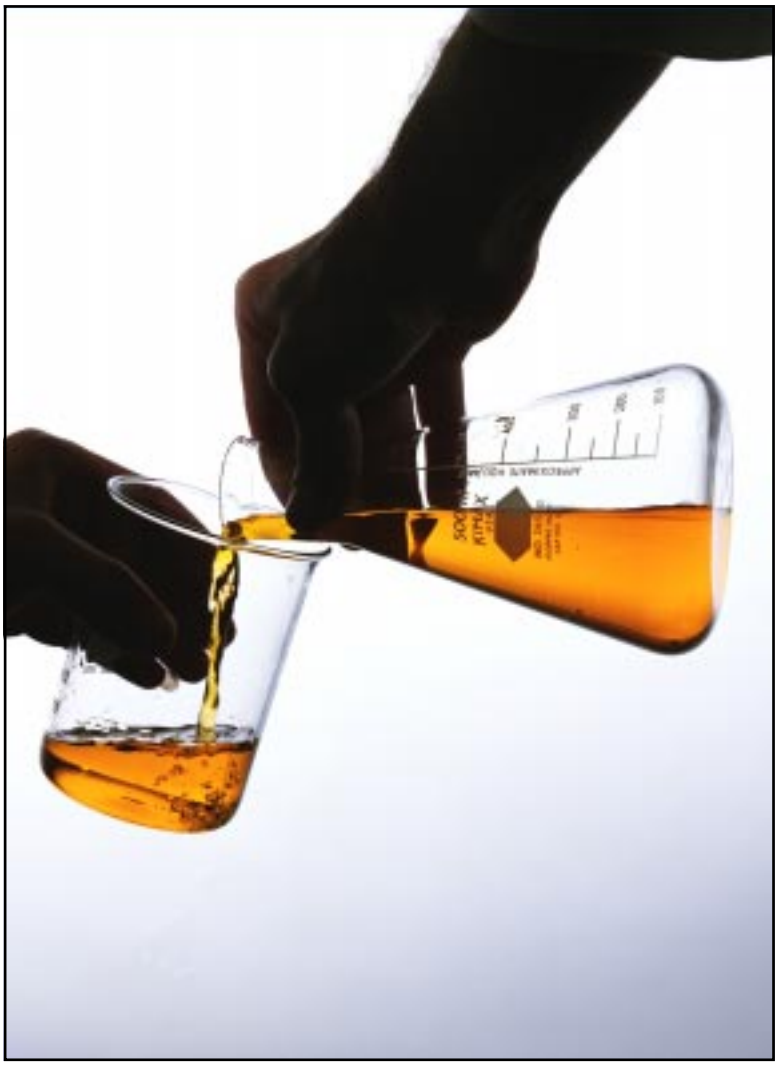
NATIONAL CENTER FOR EDUCATION STATISTICS

NAEP 1996 SCIENCE

State Report for New Hampshire



THE NATION'S REPORT CARD

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

Findings from the National Assessment of Educational Progress

What is The Nation's Report Card?

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

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

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

The National Assessment Governing Board

NATIONAL CENTER FOR EDUCATION STATISTICS

NAEP 1996 SCIENCE STATE REPORT

for

NEW HAMPSHIRE

Christine Y. O'Sullivan
Laura Jerry
Nada Ballator
Fiona Herr

In collaboration with
Audrey Champagne, Peggy Carr,
Will Pfeiffenberger, and Mistilina Sato

September 1997

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

Prepared by Educational Testing Service under a cooperative agreement with the National Center for Education Statistics.

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

Office of Educational Research and Improvement
Ramon C. Cortines
Acting Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

Education Assessment Group
Gary W. Phillips
Associate Commissioner

September 1997

SUGGESTED CITATION
O'Sullivan, C.Y., Jerry, L., Ballator, N., and Herr, F.
NAEP 1996 Science State Report for New Hampshire,
Washington, DC: National Center for Education Statistics, 1997.

FOR MORE INFORMATION
Contact:
Arnold A. Goldstein
202-219-1741

For ordering information on this report, write:

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

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

This report also is available on the World Wide Web: http://www.ed.gov/NCES/naep

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

INTRODUCTION

New Hampshire was one of the original participants in the state-level National Assessment of Educational Progress (NAEP) in 1990. Results were reported for New Hampshire in 1990 (for mathematics at grade 8), in 1992 (for mathematics at grades 4 and 8, and reading at grade 4), and again in 1994 (for reading at grade 4). In 1994, New Hampshire participated with both public and nonpublic school samples, but met the participation rate requirements[1] for publication of results only for their public school sample.

In 1996, New Hampshire again participated at grade 8 only, but with both public and nonpublic school samples. The grade 8 public school sample did not meet the guidelines for publication, due to low participation rate (see Appendix A); however, the nonpublic school sample was sufficient to meet the guidelines for publication.

The results of the NAEP 1996 science assessment in New Hampshire's nonpublic schools are presented here. The total sample size for nonpublic schools is modest, and only those results based on pre-established NAEP minimum sample sizes are reported. In this report are tables showing the demographic composition of the sample, the participation rates for sample components, and the students' average scale scores.

A full set of Appendices is included:

Appendix A	Reporting NAEP 1996 Science Results
Appendix B	The NAEP 1996 Science Assessment
Appendix C	Technical Appendix

[1] To ensure comparability across jurisdictions, NCES has established guidelines for school and student participation rates. Appendix A highlights these guidelines, which are applied separately for public and nonpublic schools. For jurisdictions failing to meet the initial school participation rate of 70 percent for either public or nonpublic schools, appropriate results are not reported. Jurisdictions that exceed the 70 percent rate but fail to meet others of these guidelines are noted in tables and figures in NAEP reports containing state-by-state results.

OVERVIEW

Monitoring the performance of students in subjects such as science is a key concern of the citizens, policy makers, and educators who direct educational reform efforts. The 1996 National Assessment of Educational Progress (NAEP) in science assessed the current level of science performance as a mechanism for informing education reform. This science assessment is the first to be constructed on a new framework, and it is also the first to be given at the state level. In 1996, New Hampshire participated in the NAEP science assessment at grade 8 but only the nonpublic schools met the participation guidelines. This report contains those results.

What Is NAEP?

The National Assessment of Educational Progress (NAEP), the "Nation's Report Card," is the only ongoing nationally representative assessment of what America's students know and can do in various academic subjects.

NAEP is a congressionally mandated project of the National Center for Education Statistics, U.S. Department of Education. In 1990 Congress authorized a voluntary state-by-state NAEP assessment. The 1990 Trial State Assessment in mathematics at grade 8 was the first state-level NAEP assessment. Since then, state-level assessments have taken place in 1992 and 1994 in reading (grade 4), in 1992 and 1996 in mathematics (grades 4 and 8), and in 1996 in science (grade 8). In 1996, 44 states, the District of Columbia, Guam, and the Department of Defense Education Activity Schools took part in the NAEP state assessment program; the state science assessment was at grade 8 only, although grades 4, 8, and 12 were assessed at the national level as usual.

Results are provided only for group performance, since NAEP is forbidden by law to report results at an individual or school level. By making information on student performance available to policy makers, educators, and the general public, NAEP is an integral part of our nation's evaluation of the conditions and progress of education.

What Was Assessed?

The science assessment was crafted to measure the content and skills specified in the science framework for the 1996 assessment.[2] Two organizing concepts underlie the science framework. First, scientific knowledge should be structured so as to make factual information meaningful. The way in which knowledge is structured should be influenced by the context in which the knowledge is being presented. Second, science performance depends on knowledge of facts, the ability to integrate this knowledge into larger constructs, and the capacity to use the tools, procedures, and reasoning processes of science to develop an increased understanding of the natural world. Thus, the framework called for the NAEP 1996 science assessment to include the following:

- Multiple-choice questions that assess students' knowledge of important facts and concepts and that probe their analytical reasoning skills;
- Constructed-response questions that explore students' abilities to explain, integrate, apply, reason about, plan, design, evaluate, and communicate scientific information; and
- Hands-on tasks that probe students' abilities to use materials to make observations, perform investigations, evaluate experimental results, and apply problem-solving skills.

The core of the science framework is organized along two dimensions. The first dimension divides science into three major fields: earth, physical, and life sciences. The second dimension defines characteristic elements of knowing and doing science: conceptual understanding, scientific investigation, and practical reasoning. Each question in the assessment is categorized as measuring one of the elements of knowing and doing within one of the fields of science (e.g., scientific investigation in the context of earth science). The framework also contains two overarching domains — the nature of science and the organizing themes of science. The nature of science encompasses the historical development of science and technology, the habits of mind that characterize science, and the methods of inquiry and problem solving. It also includes the nature of technology — specifically, design issues involving the application of science to real-world problems and associated trade-offs or compromises. The themes of science include the notions of systems and their application in the scientific disciplines, models and their functioning in the development of scientific understanding, and patterns of change as they are exemplified in natural phenomena. A fuller description of the framework is provided in Appendix B.

[2] *Science Framework for the 1996 National Assessment of Educational Progress* . (Washington, DC: National Assessment Governing Board, 1993).

Who Was Assessed?

Eighth-Grade Student Characteristics

Table 1 provides demographic profiles of the eighth-grade nonpublic school students in New Hampshire, the Northeast region, and the nation, based on data collected during the 1996 state and national science assessments. As described in Appendix A, the sample providing the state data and the sample providing the regional and national data are separate.

THE NATION'S REPORT CARD
1996 State Assessment

TABLE 1 — GRADE 8

Profile of Students in New Hampshire, the Northeast Region, and the Nation

Demographic Subgroups		Nonpublic Schools
		Percentage
RACE/ETHNICITY		
New Hampshire	White	91 (2.9)
	Black	2 (****)
	Hispanic	4 (0.9)
	Asian/Pacific Islander	2 (0.9)
	American Indian	1 (****)
Northeast	White	79 (5.7)
	Black	10 (3.3)
	Hispanic	8 (4.6)
	Asian/Pacific Islander	3 (1.4)
	American Indian	0 (****)
Nation	White	80 (2.7)
	Black	7 (1.5)
	Hispanic	9 (2.1)
	Asian/Pacific Islander	4 (0.8)
	American Indian	1 (0.2)
PARENTS' EDUCATION		
New Hampshire	Did not finish high school	1 (****)
	Graduated from high school	10 (2.1)
	Some education after high school	14 (3.9)
	Graduated from college	71 (4.4)
	I don't know	4 (1.2)
Northeast	Did not finish high school	1 (0.6)
	Graduated from high school	10 (2.1)
	Some education after high school	15 (3.8)
	Graduated from college	65 (6.4)
	I don't know	9 (1.6)
Nation	Did not finish high school	2 (0.3)
	Graduated from high school	10 (1.1)
	Some education after high school	17 (1.8)
	Graduate from college	66 (3.0)
	I don't know	6 (1.0)
GENDER		
New Hampshire	Male	51 (8.7)
	Female	49 (8.7)
Northeast	Male	51 (3.6)
	Female	49 (3.6)
Nation	Male	51 (1.8)
	Female	49 (1.8)

(continued on next page)

THE NATION'S REPORT CARD 1996 State Assessment

TABLE 1 — GRADE 8 (continued)

Profile of Students in New Hampshire, the Northeast Region, and the Nation

Demographic Subgroups		Nonpublic Schools
		Percentage
TITLE 1		
New Hampshire	Participated	0 (****)
	Did not participate	100 (****)
Northeast	Participated	14 (7.1)
	Did not participate	86 (7.1)
Nation	Participated	7 (3.6)
	Did not participate	93 (3.6)
FREE/REDUCED-PRICE LUNCH		
New Hampshire	Eligible	0 (****)
	Not eligible	37 (19.4)
	Information not available	63 (19.4)
Northeast	Eligible	12 (7.0)
	Not eligible	47 (15.2)
	Information not available	42 (16.3)
Nation	Eligible	7 (3.4)
	Not eligible	49 (7.7)
	Information not available	44 (8.2)

The standard errors of the statistics appear in parentheses. It can be said with about 95 percent confidence that, for each population of interest, the value for the entire population is within ± 2 standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference (see Appendix A for details). The percentages for Race/Ethnicity may not add to 100 percent because some students categorized themselves as "Other." **** Standard error estimates cannot be accurately determined.

