEM
M012631	1.983	(0.153)	0.788	(0.017)	0.216	(0.008)	M8	5
M013031	1.167	(0.041)	1.508	(0.029)	0.000	(0.000)	M8	9
M013131	0.952	(0.032)	1.390	(0.029)	0.000	(0.000)	M8	10
M015801	1.074	(0.060)	0.436	(0.031)	0.116	(0.015)	M7	5
M016101	1.429	(0.094)	0.481	(0.027)	0.246	(0.012)	M7	8
M017001	0.860	(0.070)	1.183	(0.032)	0.140	(0.013)	M7	18
M017801	1.198	(0.084)	-0.228	(0.064)	0.304	(0.026)	M4	5
M018901	1.207	(0.224)	2.063	(0.138)	0.157	(0.007)	M4	16
M020201	0.576	(0.019)	-2.059	(0.056)	0.000	(0.000)	M5	6
M020801	1.140	(0.048)	1.630	(0.037)	0.000	(0.000)	M5	10
M021101	0.944	(0.025)	0.157	(0.013)	0.000	(0.000)	M5	13
M023301	1.792	(0.120)	-0.459	(0.045)	0.247	(0.023)	M6	16
M023501	1.920	(0.142)	0.834	(0.015)	0.123	(0.007)	M6	18
M023601	0.895	(0.040)	-0.366	(0.055)	0.093	(0.025)	M6	19
M028531	0.981	(0.029)	-0.777	(0.022)	0.000	(0.000)	M9	16
N250201	0.668	(0.031)	-1.437	(0.124)	0.175	(0.051)	M3	8
N250901	0.333	(0.018)	-3.623	(0.256)	0.175	(0.054)	M3	4
N250902	0.829	(0.033)	-0.881	(0.069)	0.104	(0.032)	M3	5
N263501	1.368	(0.082)	0.104	(0.035)	0.214	(0.016)	M3	19

IRT PARAMETERS FOR MATHEMATICS ITEMS

ALGEBRA AND FUNCTIONS

NAEP ID	A	S.E.	B	S.E.	C	S.E.	BLOCK	ITEM
M012231	0.436	(0.027)	-3.985	(0.236)	0.148	(0.051)	M8	1
M013231	1.180	(0.116)	1.916	(0.055)	0.123	(0.006)	M8	11
M013731	0.925	(0.079)	1.520	(0.042)	0.117	(0.010)	M8	18
M016001	0.919	(0.038)	0.475	(0.029)	0.065	(0.013)	M7	7
M016801	0.949	(0.053)	1.766	(0.038)	0.040	(0.005)	M7	15
M016901	2.279	(0.000)	0.862	(0.012)	0.161	(0.000)	M7	16
M016902	1.719	(0.000)	1.170	(0.011)	0.000	(0.000)	M7	17
M018301	0.842	(0.035)	-0.411	(0.062)	0.132	(0.028)	M4	10
M018701	1.334	(0.073)	0.318	(0.030)	0.223	(0.013)	M4	14
M018801	0.840	(0.071)	1.122	(0.041)	0.277	(0.015)	M4	15
M019301	1.192	(0.089)	1.300	(0.028)	0.191	(0.008)	M4	20
M019701	0.510	(0.016)	-1.641	(0.046)	0.000	(0.000)	M5	1
M020401	0.637	(0.016)	0.029	(0.016)	0.000	(0.000)	M5	8
M021201	1.020	(0.026)	0.599	(0.013)	0.000	(0.000)	M5	14
M022101	0.739	(0.035)	-2.689	(0.126)	0.222	(0.058)	M6	3
M022401	1.098	(0.069)	-0.575	(0.082)	0.391	(0.033)	M6	6
M023201	0.998	(0.042)	-0.435	(0.050)	0.124	(0.025)	M6	15
M027131	0.843	(0.030)	-1.989	(0.076)	0.124	(0.043)	M9	2
M027531	0.627	(0.033)	-0.803	(0.126)	0.221	(0.045)	M9	6
M027731	0.864	(0.041)	0.197	(0.044)	0.117	(0.019)	M9	8
M027931	0.977	(0.022)	0.093	(0.012)	0.000	(0.000)	M9	10
M028431	0.721	(0.019)	0.786	(0.020)	0.000	(0.000)	M9	15
N255701	1.227	(0.070)	0.749	(0.023)	0.132	(0.011)	M3	23
N256101	0.925	(0.025)	-1.189	(0.023)	0.000	(0.000)	M3	7
N264701	1.544	(0.091)	0.481	(0.024)	0.186	(0.012)	M3	20

APPENDIX E

TRIAL STATE ASSESSMENT REPORTING SUBGROUPS

COMPOSITE AND DERIVED COMMON BACKGROUND VARIABLES

COMPOSITE AND DERIVED REPORTING VARIABLES

APPENDIX E

NAEP REPORTING SUBGROUPS

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 on the student file contains a value for every student and is used for gender comparisons among students.

RACE (Observed Race/Ethnicity)

The variable RACE is the race/ethnicity of the student being assessed, as reported in the school records.

DRACE (Imputed Race/Ethnicity)

The variable DRACE is an imputed definition of race/ethnicity, derived from up to three sources of information.

