Overview of Procedures Used for the NAEP 2000 Science Assessment

This document provides an overview of the NAEP 2000 science assessment's primary components—framework, development, administration, scoring, and analysis. A more extensive review of the procedures and methods used in the science assessment will be included in the forthcoming *NAEP 2000 Technical Report*.

The NAEP 2000 Science Assessment 2
The Assessment Design 6
National and State Samples 9
Asian/Pacific Islander Samples 15
Standards for State Sample Participation and Reporting of Results 16
Students with Disabilities (SD) and/or Limited English Proficient (LEP) Students 19
Data Collection and Scoring 31
Data Analysis and IRT Scaling 31
Item Mapping Procedures 34
Weighting and Variance Estimation 35
Drawing Inferences from the Results 37
Analyzing Group Differences in Averages and Percentages 37
Conducting Multiple Tests 39
NAEP Reporting Groups 41
Grade 12 Participation Rates and Motivation 44
Cautions in Interpretations 45

The NAEP 2000 Science Assessment

The design of the NAEP 2000 science assessment followed the guidelines provided in the framework developed for the 1996 assessment. While maintaining some conceptual continuity with the NAEP 1990 science assessment, the 1996 framework took 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 be used for assessing students' performance. These included:

- performance tasks that allow students to manipulate physical objects and draw scientific understanding from the materials before them;
- constructed-response questions that provide insight 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; and
- multiple-choice questions that probe students' conceptual understanding and ability to connect ideas in a scientifically sound way.

Samples of each type of task and question are available in the "NAEP Questions" section of the NAEP web site http://nces.ed.gov/nationsreportcard. The framework for the 1996 and 2000 science assessments is represented as a matrix with two dimensions: 1) fields of science (Earth, physical, and life) and 2) elements of knowing and doing science (conceptual understanding, scientific investigation, and practical reasoning). The fields-of-science dimension is used to create three subscales at each grade. Subscales are not created based on the elements of knowing and doing science. In addition there are two overarching domains that describe science: 1) nature of science and 2) themes. [These overarching domains provide additional guidance to the development of assessment questions and tasks, ensuring that the assessment also integrates the three fields of science rather than only represents three separate content areas.] Figures A.1a, A.1b, and A.1c describe, respectively, the fields of science, the elements of knowing and doing science, and the overarching domains that guided the development of the 1996 and 2000 science assessments.

Figure A.1a	Descriptions of the Three Fields of Science
Earth Science	The content of Earth science falls under the general headings of the solid Earth (lithosphere), water (hydrosphere), air (atmosphere), and the Earth in space. Topics related to the solid Earth include the composition of the earth; 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. Topics related to water include 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. Topics related to air include the composition and structure of the atmosphere (including energy transfer); the nature of weather; common weather hazards; and air quality and climate. Topics related to the Earth in space include the 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; the apparent daily motions of the Sun, the Moon, the planets and the stars; the rotation of the Earth about its axis, and the Earth's revolution around the Sun; and the tilt of the Earth's axis that produces seasonal variations in the climate.
Physical Science	The physical science component covers basic knowledge and understanding concerning the structure of the universe as well as the physical principles that operate within it. The major topics 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 includes 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 includes 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 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 covers information transfer; energy transfer for the construction of proteins; and communication among cells. Organisms covers reproduction, growth and development; life cycles; and functions and interactions of systems within organisms. Ecology focuses on the interdependence of life—populations, communities, and ecosystems.

SOURCE: National Assessment Governing Board. (2000). *Science Framework for the 1996 and 2000 National Assessment of Educational Progress.* Washington, DC: Author.

Figure A.1b	Descriptions 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 assesses students' ability to use and apply science understanding in new, real-world applications.

SOURCE: National Assessment Governing Board. (2000). *Science Framework for the 1996 and 2000 National Assessment of Educational Progress*. Washington, DC: Author.

Figure A.1c	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 and 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 require students to recognize patterns of similarity and differences and to recognize how these patterns change over time. In addition, students should be able to remember common types of patterns and transfer their understanding of a familiar pattern of change to a new and unfamiliar one.

SOURCE: National Assessment Governing Board. (2000). *Science Framework for the 1996 and 2000 National Assessment of Educational Progress*. Washington, DC: Author.

Table A.1a summarizes the distribution of assessment time across the three fields of science—Earth, physical, and life. These fields provide the basis for the content area scales. Care was taken to ensure congruence between the percentages used in the assessment (actual) and those indicated in the assessment specifications (target). The classification of items by field of science was overseen and approved by a committee of expert science educators.

Table A.1a

Distribution of assessment time by field of science: 1996 and 2000

	Earth			Physical			Life		
	Target	Actual 1996	Actual 2000	Target	Actual 1996	Actual 2000	Target	Actual 1996	Actual 2000
Grade 4	33%	33%	33%	33%	34%	33%	33%	33%	33%
Grade 8	30%	30%	31%	30%	30%	34%	40%	40%	35%
Grade 12	33%	33%	33%	33%	33%	31%	33%	34%	37%

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 and 2000 Science Assessments.

Table A.1b shows the distribution of assessment time across the second dimension: knowing and doing science. This dimension includes conceptual understanding, scientific investigation, and practical reasoning. As with the above classification of items, an expert committee of science educators oversaw the categorization of items by this dimension. In both this table and the table above, variation is evident across the two assessment years in percentages of questions within categories. Such variation is the result of releasing several blocks of questions from the 1996 assessment and replacing them with newly developed questions in 2000. In addition, one of the four hands-on blocks administered at each grade in 1996 was released,

Table A.1b

Distribution of assessment time by knowing and doing science: 1996 and 2000

	Conceptual understanding			Scientific investigation			Practical reasoning		
	Target	Actual 1996	Actual 2000	Target	Actual 1996	Actual 2000	Target	Actual 1996	Actual 2000
Grade 4	45%	45%	56%	45%	38%	27%	10%	17%	17%
Grade 8	45%	45%	59%	30%	29%	18%	25%	26%	24%
Grade 12	45%	44%	56%	30%	28%	24%	25%	28%	20%

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 and 2000 Science Assessments.

and no replacement block was developed for 2000. This resulted in a smaller proportion of scientific investigation questions at each grade in 2000 than in 1996. The reporting of changes in student performance is not affected by these variations because trend reporting is based upon the underlying scale, which uses the common blocks (i.e., those used in both assessment years), but maintains its stability even if some blocks are dropped or replaced.

The Assessment Design

One-half of the students who participated in the science assessment received a booklet containing six sections; the other half, five sections. All the booklets contained either two or three sections that were blocks of cognitive questions assessing knowledge and skills outlined in the framework. In addition, each booklet contained two sections that were sets of background questions. Each booklet had two cognitive sections containing only paper-and-pencil questions. The booklets with three blocks of cognitive questions also contained a hands-on task with related paper-and-pencil questions. The booklets with two blocks of cognitive questions did not contain a hands-on task. Thus, one-half of the students who participated in the assessment performed a hands-on task.

At each grade level there were 14 different sections or blocks of cognitive questions usually consisting of both multiple-choice and constructed-response questions.[1] Short constructed-response questions required a few words or a sentence or two for an answer (e.g., briefly stating why a potted plant can survive in a sealed container much longer than a mouse), while extended constructed-response questions generally required a paragraph or more (e.g., outlining an experiment to find the density of a metal ring). Some extended constructed-response questions also required diagrams, graphs, or calculations. It was expected that students could adequately answer the short constructed-response questions in about two to three minutes and the extended constructed-response questions in about five minutes.

Other features were built into the blocks of questions. Three of the blocks at each grade level were hands-on tasks where students were given a set of equipment and asked to conduct an investigation and answer questions relating to the investigation. One-half of the students conducted a hands-on task that was always presented as the third cognitive section. A second feature was the inclusion of theme blocks at each grade level—one assessing systems, one assessing models, and one assessing patterns of change. A theme block contains a set of questions that all focus on a particular theme, and requires students to engage more thoroughly in the topics related to that theme. For example, students were asked to make drawings and graphs based on data given about the solar system and then answer a number of questions. Theme blocks were placed randomly in the student booklets, but did not appear in every booklet. No student received more than one theme block.

[1] These 14 blocks were distributed across the student booklets in a Balanced Incomplete Block (BIB) design that is described later in this section.

The data in table A.2 display the number of questions by type and by grade level for the 1996 and 2000 assessments. Some of these questions were used at more than one grade level; thus, the sum of the questions that appear at each grade level is greater than the total number of unique questions. The total number of questions at each grade level in 2000 is up from 1996. This increase was possible because more multiple-choice questions that take less time were used in 2000. This increase in multiple-choice questions across the entire assessment was due to the fact that the blocks developed for 2000 to replace those released from the 1996 assessment contained a greater proportion of multiple-choice questions. In addition, as mentioned earlier, one of four hands-on blocks at each grade in 1996 was released and not replaced for 2000. These hands-on blocks contain only constructed-response questions. As a consequence, the total number of constructed-response questions in 2000 was less than that in 1996. It should be noted that these variations across years do not affect the ability of NAEP to report trends in students' performance across years. Trend reporting is based on those blocks that were common across the two years.