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 Science Assessment.

Schools and Students Assessed

To ensure comparability across jurisdictions, NCES has established guidelines for school and student participation rates, highlighted in Appendix A. Jurisdictions failing to meet these guidelines are noted in tables and figures in NAEP reports containing state-by-state results. For jurisdictions failing to meet the initial school participation rate of 70 percent, results are not reported.

For public schools in 1996, the weighted school participation rate before substitution was 66 percent for New Hampshire; a level of 70 percent was required for publication of public school results. New Hampshire's weighted school participation rate for nonpublic schools was 83 percent, meeting the guidelines for publication. Table 2 summarizes participation data for nonpublic schools and students sampled in New Hampshire for the 1996 state assessment program in science.[3]

In New Hampshire, 12 nonpublic schools participated in the 1996 eighth-grade science assessment. The weighted school participation rate after substitution in 1996 was 83 percent for nonpublic schools, which means that the eighth-grade students in this sample were directly representative of 83 percent of all the eighth-grade nonpublic school students in New Hampshire. Because the weighted participation rate for nonpublic schools was below 85 percent and the weighted rate after substitution was below 90 percent, New Hampshire received a notation based on Guideline 5 because samples not meeting *both* requirements are possibly subject to bias due to school nonresponse.[4]

In each school, a random sample of students was selected for the assessment. In 1996, on the basis of sample estimates, none of the eighth-grade nonpublic school population in New Hampshire was classified as having limited English proficiency (LEP), and none of the students had an Individualized Education Plan (IEP). Consequently, none were excluded on basis of LEP or IEP.[5]

In New Hampshire, 179 eighth-grade students from nonpublic schools were assessed in 1996. The weighted student participation rate was 95 percent for nonpublic schools. This means that the sample of eighth-grade students who took part in the assessment was directly representative of 95 percent of the eligible nonpublic school student population in participating schools in New Hampshire (that is, all students from the population represented by the participating schools, minus those students excluded from the assessment). The overall weighted response rate (school rate times student rate) was 79 percent for nonpublic schools. This means that the sample of students who participated in the assessment was directly representative of 79 percent of the eligible eighth-grade nonpublic school population in New Hampshire.

[3] For a detailed discussion of the NCES guidelines for sample participation, see Appendix A of this report or the *Technical Report of the NAEP 1996 State Assessment Program in Science.* (Washington, DC: National Center for Education Statistics, 1997).

[4] *Ibid.*

[5] An IEP is a plan written for a student who has been determined to be eligible for special education. The IEP typically sets forth goals and objectives for the student and describes a program of activities and/or related services necessary to achieve the goals and objectives.

THE NATION'S REPORT CARD 1996 State Assessment

TABLE 2 — GRADE 8

School and Student Participation in New Hampshire

	1996
	Nonpublic Schools
SCHOOL PARTICIPATION	
Weighted school participation rate before substitution	83%
Weighted school participation rate after substitution	83%
Number of schools originally sampled	20
Number of schools not eligible	4
Number of schools in original sample participating	12
Number of substitute schools provided	3
Number of substitute schools participating	0
Total number of participating schools	12
STUDENT PARTICIPATION	
Weighted student participation rate after makeups	95%
Number of students selected to participate in the assessment	190
Number of students withdrawn from the assessment	2
Percentage of students who were of Limited English Proficiency	0%
Percentage of students excluded from the assessment due to Limited English Proficiency	0%
Percentage of students who had an Individualized Education Plan	0%
Percentage of students excluded from the assessment due to Individualized Education Plan status	0%
Number of students to be assessed	188
Number of students assessed	179
Overall weighted response rate	79%

RESULTS FOR NONPUBLIC SCHOOLS

NAEP Science Results for New Hampshire

The NAEP 1996 state assessment program in science provides a wealth of information on the abilities and skills of the eighth-grade students in participating jurisdictions. The NAEP results presented in Table 3 are average scale scores on the NAEP science scale, representing what students in New Hampshire's nonpublic schools know and can do in science.

Reporting NAEP Science Results

The NAEP Science Scale

The NAEP 1996 science assessment spans the broad fields of science in each of the grades assessed. Because of the survey nature of the assessment and the breadth of the domain, each student participating cannot be expected to answer all the questions in the assessment since this would impose an unreasonable burden on students and their schools. Thus, each student was administered a portion of the assessment, and data were combined across students to report on the performance of eighth graders.

Student responses to the assessment questions were analyzed to determine the percentage of students responding correctly to each multiple-choice question and the percentage of students achieving each of the score categories for constructed-response questions. Item response theory (IRT) methods were used to produce scales that summarized results for each of the three fields of science (i.e., earth, physical, and life) at each grade level. An overall composite scale also was developed separately at each of grades 4, 8, and 12 by weighting the separate scales based on the relative importance of each field of science in the NAEP science framework. Results presented in this report are based on this overall composite scale, which ranges from 0 to 300.

The use of separate grade-specific reporting scales for the science assessment is consistent with the National Assessment Governing Board's 1993 policy that future NAEP assessments be developed using within-grade frameworks and that scaling be carried out within grade. Because this science assessment was based on a new framework, and no comparisons with previous NAEP science assessments were possible, a new scale was developed. The ranges of the science scales (from 0 to 300) differ by design from the 0-to-500 reporting scales used in other NAEP subject areas; they were chosen to discourage inappropriate cross-grade comparisons and to minimize confusion with other common test scales.

The national average on the science scale is 150, including both public and nonpublic school students. The national average scale score reported here (162) is for students in nonpublic schools only. (Additional details of the scaling procedures are in Appendix C of this report and in the forthcoming *NAEP 1996 Technical Report*.)

Science Achievement Levels

A companion report, being issued by the National Assessment Governing Board, will present the NAEP 1996 science results in terms of achievement levels. As authorized by the NAEP legislation and adopted by the National Assessment Governing Board, the achievement levels are based on the Board's judgments about what are reasonable performance expectations for students on the NAEP 1996 science assessment. The achievement levels for the NAEP 1996 science assessment were adopted on an interim basis, indicating that they may be revised when other information becomes available, such as the fourth- and twelfth-grade results from the Third International Mathematics and Science Study (TIMSS).

Interpreting NAEP Results

Because the students whose average science scale scores reported here are sampled, and are not the entire population of eighth graders in a jurisdiction, the numbers reported are necessarily *estimates*. As such, they are subject to a measure of uncertainty, reflected in the *standard error* of the estimate. When the percentages or average scale scores of certain groups are compared, it is essential to take the standard error into account, rather than to rely solely on observed similarities or differences. Therefore, the comparisons discussed in this report are based on *statistical tests* that consider both the magnitude of the difference between the means or percentages and the standard errors of those statistics.

The statistical tests determine whether the evidence, based on the data from the groups in the *sample*, is strong enough to conclude that the averages or percentages are really different for those groups in the *population*. If the evidence is strong (i.e., the difference is statistically significant), the report describes the group averages or percentages as being different (e.g., one group performed *higher than* or *lower than* another group). Rather than relying on the apparent magnitude of the difference between sample averages or percentages, the reader is cautioned to rely on the results of the statistical tests to determine whether those sample differences are likely to represent actual differences between the groups in the population. The statistical tests are discussed in greater detail in Appendix A.

Mapping Items on the Scale

This discussion and the item map illustration refer to eighth-grade students in the national assessment, whose scores may not resemble those of eighth-grade students in New Hampshire's nonpublic schools. Students' performance is summarized on the NAEP science scale which ranges from 0 to 300. Sample questions shown in Figure 1 illustrate the range of performance on the NAEP science scale for grade 8. Each sample question is one that is likely to be answered correctly by a student whose score is at or near the given percentile.

THE NATION'S REPORT CARD 1996 State Assessment

FIGURE 1

Sample Questions Likely to Be Answered Correctly by Grade 8 Students At or Near Selected Percentiles

Percentile	Question
10th	Find typical yearly rainfall from a graph. (104)
25th	*Explain the impact of fish death on an ecosystem.* (127)
50th	Identify the effect of acid rain. (150)
75th	Understand where earthquakes occur. (172)
90th	*Explain why lightning is seen before thunder is heard.* (194)

The value in parentheses represents the scale score attained by students who had a 65 percent probability of reaching a given level on a constructed-response question (in italic type) or a 74 percent probability of correctly answering a 4-option multiple-choice question (in regular type).
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 Science Assessment.

To illustrate the range of performance in more detail, questions from the assessment were "mapped" onto a 0 to 300 scale, as in Figure 2. The item map is a visual representation of the scale, showing selected questions in positions corresponding to their difficulty. The item map shows which questions a student of any particular ability is likely to answer correctly. The position of the question on the scale represents a dividing line: Students who attained scores greater than the score corresponding to the question's difficulty are likely to answer it correctly, while students with scores below that degree of difficulty are less likely to answer it correctly.

More specifically, students who scored below the scale score associated with a particular question had less than a 65 percent probability of earning a given amount of credit on a constructed-response question or less than a 74 percent probability of correctly answering a multiple-choice question. A small proportion of these students — those near but below the question's position on the scale — may be more likely than not to answer the question correctly (between 50 and 65 or 74 percent). Such students are not considered "able" to answer the question, since they have not achieved sufficient consistency in their responses.

Figure 2 is an item map for grade 8.[6] Multiple-choice questions are shown in regular type; constructed-response questions are in italic type.[7] An example of how to interpret the item map may be helpful. In this figure, a multiple-choice question involving interpreting a graph maps at the 136 point on the scale. This means that eighth-grade students with science scale scores at or above 136 are likely to answer this question correctly — that is, they have at least a

[6] Details on the procedures used to develop the item map are provided in the forthcoming *NAEP 1996 Technical Report*. The procedures are similar to those used in past NAEP assessments.
[7] The placement of constructed-response questions is based on (1) the "mapping" of a score of 3 on a 3-point scoring guide for short constructed-response questions and (2) the "mapping" of a score of at least 3 on a 4-point scoring guide and a score of at least 4 on a 5-point scoring guide for extended constructed-response questions.

74 percent chance of doing so.[8] Put slightly differently, this question is answered correctly by at least 74 of every 100 students scoring at or above the 136 scale-score level. Note that this does not mean that students at or above the 136 scale score always answer the question correctly or that students below the 136 scale score always answer it incorrectly.