Two common background items were used in the determination of race/ethnicity:

Common Background Item Number Two:

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 two 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 one were examined to determine race/ethnicity. Item number one read as follows:

Common Background Item Number One:

1. Which best describes you?

○ White (not Hispanic)

○ Black (not Hispanic)

○ Hispanic ("Hispanic" means someone who is Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or from some other Spanish or Hispanic background.)

○ Asian or Pacific Islander ("Asian or Pacific Islander" means someone who is Chinese, Japanese, Korean, Filipino, Vietnamese, or from some other Asian or Pacific Island background.)

○ American Indian or Alaskan Native ("American Indian or Alaskan Native" means someone who is from one of the American Indian tribes, or one of the original people of Alaska.)

○ Other (What?) ________

Students' race/ethnicity (DRACE) was then assigned to correspond with their selection. For students who filled in the sixth oval ("Other"), provided illegible information or information that could not be classified, or did not respond at all, observed race/ethnicity (RACE), if provided by the exercise administrator, was used.

Imputed race/ethnicity could not be determined for students who did not respond to background items one or two and for whom an observed race/ethnicity was not provided.

TOC (Type of community)

NAEP assigned each participating school to one of four type of categories designed to provide information about the communities in which the schools are located.

The TOC reporting categories consist of three "extreme" types of communities and one "other" type of community. Schools were placed into TOC categories on the basis of information about the type of community, the size of its population (as of the 1980 Census), and an occupational profile of residents provided by school principals before the assessment. The principals completed estimates of the percentage of students whose parents fit into each of six occupational categories.

A weighted version of TOC was created: schools were ranked in order based on principals' responses about the type of community, size of its population, and occupational profile of the students' parents. Schools were assigned to the extreme TOC categories (one, two, and three) so that ten percent of sampled students (weighted) would be enrolled in schools in each category. The remaining schools were classified as "other". The TOC categories are as follows:

TOC 1 - Extreme Rural: Students in this group attend schools in areas with a population below 10,000 where many of the students' parents are farmers or farm workers.

TOC 2 - Disadvantaged Urban: Students in this group attend schools in Metropolitan Statistical Areas where a high proportion of the students' parents are on welfare or are not regularly employed.

TOC 3 - Advantaged Urban: Students in this group attend schools in Metropolitan Statistical Areas where a high proportion of the students' parents are in professional or managerial positions.

TOC 4 - Other: Schools that did not meet the criteria for TOC categories 1, 2, or 3 were classified as "Other":

TOC 8 - Missing data

PARED (Parental education)

The variable PARED is derived from responses to two common background questions, question number four and question number five. Students were asked to indicate the extent of their mother's education (question four) 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 (question five) 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 as follows:

1 Did not finish high school.
2 Graduated from high school.
3 Some education after high school.
4 Graduated from college.
5 I don't know.
8 No response

If a student indicated the extent of education for only one parent, that value was considered to be the parental education level for the student. If a student indicated the extent of education for both parents, the higher of the two levels was included as the value for the student's parental education variable. 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 as follows[1]:

Northeast

Connecticut
Delaware
District of Columbia
Maine
Maryland
Massachusetts
New Hampshire
New Jersey
New York
Pennsylvania
Rhode Island
Vermont
Virginia[2]

Southeast

Alabama
Arkansas
Florida
Georgia
Kentucky
Louisiana
Mississippi
North Carolina
South Carolina
Tennessee
Virginia[2]
West Virginia

Central

Illinois
Indiana
Iowa
Kansas
Michigan
Minnesota
Missouri
Nebraska
North Dakota
Ohio
South Dakota
Wisconsin

West

Alaska
Arizona
California
Colorado
Hawaii
Idaho
Montana
Nevada
New Mexico
Oklahoma
Oregon
Texas
Utah
Washington
Wyoming

MODAGE Modal Age

1 Less than age 13
2 Equal to age 13
3 Greater than age 13

[1] All fifty states are listed with the states that participated in the Trial State Assessment highlighted in bold print. Territories were not assigned to a region.

[2] That 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.

COMPOSITE AND DERIVED COMMON BACKGROUND VARIABLES

Several variables are formed from the systematic combination of response values for one or more common background questions (section one in every student's booklet asked questions concerning subjects such as materials in the home, languages spoken, hours spent watching television, and after-school activities).

HOMEEN2 (Home Environment—Articles [of 4] in the Home)

The variable HOMEEN2 was created from the responses to background questions six through nine concerning articles found in the student's home (newspaper, encyclopedia, more than 25 books, and magazines). The values for this variable were derived as follows:

1 0-2 ARTICLES — The student responded to at least two questions and answered YES to two or fewer.

2 3 ARTICLES — The student answered YES to three questions.

3 4 ARTICLES — The student answered YES to four questions.

8 NO RESPONSE — The student answered fewer than two questions.

SINGLEP (How Many Parents Live at Home)

SINGLEP was created from questions 19 and 20 which asked whether the student's mother (or stepmother) and father (or stepfather) lived at home with the student. The values for SINGLEP were derived as follows:

1 2 PARENTS AT HOME — The student answered YES to both questions.