The assessment design allowed for maximum coverage of science content at grades 4, 8, and 12, while minimizing the time burden for any one student. This was accomplished through the use of matrix sampling of questions, in which representative samples of students took various portions of the entire pool of assessment questions. Individual students were required to take only a small portion of the assessment, but the aggregate results across the entire assessment allowed for broad reporting of science abilities for the targeted population.

Table A.2

Distribution of questions administered by question type: 1996 and 2000

		Grade 4 only		Grade 4 and 8 overlap		Grade 8 only		Grade 8 and 12 overlap		Grade 12 only		Total by grade	
		1996	2000	1996	2000	1996	2000	1996	2000	1996	2000	1996	2000
Grade 4	MC[1]	42	62	9	9							51	71
	SCR[2]	57	49	16	16							73	65
	ECR[3]	12	3	4	4							16	7
Grade 8	MC[1]			9	9	44	65	21	21			74	95
	SCR[2]			16	16	58	49	26	26			100	91
	ECR[3]			4	4	13	3	3	3			20	10
Grade 12	MC[1]							21	21	49	70	70	91
	SCR[2]							26	26	62	57	88	83
	ECR[3]							3	3	27	18	30	21

[1] Multiple-choice questions.
[2] Short constructed-response questions.
[3] Extended constructed-response questions.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 and 2000 Science Assessments.

In addition to matrix sampling, the Balanced Incomplete Block (BIB) design also balances the order of presentation of the blocks of questions, except for the hands-on blocks, which always appear in position three of a booklet. Furthermore, the design was set up to ensure that no student answered more than one theme-based block (though some students did not receive any). This design allows for some balancing of the impact of context and fatigue effects to be measured and reported, but makes allowance for the difficulties and disruption of administering hands-on blocks. It also takes into account the limited breadth of content coverage included in the theme blocks.[2]

Each booklet in the assessment also included two sections of student background questions. The first section, consisting of general background questions, asked students about their race/ethnicity, mother's and father's level of education, reading materials in the home, homework, school attendance, and, at grade 12, academic expectations.[3] The second section asked students questions about their science classroom activities (e.g., hands-on exercises, courses taken, and use of specialized resources such as computers).

In addition to the student assessment booklets, four other instruments provided data relating to the assessment: a teacher questionnaire, a school characteristics and policy questionnaire, a questionnaire designed to gather information about students with disabilities (SD) and/or limited English proficient (LEP) students, and a department chair/lead teacher questionnaire at grade 12.

The teacher questionnaire was administered to the science teachers of the fourth- and 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 section, on the teacher's background related to science; and the third section, on classroom information about science instruction.

The school characteristics and policy questionnaire was given to the principal or other administrator in each participating school and took about 20 minutes to complete. The questions asked about school policies, programs, facilities, and the demographic composition and background of the student body.

The SD and/or LEP student questionnaire was completed by a school staff member knowledgeable about those students who were selected to participate in the assessment and who were identified as: 1) having an Individualized Education Program (IEP) or equivalent program (for reasons other than being gifted and talented) or 2) being limited English proficient (LEP). A questionnaire was completed for each SD and/or LEP student sampled regardless of whether the student participated in the assessment. Each questionnaire took approximately three minutes to complete and asked about the student and the special programs in which he or she participated.

---

2 For further details on the booklet design, see the forthcoming *NAEP 2000 Technical Report*.

3 Some questions, such as those referring to parental education, were not asked of fourth-graders.

The department chair/lead teacher questionnaire was given to the high school science department chair or lead teacher in each participating high school. Previous to the 2000 assessment, NAEP had not attempted to collect information from teachers of twelfth-grade science, partly due to the difficulty in identifying the science teachers of assessed twelfth-graders. The questionnaire took about 20 minutes to complete. The questions asked about the certification of the teachers, science courses offered, use of computers in the classroom, teacher preparation time, and frequency of textbook replacement. As this was NAEP's first attempt to collect information from department chairs or lead teachers, an official report of those data is not currently planned. The data are available on NAEP's web site at http://nces.ed.gov/nationsreportcard through the data tool function.

National and State Samples

National Sample

The national results presented in this report are based on a nationally representative probability sample of fourth-, eighth-, and twelfth-grade students.[4] The sample was chosen using a multistage design that involved sampling 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 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 represents a portion of the population of interest. Sampling weights are needed to make valid inferences between the student samples and the respective populations from which they were drawn. Sampling weights account for disproportionate representation due to the oversampling of students who attend schools with high concentrations of black and/or Hispanic students and students who attend nonpublic schools. Among other uses, sampling weights also account for lower sampling rates for very small schools and are used to adjust for school and student nonresponse.[5]

A special feature of the 1996 and 2000 national assessments of science was the collection of data from samples of students where assessment accommodations for special-needs students were not permitted and from samples of students where accommodations for special-needs students were permitted. NAEP inclusion rules were applied, and accommodations were offered only when a student had an Individualized Education Program (IEP) because of a disability and/or was identified as being a limited English proficient student (LEP); all other students were asked to participate in the assessment under standard conditions.

[4] The student samples from American Samoa, Department of Defense Domestic Dependent Elementary and Secondary Schools, Department of Defense Dependents Schools (Overseas), Guam, and the Virgin Islands are not included in the national sample.

[5] Additional details regarding the design and structure of the national and state samples will be included in the forthcoming *NAEP 2000 Technical Report*. In addition, the reader may consult the *NAEP 1998 Technical Report* for a discussion of sampling procedures that are mostly common to all NAEP assessments.

Table A.3 shows the number of students included in the national samples for the NAEP science assessments at each grade level. For the 1996 and 2000 assessments, the table includes the number of students in the sample where accommodations were not permitted and the number of students in the sample where accommodations were permitted. The table shows that the same non-SD and/or LEP students were included in both samples in 2000; only the SD and/or LEP students differed between the two samples. The 1996 design differed somewhat, in that the two samples did not include all the same non-SD and/or LEP students. As indicated in the table, additional non-SD and/or LEP students were included in the accommodations-permitted sample.

Table A.3

National student sample size, grades 4,8, and 12 (public and nonpublic schools combined): 1996 and 2000

	1996		2000	
	Accommodations-not-permitted sample	Accommodations-permitted sample	Accommodations-not-permitted sample	Accommodations-permitted sample
Grade 4				
Non-SD and/or LEP students assessed	6,704	3,780*	15,068	
SD and/or LEP students assessed without accommodations	601	319	652	750
SD and/or LEP students assessed with accommodations	NA	174	NA	279
Total students assessed	7,305	10,977	15,720	16,097
Grade 8				
Non-SD and/or LEP students assessed	7,122	3,670*	14,905	
SD and/or LEP students assessed without accommodations	652	364	882	798
SD and/or LEP students assessed with accommodations	NA	163	NA	252
Total students assessed	7,774	11,319	15,787	15,955
Grade 12				
Non-SD and/or LEP students assessed	7,128	3,621*	14,555	
SD and/or LEP students assessed without accommodations	409	285	554	607
SD and/or LEP students assessed with accommodations	NA	75	NA	163
Total students assessed	7,537	11,109	15,109	15,325

* The 1996 accommodations-permitted sample included additional non-SD and/or LEP students.
SD = Students with Disabilities.
LEP = Limited-English-Proficient students.
NA = Not applicable. No accommodations were permitted in this sample.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 and 2000 Science Assessments.

Table A.4 provides a summary of the national school and student participation rates for the science assessment samples where accommodations were not permitted and where accommodations were permitted. Participation rates are presented for public and nonpublic schools, individually and combined. The first rate is the weighted percentage of schools participating in the assessment before substitution of demographically similar schools.[6] This rate is based only on the sample of schools that was initially selected for the assessment. The numerator of this rate is the sum of the estimated number of students represented by each initially selected school that participated in the assessment. The denominator is the sum of the estimated number of students represented by each of the initially selected schools that had eligible students enrolled.

The second school participation rate is the weighted participation rate after substitution. The numerator of this rate is the sum of the estimated number of students represented by each of the participating schools, whether originally selected or selected as a substitute for a school that chose not to participate. The denominator is the sum of the estimated number of students represented by each of the initially selected schools that had eligible students enrolled (this is the same as that for the weighted participation rate for the sample of schools before substitution). The denominator for these two rates is an estimate of the number of students eligible for the assessment, from all schools in the nation with eligible students enrolled. Because of the common denominators, the weighted participation rate after substitution is at least as great as the weighted participation rate before substitution.