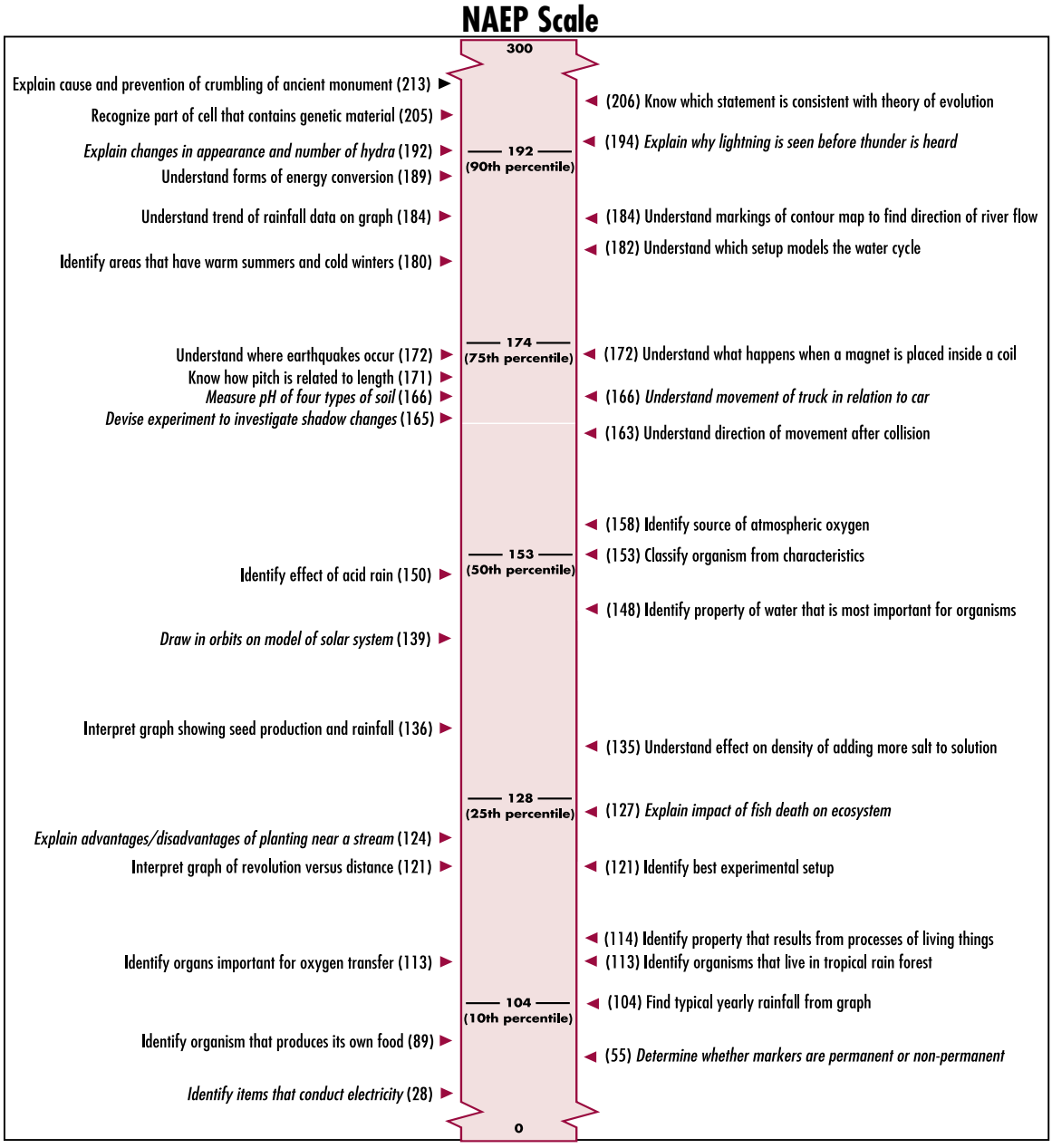
As another example, consider the constructed-response question that maps at a scale score of 194. This question concerns the differing speeds of light and sound. Scoring of responses to this question allows for partial credit by using a three-level scoring guide. Mapping a question at the 194 scale score indicates that at least 65 percent of the students performing at or above this point achieved a score of 3 ("Complete") on the question. Among students with lower scores, less than 65 percent gave complete responses to the question.

[8] For constructed-response questions, a criterion of 65 percent was used. For multiple-choice questions, the criterion was 74 percent. The use of a higher criterion for multiple-choice questions reflected the students' ability to "guess" the correct answer from among the alternatives.

THE NATION'S REPORT CARD
1996 State Assessment

FIGURE 1.2 — GRADE 8

Map of Selected Questions on the NAEP Science Scale for Grade 8



NOTE: Position of questions is approximate and an appropriate scale range is displayed for grade 8.
Italic type indicates a constructed-response question. Regular type denotes a multiple-choice question.

Each grade 8 science question was mapped onto the NAEP 0-to-300 science scale. The position of the question on the scale represents the scale score attained by students who had a 65 percent probability of reaching a given score level on a constructed-response question or a 74 percent probability of correctly answering a 4-option multiple-choice question. Only selected questions are presented. Percentiles of scale score distribution are referenced on the map.

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 Science Assessment.

Results for New Hampshire

The table below shows not only the average scale scores for students but also the distribution of their average scores at five selected percentiles. The distribution of the scores through these percentiles encourages the reader to consider the performance of the students in the various groupings (such as state, region, or nation) as overlapping ranges of heterogeneous performance, rather than as a simple monolithic average. As an example, consider the results below which show that, for New Hampshire's nonpublic schools, students at the twenty-fifth percentile scored 161, while the average scale score for the Northeast is 160. This means that at least seventy-five percent of New Hampshire's nonpublic school students achieved scores higher than the average scale score for nonpublic school students in the Northeast.

- In New Hampshire, the average science scale score of students attending nonpublic schools (176) was higher than that of nonpublic school students across the nation (162).

THE NATION'S REPORT CARD 1996 State Assessment

TABLE 3 — GRADE 8

Distribution of Science Scale Scores for Students in Nonpublic Schools

	Average Scale Score	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile
Nonpublic Schools						
1996 New Hampshire	176 (3.6)	148 (4.1)	161 (4.6)	176 (2.9)	190 (6.5)	204 (5.4)
Northeast	160 (4.9)!	118 (5.8)!	140 (4.7)!	163 (9.1)!	181 (2.6)!	200 (3.8)!
Nation	162 (2.5)	123 (8.1)	143 (3.1)	164 (3.0)	182 (2.8)	199 (2.1)

The NAEP science scale ranges from 0 to 300. The standard errors of the statistics appear in parentheses. It can be said with about 95 percent confidence that, for each population of interest, the value for the entire population is within ± 2 standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference (see Appendix A for details).
! Interpret with caution — the nature of the sample does not allow accurate determination of the variability of this statistic.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 Science Assessment.

This report and the reports for other participating jurisdictions will be available on the NCES Web site (http://www.ed.gov/NCES/naep). A separate report describes additional eighth-grade science assessment results for the nation and the states, as well as the national results for grades 4 and 12.[9]

The Appendices that follow cover the NAEP 1996 science assessment for all jurisdictions.

[9] O'Sullivan, C.Y., C.M. Reese, and J. Mazzeo. *NAEP 1996 Science Report Card for the Nation and the States* . (Washington, DC: National Center for Education Statistics, 1997).

APPENDIX A

Reporting NAEP 1996 Science Results

A.1 Participation Guidelines

As was discussed in the Introduction, unless the overall participation rate for a jurisdiction is sufficiently high, the assessment results for that jurisdiction may be subject to appreciable nonresponse bias. Moreover, even if the overall participation rate is high, significant nonresponse bias may exist if the nonparticipation that does occur is heavily concentrated among certain types of schools or students. The following guidelines concerning school and student participation rates in the state assessment program were established to address four significant ways in which nonresponse bias could be introduced into the jurisdiction sample estimates.

The first three guidelines describe the determination of whether a jurisdiction is eligible to have its results published. Guidelines 4-11 describe conditions under which a jurisdiction's published results will include a notation. Such a notation would indicate the possibility of bias in particular results, due to nonresponse from segments of the sample. Note that in order for a jurisdiction's results to be published without notations, that jurisdiction must comply with all guidelines. (A thorough discussion of the NAEP participation guidelines can be found in the *Technical Report of the NAEP 1996 State Assessment Program in Science*.)

Guidelines on the Publication of NAEP Results

Guideline 1 — Publication of Public School Results

A jurisdiction will have its public school results published in the *NAEP 1996 Science Report Card* (or in other reports that include all state-level results) if and only if its weighted participation rate for the initial sample of public schools is greater than or equal to 70 percent. Similarly, a jurisdiction will receive a separate *NAEP 1996 Science State Report* if and only if its weighted participation rate for the initial sample of public schools is greater than or equal to 70 percent.

Guideline 2 — Publication of Nonpublic School Results

A jurisdiction will have its nonpublic school results published in the *NAEP 1996 Science Report Card* (or in other reports that include all state-level results) if and only if its weighted participation rate for the initial sample of nonpublic schools is greater than or equal to 70 percent **AND** meets minimum sample size requirements.[1] A jurisdiction eligible to receive a separate *NAEP 1996 Science State Report* under guideline 1 will have its nonpublic school results included in that report if and only if that jurisdiction's weighted participation rate for the initial sample of nonpublic schools is greater than or equal to 70 percent **AND** meets minimum sample size requirements. If a jurisdiction meets guideline 2 but fails to meet guideline 1, a separate *NAEP 1996 Science State Report* will be produced containing only nonpublic school results.

Guideline 3 — Publication of Combined Public and Nonpublic School Results

A jurisdiction will have its combined results published in the *NAEP 1996 Science Report Card* (or in other reports that include all state-level results) if and only if both guidelines 1 and 2 are satisfied. Similarly, a jurisdiction eligible to receive a separate *NAEP 1996 Science State Report* under guideline 1 will have its combined results included in that report if and only if guideline 2 is also met.

Guidelines for Notations of NAEP Results

Guideline 4 — Notation for Overall Public School Participation Rate

A jurisdiction that meets guideline 1 will receive a notation if its weighted participation rate for the initial sample of public schools was below 85 percent **AND** the weighted public school participation rate after substitution was below 90 percent.

Guideline 5 — Notation for Overall Nonpublic School Participation Rate

A jurisdiction that meets guideline 2 will receive a notation if its weighted participation rate for the initial sample of nonpublic schools was below 85 percent **AND** the weighted nonpublic school participation rate after substitution was below 90 percent.

[1] Minimum participation size requirements for reporting nonpublic school data consist of two components: (1) a school sample size of six or more participating schools and (2) an assessed student sample size of at least 62.

Guideline 6 — Notation for Strata-Specific Public School Participation Rate

A jurisdiction that is not already receiving a notation under guideline 4 will receive a notation if the sample of public schools included a class of schools with similar characteristics that had a weighted participation rate (after substitution) of below 80 percent, and from which the nonparticipating schools together accounted for more than five percent of the jurisdiction's total weighted sample of public schools. The classes of schools from each of which a jurisdiction needed minimum school participation levels were determined by degree of urbanization, minority enrollment, and median household income of the area in which the school is located.

Guideline 7 — Notation for Strata-Specific Nonpublic School Participation Rate

A jurisdiction that is not already receiving a notation under guideline 5 will receive a notation if the sample of nonpublic schools included a class of schools with similar characteristics that had a weighted participation rate (after substitution) of below 80 percent, and from which the nonparticipating schools together accounted for more than five percent of the jurisdiction's total weighted sample of nonpublic schools. The classes of schools from each of which a jurisdiction needed minimum school participation levels were determined by type of nonpublic school (Catholic versus non-Catholic) and location (metropolitan versus nonmetropolitan).

Guideline 8 — Notation for Overall Student Participation Rate in Public Schools

A jurisdiction that meets guideline 1 will receive a notation if the weighted student response rate within participating public schools was below 85 percent.

Guideline 9 — Notation for Overall Student Participation Rate in Nonpublic Schools

A jurisdiction that meets guideline 2 will receive a notation if the weighted student response rate within participating nonpublic schools was below 85 percent.

Guideline 10 — Notation for Strata-Specific Student Participation Rates in Public Schools

A jurisdiction that is not already receiving a notation under guideline 8 will receive a notation if the sampled students within participating public schools included a class of students with similar characteristics that had a weighted student response rate of below 80 percent, and from which the nonresponding students together accounted for more than five percent of the jurisdiction's weighted assessable public school student sample. Student groups from which a jurisdiction needed minimum levels of participation were determined by the age of the student, whether or not the student was classified as a student with a disability (SD) or of limited English proficiency (LEP), and the type of assessment session (monitored or unmonitored), as well as school level of urbanization, minority enrollment, and median household income of the area in which the school is located.

Guideline 11 — Notation for Strata-Specific Student Participation Rates in Nonpublic Schools

A jurisdiction that is not already receiving a notation under guideline 9 will receive a notation if the sampled students within participating nonpublic schools included a class of students with similar characteristics that had a weighted student response rate of below 80 percent, and from which the nonresponding students together accounted for more than five percent of the jurisdiction's weighted assessable nonpublic school student sample. Student groups from which a jurisdiction needed minimum levels of participation were determined by the age of the student, whether or not the student was classified as a student with a disability (SD) or of limited English proficiency (LEP), and the type of assessment session (monitored or unmonitored), as well as type and location of school.

A.2 NAEP Reporting Groups

The NAEP state assessment program provides results for groups of students defined by shared characteristics — region of the country, gender, race/ethnicity, parental education, type of school, and participation in federally funded Title I programs and the free/reduced-price lunch component of the National School Lunch Program. Based on criteria described later in this appendix, results are reported for subpopulations only when sufficient numbers of students and adequate school representation are present. For public school students, there must be at least 62 students in a particular subgroup from at least 5 primary sampling units (PSUs).[2] For nonpublic school students, the minimum requirement is 62 students in a particular subgroup from at least 6 different schools. However, the data for all students, regardless of whether their subgroup was reported separately, were included in computing overall results for New Hampshire. Definitions of the subpopulations referred to in this report are presented on the following pages.