2 1 PARENT AT HOME — The student answered YES to question 19 and NO to question 20, or YES to question 20 and NO to question 19.

3 NEITHER AT HOME — The student answered NO to both questions.

8 NO RESPONSE — The student did not respond to one or both questions.

9 MULT. — The student filled in more than one oval for one or both questions.

IEP/LEP (Did the student either have an Individual Education Plan or was the student classified as Limited English Proficient)

The IEP/LEP variable was created from both the IEP variable and the LEP variable. The values of IEP/LEP are as follows:

1 = Either (the student either had an Individual Education Plan or was classified as Limited English Proficient)

2 = Neither (the student did not have an Individual Education Plan and was not classified as Limited English Pro

SUBJECT-SPECIFIC COMPOSITE AND DERIVED REPORTING VARIABLES

CALCUSE (Calculator-Usage Index)

CALCUSE was created from noncognitive questions included in mathematics blocks MH and MI. Students were provided a scientific calculator to use in answering the cognitive questions in those two blocks. A noncognitive question, which asked students to indicate whether or not they had used the calculator on the immediately preceding cognitive question, immediately followed each cognitive question.

The cognitive items in blocks MH (18 items) and MI (20 items) were classified into one of three categories -- calculator-active, calculator-inactive, and calculator-neutral. Calculator-active items required the use of a calculator for their solution. Calculator-neutral items could be solved with or without a calculator. Calculator-inactive items posed problems for which use of a calculator was inappropriate. Block MH contained seven calculator-inactive items, three calculator-active items, and eight calculator-neutral items. Block MI contained ten calculator-inactive items, five calculator-active items, and five calculator-neutral items. Blocks MH and MI each appeared in a total of three test booklets. However, one booklet contained both blocks MH and MI. Therefore, at least one calculator block of items appeared in five of the seven assessment booklets.

For those students assigned a booklet containing a block of calculator items, the calculator- usage variable was derived from the noncognitive questions that followed the calculator-inactive and calculator-active items only. Therefore, the calculator-usage index for students assigned a booklet containing only block MH was based on 10 items, the calculator-usage index for students assigned a booklet containing only block MI was based on 15 items, and the calculator-usage index for students assigned a booklet containing both blocks MH and MI was based on 25 items.

CALCUSE had two levels, *high* and *other*, defined as follows:

HIGH	Students who used the calculator appropriately (i.e., used it for the calculator-active items and did not use it for the calculator-inactive items) at least 85 percent of the time and indicated they had used the calculator for at least half of the calculator-active items they were presented.
OTHER	Students who did not use the calculator appropriately at least 85 percent of the time or indicated that they had used the calculator for less than half of the calculator- active items they were presented.

The percentage of appropriate calculator usage was determined using only those noncognitive items which were answered by the student. Omitted noncognitive items were not included as part of the denominator in calculating the percentage of appropriate calculator use.

PERCMAT (Students' Perception of Mathematics)

PERCMAT was created from questions 17-21 in the mathematics background questionnaire. Those questions asked the students about their perceptions of each of the five statements. The statements were:

17. I like mathematics
18. Almost all people use mathematics in their jobs.
19. I am good in mathematics.
20. Mathematics is more for boys than for girls.
21. Mathematics is useful for solving everyday problems.

For each question, the student could respond as follows:

1. STRONGLY AGREE
2. AGREE
3. UNDECIDED
4. DISAGREE
5. STRONGLY DISAGREE

To derive PERCMAT the categories for question 20 were recoded (category five became category one, four became two, two became four, and one became five). Then, for each question, categories three, four, and five (undecided, disagree, and strongly disagree) were combined (new category three). PERCMAT was determined by adding the values for the five questions and dividing by five to obtain a mean. Then the mean was recoded as follows:

1 - 1.67	=	1 STRONGLY AGREE
1.68 - 2.33	=	2 AGREE
2.34 - 3	=	3 UNDECIDED, DISAGREE, OR STRONGLY DISAGREE

The student had to answer at least one of the five questions to get a value for PERCMAT.

TCERTIF (Type of Teaching Certificate)

Questions six through ten in part one of the teacher questionnaire were combined to produce TCERTIF. TCERTIF has three values and the following rules were followed to determine it.

1 MATHEMATICS (MIDDLE SCHOOL OR SECONDARY)

yes for question eight or nine (middle school, junior high school, or secondary school mathematics certification)

2 EDUCATION (ELEMENTARY OR MIDDLE SCHOOL)

yes for question six or seven (elementary education (general) or middle/junior high school education (general) certification) and no for question eight and nine

3 ELSE (NONE OR OTHER TYPE OF CERTIFICATION)

TUNDMAJ (Undergraduate major)

Question 12 was used to determine TUNDMAJ as follows:

1 MATHEMATICS yes for undergraduate major: mathematics

2 EDUCATION yes for undergraduate major: education and no for undergraduate major: mathematics

3 OTHER yes for undergraduate major: other and no for undergraduate major: mathematics or education

TGRDMAJ (Graduate major)

Question 13 was used to determine TGRDMAJ as follows:

1 MATHEMATICS yes for graduate major: mathematics

2 EDUCATION yes for graduate major: education and no for graduate major: mathematics

3 OTHER yes for graduate major: other or no graduate education and no for graduate major: mathematics or education

TMATCRS (Number of mathematics areas in which courses were taken)

TMATCHR was derived from questions 20 - 24, 26, and 27 in part one of the teacher questionnaire. Those questions asked how many courses the teacher had taken in a variety of areas. TMATCHR was derived by obtaining a count of the number of times (out of seven) that the teacher responded to number-of-courses category "One," "Two," or "Three or more". This resulted in a variable whose range was 0-7. Then the levels of TMATCRS were defined as follows:

1	ZERO TO THREE COURSES
2	FOUR TO FIVE COURSES
3	SIX TO SEVEN COURSES

The teacher had to answer at least one question to receive a value for TMATCHR.