Also presented in table A.4 are weighted student participation rates. The numerator of this rate is the sum across all students assessed (in either an initial session or a makeup session) of the number of students that each represents. The denominator of this rate is the sum of the number of students represented in the sample, across all eligible sampled students in participating schools. The overall participation rate is calculated as the product of the weighted percentage of school participation before (or after) substitution, and the weighted percentage of student participation after makeup sessions.

For the grade 12 national sample, where school and student response rates did not meet NCES standards, an extensive analysis was conducted that examined, among other factors, the potential for nonresponse bias at both the school and student level. No evidence of any significant potential for either school or student nonresponse bias was found. Results of these analyses, as well as nonresponse bias analyses for the grade 4 and grade 8 national samples, will be included in the forthcoming *NAEP 2000 Technical Report*.

6 The initial base sampling weights were used in weighting the percentages of participating schools and students. An attempt was made to preselect (before field processes began) a maximum of two substitute schools for each sampled public school (one in-district and one out-of-district) and each sampled Catholic school, and one for each sampled nonpublic school (other than Catholic). To minimize bias, a substitute school resembled the original selection as much as possible on affiliation, estimated number of grade-eligible students, and minority composition.

Table A.4

National school and student participation rates for public schools, nonpublic schools, and public and nonpublic schools combined, grades 4, 8, and 12: 2000

	Weighted school participation			Samples where accommodations were not permitted				Samples where accommodations were permitted			
				Student participation		Overall participation rate		Student participation		Overall participation rate	
	Percentage before substitution	Percentage after substitution	Total number of schools	Weighted percentage student participation	Total number of students assessed	Before substitution	After substitution	Weighted percentage student participation	Total number of students assessed	Before substitution	After substitution
Grade 4											
Public	85	88	414	96	9,144	81	85	95	9,484	81	84
Nonpublic	85	88	363	96	6,576	82	85	96	6,613	82	85
Combined	85	88	777	96	15,720	81	85	96	16,097	81	85
Grade 8											
Public	83	85	385	92	9,443	76	78	91	9,617	76	78
Nonpublic	81	84	366	96	6,344	77	81	96	6,338	77	81
Combined	82	85	751	92	15,787	76	78	92	15,955	76	78
Grade 12											
Public	78	82	243	75	8,562	58	61	75	8,727	58	61
Nonpublic	73	80	307	89	6,547	65	71	89	6,598	65	71
Combined	77	82	550	76	15,109	59	62	76	15,325	59	62

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Science Assessment.

State Samples

The results of the 2000 state assessment program in science provided in this report are based on state-level samples of fourth- and eighth-grade public school students independent of the national samples. The samples were selected using a two-stage sample design that first selected schools within participating jurisdictions and then students within schools. As with the national samples, the jurisdiction samples were weighted to allow for valid inferences about the populations of interest. Tables A.5a and A.5b contain the unweighted number of participating schools and students, as well as weighted school and student participation rates for state samples where accommodations were not permitted and where accommodations were permitted. Participation rates for the states were calculated the same way rates were computed for the nation.

Table A.5a

State school and student participation rates, grade 4 (public schools only): 2000

	Weighted school participation			Samples where accommodations were not permitted				Samples where accommodations were permitted			
						Overall participation rate				Overall participation rate	
	Percentage before substitution	Percentage after substitution	Total number of schools	Weighted percentage student participation	Total number of students assessed	Before substitution	After substitution	Weighted percentage student participation	Total number of students assessed	Before substitution	After substitution
Nation	85	88	414	96	9,144	81	85	95	9,484	81	84
Alabama	87	94	109	96	2,526	83	91	96	2,552	83	91
Arizona	87	87	95	93	2,080	81	81	93	2,068	81	81
Arkansas	85	85	93	95	2,175	81	81	95	2,214	81	81
California †	76	76	81	94	1,682	72	72	94	1,714	71	71
Connecticut	100	100	107	96	2,493	96	96	95	2,550	95	95
Georgia	99	99	107	95	2,640	94	94	94	2,687	94	94
Hawaii	98	98	106	95	2,425	94	94	95	2,439	93	93
Idaho †	75	75	78	95	1,717	71	71	95	1,750	71	71
Illinois †	73	73	77	95	1,596	70	70	95	1,671	70	70
Indiana †	70	70	78	95	1,812	66	66	95	1,870	67	67
Iowa †	71	71	89	96	1,887	68	68	95	1,951	67	67
Kentucky	92	94	105	95	2,248	87	89	95	2,311	87	89
Louisiana	100	100	108	95	2,452	95	95	95	2,538	95	95
Maine †	85	85	107	95	2,094	81	81	94	2,184	81	81
Maryland	100	100	110	95	2,648	95	95	94	2,737	94	94
Massachusetts	99	99	106	95	2,274	94	94	95	2,351	94	94
Michigan †	71	83	83	94	1,875	67	78	94	1,922	67	78
Minnesota †	83	83	78	95	1,853	79	79	95	1,894	78	78
Mississippi	98	98	106	95	2,776	93	93	95	2,799	93	93
Missouri	96	96	103	95	2,367	91	91	94	2,473	91	91
Montana †	76	77	67	95	1,176	72	74	95	1,201	72	74
Nebraska	96	96	73	94	1,289	90	90	95	1,315	91	91
Nevada	100	100	109	94	2,526	94	94	94	2,619	94	94
New Mexico	93	93	98	94	1,895	87	87	94	1,999	87	87
New York †	72	72	79	93	1,764	67	67	93	1,848	67	67
North Carolina	100	100	108	95	2,374	95	95	95	2,482	95	95
North Dakota	89	89	129	96	2,338	86	86	97	2,400	86	86
Ohio †	82	82	85	93	1,887	76	76	93	1,922	76	76
Oklahoma	99	99	120	95	2,377	93	93	94	2,475	93	93
Oregon †	73	74	79	94	1,625	69	70	95	1,686	69	70
Rhode Island	100	100	110	95	2,395	95	95	95	2,500	95	95
South Carolina	97	97	103	96	2,448	93	93	96	2,495	93	93
Tennessee	97	97	105	95	2,496	92	92	95	2,522	92	92
Texas	97	99	100	96	2,125	93	95	96	2,229	93	95
Utah	100	100	110	95	2,652	95	95	95	2,694	95	95
Vermont †	75	75	66	95	1,237	71	71	95	1,312	71	71
Virginia	100	100	108	96	2,502	96	96	96	2,615	96	96
West Virginia	100	100	126	95	2,522	95	95	95	2,639	95	95
Wisconsin †	65	67	69	95	1,393	62	64	96	1,474	62	64
Wyoming	100	100	93	95	1,745	95	95	95	1,821	95	95
Other Jurisdictions											
American Samoa	100	100	17	93	453	93	93	93	475	93	93
DDESS	100	100	39	95	1,295	95	95	96	1,300	96	96
DoDDS	100	100	84	95	2,790	95	95	96	2,825	96	96
Guam	96	96	23	95	996	90	90	95	1,064	91	91
Virgin Islands	100	100	22	96	690	96	96	96	698	96	96

† Indicates that the jurisdiction did not meet one or more of the guidelines for school participation.
DDESS: Department of Defense Domestic Dependent Elementary and Secondary Schools.
DoDDS: Department of Defense Dependents Schools (Overseas).
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Science Assessment.

Table A.5b

State school and student participation rates, grade 8 (public schools only): 2000