Region

Results are reported for four regions of the nation: Northeast, Southeast, Central, and West. The states included in each region are shown in Figure A.1. All 50 states and the District of Columbia are listed. Territories and the two Department of Defense Education Activity jurisdictions were not assigned to a region.

Regional results are based on national assessment samples, not on aggregated state assessment program samples. Thus, the regional results are based on a *different* and *separate* sample from that used to report the state results.

THE NATION'S REPORT CARD 1996 State Assessment

FIGURE A.1

Regions of the Country

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

Note: The part of Virginia that is included in the Washington, DC, metropolitan area is included in the Northeast region; the remainder of the state is in the Southeast region.

[2] For the State Assessment Program, a PSU is most often a single school; for the national assessment, a PSU is a selected geographic region (a county, group of counties, or metropolitan statistical area).

Gender

Results are reported separately for males and females.

Race/Ethnicity

The racial/ethnic results presented in this report attempt to provide a clear picture based on several sources. The race/ethnicity variable is an imputed definition of race/ethnicity derived from up to three sources of information. This variable is used for race/ethnicity subgroup comparisons. Two questions from the student demographics questionnaire were used in the determination of derived race/ethnicity:

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 this question 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 question, or provided information that was illegible or could not be classified, responses to the question below were examined in an effort to determine race/ethnicity.

Which best describes you?

- ° White (not Hispanic)
- ° Black (not Hispanic)
- ° Hispanic ("Hispanic" means someone who is from a Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or other Spanish or Hispanic background.)
- ° Asian or Pacific Islander ("Asian or Pacific Islander" means someone who is from a Chinese, Japanese, Korean, Filipino, Vietnamese, or 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 (specify) _____

Students' race/ethnicity was then assigned on the basis of their response. For students who filled in the sixth oval ("Other") or provided illegible information or information that could not be classified, or did not respond at all, race/ethnicity was assigned as determined by school records.[3]

Derived race/ethnicity could not be determined for students who did not respond to either of the demographic questions and for whom a race/ethnicity designation was not provided by the school.

The details of how race/ethnicity classifications are derived is presented so that the readers can determine the usefulness of the results for their particular uses. It should be noted that a nonnegligible number of students indicated a Hispanic background (e.g., Puerto Rican or Cuban) **and** indicated that a racial/ethnic category other than Hispanic best described them. These students were classified as Hispanic according to the rules described above. Also, information from the schools did not always correspond to students' descriptions of themselves.

Parents' Highest Level of Education

The variable representing level of parental education is derived from responses to two questions from the set of general background questions. Students were asked to indicate the extent of their mothers' education:

How far in school did your mother go?

- 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 a similar question about their fathers' education level:

How far in school did your father go?

- 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.

[3] The procedure for assigning race/ethnicity was modified for Hawaii. See the *Technical Report for the NAEP 1996 State Assessment Program in Science* for details.

This information was combined into one parental education reporting variable through the following procedure. If a student indicated the extent of education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. For students who did not know the level of education for both parents or did not know the level for one parent and did not respond for the other, the parental education level was classified as "I don't know." If the student did not respond for either parent, the student was recorded as having provided no response.

It should be noted that, nationally, approximately one-tenth of eighth graders reported not knowing the education level of either of their parents.

Type of School

Samples for the 1996 state assessment program were expanded to include students attending nonpublic schools (Catholic schools and other religious and private schools) in addition to students attending public schools. The expanded coverage was instituted for the first time in 1994. Samples for the 1990 and 1992 Trial State Assessment programs had been restricted to public school students only. For those jurisdictions meeting pre-established participation rate standards (see earlier section of this appendix), separate results are reported for public schools, for nonpublic schools, and for the combined public and nonpublic school samples. The combined sample for each jurisdiction also contains students attending Bureau of Indian Affairs (BIA) schools and Department of Defense Domestic Dependent Elementary and Secondary Schools (DDESS) in that jurisdiction. These two categories of schools are not included in either the public or nonpublic school samples.

Note that eighth graders in the DDESS and Department of Defense Dependents Schools (DoDDS)[4] were assessed in 1996 as separate jurisdictions and reported as jurisdictions with public school samples only.

[4] The Department of Defense Dependents Schools (DoDDS) refers to overseas schools (i.e., schools outside the United States). Department of Defense Domestic Dependent Elementary and Secondary Schools (DDESS) refers to domestic schools (i.e., schools in the United States). DoDDS and DDESS fourth grades were also assessed in science, for a special report.

Title I Participation

On the basis of available school records, students were classified either as currently participating in a Title I program or receiving Title I services, or as not receiving such services. The classification only refers to the school year when the assessment was administered (i.e., the 1995—96 school year) and is not based on participation in previous years. If the school did not offer any Title I programs or services, all students in that school were classified as not participating.

Free/Reduced-Price School Lunch Program Eligibility

On the basis of available school records, students were classified either as currently eligible for the Department of Agriculture's free/reduced-price lunch program or not. The classification refers only to the school year when the assessment was administered (i.e., the 1995—96 school year) and is not based on eligibility in previous years. If the school did not participate in the program or if school records were not available, all students in that school were classified as "Information not available."

A.3 Guidelines for Analysis and Reporting

This report describes science performance for eighth graders and compares the results for various groups of students within this population — for example, those who have certain demographic characteristics or who responded to a specific background question in a particular way. The report examines the results for individual demographic groups and individual background questions. It does not include an analysis of the relationships among combinations of these subpopulations or background questions.

Drawing Inferences from the Results

Because the percentages of students in these subpopulations and their average scale scores are based on samples — rather than on the entire population of eighth graders in a jurisdiction — the numbers reported are necessarily *estimates*. As such, they are subject to a measure of uncertainty, reflected in the *standard error* of the estimate. When the percentages or average scale scores of certain groups are compared, it is essential to take the standard error into account, rather than to rely solely on observed similarities or differences. Therefore, the comparisons discussed in this report are based on *statistical tests* that consider both the magnitude of the difference between the averages or percentages and the standard errors of those statistics.

One of the goals of the science state assessment program is to estimate scale score distributions and percentages of students in the categories described in A.2 for the overall populations of eighth-grade students in each participating jurisdiction based on the particular samples of students assessed. The use of *confidence intervals*, based on the standard errors, provides a way to make inferences about the population average scale scores and percentages in a manner that reflects the uncertainty associated with the sample estimates. An estimated sample average scale score ± 2 standard errors approximates a *95 percent confidence interval* for the corresponding population average or percentage. This means that one can conclude with approximately 95 percent confidence that the average scale score of the entire population of interest (e.g., all eighth-grade students in public schools in a jurisdiction) is within ± 2 standard errors of the sample average.

As an example, suppose that the average science scale score of the students in a particular jurisdiction's eighth-grade sample were 156 with a standard error of 1.2. A 95 percent confidence interval for the population average would be as follows:

$$\text{Average} \pm 2 \text{ standard errors} = 156 \pm 2 \times (1.2) = 156 \pm 2.4 =$$

$$156 - 2.4 \text{ and } 156 + 2.4 = (153.6, 158.4)$$

Thus, one can conclude with 95 percent confidence that the average scale score for the entire population of eighth-grade students in public schools in that jurisdiction is between 153.6 and 158.4.

Similar confidence intervals can be constructed for percentages, *if the percentages are not extremely large or extremely small*. For extreme percentages, confidence intervals constructed in the above manner may not be appropriate, and accurate confidence intervals can be constructed only by using procedures that are quite complicated.

Extreme percentages, defined by both the magnitude of the percentage and the size of the sample from which it was derived, should be interpreted with caution. (The forthcoming *Technical Report of the NAEP 1996 State Assessment Program in Science* contains a more complete discussion of extreme percentages.)

Analyzing Subgroup Differences in Averages and Percentages

The statistical tests determine whether the evidence, based on the data from the groups in the *sample*, is strong enough to conclude that the averages or percentages are actually different for those groups in the *population*. If the evidence is strong (i.e., the difference is statistically significant), the report describes the group averages or percentages as being different (e.g., one group performed *higher than* or *lower than* another group), regardless of whether the sample averages or sample percentages appear to be about the same or not. If the evidence is not sufficiently strong (i.e., the difference is not statistically significant), the averages or percentages are described as being *not significantly different* — again, regardless of whether the sample averages or sample percentages appear to be about the same or widely discrepant. The reader is cautioned to rely on the results of the statistical tests rather than on the apparent magnitude of the difference between sample averages or percentages when determining whether those sample differences are likely to represent actual differences between the groups in the population.

In addition to the overall results, this report presents outcomes separately for a variety of important subgroups. Many of these subgroups are defined by shared characteristics of students, such as their gender or race/ethnicity. Other subgroups are defined by the responses of the assessed students' science teachers to questions in the science teacher questionnaire.

In Chapter 1 of this report, differences between the jurisdiction and the nation were tested for overall science scale score and for each of the fields of science. In Chapter 2, significance tests were conducted for the overall scale score for each of the subpopulations. In Chapters 3 through 6, comparisons were made across subgroups for responses to various background questions.

As an example of comparisons across subgroups, consider the question: *Do students who reported discussing studies at home almost every day exhibit higher average science scale scores than students who report never or hardly ever doing so?*

To answer the question posed above, begin by comparing the average science scale score for the two groups being analyzed. If the average for the group that reported discussing their studies at home almost every day is higher, it may be tempting to conclude that that group does have a higher science scale score than the group that reported never or hardly ever discussing their studies at home. However, even though the averages differ, there may be no real difference in performance between the two groups in the population because of the uncertainty associated with the estimated average scale scores of the groups in the sample. Remember that the intent is to make a statement about the entire population, not about the particular sample that was assessed. The data from the sample are used to make inferences about the population as a whole.

As discussed in the previous section, each estimated sample average scale score (or percentage) has a degree of uncertainty associated with it. It is therefore possible that if all students in the population (rather than a sample of students) had been assessed or if the assessment had been repeated with a different sample of students or a different, but equivalent, set of questions, the performances of various groups would have been different. Thus, to determine whether there is a *real* difference between the average scale score (or percentage of students with a certain attribute) for two groups in the population, an estimate of the degree of uncertainty associated with the difference between the scale score averages or percentages of those groups must be obtained for the sample. This estimate of the degree of uncertainty — called *the standard error of the difference* between the groups — is obtained by taking the square of each group's standard error, summing these squared standard errors, and then taking the square root of this sum.

In a manner similar to that in which the standard error for an individual group average or percentage is used, the *standard error of the difference* can be used to help determine whether differences between groups in the population are real. The difference between the mean scale score or percentage of the two groups — 2 *standard errors of the difference* — represents an approximate 95 percent confidence interval. If the resulting interval includes zero, there is insufficient evidence to claim a real difference between groups in the population. If the interval does not contain zero, the difference between groups is *statistically significant* (different) at the 0.05 level.

As another example, to determine whether the average science scale score of eighth-grade males is higher than that of eighth-grade females in a particular jurisdiction's public schools, suppose that the sample estimates of the average scale scores and standard errors for males and females were as follows:

Group	Average Scale Score	Standard Error
Males	148	0.9
Females	146	1.1

The difference between the estimates of the average scale scores of males and females is two points (148 — 146). The standard error of this difference is

$$\sqrt{0.9^2 + 1.1^2} = 1.4$$

Thus, an approximate 95 percent confidence interval for this difference is

Mean difference ± 2 standard errors of the difference =

$2 \pm 2 \times (1.4) = 2 \pm 2.8 = 2 - 2.8$ and $2 + 2.8 = (-0.8, 4.8)$

The value zero is within this confidence interval, which extends from -0.8 to 4.8 (i.e., zero is between -0.8 and 4.8). Thus, there is insufficient evidence to claim a difference in average science scale score between the populations of eighth-grade males and females in public schools in the hypothetical jurisdiction.