TEMPHNO (Teacher's emphasis in numbers and operations)

TEMPHNO was derived from questions 16 through 20 in part two of the teacher questionnaire. The variable was derived by first for each question, combining categories three and four (little emphasis and none) and having the value for that category be three. Then the mean of the values for questions 16 through 20 was obtained and recoded as follows:

1 - 1.67	1	HEAVY EMPHASIS
1.68 - 2.33	2	MODERATE EMPHASIS
2.34 - 3	3	LITTLE OR NO EMPHASIS

The teacher had to answer at least one question to receive a value for TEMPHNO.

TEMPHPS (Teacher's emphasis in data analysis, probability, and statistics)[3]

TEMPHPS was derived from questions 23 through 24 in part two of the teacher questionnaire. The variable was derived by first for each question, combining categories three and four (little emphasis and none) and having the value for that category be three. Then the mean of the values for questions 23 and 24 was obtained and recoded as follows:

1 - 1.67	1	HEAVY EMPHASIS
1.68 - 2.33	2	MODERATE EMPHASIS
2.34 - 3	3	LITTLE OR NO EMPHASIS

The teacher had to answer at least one question to receive a value for TEMPHPS.

[3] The derivation of the teacher's emphasis in measurement, in geometry, and in algebra and functions is not given because each was based on only one question in the teacher questionnaire.

School Ranking Variables

A mean mathematics composite score was calculated for each school using the students' sampling weights. The schools were ordered from highest (rank=1) to lowest (rank=number of schools in the jurisdiction) on the basis of the schools' mean composite mathematics score. The following variables were created

SRANKM	School Rank
SNSCHM	Number of Schools Ranked
SMEANM	School Mean Score

These variables were used in partitioning the schools in each state into three equal groups based on their ranking (e.g. highest one-third, middle one-third, and lowest one-third).

APPENDIX F

THE NAEP SCALE ANCHORING PROCESS FOR THE 1990 MATHEMATICS ASSESSMENT

APPENDIX F

The NAEP Scale Anchoring Process for the 1990 Mathematics Assessment[1]

Introduction

Beginning with the 1984 assessments, NAEP has generally reported students' subject area proficiency on 0 to 500 scales. These scales are used to report achievement for students at the various grades or ages assessed, including differences between performance from assessment to assessment for the nation and various subpopulations of interest. To date, NAEP has used item response theory techniques to develop proficiency scales for reading, mathematics, science, U.S. history, and civics.

Although average proficiency is an efficient summary measure, some of the most interesting NAEP results are those based on performance differences for different points in the scale distributions. To provide an interpretation for both the average results (What does a 306 on the 0 to 500 scale actually mean?) and changes in performance distributions (What does it mean that fewer students are reaching level 250?), NAEP invented a scale anchoring process to describe the characteristics of student performance at various levels along the scales--typically, at levels 200, 250, 300, and 350. The descriptions of student performance are presented in the reports accompanied by the percentages of students performing at or above the various scale levels.

Because of recent changes in NAEP, the purpose of this paper is to describe the updated scale anchoring process as it was carried out for NAEP's newly developed 1990 mathematics scale. In 1988, Congress added a new dimension to NAEP in the form of the Trial State Assessment Program in mathematics at grade 8 in 1990, and in mathematics at grades 4 and 8 as well as reading at grade 4 in 1992. Because NAEP's 1990 mathematics assessment was designed to yield state reports in addition to national reports, the assessment development process was expanded to provide for a new assessment that could be used to report trends into the future. Also, state representatives, as well as a congressionally-mandated Independent State Review Panel, have been involved in monitoring every stage of the assessment from objectives development through reporting plans. Thus, the 1990 mathematics assessment has many new features, including an updated approach to scale anchoring.

In brief, NAEP's scale anchoring procedure for the 1990 mathematics assessment was based on comparing item level performance by students at four levels on the 0 to 500 overall mathematics proficiency scale--levels 200, 250, 200, and 350. This analysis delineated four sets of anchor items that discriminated between adjacent performance levels on the scale. The four sets of empirically derived anchor items were studied by a panel of distinguished mathematics educators, who carefully considered and articulated the types of knowledge, skills, and reasoning abilities demonstrated by correct responses to the items in each set. The 19 panelists and NAEP staff involved in the process worked first in two independent groups to develop

[1] This appendix is from a paper written by Ina V.S. Mullis and presented at the annual meeting of the American Educational Research Association, Chicago, Illinois, April, 1991.

descriptions. As might be expected, the two sets of descriptions were quite similar, but not identical. Thus, the panelists subsequently met as a whole to review both sets of descriptions and decide how best to present the combined view of the entire group. In NAEP's 1990 mathematics report, the descriptions will be supported by all the anchor items available for public-release (some in the body of the report and some in an Appendix). For each grade level at which the item was administered, each item will be accompanied by its p-value for the total population assessed and the p-values for each anchor level. The various steps in the procedure are detailed in the remainder of the paper.