	Weighted school participation			Samples where accommodations were not permitted				Samples where accommodations were permitted			
						Overall participation rate				Overall participation rate	
	Percentage before substitution	Percentage after substitution	Total number of schools	Weighted percentage student participation	Total number of students assessed	Before substitution	After substitution	Weighted percentage student participation	Total number of students assessed	Before substitution	After substitution
Nation	83	85	385	92	9,443	76	78	91	9,617	76	78
Alabama	82	92	102	94	2,400	77	86	93	2,382	77	86
Arizona †	76	76	80	91	1,783	69	69	91	1,822	69	69
Arkansas	87	87	92	92	2,115	80	80	92	2,140	80	80
California †	72	72	76	93	1,650	67	67	93	1,723	67	67
Connecticut	100	100	104	91	2,506	91	91	91	2,551	91	91
Georgia	99	99	102	92	2,550	91	91	92	2,578	91	91
Hawaii	91	91	50	90	2,268	82	82	91	2,285	83	83
Idaho †	78	78	63	93	1,973	73	73	93	2,003	73	73
Illinois †	75	75	80	94	1,753	70	70	93	1,808	70	70
Indiana †	73	73	76	93	1,878	68	68	93	1,904	68	68
Kentucky	94	95	96	94	2,303	89	90	94	2,383	89	90
Louisiana	100	100	104	91	2,373	91	91	90	2,393	90	90
Maine †	83	85	86	94	2,156	78	79	94	2,254	78	79
Maryland	97	97	103	89	2,336	86	86	89	2,434	87	87
Massachusetts	99	99	99	93	2,277	92	92	92	2,389	91	91
Michigan †	72	81	86	91	2,024	65	74	91	2,047	65	73
Minnesota †	73	73	59	92	1,435	68	68	92	1,458	68	68
Mississippi	98	98	101	93	2,495	91	91	93	2,514	91	91
Missouri	92	94	104	93	2,320	86	88	93	2,415	86	87
Montana †	73	74	62	92	1,692	68	69	93	1,745	68	69
Nebraska	98	98	87	91	1,898	90	90	90	1,863	89	89
Nevada	100	100	64	92	2,694	92	92	91	2,733	91	91
New Mexico	91	91	85	89	1,903	81	81	89	1,981	81	82
New York †	71	71	74	89	1,616	63	63	89	1,697	63	63
North Carolina	98	98	103	92	2,342	90	90	91	2,452	90	90
North Dakota	91	91	93	93	2,194	84	84	92	2,221	84	84
Ohio	91	91	88	92	2,122	83	83	91	2,169	83	83
Oklahoma	100	100	114	92	2,452	92	92	93	2,515	93	93
Oregon †	74	74	78	90	1,751	67	67	90	1,780	67	67
Rhode Island	100	100	52	91	2,360	91	91	90	2,440	90	90
South Carolina	91	92	95	93	2,298	85	86	93	2,336	85	86
Tennessee	90	92	97	91	2,227	82	83	91	2,257	82	84
Texas	91	94	100	93	2,302	85	88	92	2,331	84	87
Utah	100	100	95	92	2,446	92	92	92	2,475	92	92
Vermont †	80	80	74	93	1,966	74	74	92	2,021	74	74
Virginia	100	100	105	91	2,435	91	91	90	2,508	90	90
West Virginia	100	100	102	93	2,436	93	93	92	2,567	92	92
Wisconsin †	66	75	80	91	1,811	61	68	91	1,883	60	68
Wyoming	100	100	64	93	2,560	93	93	93	2,575	93	93
Other Jurisdictions											
American Samoa	96	96	16	97	445	93	93	97	471	93	93
DDESS	100	100	14	94	650	94	94	95	701	95	95
DoDDS	100	100	53	94	1,962	94	94	94	1,999	94	94
Guam	100	100	7	90	945	90	90	90	921	90	90
Virgin Islands †	100	100	7	90	606	90	90	89	619	89	89

† Indicates that the jurisdiction did not meet one or more of the guidelines for school participation.
DDESS: Department of Defense Domestic Dependent Elementary and Secondary Schools.
DoDDS: Department of Defense Dependents Schools (Overseas).
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Science Assessment.

Asian/Pacific Islander Samples

National scale score and achievement-level results for fourth-grade Asian/Pacific Islander students in 2000 are not reported. Table A.6 contains average science scale score estimates, and their standard errors, for the nation and for the Asian/Pacific Islander subgroup for the 1996 and 2000 assessment years. In 2000, the average scale score for Asian/Pacific Islanders at grade 4 was 8 points higher than in 1996. However, this cross-year difference was not statistically significant.

It is important to note that all NAEP results are estimates and are subject to some degree of sampling variability. If different samples of schools or students had been obtained, results for some subgroups would be higher than reported here and some would be lower. In most subgroups, particularly large subgroups or subgroups for which special sampling procedures are employed, estimates of performance are likely to remain similar from one sample to another. However, the national population of Asian/Pacific Islander students is small (about 3 percent of the national population), heterogeneous with respect to academic achievement, and highly clustered in certain locations and schools. These factors are associated with large sampling variability in survey results and are reflected in the large standard errors associated with performance estimates for this subgroup. Furthermore, the sampling plan for the national assessment does not include explicit stratification procedures designed to mitigate these factors. The occurrence of the large, but statistically nonsignificant, change in the 2000 grade 4 Asian/Pacific Islander results was a likely consequence of these factors: 1) the heterogeneous nature of the Asian/Pacific Islander population; 2) the current NAEP sampling design; and 3) the sample sizes that were assessed.

Table A.6

Average science scale scores for the Asian/Pacific Islander subgroup, grade 4 (public and nonpublic schools combined): 1996 and 2000

	1996		2000	
	Percentage	Average score	Percentage	Average score
All students at grade 4	100	150 (0.8)	100	150 (0.7)
Asian/ Pacific Islander at grade 4	3 (0.2)	151 (3.6)	3 (0.2)	159 (4.1)

NOTE: The standard errors of the estimated percentages and average scale scores appear in parentheses.
Results are based on administration procedures that did not permit accommodations.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 and 2000 Science Assessments.

Standards for State Sample Participation and Reporting of Results

In carrying out the 2000 state assessment program, the National Center for Education Statistics (NCES) established participation rate standards that jurisdictions were required to meet in order for their results to be reported. NCES also established additional standards that required the annotation of published results for jurisdictions whose sample participation rates were low enough to raise concerns about their representativeness. The NCES guidelines used to report results in the state assessments, and the guidelines for notation when there is some risk of nonresponse bias in the reported results, are presented in this section.

Guideline 1

The publication of NAEP results

The conditions that will result in the publication of a jurisdiction's results are presented below.

Guideline 1—Publication of Public School Results

A jurisdiction will have its public school results published in the 2000 NAEP Science results (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 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.

Discussion: If a jurisdiction's public school participation rate for the initial sample of schools is below 70 percent, there is a substantial possibility that bias will be introduced into the assessment results. This possibility remains even after making statistical adjustments to compensate for school nonparticipation. There remains the likelihood that, in aggregate, the substitute schools are sufficiently dissimilar from the originals that they are replacing and represent too great a proportion of the population to discount such a difference. Similarly, the assumptions underlying the use of statistical adjustments to compensate for nonparticipation are likely to be significantly violated if the initial response rate falls below the 70 percent level. Guideline 1 takes this into consideration. This guideline is congruent with current NAGB policy, which requires that data for jurisdictions that do not have a 70 percent before-substitution participation rate be reported "in a different format," and with the Education Information Advisory Committee (EIAC) resolution, which calls for data from such jurisdictions not to be published.

The following guidelines concerning school and student participation rates in the NAEP state assessment program were established to address four significant ways in which nonresponse bias could be introduced into the jurisdiction sample estimates. Presented on the following pages are the conditions that will result in a jurisdiction's receiving a notation in the 2000 reports. Note that in order for a jurisdiction's results to be published with no notations, that jurisdiction must satisfy all guidelines.

Guideline 2

Reporting school and student participation rates with possible bias due to school nonresponse

Guideline 2–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.

Discussion: For jurisdictions that did not use substitute schools, the participation rates are based on participating schools from the original sample. In these situations, the NCES standards specify weighted school participation rates of at least 85 percent to guard against potential bias due to school nonresponse. Thus the first part of these guidelines, referring to the weighted school participation rate for the initial sample of schools, is in direct accordance with NCES standards.

To help ensure adequate sample representation for each jurisdiction participating in the NAEP 2000 state assessments, NAEP provided substitutes for nonparticipating public schools. For jurisdictions that used substitute schools, the assessment results will be based on the student data from all schools participating from both the original sample and the list of substitutes (unless both an initial school and its substitute eventually participated, in which case only the data from the initial school will be used).

The NCES standards do not explicitly address the use of substitute schools to replace initially selected schools that decide not to participate in the assessment. However, considerable technical consideration was given to this issue. Even though the characteristics of the substitute schools were matched as closely as possible to the characteristics of the initially selected schools, substitution does not entirely eliminate bias due to the nonparticipation of initially selected schools. Thus, for the weighted school participation rates including substitute schools, the guidelines were set at 90 percent.

If a jurisdiction meets either standard (i.e., 85 percent or higher prior to substitution or 90 percent or higher after substitution), there will be no notation for the relevant overall school participation rate.

Guideline 3

Important segments of the jurisdiction's student population that must be adequately represented to avoid possible nonresponse bias

Guideline 3–Notation for Strata-Specific Public School Participation Rates

A jurisdiction that is not already receiving a notation under Guideline 2 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.

Discussion: The NCES standards specify that attention should be given to the representativeness of the sample coverage. Thus, if some important segment of the jurisdiction's population is not adequately represented, it is of concern, regardless of the overall participation rate.

If nonparticipating schools are concentrated within a particular class of schools, the potential for substantial bias remains, even if the overall level of school participation appears to be satisfactory. Nonresponse adjustment cells for public schools have been formed within each jurisdiction, and the schools within each cell are similar with respect to minority enrollment, degree of urbanization, and/or median household income, as appropriate for each jurisdiction.

If the weighted response rate, after substitution, for a single adjustment cell falls below 80 percent, and more than five percent (weighted) of the sampled schools are nonparticipants from such a cell, the potential for nonresponse bias is too great. This guideline is based on the NCES standard for stratum-specific school response rates.