Throughout this report, when the average scale scores or percentages for two groups were compared, procedures like the one described above were used to draw the conclusions that are presented in the text.[5] If a statement appears in the report indicating that a particular group had a *higher* (or *lower*) average scale score than a second group, the 95 percent confidence interval for the difference between groups did not contain zero. 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. A procedure based on effect sizes was used. Statistically significant differences that are rather small are described in the text as *somewhat higher* or *somewhat lower*. When a statement indicates that the average scale score or percentage of some attribute was *not significantly different* for two groups, the confidence interval included zero, and thus no difference could be assumed between the groups. The reader is cautioned to avoid drawing conclusions solely on the basis of the magnitude of the difference. A difference between two groups in the sample that appears to be slight may represent a statistically significant difference in the population because of the magnitude of the standard errors. Conversely, a difference that appears to be large may not be statistically significant.

The procedures described in this section, and the certainty ascribed to intervals (e.g., a 95 percent confidence interval), are based on statistical theory that assumes that only one confidence interval or test of statistical significance is being performed. However, in each chapter of this report, many different groups are being compared (i.e., multiple sets of confidence intervals are being calculated). In sets of confidence intervals, statistical theory indicates that the certainty associated with the entire set of intervals is less than that attributable to each individual comparison from the set if considered individually. To hold the certainty level for the set of comparisons at a particular level (e.g., 0.95), modifications (called multiple comparison procedures) must be made to the methods described in the previous section. One such procedure — the *Bonferroni* method — was used in the analyses described in this report to form confidence intervals for the differences between groups whenever sets of comparisons were considered.[6] Using this method, the confidence intervals in the text that are based on sets of comparisons are more conservative than those described on the previous pages. In other words, some comparisons that were individually statistically significant using the methods previously described may not be statistically significant when the Bonferroni method was used to take the number of related comparisons into account.

[5] The procedure described above (especially the estimation of the standard error of the difference) is, in a strict sense, appropriate only when the statistics being compared come from independent samples. For certain comparisons in the report, the groups were not independent. In those cases, a different (and more appropriate) estimate of the standard error of the difference was used.

[6] Miller, R.G. *Simultaneous Statistical Inference*. (New York: McGraw-Hill, 1966).

Most of the multiple comparisons in this report pertain to relatively small sets or "families" of comparisons. For example, when comparisons were discussed concerning students' reports of parental education, six comparisons were conducted — all pairs of the four parental education levels. In these situations, Bonferroni procedures were appropriate. However, the maps in Chapter 1 of this report display comparisons between New Hampshire and all other participating jurisdictions. The "family" of comparisons in this case was as many as 46. To control the certainty level for a large family of comparisons, the False Discovery rate (FDR) criterion[7] was used. Unlike the Bonferroni procedures which control the familywise error rate (i.e., the probability of making even one false rejection in the set of comparisons), the Benjamini and Hochberg (BH) approach using the FDR criterion controls the expected proportion of falsely rejected hypotheses as a proportion of all rejected hypotheses. Bonferroni procedures may be considered conservative for large families of comparisons.[8] In other words, using the Bonferroni method would produce more statistically nonsignificant comparisons than using the BH approach. Therefore, the BH approach is potentially more powerful for comparing New Hampshire to all other participating jurisdictions. A more detailed description of the Bonferroni and BH procedures appears in the *Technical Report of the NAEP 1996 State Assessment Program in Science*.

Statistics with Poorly Estimated Standard Errors

Not only are the averages and percentages reported in NAEP subject to uncertainty, but their standard errors are as well. In certain cases, typically when the standard error is based on a small number of students or when the group of students is enrolled in a small number of schools, the amount of uncertainty associated with the standard errors may be quite large. Throughout this report, estimates of standard errors subject to a large degree of uncertainty are followed by the symbol "!". In such cases, the standard errors — and any confidence intervals or significance tests involving these standard errors — should be interpreted cautiously. Further details concerning procedures for identifying such standard errors are discussed in the *Technical Report of the NAEP 1996 State Assessment Program in Science*.

[7] Benjamini, Y. and Y. Hochberg. "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society, Series B,* 57(1). (pp. 289-300, 1994).

[8] Williams, V.S.L., L.V. Jones, and J.W. Tukey. *Controlling Error in Multiple Comparisons, with Special Attention to the National Assessment of Educational Progress*. (Research Triangle Park, NC: National Institute of Statistical Sciences, December 1994).

Minimum Subgroup Sample Sizes

Results for science performance and background variables were tabulated and reported for groups defined by gender, race/ethnicity, parental education, type of school, and participation in federally funded Title I programs and the free/reduced-price school lunch component of the National School Lunch Program. NAEP collects data for five racial/ethnic subgroups (White, Black, Hispanic, Asian/Pacific Islander, and American Indian/Alaskan Native) and four levels of parents' education (Graduated From College, Some Education After High School, Graduated From High School, and Did Not Finish High School) plus the category "I Don't Know."

In many jurisdictions, and for some regions of the country, the number of students in some of these groups was not sufficiently high to permit accurate estimation of performance and/or background variable results. As a result, data are not provided for the subgroups with students from very few schools or for the subgroups with very small sample sizes. For results to be reported for any state assessment subgroup, public school results must represent at least 5 primary sampling units (PSUs) and nonpublic school results must represent at least 6 schools. For results to be reported for any national assessment subgroup, at least 5 PSUs must be represented in the subgroup. In addition, a minimum sample of 62 students per subgroup is required. For statistical tests pertaining to subgroups, the sample size for both groups has to meet the minimum sample size requirements.

The minimum sample size of 62 was determined by computing the sample size required to detect an effect size of 0.5 total-group standard deviation units with a probability of 0.8 or greater. The effect size of 0.5 pertains to the *true* difference between the average scale score of the subgroup in question and the average scale score for the total eighth-grade public school population in the jurisdiction, divided by the standard deviation of the scale score in the total population. If the *true* difference between subgroup and total group mean is 0.5 total-group standard deviation units, then a sample size of at least 62 is required to detect such a difference with a probability of 0.8. Further details about the procedure for determining minimum sample size appear in the *Technical Report of the NAEP 1996 State Assessment Program in Science*.

Describing the Size of Percentages

Some of the percentages reported in the text of the report are given qualitative descriptions. For example, the number of students currently taking a biology class might be described as "relatively few" or "almost all," depending on the size of the percentage in question. Any convention for choosing descriptive terms for the magnitude of percentages is to some degree arbitrary. The descriptive phrases used in the report and the rules used to select them are shown below.

Percentage	Descriptive Term Used in Report
$p = 0$	None
$0 < p \leq 8$	A small percentage
$8 < p \leq 13$	Relatively few
$13 < p \leq 18$	Less than one fifth
$18 < p \leq 22$	About one fifth
$22 < p \leq 27$	About one quarter
$27 < p \leq 30$	Less than one third
$30 < p \leq 36$	About one third
$36 < p \leq 47$	Less than half
$47 < p \leq 53$	About half
$53 < p \leq 64$	More than half
$64 < p \leq 71$	About two thirds
$71 < p \leq 79$	About three quarters
$79 < p \leq 89$	A large majority
$89 < p < 100$	Almost all
$p = 100$	All

APPENDIX B

The NAEP 1996 Science Assessment

The science framework for the 1996 National Assessment of Educational Progress was produced under the auspices of the National Assessment Governing Board through a consensus process. The consensus process, managed by the Council of Chief State School Officers, with the National Center for Improving Science Education and the American Institutes for Research, developed the framework over a ten-month period between October 1990 and August 1991. The following factors guided the process for developing consensus on the science framework:[1]

- The active participation of individuals such as curriculum specialists, science teachers, science supervisors, state supervisors, administrators, individuals from business and industry, government officials, and parents;
- The representation of what is considered essential learning in science, and the recommendation of innovative assessment techniques to probe the critical abilities and content areas;
- The recognition of the lack of agreement on such things as common scope of instruction and sequence, components of scientific literacy, important outcomes of learning, and the nature of overarching themes in science.

While maintaining some conceptual continuity with the 1990 NAEP Science Assessment, the 1996 framework takes into account the current reforms in science education, as well as documents such as the science framework used for the 1991 International Assessment of Educational Progress. In addition, the Framework Steering Committee recommended that a variety of strategies, including the following, be used for assessing students' performance.[2]

[1] *Science Framework for the 1996 National Assessment of Educational Progress*. (Washington, DC: National Assessment Governing Board, 1993).

[2] *Ibid.*

- Performance tasks that allow students to manipulate physical objects and draw scientific understanding from the materials before them

- Constructed-response questions that provide insights into students' levels of understanding and ability to communicate in the sciences as well as their ability to generate, rather than simply recognize, information related to scientific concepts and their interconnections

- Multiple-choice items that probe students' conceptual understanding and ability to connect ideas in a scientifically sound way

B.1 Percentage of Assessment Time by Domain

The framework for the 1996 science assessment can be described as a two-dimensional matrix. The three fields of science (earth, physical, and life) make up the first dimension and ways of knowing and doing science (conceptual understanding, scientific investigation, and practical reasoning) make up the second dimension. Every question or task in the assessment is classified according to the two major dimensions. There are also two overarching domains — nature of science (that includes nature of technology) and themes (systems, models, and patterns of change).

In addition to describing the content of the assessment, the framework also recommends what percentage of time should be devoted to each field of science, each way of knowing and doing science, the nature of science, and themes.

In this section, each figure describes an element of the framework, and is followed by a table showing the *actual* distribution of assessment time as well as the distribution *recommended* by the framework. Care was taken to ensure congruence between the proportions actually used in the assessment and those recommended in the assessment specifications. Note that the tables represent all three grades assessed nationally; only grade 8 was assessed at the state level.

Figure B.1 describes the fields of science and Table B.1 shows the actual and recommended distribution of assessment time across each field. The ways of knowing and doing science are outlined in Figure B.2. The distribution of assessment time for this dimension, both actual and recommended, is depicted in Table B.2.

THE NATION'S REPORT CARD 1996 State Assessment

FIGURE B.1

Description of the Three Fields of Science

Earth Science

The earth science content assessed centers on objects and events that are relatively accessible or visible. The concepts and topics covered are solid Earth (lithosphere), water (hydrosphere), air (atmosphere), and the Earth in space. The solid Earth consists of composition; forces that alter its surface; the formation, characteristics and uses of rocks; the changes and uses of soil; natural resources used by humankind; and natural forces within the Earth. Concepts and topics related to water consist of the water cycle; the nature of oceans and their effects on water and climate; and the location of water, its distribution, characteristics, and effect of and influence on human activity. The air is broken down into composition and structure of the atmosphere (including energy transfer); the nature of weather; common weather hazards; and air quality and climate. The Earth in space consists of setting of the Earth in the solar system; the setting and evolution of the solar system in the universe; tools and technology that are used to gather information about space; apparent daily motions of the Sun, the Moon, the planets and the stars; rotation of the Earth about its axis, and the Earth's revolution around the Sun; and tilt of the Earth's axis that produces seasonal variations in the climate.

Physical Science

The physical science component relates to basic knowledge and understanding concerning the structure of the universe as well as the physical principles that operate within it. The major sub-topics probed are matter and its transformations, energy and its transformations, and the motion of things. Matter and its transformations are described by diversity of materials (classification and types and the particulate nature of matter); temperature and states of matter; properties and uses of material (modifying properties, synthesis of materials with new properties); and resource management. Energy and its transformations involve different forms of energy; energy transformations in living systems, natural physical systems, and artificial systems constructed by humans; and energy sources and use, including distribution, energy conversion, and energy costs and depletion. Motion is broken down into an understanding of frames of reference; force and changes in position and motion; action and reaction; vibrations and waves as motion; general wave behavior; electromagnetic radiation; and the interactions of electromagnetic radiation with matter.

Life Science

The fundamental goal of life science is to attempt to understand and explain the nature and function of living things. The major concepts assessed in life science are change and evolution, cells and their functions (not at grade 4), organisms, and ecology. Change and evolution includes diversity of life on Earth; genetic variation within a species; theories of adaptation and natural selection; and changes in diversity over time. Cells and their functions consists of information transfer; energy transfer for the construction of proteins; and communication among cells. Organisms are described by reproduction, growth and development; life cycles; and functions and interactions of systems within organisms. The topic of ecology centers on the interdependence of life — populations, communities, and ecosystems.

SOURCE: *Science Framework for the 1996 National Assessment of Educational Progress*. (Washington, DC: National Assessment Governing Board, 1993).