The Scale Anchoring Analysis

NAEP's scale anchoring is grounded in an empirical process whereby the scaled assessment results are analyzed to delineate sets of items that discriminate between adjacent performance levels on the scale. For the 1990 mathematics assessment these levels were 200, 250, 300, and 350. That is, for these four levels, items were identified that were likely to be answered correctly by students performing at a particular level on the scale and much less likely to be answered correctly by students performing at the next lower level.

To provide a sufficient pool of respondents, in identifying anchor items, students at level 200 were defined as those whose estimated mathematics proficiency was between 187.5 and 212.5, students at 250 were defined as those with estimated proficiency between 237.5 and 262.5, those at 300 had estimated proficiencies between 287.5 and 312.5, and those at 350 between 337.5 and 362.5. In theory, proficiency levels above 350 or below 200 could have been defined, however, so few students in the assessment performed at the extreme ends of the scale that it was not possible to do so.

The 1990 mathematics scale anchoring analysis was based on the scaled proficiency results for fourth, eighth, and twelfth graders participating in the 1990 assessment. As illustrated below, for each item in the NAEP assessment, ETS determined the weighted percentage and raw frequency for students at each of the four scale levels correctly answering the item. This was done for each of the grade levels at which the item was administered, and for the grade levels combined, if the item was administered at more than one grade level. For example, regardless of the grade level, the data for each item were analyzed as shown in the following sample.

Sample Scale Anchoring Results

Scale Point	200	250	300	350
Weighted P-value	0.49	0.85	0.96	0.98
Raw Frequency	902	1555	1271	276

It should be noted that the percentages of students answering the item correctly at the four scale levels differ from the over all p-value for the total sample at any one grade level. Although the p-values for the total sample are also provided as part of the scale anchoring analysis.

As described below, criteria were applied to the scale-level results and an analysis conducted to delineate the items that discriminated between scale levels. Because it was the lowest level being defined, level 200 did not have to be analyzed in terms of the next lower level, but only for the percentage of students at that level answering the item correctly. More specifically, for an item to anchor at level 200:

1) The p-value for students at level 200 had to be greater than or equal to 0.65, and

2) the calculation of the p-value at that level had to have been based on at least 100 students.

As an example,

Level 200 Anchor Item Results				
Scale Point	200	250	300	350
Weighted P-value	0.65	0.89	0.98	1.00
Raw Frequency	116	706	510	23

For an item to anchor at the remaining levels, additional criteria had to be met. For example, to anchor at level 250:

1) The p-value for students at level 250 had to be greater than or equal to 0.65;

2) the p-value for students at level 200 had to be less than or equal to 0.50;

3) the difference between the two p-values had to be at least 0.30; and

4) the calculations of the p-values at both levels 200 and 250 had to have been based on at least 100 students.

The following data set illustrates the results for a level 250 anchor item:

Level 250 Anchor Item Results				
Scale Point	200	250	300	350
Weighted P-value	0.38	0.75	0.89	0.98
Raw Frequency	247	569	509	83

The same principles were used to identify anchor items at levels 300 and 350. For example,

1) The p-value at the anchor level had to be greater than or equal to 0.65;

2) the p-value at the adjacent lower level had to be less than or equal to 0.50;

3) the differences between the p-values had to be greater than or equal to 0.30; and

4) the p-values at the adjacent levels being considered had to have been based on at least 100 students.

For example, the following results were obtained for an item anchoring at level 300:

Level 300 Anchor Item Results				
Scale Point	<u>200</u>	<u>250</u>	<u>300</u>	<u>350</u>
Weighted P-value	0.11	0.28	0.83	1.00
Raw Frequency	134	670	512	52

The results below are for an item anchoring at level 350:

Level 350 Anchor Item Results				
Scale Point	<u>200</u>	<u>250</u>	<u>300</u>	<u>350</u>
Weighted P-value	0.00	0.22	0.37	0.94
Raw Frequency	50	324	585	241

In summary, for any given anchor item, the students at the anchor level are likely to answer the item correctly ($p\geq.65$) and the students at the next lower level are less likely to answer the item correctly ($p\geq.30$) and somewhat unlikely to answer the item correctly ($p\leq.50$). Collectively, as identified through this procedure, the 1990 NAEP mathematics items at each anchor level represented advances in students' understandings from one level to the next--mathematical areas where students at that level were more likely to answer items correctly than were students at the next lower level.