Guideline 4

Possible student nonresponse bias

Guideline 4—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.

Discussion: This guideline follows the NCES standard of 85 percent for overall student participation rates. The weighted student participation rate is based on all eligible students from initially selected or substitute schools who participated in the assessment in either an initial session or a makeup session. If the rate falls below 85 percent, the potential for bias due to students' nonresponse is too great.

Guideline 5

Possible nonresponse bias from inadequately represented strata

Guideline 5—Notation for Strata-Specific Student Participation Rates in Public Schools

A jurisdiction that is not already receiving a notation under Guideline 4 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 5 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 or grade 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),[7] as well as school level of urbanization, minority enrollment, and median household income of the area in which the school is located.

Discussion: This guideline addresses the fact that if nonparticipating students are concentrated within a particular class of students, the potential for substantial bias remains, even if the overall student participation level appears to be satisfactory. Student nonresponse adjustment cells have been formed using the school-level nonresponse adjustment cells, together with the student's age and the nature of the assessment session (unmonitored or monitored).

If the weighted response rate for a single adjustment cell falls below 80 percent, and more than five percent (weighted) of the invited students who do not participate in the assessment are from such a cell, the potential for nonresponse bias is too great. This guideline is based on the NCES standard for stratum-specific student response rates.

[7] In the state assessments, 25 percent of the administration sessions were observed by quality control monitors.

At both fourth- and eighth-grade, one state, Wisconsin, failed to meet the initial public school participation rate of 70 percent and, at eighth grade, the Virgin Islands failed to meet this standard. Results for these jurisdictions are not included with the findings reported for the state NAEP 2000 science assessment.

At grade 4, there were 12 jurisdictions (California, Idaho, Illinois, Indiana, Iowa, Michigan, Minnesota, Montana, New York, Ohio, Oregon, and Vermont) that failed to meet the required weighted participation rate of 85 percent for the initial sample of schools and their weighted school sample rate after substitution was below 90 percent. At grade 8, 12 jurisdictions (Arizona, California, Idaho, Illinois, Indiana, Maine, Michigan, Minnesota, Montana, New York, Oregon, and Vermont) failed to meet this guideline as well. At grade 4, Maine failed to meet Guideline 3 indicating that 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 5 percent of the jurisdiction's total weighted sample of public schools. Results for each of these states at the appropriate grade level are shown with a notation indicating possible bias related to nonresponse.

Students with Disabilities (SD) and/or Limited English Proficient (LEP) Students

It is NAEP's intent to assess all selected students from the target population. Therefore, every effort is made to ensure that all selected students who are capable of participating in the assessment are assessed. Some students sampled for participation in NAEP can be excluded from the sample according to carefully defined criteria. These criteria were revised in 1996 to communicate more clearly a presumption of inclusion except under special circumstances. According to these criteria, students with Individualized Education Programs (IEPs) were to be included in the NAEP assessment except in the following cases:

1) The school's IEP team determined that the student could not participate, OR,
2) The student's cognitive functioning was so severely impaired that she or he could not participate, OR,
3) The student's IEP required that the student had to be tested with an accommodation or adaptation and that the student could not demonstrate his or her knowledge without that accommodation.[8]

All LEP students receiving academic instruction in English for three years or more were to be included in the assessment. Those LEP students receiving instruction in English for fewer than three years were to be included unless school staff judged them to be incapable of participating in the assessment in English.

Participation of SD and/or LEP Students in the Two NAEP Samples

Testing all sampled students is the best way for NAEP to ensure that the statistics generated by the assessment are as representative as possible of the performance of the entire national population and the populations of participating jurisdictions. However, all groups of students include

[8] As described in the following section, a second sample in the 1996 national and the 2000 national and state assessments was assessed that included students who required and were provided with accommodations.

certain proportions that cannot be tested in large-scale assessments (such as students who have profound mental disabilities), or who can only be tested through the use of "accommodations" such as extra time, one-on-one administration, or use of magnifying equipment. Some students with disabilities and some LEP students cannot show on a test what they know and can do unless they are provided accommodations. When such accommodations are not allowed, students requiring such adjustments are often excluded from large-scale assessments such as NAEP. This phenomenon has become more common in the last decade and gained momentum with the passage of the Individuals with Disabilities Education Act (IDEA), which led schools and states to identify increasing proportions of students as needing accommodations on assessments to best show what they know and can do.[9] Furthermore, Section 504 of the Rehabilitation Act of 1973 requires that, when students with disabilities are tested, schools must provide them with appropriate accommodations so that the test results accurately reflect what the students know and are able to do.[10] In addition, as the proportion of English language learners in the population has increased, some states have started offering accommodations, such as translated versions of assessments or the use of bilingual dictionaries as part of assessments.

Before 1996, NAEP did not allow any testing under nonstandard conditions (i.e., accommodations were not permitted). At that time, NAEP samples were able to include almost all sampled students in "standard" assessment sessions. However, as the influence of IDEA grew more widespread, the failure to provide accommodations led to increasing levels of exclusion in the assessment. Such increases posed two threats to the program: 1) they threatened the stability of trend lines (because excluding more students in one year than the next might lead to apparent rather than real gains), and 2) they made NAEP samples less than optimally representative of target populations.

NAEP reacted to this challenge by adopting a multipart strategy. It became clear that to ensure that NAEP samples were as inclusive as possible, the program had to move toward allowing the same assessment accommodations that were afforded students in state and district testing programs. However, allowing accommodations represents a change in testing conditions that may affect trend. Therefore, beginning with the 1996 national assessments and the 1998 state assessments, NAEP has assessed a series of parallel samples of students. In one set of samples, testing accommodations were not permitted; this has allowed NAEP to maintain the measurement of achievement trends on an assessment that was, throughout its existence, administered under common conditions. In addition to the samples where accommodations were not permitted, parallel samples in which accommodations were permitted were also assessed. By having two overlapping samples and two sets of related data points, NAEP could

9 Office of Special Education Programs. (1997). *Nineteenth annual report to Congress on the implementation of the individuals with disabilities education act*. Washington, DC: U. S. Department of Education.

10 Section 504 of the Rehabilitation Act of 1973 is a civil rights law designed to prohibit discrimination on the basis of disability in programs and activities, including education, that receive federal financial assistance.

meet two core program goals.[11] First, data trends could be maintained. Second, parallel trend lines could be set in ways that ensure that in future years the program will be able to use the most inclusive practices possible and mirror the procedures used by most state and district assessments. Beginning in 2002, NAEP will use only the more inclusive samples in which assessment accommodations are permitted.

In science, national and state data from 1996 and 2000 are reported for the sample in which accommodations were not permitted. National data for the second sample, in which accommodations were permitted, are reported at all grades for 1996 and 2000. State data on this more inclusive sample are reported for 2000 only.

In order to make it possible to evaluate both the impact of increasing exclusion rates in some jurisdictions and differences between jurisdictions, complete data on exclusion in both assessment years are included in this appendix. Since the exclusion rates may affect trend measurement within a jurisdiction, readers should consider the magnitude of exclusion rate changes when interpreting score changes in jurisdictions. In addition, different rates of exclusion may influence the meaning of state comparisons. Thus, exclusion data should be reviewed in this context as well.

Participation rates across the assessment years for students with disabilities (SD) and/or limited English proficient (LEP) students for the national sample where accommodations were not permitted are presented in table A.7. The data in this table include the percentages of students *identified* as SD and/or LEP, the percentage of students *excluded*, and the percentage of *assessed* SD and/or LEP students. Tables A.8a and A.8b show similar information by jurisdiction for grades 4 and 8 (only 2000 data are presented for grade 4 since there was no fourth-grade state science assessment in 1996). Participation rates for the national sample where accommodations were permitted are presented in table A.9, and state results where accommodations were permitted are shown in tables A.10a and A.10b. The data in these tables include the percentages of students *identified* as SD and/or LEP, the percentage of students *excluded*, the percentage of *assessed* SD and/or LEP students, the percentage *assessed without accommodations*, and the percentage *assessed with accommodations*. Expanded state-level data are available on the NAEP web site (http://nces.ed.gov/nationsreportcard) that break out these percentages for SD students and LEP students separately.

In the 2000 accommodations-not-permitted national sample, 7 percent of students at grades 4 and 8 and 4 percent of students at grade 12 were excluded from the assessment. The comparable percentages in the 2000 accommodations-permitted national sample were 4 percent at grade 4, 3 percent at grade 8, and 2 percent at grade 12. This comparison would suggest that allowing accommodations did help to decrease the percentage of students excluded from the assessment. A similar pattern is evident in the various jurisdictions that participated in the 2000 state assessment. Across the jurisdictions, the percentage of students excluded in the accommodations-not-permitted sample ranged from 4 to 15 percent at grade 4, and from 4 to 14 percent at grade 8. In the

[11] The two samples are described as "overlapping" because in 2000 the same group of non-SD and/or LEP students were included in both samples. In 1996, all of the non-SD and/or LEP students in the sample that did not permit accommodations were included in the analysis of results for the sample that did permit accommodations, with the inclusion of additional non-SD and/or LEP students selected for the accommodations-permitted sample only.