THE NATION'S REPORT CARD 1996 State Assessment

TABLE B.1

Distribution of Assessment Time by Field of Science

	Earth		Physical		Life	
	Actual	**Recommended**	**Actual**	**Recommended**	**Actual**	**Recommended**
Grade 4	33%	33%	34%	33%	33%	33%
Grade 8	30%	30%	30%	30%	40%	40%
Grade 12	33%	33%	33%	33%	34%	33%

THE NATION'S REPORT CARD 1996 State Assessment

FIGURE B.2

Description of Knowing and Doing Science

Conceptual Understanding

Conceptual understanding includes the body of scientific knowledge that students draw upon when conducting a scientific investigation or engaging in practical reasoning. Essential scientific concepts involve a variety of information including facts and events the student learns from science instruction and experiences with the natural environment and scientific concepts, principles, laws, and theories that scientists use to explain and predict observations of the natural world.

Scientific Investigation

Scientific investigation probes students' abilities to use the tools of science, including both cognitive and laboratory tools. Students should be able to acquire new information, plan appropriate investigations, use a variety of scientific tools, and communicate the results of their investigations.

Practical Reasoning

Practical reasoning probes students' ability to use and apply science understanding in new, real-world applications.

SOURCE: *Science Framework for the 1996 National Assessment of Educational Progress*. (Washington, DC: National Assessment Governing Board, 1993).

THE NATION'S REPORT CARD 1996 State Assessment

TABLE B.2

Distribution of Assessment Time by Knowing and Doing Science

	Conceptual Understanding		Scientific Investigation		Practical Reasoning	
	Actual	**Recommended**	**Actual**	**Recommended**	**Actual**	**Recommended**
Grade 4	45%	45%	38%	45%	17%	10%
Grade 8	45%	45%	29%	30%	26%	25%
Grade 12	44%	45%	28%	30%	28%	25%

The two overarching dimensions are described and accounted for by Figure B.3 and Table B.3, which describe the nature of science and the themes that transcend the scientific disciplines.

THE NATION'S REPORT CARD 1996 State Assessment

FIGURE B.3

Description of Overarching Domains

The Nature of Science

The nature of science incorporates the historical development of science and technology, the habits of mind that characterize these fields, and methods of inquiry and problem-solving. It also encompasses the nature of technology that includes issues of design, application of science to real-world problems, and trade-offs or compromises that need to be made.

Themes

Themes are the "big ideas" of science that transcend the various scientific disciplines and enable students to consider problems with global implications. The NAEP science assessment focuses on three themes: systems, models, and patterns of change.

- Systems are complete, predictable cycles, structures or processes occurring in natural phenomena. Students should understand that a system is an artificial construction created to represent, or explain a natural occurrence. Students should be able to identify and define the system boundaries, identify the components and their interrelationships and note the inputs and outputs to the system.

- Models of objects and events in nature are ways to understand complex or abstract phenomena. As such they have limits and involve simplifying assumptions but also possess generalizability and often predictive power. Students need to be able to distinguish the idealized model from the phenomenon itself and to understand the limitations and simplified assumptions that underlie scientific models.

- Patterns of change involve students' recognition of patterns of similarity and differences, and recognize how these patterns change over time. In addition, students should have a store of common types of patterns and transfer their understanding of a familiar pattern of change to a new and unfamiliar one.

SOURCE: *Science Framework for the 1996 National Assessment of Educational Progress*. (Washington, DC: National Assessment Governing Board, 1993).

THE NATION'S REPORT CARD 1996 State Assessment

TABLE B.3

Distribution of Assessment Time by Overarching Domains

	Nature of Science		Themes	
	Actual	**Recommended**	**Actual***	**Recommended**
Grade 4	19%	≥15%	53%	33%
Grade 8	21%	≥15%	49%	50%
Grade 12	31%	≥15%	55%	50%

* Several of the hands-on tasks were classified as themes.

SOURCE: *Science Framework for the 1996 National Assessment of Educational Progress*. (Washington, DC: National Assessment Governing Board, 1993).

B.2 The Assessment Design

The state science assessment used booklets that were identical to those used at grade 8 for the national assessment. Each student in the state assessment program in science received a booklet containing six sections. Three of these sections were blocks[3] of cognitive questions that assessed the knowledge and skills outlined in the framework, and the other three sections were sets of background questions. Two of the three cognitive sections were paper-and-pencil, and the third section consisted of a hands-on task with related questions. In the state assessment at grade 8, students were allowed 30 minutes to complete each cognitive block. (For the national assessment, students at grades 8 and 12 were allowed 30 minutes, while students at grade 4 were given cognitive blocks that each required 20 minutes to complete.)

At each grade level there were 15 different sections or blocks of cognitive questions, but each student's booklet contained only three of these blocks of items. Every block consisted of both multiple-choice and constructed-response questions. Short constructed-response questions required a few words or a sentence or two for an answer (e.g., briefly stating how nutrients move from the digestive system to the tissues) while the extended constructed-response questions generally required a paragraph or more (e.g., outlining an experiment to test the effect of increasing the amount of available food on the rate of increase of the hydra population). Some constructed-response questions also required diagrams, graphs, or calculations. It was expected that students could adequately answer the short constructed-response questions in about 2 to 3 minutes and the extended constructed-response questions in about 5 minutes.

[3] "Blocks" are collections of questions grouped, in part, according to the amount of time required to answer them.

Other features were built into the blocks of cognitive questions. Four of the blocks were hands-on tasks in which students were given a set of equipment and asked to conduct an investigation and answer questions relating to the investigation. Every student was assessed on one of these four blocks. A second feature was the inclusion of three theme blocks — one assessing systems, one assessing models, and one assessing patterns of change. For example, students were shown a simplified model of part of the Solar System with a brief description, and then asked a number of questions based on this scenario. Theme blocks were randomly placed in booklets, but not in all booklets. No student received more than one theme block.

Each booklet in the assessment also included three sets of student background questions. The first, consisting of general background questions, asked students about such things as mother's and father's level of education, reading materials in the home, homework, and school attendance. The second, consisting of science background questions, asked students questions about their classroom learning activities such as hands-on exercises, courses taken, use of specialized resources such as computers, and views on the utility and value of science. Students were given five minutes to complete each of these questionnaires. The third set contained five questions about students' motivation to do well on the assessment, their perception of the difficulty of the assessment, and their familiarity with the types of cognitive questions asked. This section took three minutes or less to complete.

Using information gathered from the field test, the booklets were carefully constructed to balance time requirements for the question types in each block. For more information on the design of the assessment, the reader is referred to Appendix C.

B.3 Usage of Question Types

The data in Table B.4 reflect the number of questions by type and by grade level for the 1996 assessment. One hundred and sixty-five multiple-choice (MC), 219 short constructed-response (SCR), and 59 extended constructed-response (ECR) questions make up the assessment, giving a total of 443 unique questions in the pool. Some of these questions were used at more than one grade level; thus, the sum at each grade level is greater than the total number of unique questions. For the state assessment program at grade 8, students responded to subsets (determined by booklet) of 74 multiple-choice questions, 100 short constructed-response questions, and 20 extended constructed-response tasks.

THE NATION'S REPORT CARD 1996 State Assessment

TABLE B.4

Distribution of Items by Question Type

	Grade 4			Grade 8			Grade 12		
	MC	**SRC**	**ERC**	**MC**	**SRC**	**ERC**	**MC**	**SRC**	**ERC**
Grade 4 only	42	57	12						
Grades 4 & 8 overlap	9	16	4	9	16	4			
Grade 8 only				44	58	13			
Grades 8 & 12 overlap				21	26	3	21	26	3
Grade 12 only							49	62	27
TOTAL by grade	51	73	16	74	100	20	70	88	30

MC — multiple-choice questions; SRC — short constructed-response questions; ERC — extended constructed-response questions

APPENDIX C

Technical Appendix: The Design, Implementation, and Analysis of the 1996 State Assessment Program in Science

C.1 Overview

The purpose of this appendix is to provide technical information about the 1996 state assessment program in science. It describes the design of the assessment and gives an overview of the steps used to implement the program, from the planning stages through the analysis of the data.

This appendix is one of several documents that provide technical information about the 1996 state assessment program. Readers interested in more details are referred to the *Technical Report of the NAEP 1996 State Assessment Program in Science*. Theoretical information about the models and procedures used in NAEP can be found in the special NAEP-related issue of the *Journal of Educational Statistics* (Summer 1992/Volume 17, Number 2) as well as previous national technical reports.

Educational Testing Service (ETS) was awarded the cooperative agreement for the 1996 NAEP programs, including the state assessment program. ETS was responsible for overall management of the programs as well as for development of the overall design, the cognitive questions and questionnaires, data analysis, and reporting. National Computer Systems (NCS) was a subcontractor to ETS on both the national and state NAEP programs. NCS was responsible for printing, distributing, and receiving all assessment materials, and for scanning and scoring the assessments. The National Center for Education Statistics (NCES) awarded a separate cooperative agreement to Westat, Inc., for handling all aspects of sampling and field operations for the national and state assessments for 1996.

Organization of the Technical Appendix

This appendix has the following organization:

- Section C.2 provides an overview of the design of the 1996 state assessment program in science.
- Section C.3 discusses the partially-balanced incomplete block (PBIB) spiral design used to assign cognitive questions to assessment booklets and assessment booklets to students.
- Section C.4 outlines the sampling design used for the 1996 state assessment program.
- Section C.5 summarizes Westat's field administration procedures.
- Section C.6 describes the flow of the data from receipt at NCS through data entry and professional scoring.
- Section C.7 summarizes the procedures used to weight the assessment data and to obtain estimates of the sampling variability of subpopulation estimates.
- Section C.8 describes the initial analyses performed to verify the quality of the data.
- Section C.9 describes the item response theory scales and the overall science composite scale created for the final analyses of the state assessment program data.
- Section C.10 provides an overview of the linking of the scaled results from the state assessment program in science to those from the national assessment.

C.2 Design of the NAEP 1996 State Assessment Program in Science

The design for the state assessment program in science included the following major aspects:

- Participation at the jurisdiction level was voluntary, except for a few jurisdictions for which NAEP has been mandated by the state legislature.
- Students from public and nonpublic schools were assessed. Nonpublic schools included Catholic schools, other religious schools, and private schools. Separate representative samples of public and nonpublic schools were selected in each participating jurisdiction and students were randomly sampled within schools. The size of a jurisdiction's nonpublic school samples was proportional to the percentage of students in that jurisdiction attending such schools.

- The eighth-grade science assessment instruments used for the state assessment program and the national assessment consisted of 15 blocks of questions, of which 4 were hands-on tasks. Each block could contain a mixture of question types — constructed-response or multiple-choice — that was determined by the nature of the task. In addition, the constructed-response questions were of two types: *short constructed-response* questions required students to respond to a question with a few words or a few sentences, while *extended constructed-response* questions required students to respond to a question with a paragraph or more, sometimes including graphs or calculations. The hands-on tasks were similar to laboratory exercises. Each student was given 2 of the 11 cognitive blocks of questions, and one of the four hands-on blocks.

- A complex form of matrix sampling called a partially balanced incomplete block (PBIB) spiraling design was used. With PBIB spiraling, students in an assessment session received different booklets containing 3 of the 15 blocks. This provided for greater science content coverage without imposing an undue testing burden by administering an identical set of questions to each student.

- Sets of background questions given to the students, the students' science teachers, and the principals or other school administrators provided a variety of contextual information. The background questionnaires for the state assessment program were identical to those used in the national eighth-grade assessment.

- The total assessment time for each student was approximately two hours, including cleanup and collection of materials from hands-on tasks. Each assessed student was assigned a science booklet that contained 3 of the 15 blocks of science questions requiring 30 minutes each (including a hands-on task block in the last position), followed by a 5-minute general background questionnaire, a 5-minute science background questionnaire, and a 3-minute motivation questionnaire. Thirty-seven different booklets were assembled.

- The assessments were administered in the five-week period between January 29 and March 4, 1996. One-fourth of the schools in each jurisdiction were assessed each week throughout the first four weeks. Because of the severe weather throughout much of the country, the fifth week was used for regular testing as well as for makeup sessions.