Preparing for the Mathematics Item Anchoring Panel Meeting

The analysis procedures described above yielded 35 questions that anchored at level 200, 30 questions at level 250, 48 questions at level 300, and 30 questions at level 350. Additionally, to provide some information for cross-referencing purposes, items that "almost anchored" were also identified. These items would have fulfilled all the criteria, except that one of the p-values under consideration was less than 0.05 different from the criterion value. This included items that, because of rounding, had results that appeared to meet the criteria, but were not identified in the analysis. This procedure yielded some additional items at each score point (level 200--8 items, level 250--16 items, level 300--16 items, level 350--13 items) that could be used for further context in developing descriptions.

In preparation for use by the scale anchoring panelists, the items were placed in notebooks by section in the following order: anchored at 200, almost anchored at 200, anchored at 250, almost anchored at 250, anchored at 300, etc. Again, for further cross-referencing purposes, the remaining items in the assessment were also included in the notebook under the "did not anchor" heading. Each item was accompanied by its scoring guide (for open-ended items) and by the full anchoring documentation; that is, the anchoring information for each grade level at which an item was administered, the anchoring information across grades, the p-value for the total population of respondents at each grade level, and the mathematics content-area and process classifications for the items.

As described in *Mathematics Objectives, 1990 Assessment* the mathematics assessment was designed to measure five content areas, each with three ability levels. To ensure that the anchoring performance descriptions tied back to the assessment specifications, within anchor level sections, the items in the notebooks were sorted by the five content areas--numbers and operations; measurement; geometry; data analysis, probability, and statistics; and algebra and functions. Within content area, the items were sorted by ability level--procedural knowledge, conceptual understanding, and problem-solving.

The Scale Anchoring Panel

Twenty mathematics educators were invited to participate in the anchoring process. They represented teachers at the various grade levels involved, state mathematics supervisors from several of the 38 states (including Washington, D.C.) participating in the Trial State Assessment, large-city mathematics curriculum coordinators, and college mathematics professors and researchers. The group was also balanced by region of the country, race/ethnicity, and gender. One panelist was unable to attend at the last minute, resulting in 19 participants (See Appendix A for a list of the participants.

The Process for Developing the Descriptions

The two-and-one-half day anchoring meeting began in the afternoon of the first day, during which time panelists were thoroughly briefed in the anchoring process and given their assignment. Which was, with the objectives for the 1990 mathematics assessment as a reference, to use the information in the anchor item notebooks to describe the mathematical knowledge, understandings, and problem-solving abilities demonstrated by the students at each anchor level in each of the five content areas. Based on the items anchoring at each anchor level (cross-referenced with "almost anchored" and "did not anchor" items), the panelists were asked to draft a description of achievement at each level in one-half page or less.

The meeting was structured so that the entire second day was devoted to the panelists working with staff in two independent groups to accomplish this task. In each of the independent groups, panelists and staff worked together to analyze the knowledge, skills, and reasoning abilities required by each item. Lists were developed portraying these for each mathematics content area at each anchor level. Based on these question by question analyses prominently displayed around the room on poster paper, each group of panelists then drafted a description of performance for each anchor level. The two sets of draft descriptions can be found in a later section of this appendix.

On the third day, the panelists and staff met as a whole to combine the two independently derived sets of descriptions. They also worked on developing short "titles" or descriptors for each category, and selecting example items to accompany the anchor level descriptions. Finally, the panelists were asked to discuss and indicate where the material described at the four levels might generally occur in the typical K-12 curriculum.

Both groups agreed that the two drafts were very similar and that with some final review and editing, either set would have appropriately described the anchor item information. However, they did like the benefit of the cross-validation process and the fact that more people were able to participate in the process. As the group worked through the two descriptions, they identified preferences for some parts of each of the descriptions, resolved some issues, and made some formatting decisions. The combined view was checked by staff against the anchoring data, edited, and sent to the panelists for final review. The final draft of the descriptions is presented in Figure 1.

FIGURE 1

Description of Mathematics Proficiency at Four Levels on the NAEP Scale

Level 200--Simple Additive Reasoning and Problem-Solving with Whole Numbers

Students at this level have some degree of understanding of simple quantitative relationships involving whole numbers. They can solve simple addition and subtraction problems with and without regrouping. Using a calculator, they can extend these abilities to multiplication and division problems. These students can identify solutions to one-step word problems and select the greatest four-digit number from a list.

In measurement, these students can read a ruler as well as common weight and graduated scales. They also can make volume comparisons based on visualization and determine the value of coins. In geometry, these students can recognize simple figures. In data analysis, they are able to read simple bar graphs. In the algebra dimension, these students can recognize translations of word problems to numerical sentences and extend simple pattern sequences.

Level 250--Simple Multiplicative Reasoning and Two-Step Problem-Solving

Students at this level have extended their understanding of quantitative reasoning with whole numbers from additive to multiplicative settings. They can solve routine one-step multiplication and division problems involving remainders and two-step addition and subtraction problems involving money. Using a calculator, they can identify solutions to other elementary two-step word problems. In these basic problem-solving situations, they can identify missing or extraneous information and have some knowledge of when to use computational estimation. They have a rudimentary understanding of such concepts as whole number place value, "even," factor," and "multiple."