Table A.7

Percentage of students identified as SD and/or LEP where accommodations were not permitted (public and nonpublic schools combined): 1996 and 2000

	1996		2000	
	Number of students sampled	Weighted percentage of students	Number of students sampled	Weighted percentage of students
Grade 4				
SD and/or LEP students				
Identified	1,357	16	1,248	14
Excluded	756	8	596	7
Assessed	601	7	652	7
SD students only				
Identified	773	11	782	10
Excluded	425	6	453	6
Assessed	348	5	329	4
LEP students only				
Identified	654	5	557	5
Excluded	393	3	225	2
Assessed	261	2	332	3
Grade 8				
SD and/or LEP students				
Identified	1,078	12	1,728	14
Excluded	426	4	846	7
Assessed	652	7	882	8
SD students only				
Identified	763	10	1,306	12
Excluded	314	4	711	6
Assessed	449	6	595	6
LEP students only				
Identified	373	3	530	4
Excluded	156	1	217	1
Assessed	217	2	313	2
Grade 12				
SD and/or LEP students				
Identified	834	8	1,066	9
Excluded	425	3	512	4
Assessed	409	4	554	5
SD students only				
Identified	530	5	843	8
Excluded	321	3	449	4
Assessed	209	3	394	4
LEP students only				
Identified	340	3	282	2
Excluded	136	1	111	1
Assessed	204	2	171	1

SD = Students with Disabilities.
LEP = Limited-English-Proficient students.
NOTE: Within each grade level, the combined SD and/or LEP portion of the table is not a sum of the separate SD and LEP portions because some students were identified as both SD and LEP. Such students would be counted separately in the bottom portions, but counted only once in the top portion.
Within each portion of the table, percentages may not sum properly due to rounding.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 and 2000 Science Assessments.

Table A.8a

State percentage of students identified as SD and/or LEP where accommodations were not permitted, grade 4 (public schools only): 2000

	Identified	Excluded	Assessed
Nation	16	8	8
Alabama	12	6	7
Arizona	24	11	12
Arkansas	13	6	6
California †	33	11	22
Connecticut	15	10	5
Georgia	11	8	4
Hawaii	19	9	10
Idaho †	16	6	10
Illinois †	16	9	7
Indiana †	12	7	5
Iowa †	14	10	5
Kentucky	12	8	4
Louisiana	16	8	8
Maine †	18	11	7
Maryland	13	9	3
Massachusetts	20	11	9
Michigan †	11	9	2
Minnesota †	16	7	9
Mississippi	6	4	2
Missouri	15	10	5
Montana †	13	5	7
Nebraska	16	6	11
Nevada	20	11	9
New Mexico	30	13	17
New York †	17	13	4
North Carolina	17	14	2
North Dakota	14	6	7
Ohio †	12	10	2
Oklahoma	20	10	10
Oregon †	18	8	10
Rhode Island	23	12	11
South Carolina	17	8	9
Tennessee	11	4	7
Texas	26	15	11
Utah	14	7	7
Vermont †	15	11	5
Virginia	15	10	5
West Virginia	13	10	3
Wisconsin †	20	13	7
Wyoming	14	6	8
Other Jurisdictions			
American Samoa	17	15	2
DDESS	11	7	4
DoDDS	11	5	6
Guam	26	10	17
Virgin Islands	7	5	2

† Indicates that the jurisdiction did not meet one or more of the guidelines for school participation.
SD = Students with Disabilities.
LEP = Limited-English-Proficient students.
DDESS: Department of Defense Domestic Dependent Elementary and Secondary Schools.
DoDDS: Department of Defense Dependents Schools (Overseas).
NOTE: Percentages may not sum properly due to rounding.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Science Assessment.

Table A.8b

State percentage of students identified as SD and/or LEP where accommodations were not permitted, grade 8 (public schools only): 1996 and 2000

	1996			2000		
	Identified	Excluded	Assessed	Identified	Excluded	Assessed
Nation	13	5	8	16	7	8
Alabama	13	8	5	13	4	8
Arizona †	15	6	9	18	9	9
Arkansas	12	7	5	15	8	7
California †	21	9	12	26	9	16
Connecticut	15	9	6	14	9	5
Georgia	10	5	5	11	7	4
Hawaii	13	5	7	20	8	12
Idaho †	—	—	—	14	5	9
Illinois †	—	—	—	15	11	5
Indiana †	11	6	5	11	6	5
Kentucky	9	4	5	13	9	3
Louisiana	11	6	5	13	5	8
Maine	13	7	6	16	9	7
Maryland	12	6	7	14	10	4
Massachusetts	17	7	10	20	13	7
Michigan †	10	5	4	11	8	3
Minnesota †	11	4	7	15	5	10
Mississippi	11	6	5	8	5	3
Missouri	13	6	7	13	8	5
Montana †	9	3	6	13	6	6
Nebraska	11	4	7	15	4	11
Nevada	13	9	5	14	9	6
New Mexico	20	9	11	26	13	13
New York †	15	9	6	18	14	4
North Carolina	10	5	5	15	12	2
North Dakota	9	2	7	13	4	9
Ohio	—	—	—	11	8	3
Oklahoma	—	—	—	14	8	7
Oregon †	12	5	7	17	6	11
Rhode Island	17	7	10	19	10	9
South Carolina	10	7	4	14	8	6
Tennessee	12	4	8	14	6	8
Texas	17	8	9	19	9	11
Utah	9	4	5	12	6	6
Vermont †	14	6	8	19	11	9
Virginia	12	7	6	15	10	5
West Virginia	12	7	5	16	11	4
Wisconsin †	11	7	4	15	9	6
Wyoming	10	4	6	12	4	8
Other Jurisdictions						
American Samoa	—	—	—	15	12	3
DDESS	10	6	3	15	13	3
DoDDS	8	3	5	8	4	4
Guam	9	7	2	17	5	12

† Indicates that the jurisdiction did not meet one or more of the guidelines for school participation in 2000.
— Indicates that the jurisdiction did not participate.
SD = Students with Disabilities.
LEP = Limited-English-Proficient students.
DDESS: Department of Defense Domestic Dependent Elementary and Secondary Schools.
DoDDS: Department of Defense Dependents Schools (Overseas).
NOTE: Percentages may not sum properly due to rounding.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 and 2000 Science Assessments.

Table A.9

Percentage of students identified as SD and/or LEP where accommodations were permitted (public and nonpublic schools combined): 1996 and 2000

		1996		2000	
		Number of students sampled	Weighted percentage of students	Number of students sampled	Weighted percentage of students
Grade 4					
SD and/or LEP students	Identified	820	16	1,427	16
	Excluded	327	6	398	4
	Assessed	493	10	1,029	12
	Assessed without accommodations	319	6	750	8
	Assessed with accommodations	174	4	279	4
SD students only	Identified	496	12	860	11
	Excluded	145	4	257	3
	Assessed	351	8	603	8
	Assessed without accommodations	192	4	367	5
	Assessed with accommodations	159	4	236	3
LEP students only	Identified	370	5	649	6
	Excluded	196	2	193	1
	Assessed	174	3	456	4
	Assessed without accommodations	138	2	402	4
	Assessed with accommodations	36	1	54	1
Grade 8					
SD and/or LEP students	Identified	850	11	1,468	13
	Excluded	323	4	418	3
	Assessed	527	7	1,050	9
	Assessed without accommodations	364	5	798	7
	Assessed with accommodations	163	3	252	2
SD students only	Identified	604	8	1,094	10
	Excluded	244	3	354	3
	Assessed	360	6	740	7
	Assessed without accommodations	223	3	511	5
	Assessed with accommodations	137	2	229	2
LEP students only	Identified	283	3	427	3
	Excluded	101	1	88	1
	Assessed	182	2	339	2
	Assessed without accommodations	147	1	303	2
	Assessed with accommodations	35	▲	36	▲
Grade 12					
SD and/or LEP students	Identified	596	7	1,065	9
	Excluded	236	3	295	2
	Assessed	360	5	770	7
	Assessed without accommodations	285	4	607	5
	Assessed with accommodations	75	1	163	2
SD students only	Identified	395	5	726	7
	Excluded	203	2	257	2
	Assessed	192	3	469	5
	Assessed without accommodations	119	2	329	3
	Assessed with accommodations	73	1	140	2
LEP students only	Identified	228	2	385	2
	Excluded	52	▲	75	▲
	Assessed	176	2	310	2
	Assessed without accommodations	171	2	286	2
	Assessed with accommodations	5	▲	24	▲

▲ Percentage is between 0.0 and 0.5.
SD = Students with Disabilities.
LEP = Limited-English-Proficient students.
NOTE: Within each grade level, the combined SD and/or LEP portion of the table is not a sum of the separate SD and LEP portions because some students were identified as both SD and LEP. Such students would be counted separately in the bottom portions but counted only once in the top portion.
Within each portion of the table, percentages may not sum properly due to rounding.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 and 2000 Science Assessments.