- Data collection was, by law, the responsibility of each participating jurisdiction. Security and uniform assessment administration were high priorities. Extensive training of state assessment personnel was conducted to assure that the assessment would be administered under standard, uniform procedures. For jurisdictions that had participated in previous NAEP state assessments, 25 percent of both public and nonpublic school assessment sessions were monitored by Westat staff. For the jurisdictions new to NAEP, 50 percent of both public and nonpublic school sessions were monitored.

C.3 Assessment Instruments

The *student assessment booklets* contained six sections and included both cognitive and noncognitive questions. The assembly of cognitive questions into booklets and their subsequent assignment to assessed students were determined by a matrix sampling design using a variant of a balanced incomplete block design (BIB), with spiraled administration. Each assessed student received a booklet containing 3 of the 15 cognitive blocks according to a design that ensured that each block was administered to a representative sample of students within each jurisdiction. The third cognitive block was always one of the four hands-on blocks; this requirement meant that the BIB was partially balanced (PBIB).

In addition to two 30-minute sections of cognitive questions and the 30-minute performance task section, each booklet included two 5-minute sets of general and science background questions designed to gather contextual information about students, their experiences in science, and their attitudes toward the subject, and one 3-minute section of motivation questions designed to gather information about the student's level of motivation while taking the assessment.

In addition to the student assessment booklets, three other instruments provided data relating to the assessment: a science teacher questionnaire, a school characteristics and policies questionnaire, and an SD/LEP student questionnaire (for students categorized as students with disabilities or with limited English proficiency).

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

The *school characteristics and policies questionnaire* was given to the principal or other administrator in each participating school and took about 20 minutes to complete. The questions asked about the principal's background and experience, school policies, programs, and facilities, and the demographic composition and background of the students and teachers.

The *SD/LEP student questionnaire* was completed by the staff member most familiar with any student selected for the assessment who was classified in either of two ways: students with disabilities (SD) had an Individualized Education Plan (IEP) or equivalent special education plan (for reasons other than being gifted and talented); students with limited English proficiency were classified as LEP students. The questionnaire took approximately three minutes to complete and asked about the student and the special programs in which the student participated. It was completed for all selected SD or LEP students regardless of whether or not they participated in the assessment. Selected SD or LEP students participated in the assessment if they were determined by the school to be able to participate, considering the terms of their IEP and accommodations provided by the school or by NAEP.

C.4 The Sampling Design

The sampling design for NAEP is complex, in order to minimize burden on schools and students while maximizing the utility of the data. For further details see the *Technical Report for the NAEP 1996 State Assessment Program in Science*. The target populations for the state assessment program in science consisted of eighth-grade students enrolled in either public or nonpublic schools. The representative samples of public school eighth graders assessed in the state assessment program came from about 100 schools (per grade) in each jurisdiction. If a jurisdiction had fewer than 100 public schools with a particular grade, all or almost all schools were asked to participate. If a jurisdiction had smaller numbers of students in each school than expected, more than 100 schools were selected for participation. The nonpublic school samples differed in size across the jurisdictions, with the number of schools selected proportional to the nonpublic school enrollment within each jurisdiction. Typically, about 25 nonpublic schools were included for each jurisdiction. The school samples in each state were designed to produce aggregate estimates for the jurisdiction and for selected subpopulations (depending upon the size and distribution of the various subpopulations within the jurisdiction) and also to enable comparisons to be made, at the jurisdiction level, between administration of assessment tasks with monitoring and without monitoring. The public schools were stratified by urbanization, percentage of Black and Hispanic students enrolled, and median household income within the ZIP code area of the school. The nonpublic schools were stratified by type of control (Catholic, private/other religious, other nonpublic), metropolitan status, and enrollment size per grade.

The national and regional results are based on nationally representative samples of eighth-grade students. The samples were selected using a complex multistage sampling design involving the sampling of students from selected schools within selected geographic areas across the country. The sample design had the following stages:

(1) selection of geographic areas (a county, group of counties, or a metropolitan statistical area);

(2) selection of schools (public and nonpublic) within the selected areas; and

(3) selection of students within selected schools.

Each selected school that participated in the assessment, and each student assessed, represent a portion of the population of interest. To make valid inferences from student samples to the respective populations from which they were drawn, sampling weights are needed. Discussions of sampling weights and how they are used in analyses are presented in sections C.7 and C.8.

The state results provided in this report are based on state-level samples of eighth-grade students. The samples of both public and nonpublic school students were selected based on a two-stage sample design that entailed selecting students within schools. The first-stage samples of schools were selected with a probability proportional to the eighth-grade enrollment in the schools. Special procedures were used for jurisdictions with many small schools and for jurisdictions with a small number of schools. As with the national samples, the state samples were weighted to allow for valid inferences about the populations of interest.

The results presented for a particular jurisdiction are based on the representative sample of students who participated in the 1996 state assessment program. The results for the nation and regions of the country are based on the nationally and regionally representative samples of students who were assessed as part of the national NAEP program. Using the national and regional results from the 1996 national assessment was necessary because of the voluntary nature of the state assessment program. Because not every state participated in the program, the aggregated data across states did not necessarily provide representative national or regional results.

In most jurisdictions, up to 30 students were selected from each school, with the aim of providing an initial sample size of approximately 3,000 public school students per jurisdiction for the eighth grade. The student sample size of 30 for each school was chosen to ensure that at least 2,000 public school students participated from each jurisdiction, allowing for school nonresponse, exclusion of students, inaccuracies in the measures of enrollment, and student absenteeism from the assessment. In jurisdictions with fewer schools, larger numbers of students per school were often required to ensure initial samples of roughly 3,000 students. In certain jurisdictions, all eligible eighth graders were targeted for assessment. Jurisdictions were given the option to reduce the expected student sample size in order to reduce testing burden and the number of multiple-testing sessions for participating schools. At grade 8, four jurisdictions (Alaska, Delaware, Hawaii, and Rhode Island) elected to exercise this option. Using this option can involve compromises such as higher standard errors and accompanying loss of precision.

In order to provide for wider inclusion of students with disabilities and limited English proficiency, the 1996 state assessments both in mathematics and science involved dividing the sample of students at each grade level into two subsamples, referred to as S1 and S2. S1 provided continuity with the 1992 mathematics assessment and thus allowed for the reporting of performance over time by using the same exclusion criteria for students with disabilities and limited English proficiency as was used in that assessment. S2 provided for wider inclusion of students with disabilities and limited English proficiency by incorporating new exclusion rules.

The NAEP 1996 science assessment was developed using a new framework, and therefore does not include reporting of performance over time. However, in order to make the sample design identical for both subjects at the state level, both S1 and S2 were included. For further discussion, see the *NAEP 1996 Science Report Card*.

The 1996 national assessment in science used only the more inclusive S2 guidelines for student participation. The national assessments in mathematics and science both involved an additional subsample, S3, in which accommodations were provided for certain students with disabilities or limited English proficiency, again in order to make NAEP more inclusive.

For the national science assessment, scaling and analysis procedures (discussed in sections C.8 through C.10) were applied to all assessed students from S2. For the state science assessment, scaling and analysis procedures were applied to a combination of all assessed students from S2 and students who were **not** identified as SD or LEP from S1. This combination of segments of the S1 and S2 subsamples maximized the usefulness of available data while allowing for comparisons to the student population in the national sample. This combination, referred to as the "reporting sample," was the sample used to link the state science assessment to the national assessment (see Section C.10), as well as for scaling and reporting.

Additional analyses will be conducted on the national samples to study the effects of changing the exclusion rules and allowing the use of accommodations. Preliminary discussion can be found in the *NAEP 1996 Science Report Card* and the *NAEP 1996 Mathematics Report Card*; more detailed discussion will follow in future NAEP publications.

C.5 Field Administration

Administering the 1996 program required collaboration among staff in the participating jurisdictions and schools and the NAEP contractors, especially Westat, the field administration contractor.

Each jurisdiction volunteering to participate in the 1996 state assessment program appointed a state coordinator to serve as liaison between NAEP staff and the participating schools. In addition, Westat hired and trained a supervisor for each jurisdiction and six field managers who worked with groups of jurisdictions. The state supervisors worked with the state coordinators, overseeing assessment activities, training school district personnel to administer the assessment, and coordinating quality control monitoring efforts. Each field manager worked with the state coordinators from seven to eight jurisdictions and the state supervisors assigned to those jurisdictions. An assessment administrator prepared and conducted the assessment session in one or more schools. These individuals were usually school or district staff and were trained by Westat. Westat also hired and trained three to five quality control monitors in each jurisdiction. For jurisdictions that had previously participated in the state assessment program, 25 percent of the public and nonpublic school sessions were monitored. For jurisdictions new to the program, 50 percent of all sessions were monitored. The assessment sessions were conducted during a five-week period beginning in late January 1996.

C.6 Materials Processing, Professional Scoring, and Database Creation

Upon completion of each assessment session, school personnel shipped the assessment booklets and forms to NCS for professional scoring, entry into computer files, and checking. The files were then sent to ETS for creation of the database.

After NCS received all appropriate materials from a school, they were forwarded to the professional scoring area where the responses to the constructed-response question were evaluated by trained staff using guidelines prepared by ETS. Each constructed-response question had a unique scoring guide that defined the criteria to be used in evaluating students' responses. The extended constructed-response questions were evaluated with four- or five-level rubrics. Some of the short constructed-response questions were rated according to three-level rubrics that permit partial credit to be given; other short constructed-response questions were scored as either acceptable or unacceptable.

For the national science assessment and the state assessment program in science, over 4.1 million constructed responses were scored. This figure includes rescoring to monitor interrater reliability. The overall percentage of agreement between scorers for the reliability sample was 93 percent for the tasks in the cognitive blocks and 95 percent for the hands-on tasks.

Data transcription and editing procedures were used to generate the disk and tape files containing various assessment information, including the sampling weights required to make valid statistical inferences about the population from which the state assessment program sample was drawn. Prior to analysis, the data from these files underwent a quality control check at ETS. The files were then merged into a comprehensive, integrated database.

C.7 Weighting and Variance Estimation

A complex sample design was used to select the students who were assessed in each of the participating jurisdictions. The properties of a sample selected through 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. Therefore, the properties of the sample for the complex state assessment program design were taken into account during the analysis of the assessment data.

One way that the properties of the sample design were addressed was by using sampling weights to account for the fact that the probabilities of selection were not identical for all students. All population and subpopulation characteristics based on the state assessment program data used sampling weights in their estimation. These weights included adjustments for school and student nonresponse.

Not only must appropriate estimates of population characteristics be derived, but appropriate measures of the degree of uncertainty must be obtained for those statistics. One component of uncertainty results from sampling variability, which is a measure of the dependence of the results on the particular sample of students actually assessed. Because of the effects of cluster selection (schools are selected first, then students are selected within those schools), observations made on different students cannot be assumed to be independent of each other (and, in fact, are generally positively correlated). As a result, classical variance estimation formulas will produce incorrect results. Thus, a jackknife variance estimation procedure that accounts for the characteristics of the sample was used for all analyses.

Jackknife variance estimation provides a reasonable measure of uncertainty for any statistic based on values observed without error. Statistics such as the percentage of students correctly answering a given question meet this requirement, but other statistics based on estimates of student science performance, such as the average science scale score of a subpopulation, do not. Because each student typically responds to relatively few questions from a particular field of science (e.g., physical or life science), a nontrivial amount of imprecision exists in the measurement of the scale score of a given student. This imprecision adds another component of variability to statistics based on estimates of individual performance.

C.8 Preliminary Data Analysis

After the computer files of student responses were received and merged into an integrated database, all cognitive and noncognitive questions were subjected to an extensive item analysis. For each cognitive question, this analysis yielded the number of respondents, the percentage of responses in each category, the percentage who omitted the question, the percentage who did not reach the question, and the correlation between the question score and the block score. In addition, the item analysis program provided summary statistics for each block of cognitive questions, including a reliability (internal consistency) coefficient. These analyses were used to check the scoring of the questions, to verify that the difficulty level of the questions was appropriate, and to ensure that students had received adequate time to complete the assessment. The results were reviewed by knowledgeable project staff in search of aberrations that might signal unusual results or errors in the database.