In measurement, these students can use a ruler to measure objects, convert units within a system when the conversions require multiplication, and recognize a numerical expression solving a measurement word problem. In geometry, they demonstrate an initial understanding of basic terms and properties, such as parallelism and symmetry. In data analysis, they can complete a bar graph, sketch a circle graph, and use information from graphs to solve simple problems. They are beginning to understand the relationship between proportion and probability. In algebra, they are beginning to deal informally with a variable through numerical substitution in the evaluation of simple expressions.

FIGURE 1 (continued)

Level 300--Reasoning and Problem-Solving Involving Fractions, Decimals, Percents, Elementary Geometric Properties, and Simple Algebraic Manipulations

Students at this level are able to represent, interpret, and perform simple operations with fractions and decimal numbers. They are able to locate fractions and decimals on number lines, simplify fractions, and recognize the equivalence between common fractions and decimals, including pictorial representations. They can interpret the meaning of percents less than and greater than 100 and apply the concepts of percentages to solve simple problems. These students demonstrate some evidence of using mathematical notation to interpret expressions, including those with exponents and negative integers.

In measurement, these students can find the perimeters and areas of rectangles, recognize relationships among common units of measure, and use proportional relationships to solve routine problems involving similar triangles and scale drawings. In geometry, they have some mastery of the definitions and properties of geometric figures and solids.

In data analysis, these students can calculate averages, select and interpret data from tabular displays, pictographs, and line graphs, compute relative frequency distributions, and have a beginning understanding of sample bias. In algebra, they can graph points in the Cartesian plane and perform simple algebraic manipulations such as simplifying an expression by collecting like terms, identifying the solution to open linear sentences and inequalities by substitution, and checking and graphing an interval representing a compound inequality when it is described in words. They can determine and apply a rule for simple functional relations and extend a numerical pattern.

Level 350--Reasoning and Problem-Solving Involving Geometric Relationships, Algebraic Equations, and Beginning Statistics and Probability

Students at this level have extended their knowledge of number and algebraic understanding to include some properties of exponents. They can recognize scientific notation on a calculator and make the transition between scientific notation and decimal notation. In measurement, they can apply their knowledge of area and perimeter of rectangles and triangles to solve problems. They can find the circumferences of circles and the surface areas of solid figures. In geometry, they can apply the Pythagorean theorem to solve problems involving indirect measurement. These students also can apply their knowledge of the properties of geometric figures to solve problems, such as determining the slope of a line.

In data analysis, these students can compute means from frequency tables, and determine the probability of a simple event. In algebra, they can identify an equation describing a linear relation provided in a table and solve literal equations and a system of two linear equations. They are developing an understanding of linear functions and their graphs, as well as functional notation, including the composition of functions. They can determine the nth term of a sequence and give counter examples to disprove an algebraic generalization.

Reporting the Anchor Item Results

Because some items are kept secure to use in future assessments to measure trends in performance across time, not all of the anchor items can be shown in NAEP reports. However, the panelists decided that in addition to selecting seven or eight items (at least one from each of the five mathematics content areas, if possible) to accompany the descriptions in the main body of the report being prepared to discuss the results for the 1990 state and national mathematics assessments, the remaining anchor items available for public release should be contained in the appendix to the report. However, an additional five to 17 items per anchor level will be contained in the report of NAEP's 1990 state and national assessments.

Further, it was decided that each anchor item in the report should, for each grade level at which it was assessed, be accompanied by the overall percentage of success on the item as well as the anchor level information for each grade at which it was assessed. This should help prevent confusions between the percentages of success on the individual anchor items illustrating particular levels on the scale and the percentage of students who perform at or above each scale level. The anchor level descriptions, the accompanying sample items, and the appendix have been incorporated into the report, which is currently undergoing widespread review by the state representatives, NCES, and NAGB.

Draft Descriptions Prepared Independently by the Two Groups of Panelists

Group A

DRAFT DESCRIPTION

LEVEL 200

Students at this level have a beginning intuitive understanding of quantitative relationships among whole numbers, particularly in the area of additive reasoning. They can read and interpret basic mathematical symbols, add and subtract whole numbers without a calculator, perform straightforward multi-operations problems with a calculator, and compare four digit whole numbers. They can identify models that represent concepts, including region models of fractions. They can use addition and subtraction to solve one-step story problems and find the solutions to simple number sentences. They can read weight and volume scales, determine the value of coins, and read a ruler. They have a beginning knowledge of symmetry and can extend simple geometric patterns. They can read bar graphs and locate the coordinates on a grid.

LEVEL 250

Students at this level are developing their understanding of the quantitative relationships among whole numbers, to include multiplicative reasoning. They can select from among the four basic operations to solve one-step word problems, including some division problems requiring interpretation of remainders.
They can use addition and subtraction to solve two-step word problems, some of which deal with money and apply their understanding of whole number place value. They can convert units of measure, use their understanding of multiplication to solve simple number sentences, and analyze simple problem-situations to determine extraneous or missing information. They can measure with a ruler and have a beginning understanding of basic geometric terms. They can complete bar graphs and pie charts, as well as use the information from graphs and scales to solve problems.
They have an initial understanding of basic probability concepts and can evaluate simple algebraic expressions.