Table A.10a

State percentage of students identified as SD and/or LEP where accommodations were permitted, grade 4 (public schools only): 2000

	Identified SD and/or LEP		Assessed SD and/or LEP			All students assessed under standard conditions
	Total	Excluded	Total	Assessed under standard conditions	Assessed with accommodations	
Nation	18	5	13	9	4	91
Alabama	12	4	8	6	3	93
Arizona	24	6	18	11	7	87
Arkansas	13	4	9	5	4	92
California †	33	5	28	19	9	86
Connecticut	15	5	10	5	5	90
Georgia	11	3	8	5	3	93
Hawaii	19	8	11	9	2	89
Idaho †	16	2	13	8	6	92
Illinois †	16	4	12	6	6	90
Indiana †	12	3	8	4	4	92
Iowa †	14	3	12	4	7	90
Kentucky	12	4	9	4	5	91
Louisiana	16	2	13	2	11	86
Maine †	18	4	14	5	8	87
Maryland	13	3	9	4	6	91
Massachusetts	20	4	16	6	10	87
Michigan †	11	3	8	4	3	93
Minnesota †	16	3	13	6	7	90
Mississippi	6	2	4	2	2	95
Missouri	15	1	13	5	8	90
Montana †	13	3	10	5	5	93
Nebraska	16	5	12	8	3	92
Nevada	20	7	14	9	5	89
New Mexico	30	6	23	17	7	87
New York †	17	4	12	3	10	86
North Carolina	17	6	11	4	8	87
North Dakota	14	1	12	8	4	94
Ohio †	12	4	8	3	5	91
Oklahoma	20	4	16	11	5	91
Oregon †	18	4	14	7	7	90
Rhode Island	23	4	19	9	10	86
South Carolina	17	5	11	7	5	90
Tennessee	11	2	9	7	2	96
Texas	26	8	18	14	5	87
Utah	14	4	10	6	4	92
Vermont †	15	3	13	4	9	88
Virginia	15	5	10	5	5	90
West Virginia	13	3	10	3	7	90
Wisconsin †	20	5	16	6	10	85
Wyoming	14	1	13	6	7	92
Other Jurisdictions						
American Samoa	17	7	10	10	0	93
DDESS	11	5	7	3	4	92
DoDDS	11	2	8	4	4	94
Guam	26	6	20	15	6	88
Virgin Islands	7	4	2	2	0	96

† Indicates that the jurisdiction did not meet one or more of the guidelines for school participation.
SD = Students with Disabilities.
LEP = Limited-English-Proficient students.
DDESS: Department of Defense Domestic Dependent Elementary and Secondary Schools.
DoDDS: Department of Defense Dependents Schools (Overseas).
NOTE: Percentages may not sum properly due to rounding.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Science Assessment.

Table A.10b

State percentage of students identified as SD and/or LEP where accommodations were permitted, grade 8 (public schools only): 2000

	Identified SD and/or LEP		Assessed SD and/or LEP			All students assessed under standard conditions
	Total	Excluded	Total	Assessed under standard conditions	Assessed with accommodations	
Nation	14	4	10	7	3	93
Alabama	13	5	7	7	1	94
Arizona †	18	4	14	9	5	92
Arkansas	15	5	10	7	3	92
California †	26	4	22	18	4	92
Connecticut	14	6	8	6	3	91
Georgia	11	4	7	4	3	93
Hawaii	20	5	15	13	2	93
Idaho †	14	3	11	8	3	94
Illinois †	15	5	10	6	4	90
Indiana †	11	3	9	5	4	93
Kentucky	13	3	10	5	5	92
Louisiana	13	3	10	6	5	93
Maine †	16	2	13	7	6	91
Maryland	14	3	11	7	4	93
Massachusetts	20	4	16	8	8	88
Michigan †	11	5	6	4	2	93
Minnesota †	15	4	11	10	2	95
Mississippi	8	4	4	2	1	94
Missouri	13	2	11	6	5	93
Montana †	13	1	11	8	3	95
Nebraska	15	4	11	10	1	95
Nevada	14	4	10	7	3	93
New Mexico	26	6	20	18	3	91
New York †	18	7	11	3	8	85
North Carolina	15	5	10	4	5	90
North Dakota	13	1	12	8	4	94
Ohio	11	4	8	4	4	92
Oklahoma	14	4	11	9	2	95
Oregon †	17	4	13	9	4	92
Rhode Island	19	4	15	12	3	93
South Carolina	14	6	8	7	1	93
Tennessee	14	4	10	10	1	95
Texas	19	6	13	11	2	92
Utah	12	3	9	6	3	94
Vermont †	19	3	17	10	6	91
Virginia	15	5	10	5	5	89
West Virginia	16	3	13	5	8	89
Wisconsin †	15	4	11	7	5	92
Wyoming	12	1	11	8	3	96
Other Jurisdictions						
American Samoa	15	3	12	10	2	96
DDESS	15	2	13	8	5	93
DoDDS	8	1	7	5	2	97
Guam	17	9	8	4	4	87

† Indicates that the jurisdiction did not meet one or more of the guidelines for school participation.
SD = Students with Disabilities.
LEP = Limited-English-Proficient students.
DDESS: Department of Defense Domestic Dependent Elementary and Secondary Schools.
DoDDS: Department of Defense Dependents Schools (Overseas).
NOTE: Percentages may not sum properly due to rounding.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Science Assessment.

accommodations-permitted sample, the percentages of students excluded ranged from 1 to 8 percent at grade 4, and from 1 to 9 percent at grade 8. As with the national exclusion rates, most states and jurisdictions excluded a smaller percentage of students when accommodations were permitted.

Investigating the Effects of Exclusion Rates on Assessment Results

As indicated by the data in the previous section, exclusion rates have tended to increase across assessment years in the samples that did not permit accommodations, particularly within certain states. In considering the effects of exclusion rates on assessment results, at least two major issues become evident. First, if exclusion rates vary substantially across assessment years, then the ability to report trends (i.e., compare results between years) may be threatened by the fact that the results from different years are based on different proportions of the population. Second, the variation in exclusion rates among states and jurisdictions may threaten the comparison of state-by-state results within a given year, again because the results for different states or jurisdictions are based on different proportions of the populations.

As a consequence, NCES investigated the possibility of establishing criteria for including cautionary notations based on excessive or increased exclusion rates (similar to those based on overall participation rates) in the reporting of national and state-by-state results. This investigation, however, did not reveal a consistent relationship between levels of exclusion, or degrees of change in inclusion rates, and overall results. There were several reasons for this.

First of all, real demographic differences influence exclusion rates in states and, thus, some differences may be unavoidable. Second, program research conducted by NCES and Educational Testing Service (ETS) was unable to identify a particular level of exclusion increase that seemed to affect scores. Third, since excluded students were not tested, NAEP has no direct information about how those students *would* have done had they been tested. Given these realities and uncertainties, the best approach seemed to be to supply all data about student exclusion and allow readers to consider it as they interpret the achievement data. However, it is important to remember that the main solutions to this issue lie not in flagging results, but in ensuring that all sampled students participate in assessments. The new, more inclusive samples that will become NAEP's main samples in 2002 are intended to accomplish this goal.

The move to more inclusive samples, however, will not be a perfect solution. For example, even within the context of the samples in which accommodations are permitted, there is still some student exclusion (albeit at a far lower level, as the data in tables A.8 and A.9a/b show). In addition, the assessment accommodations may not have an entirely neutral impact on scores. In other words, it is possible that changes in the percentages of students receiving assessment accommodations may influence scores. It is also possible that differences in state and local accommodations policies will affect state comparisons.

Because of these remaining issues, NCES has funded several major research studies. These activities have been organized around two distinct questions. First, as was

mentioned above, some students are excluded from even the more inclusive NAEP. Therefore, NCES has funded research into ways excluded students might be included in the estimation of scores for overall populations. In other words, research is being conducted to investigate weighting procedures that might be used to ensure the final NAEP estimates include data for all students in a sampled population. There are two general approaches that have been investigated. The first is an idea championed by Albert Beaton of Boston College. Beaton recommends making a simple assumption about excluded students: he would assume that, had these students been tested, they would have performed below some predefined level (for example, the median score or the lowest score in the *basic* achievement range). This statistic (whether median or some other level) would be adjusted to take account of excluded students.