The question and block-level analyses were conducted using rescaled versions of the final sampling weights provided by Westat (see Section C.7). The rescaling was implemented for each jurisdiction. The sum of the sampling weights for the public school students within each jurisdiction was constrained to be equal. The same transformation was applied to the weights of the nonpublic school students in that jurisdiction. The sum of the weights for each of the Department of Defense (DoDEA) samples (i.e., DDESS and DoDDS) was constrained to equal the same value as the public school students in other jurisdictions. Using rescaled weights does not alter the value of statistics calculated separately within each jurisdiction. However, for statistics obtained from samples that combine students from different jurisdictions, using rescaled weights results in a roughly equal contribution of each jurisdiction's data to the final value of the estimate. Equal contribution of each jurisdiction's data to the results of the item response theory (IRT) scaling was viewed as a desirable outcome. The original final sampling weights provided by Westat were used in reporting.

Additional analyses that compared the data from the monitored sessions with those from the unmonitored sessions were conducted to determine the comparability of the assessment data from the two types of administrations. Differential item functioning (DIF) analyses were carried out using the national assessment data. DIF analyses identified questions that were differentially difficult for various subgroups, so that these questions could be re-examined for their fairness and their appropriateness for inclusion in the scaling process.

C.9 Scaling the Assessment Questions

The primary analysis and reporting of the results from the state assessment program used item response theory (IRT) scale-score models. Scaling models quantify a respondent's tendency to provide correct answers to the domain of questions that contribute to a scale as a function of a parameter called performance, estimated by a scale score. The scale scores can be viewed as a summary measure of performance across the domain of questions that make up the scale. Three distinct IRT models were used for scaling: three-parameter logistic models for multiple-choice questions; two-parameter logistic models for short constructed-response questions that were scored correct or incorrect; and generalized partial credit models for short and extended constructed-response questions that were scored on a multipoint scale (i.e., greater than two levels).

Three distinct scales were created for the state assessment program in science to summarize eighth-grade students' abilities according to the three defined fields of science (earth, physical, and life). These scales were defined identically to, but separately from, those used for the scaling of the national NAEP eighth-grade science data. Although the questions composing each scale were identical to those used in the national assessment program, the item parameters for the state assessment program scales were estimated from combined public school data from the jurisdictions participating in the state assessment program.[1] Item parameter estimation was carried out on an item calibration subsample. The calibration subsample consisted of a sample drawn from approximately 25 percent sample of all available public school data. To ensure equal representation in the scaling process, each jurisdiction contributed the same number of students to the item calibration sample. Within each jurisdiction, 25 percent of the calibration sample was taken from monitored administrations while the remaining 75 percent came from unmonitored administrations.

Within each scale, the estimates of the empirical item characteristic functions were compared with the theoretical curves to determine how well the IRT model fit the observed data. For correct-incorrect questions, nonmodel-based estimates of the expected proportions of correct responses to each question for students with various levels of scale proficiency were compared with the fitted item response curve. For the short and extended partial-credit constructed-response questions, the comparisons were based on the expected proportions of students with various levels of scale proficiency who achieved each score level. In general, the scaling models fit the question-level results well.

[1] For the creation of scales, schools from the DoDEA jurisdictions are considered nonpublic, so the responses from these students were not included in the item calibration sample.

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

The 1996 science assessment was developed using a new framework. Because it was not appropriate to compare results from the 1996 assessment to those of previous NAEP science assessments, no attempt was made to link or align scores on the new assessment to those of previous assessments. Therefore, it was necessary to establish a new scale for reporting. Earlier NAEP assessments (such as the current mathematics assessment and the 1994 reading assessment) were developed with a cross-grade framework, in which the trait being measured is conceptualized as cumulative across the grades of the assessment. This concept was reflected in the scaling. The score scales developed for these assessments were cross-grade scales on a single 0-500 scale for all three grades in the assessment.

In 1993, the National Assessment Governing Board (NAGB) determined that future NAEP assessments should be developed using within-grade frameworks. This removes the constraint that the trait being measured is cumulative, and there is no need for overlap of questions across grades. Consistent with this view, NAGB also declared that scaling be performed within-grade. Any items which happened to be the same across grades in the assessment were scaled separately for each grade, thus allowing common items, potentially, to function differently in the separate grades. The 1994 NAEP history and geography assessments were developed and scaled within-grade. After scaling, the scales were aligned so that grade 8 had a higher mean than did grade 4, and grade 12 had a higher mean than grade 8. The results were reported on a final 0-500 scale that looked similar to those used in mathematics and reading, in spite of the differences in development and scaling. This definition of the reporting scale was a source of potential confusion and misinterpretation.

The 1996 science assessment was also developed and scaled using within-grade procedures. A new reporting metric was adopted to differ from the 0-to-500 reporting scales used in other NAEP subject areas in order to minimize confusion with other common test scales and to discourage cross-grade comparisons. For each grade in the national assessment, the mean for each field of science was set at 150 and the standard deviation was set at 35. First, the reporting metric was developed using data from the national assessment program; the results for the state assessment program were then linked to that scale using procedures described in Section C.10.

In addition to the plausible values for each scale, a composite of the three fields of science scales was created as a measure of overall science performance; as for the individual fields of science scales, the mean of the composite scale was set to 150 with a standard deviation of 35.[2] This composite was a weighted average of the plausible values for the three fields of science scales. The scales were weighted proportionally to the relative importance assigned to each field of science in the science framework (see Table B.1). The definition of the composite for the state assessment program was identical to that used for the national eighth-grade science assessments.

C.10 Linking the State Results to the National Results

A major purpose of the state assessment program was to allow each participating jurisdiction to compare its 1996 results with those for the nation as a whole and with those for the region of the country where it is located. For meaningful comparisons to be made between each jurisdiction and the relevant national sample, results from these two assessments had to be expressed in terms of a similar system of scale units.

The results from the state assessment program were linked to those from the national assessment through linking functions determined by comparing the results for the aggregate of all students assessed in the state assessment program with the results for eighth-grade students within the National Linking Sample of the national NAEP. The National Linking Sample of the national NAEP is a representative sample of the population of all grade-eligible public school students within the aggregate of 43 participating states and the District of Columbia. (Guam and the two DoDEA jurisdictions were not included in the National Linking Sample.) Specifically, the National Linking Sample for science consisted of all eighth-grade students in public schools in the states and the District of Columbia who were assessed in the national cross-sectional science assessment.

A linear equating within each field of science scale was used to link the results of the state assessment program to the national assessment. For each scale, the adequacy of the linear equating was evaluated by comparing the distribution of science scale scores based on the aggregation of all assessed students at each grade from the participating states and the District of Columbia with the equivalent distribution based on the students in the National Linking Sample. In the estimation of these distributions, the students were weighted to represent the target population of public school students in the specified grade in the aggregation of the states and the District of Columbia. If a linear equating were adequate, the distribution for the aggregate of states and the District of Columbia and that for the National Linking Sample would have, to a close approximation, the same shape in terms of the skewness, kurtosis, and higher moments of the distributions. The only differences in the distributions allowed by linear equating would be in the means and variances. Generally, this has been found to be the case.

Thus, each field of science scale was linked by matching the scale mean and standard deviation of the scale scores across all students in the state assessment (excluding Guam and the two DoDEA jurisdictions) to the corresponding mean and standard deviation across all students in the National Linking Sample.

[2] The national average of students in public and nonpublic schools combined is 150. The national average seen in the tables in this report is based on the average for public schools only (148).

ACKNOWLEDGMENTS

This report is the culmination of the effort of many individuals who contributed their considerable knowledge, experience, and creativity to the NAEP 1996 science assessment. The NAEP 1996 science assessment was a collaborative effort among staff from the National Center for Education Statistics (NCES), the National Assessment Governing Board (NAGB), Educational Testing Service (ETS), Westat, Inc., and National Computer Systems (NCS). In addition, the program benefited 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. Most importantly, NAEP is grateful to the over 109,000 students and the teachers and administrators from 4,400 schools in 47 jurisdictions who made the science assessment possible.

The NAEP 1996 science assessment was funded through NCES, in the Office of Educational Research and Improvement of the U.S. Department of Education. The Commissioner of Education Statistics, Pascal D. Forgione, and the NCES staff — Sue Ahmed, Peggy Carr, Arnold Goldstein, Steven Gorman, Larry Ogle, Gary W. Phillips, Sharif Shakrani, Maureen Treacy — worked closely and collegially with the authors to produce this report. The authors were also provided guidance by the members of the National Assessment Governing Board and NAGB staff. In particular, the authors are indebted to Arnold Goldstein of NCES for his daily efforts to coordinate the activities of the many people who contributed to this report.

The NAEP project at ETS is housed in the Center for the Assessment of Educational Progress under the direction of Paul Williams. The NAEP 1996 assessments were directed by Stephen Lazer and John Mazzeo. Tom Corley, Lee Jones, Tim Ligget, Beth Nichols, Christine O'Sullivan, Amy Pearlmutter, Will Pfeiffenberger, Mario Yepes-Baraya, and Ann Marie Zolandz worked with the Science Instrument Development committee to develop the assessment instrument. Sampling and data collection activities were conducted by Westat under the direction of Rene Slobasky, Nancy Caldwell, Keith Rust, and Dianne Walsh. Printing, distribution, scoring, and processing activities were conducted by NCS under the direction of Brad Thayer, Patrick Bourgeacq, Charles Brungardt, Jay Happel, Mathilde Kennel, Linda Reynolds, and Brent Studer.

The complex statistical and psychometric activities necessary to report results for the NAEP 1996 Science Assessment were directed by Nancy Allen, John Barone, James Carlson, and Juliet Shaffer. The analyses presented in this report were led by John Donoghue and Steven Isham with assistance from Spencer Swinton, Lois Worthington, Inge Novatkoski, Kate Pashley, David Freund, and Norma Norris.

Laura Jerry was responsible for the development and creation of the computer-generated reports, with assistance from Xiaohui Wang, Laura Jenkins, Phillip Leung, Inge Novatkoski, Bruce Kaplan, and Alfred Rogers. Two of the reports were skillfully produced by Karen Damiano. A large group of NAEP staff at ETS checked the data, text, and tables. Debbie Kline coordinated the technical appendices. The overall production efforts were completed by Carol Errickson, Barbette Tardugno, Loretta Casalaina, Kelly Gibson (cover design), Sharon Davis-Johnson, and Alice Kass. Editorial assistance was provided by Walt Brower and by John Calderone of Aspen Systems. The World Wide Web version of the state reports was produced by Philip Leung and Pat O'Reilly with assistance from Debbie Kline, Craig Pizzuti, and Christine Zelenak.

Many thanks are due to the numerous reviewers, both internal and external to NCES and ETS. The comments and critical feedback of the following reviewers are reflected in this report: Sue Ahmed, Peggy Carr, Mary Frase, Arnold Goldstein, Andrew Kolstad, Michael Ross, and Shi-Chang Wu of NCES; Rolf Blank of CCSSO; Audrey Champagne of the State University of New York in Albany; Michelle Leon of the Connecticut Department of Education; Will Pfeiffenberger of ETS; Senta Raizen of the National Center for Improving Science Education; and Mistilina Sato of Stanford University.

NAEP 1996 Science Instrument Development Committee

An Instrument Development Committee was convened to oversee the development of items and scoring rubrics. Committee members wrote assessment exercises and ensured that the instrument adhered to the assessment framework and specifications. In addition, the committee made certain that the instrument was developmentally appropriate for each grade and that it was relevant to curricular and instructional goals. The members are to be commended for their diligence and dedication to the lengthy process of producing the instrument:

Gail Baxter, University of Michigan
Ron Bonnstetter, University of Nebraska
Audrey Champagne, State University of New York at Albany
Richard Clark, Minnetonka, Minnesota
Sally Crissman, Shady Hill School, Cambridge, Massachusetts
Pat Dung, Los Angeles Educational Partnership
Michael Johnson, Science Skills Center
Michael Jojola, Isleta, New Mexico
Clifton Poodry, University of California at Santa Cruz
Senta Raizen, National Center for Improving Science Education
Douglas Reynolds, Rensselaer, New York
Realista Rodriguez, Stuart High School, Falls Church, Virginia
Mistilina Sato, Stanford University
Gerald Weaver, University City High School, Philadelphia, Pennsylvania
Mary Louise Bellamy, National Association of Biology Teachers

NCES 97–499