LEVEL 300

Students at this level demonstrate a beginning understanding of the relationships between fractions, decimals, and percents. For example they can locate fractions and decimals on number lines, reduce fractions, and recognize the equivalence between common fractions and decimals, including picture representations. They can interpret the meaning of percents less than and greater than 100 and apply the concepts of simple percentages to solve word problems. They show some indications of proportional reasoning and an extended ability to read mathematical symbols, including negative numbers and exponents. They can find the perimeter and area of rectangles in simple situations, recognize relationships among common units of measure, and use proportions to solve problems, including scale drawings and similar triangles. They understand the definitions and properties of geometric figures and can use visionalization skills with two- and three-dimensional figures. When given a set of data, they can compute the mean. They also can identify the probability of a simple event and have a beginning understanding of bias in sampling. They have an expanded facility in reading a variety of tables and graphs, including line graphs and pictographs. Students can identify a solution or solution sets and graph the solutions of simple linear inequalities. They can collect like terms in a simple algebraic expression and evaluate multiplicative algebraic expressions that include integers. They can find and apply the rule for functional relations and extend a numerical pattern. They can identify coordinates of a point and plot the point on a coordinate grid. They have some familiarity with algebraic identities.

LEVEL 350

Students at this level can recognize scientific notation on a calculator and transfer from scientific to regular notation. They can apply their knowledge of area and perimeter or rectangles (including squares) and triangles to solve problems. They can find the surface areas of solid figures, and apply their knowledge of area and circumference of circles to solve problems. They are familiar with the concept of precision in measurement. they can apply the pythagorean theorem to solve problems. They can also apply their knowledge of the properties of geometric figures to solve problems, such as determining the slope of a line, identifying the line of symmetry in a rotated figure, and identifying perpendicular line segments embedded in two-dimensional figures. They can compute weighted means from frequency tables, use a sample space to determine the probability of an event, and construct a sample space for a simple event. Students can identify an equation to describe a linear relation given in a table. They can solve a literal equation and a system of linear equations. They can simplify expressions involving powers of ten. They are developing an understanding of functions and their graphs, as well as functional notation, including composition of functions. They can determine the nth term of a sequence and give counterexamples to disprove a generalization.

Group B

DRAFT DESCRIPTION

LEVEL 200

Learners at this level can solve simple addition and subtraction problems with and without regrouping. Using a calculator, their problem-solving abilities extend to simple multiplication and division settings. They are able to solve one-step word problems involving translation from verbal to numerical form as well as interpret place value to order whole numbers. Using models, they are able to recognize fractions.

Students are able to identify common symmetrical figures. In measurement they can read a variety of scales, including the direct reading of a ruler. These learners also have some sense of gross measurement based on visualization. In data interpretation, they are able to read data from a bar graph. Given a visual shape pattern, they are able to recognize and extend the patterns. They are also capable of solving open sentences with missing addends.

LEVEL 250

Learners at this level can solve one-step multiplication and division whole number translation problems without calculators and most forms of one- and two-step whole number translation problems involving any operation with a calculator. They are able to handle decimal problems involving using money and apply place value concepts to decimal settings. the number concepts of factor, multiple, even, and odd are familiar, and whole number estimation skills are developing.

Students' measurement skills include the ability to use a ruler to measure objects, convert simple unit measures within a system, and translate verbal measurement descriptions to numerical representations in application problems. In geometry, students can draw a line of symmetry for common figures and demonstrate basic understanding of two- and three-dimensional shapes by relating vocabulary and elementary properties of shapes and solids in real-wold contexts.

In data representation, they can sketch and interpret bar graphs and circle graphs. They also have an elementary understanding of the relationship of proportion and chance. In algebraic settings, these learners can solve open sentences involving subtraction. They are beginning to be able to deal informally with the concept of variable through substitution in the evaluation of expressions.

LEVEL 300

Learners at this level are able to interpret, represent, and operate with fractions and decimal numbers. Their knowledge of percent includes both percents greater than and less than 100% and they are able to perform multi-step problems involving simple calculations with percent. There is evidence of the beginning of proportional reasoning at this level.

These learners have use of exponential notation and are capable of performing simple algebraic manipulations such as simplifying an expression by collecting like terms, solving open linear sentences and inequalities by substitution, and checking and graphing an interval representing a compound inequality when it is described in words.

Students at this level also have the ability to operate with integers and graph points on the Cartesian plane. There is the emergence of students' ability to identify, establish, and apply simple functional relationships.

Learners at level 300 are also able to both calculate an average and use an average value to discuss a population total. They are capable of selecting and interpreting data from a tabular display, pictographs, and two-group comparison graphs. Their understanding of probability includes the calculation of relative frequency probabilities and relating such information to models. Some simple understanding of sample bias is also present.

LEVEL 350

Learners at this level have extended their knowledge of number and algebraic understanding to include exponential representations, including properties of exponents, both on paper and with calculators. They have command of percent in all forms, including markup and discount problems. These learners can also generate required terms to extend or describe patterns in linear sequences or establish a general formula. Other evidence suggests they have considerable understanding of functional notation and the ability to represent and interpret situations involving the graphs of linear functions. Their manipulation skills include the ability to solves a system of linear equations.

Students at level 350 are able to calculate group averages from a grouped frequency table as well as create the sample space for and calculate the probability of events involving more than one object.

INDEX

DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA

249133