The second approach to obtaining full population estimates has been recommended by Donald McLaughlin of the American Institutes for Research (AIR). His approach involves using background data about excluded students to estimate how they, as a group, would have performed had they been assessed. This approach is based on different and stronger assumptions than Beaton's. It would have the advantage of allowing NAEP to continue to report all the types of statistics currently in use (including average scores).

The results from an initial examination of the 1996 and 2000 NAEP science data using McLaughlin's approach indicated that the reported average score gains from 1996 to 2000 in many jurisdictions would be somewhat smaller if full-population estimates were used. This is apparently due to the increase in exclusion rates between years within these states. It should be noted that using such full-population estimates may not only alter the estimates of score gains, but may also alter the rank ordering of states within a given year.

NCES has not yet judged either statistical adjustment approach ready for operational use. Therefore, these "full population reporting" approaches may or may not be used in future years. Results of the studies produced by McLaughlin may be obtained from NCES, as can copies of an Educational Testing Service (ETS) study that implemented Beaton's methodology.

In addition to full population reporting research, NCES has commissioned studies of the impact of assessment accommodations on overall scores. Specifically, ETS has conducted differential item functioning (DIF) studies of items assessed with accommodation in the 1996 assessment.[12] In these studies, ETS researchers found little evidence that accommodations changed the functioning of test questions.

Types of Accommodations Permitted

Table A.11 displays the number and the percentages of SD and/or LEP students assessed with the variety of available accommodations. It should be noted that students assessed with accommodations typically received some combination of accommodations. The numbers and percentages presented in the table reflect only

[12] For information on DIF studies of items assessed with accommodations in the 1996 mathematics and science assessments, see Mazzeo, J. M., Carlson, J. E., Voelkl, K. E., & Lutkus, A. D. (1999). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

the primary accommodation provided. For example, students assessed in small groups (as compared to standard NAEP sessions of about 30 students) usually received extended time. In one-on-one administrations, students often received assistance in recording answers and were afforded extra time. Extended time was considered the primary accommodation only when it was the sole accommodation provided.

Table A.11

Percentage of students in national sample identified as SD and/or LEP by type of accommodation where accommodations were permitted (public and nonpublic schools combined): 1996 and 2000

	Grade 4				Grade 8				Grade 12			
	1996		2000		1996		2000		1996		2000	
	Number of students sampled	Weighted percentage of students	Number of students sampled	Weighted percentage of students	Number of students sampled	Weighted percentage of students	Number of students sampled	Weighted percentage of students	Number of students sampled	Weighted percentage of students	Number of students sampled	Weighted percentage of students
SD and/or LEP students												
Science glossary	0	0.00	37	0.45	13	0.10	13	0.11	0	0.00	2	0.01
Bilingual dictionary	NA	NA	0	0.00	NA	NA	2	0.01	NA	NA	10	0.11
Glossary/dictionary	16	0.23	NA	NA	14	0.15	NA	NA	2	0.02	NA	NA
Large-print book	0	0.00	1	0.01	0	0.00	2	0.04	0	0.00	0	0.00
Extended time	28	0.69	50	0.56	29	0.47	54	0.35	30	0.32	64	0.51
Read aloud	17	0.56	17	0.29	10	0.19	22	0.24	3	0.07	4	0.06
Small group	99	2.37	137	1.69	89	1.66	140	1.54	26	0.35	68	0.93
One-on-one	11	0.22	35	0.69	7	0.08	11	0.11	12	0.18	8	0.10
Scribe/computer	NA	NA	0	0.00	NA	NA	5	0.08	NA	NA	4	0.03
Other	3	0.07	2	0.01	1	0.01	3	0.04	2	0.02	3	0.04
SD students only												
Science glossary	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00
Bilingual dictionary	NA	NA	0	0.00	NA	NA	1	0.00	NA	NA	0	0.00
Glossary/dictionary	1	0.02	NA	NA	1	0.01	NA	NA	0	0.00	NA	NA
Large-print book	0	0.00	1	0.01	0	0.00	2	0.04	0	0.00	0	0.00
Extended time	28	0.69	49	0.55	29	0.47	52	0.34	30	0.32	54	0.47
Read aloud	17	0.56	17	0.29	10	0.19	18	0.19	3	0.07	4	0.06
Small group	99	2.37	131	1.64	89	1.66	137	1.52	26	0.35	68	0.93
One-on-one	11	0.22	35	0.69	7	0.08	11	0.11	12	0.18	8	0.10
Scribe/computer	NA	NA	0	0.00	NA	NA	5	0.08	NA	NA	4	0.03
Other	3	0.07	2	0.01	1	0.01	3	0.04	2	0.02	2	0.03
LEP students only												
Science glossary	0	0.00	36	0.44	13	0.10	13	0.11	0	0.00	2	0.01
Bilingual dictionary	NA	NA	0	0.00	NA	NA	2	0.01	NA	NA	10	0.11
Glossary/dictionary	16	0.23	NA	NA	14	0.15	NA	NA	2	0.02	NA	NA
Large-print book	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Extended time	3	0.04	1	0.01	2	0.03	3	0.01	2	0.02	10	0.04
Read aloud	2	0.08	1	0.00	4	0.06	6	0.06	0	0.00	0	0.00
Small group	14	0.18	16	0.18	2	0.02	11	0.09	1	0.02	0	0.00
One-on-one	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	1	0.00
Scribe/computer	NA	NA	0	0.00	NA	NA	0	0.00	NA	NA	0	0.00
Other	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	1	0.01

SD = Students with Disabilities. LEP = Limited-English-Proficient students.
NA = Not Applicable. Accommodation was not offered.
NOTE: The combined SD and/or LEP portion of the table is not a sum of the separate SD and LEP portions because some students were identified as both SD and LEP. Such students would be counted separately in the bottom portions, but counted only once in the top portion.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 and 2000 Science Assessments.

Data Collection and Scoring

The 2000 science assessment was conducted from January through March 2000, with some makeup sessions in early April. As with all NAEP assessments, data collection for the 2000 assessment was conducted by a trained field staff. For the national assessment, this was accomplished by staff from Westat, Inc.

For the state assessment, testing sessions were conducted and administered by employees of state and local educational agencies and institutions. These employees were carefully trained in assessment procedures by Westat. In addition, Westat employed quality control monitors who observed 25 percent of the sessions in state assessments.

Materials from the 2000 assessment were shipped to National Computer Systems, where trained staff evaluated the responses to the constructed-response questions using scoring rubrics or guides prepared by Educational Testing Service. Each constructed-response question had a unique scoring rubric that defined the criteria used to evaluate students' responses. The extended constructed-response questions were evaluated with four- and five-level rubrics, and many of the short constructed-response questions were rated according to three-level rubrics that permitted partial credit. Other short constructed-response questions were scored as either acceptable or unacceptable.

For the 2000 science assessment, approximately 4.5 million constructed responses were scored. This number includes rescoring to monitor inter-rater reliability. The within-year average percentage of exact agreement for the 2000 national reliability sample was 95 percent at grade 4, 96 percent at grade 8, and 96 percent at grade 12.

Data Analysis and IRT Scaling

Subsequent to the professional scoring, all information was transcribed to the NAEP database at ETS. Each processing activity was conducted with rigorous quality control. After the assessment information was compiled in the database, the data were weighted according to the population structure. The weighting for the national sample reflected the probability of selection for each student as a result of the sampling design, adjusted for nonresponse. Through post-stratification, the weighting assured that the representation of certain subpopulations corresponded to figures from the U.S. Census and the Current Population Survey.[13]

The procedure used for sample weighting in the state assessments is similar to that used in national samples. However, there are two important differences. First, because there is no oversampling of high-minority schools in state samples, the weighting process does not need to adjust for such a procedure. Second, Current Population Survey target totals are not available or stable on a state-by-state basis. Therefore, the post-stratification process described above is not utilized in the state program.

[13] These procedures are described more fully in the "Weighting and Variance Estimation" section later in this document. For additional information about the use of weighting procedures, see the forthcoming *NAEP 2000 Technical Report*. In addition, the reader may consult the *NAEP 1998 Technical Report* for a discussion of weighting procedures that are common to all NAEP assessments.

Analyses were then conducted to determine the percentages of students who gave various responses to each cognitive and background question. In determining these percentages for the cognitive questions, a distinction was made between missing responses at the end of a block (i.e., missing responses subsequent to the last question the student answered) and missing responses prior to the last observed response. Missing responses before the last observed response were considered intentional omissions. In analysis, omitted responses to multiple-choice items were scored as fractionally correct.[14] For constructed-response items, omitted responses were placed into the lowest score category. Missing responses at the end of the block were considered "not reached" and treated as if the questions had not been presented to the student. In calculating response percentages for each question, only students classified as having been presented the question were included in the denominator of the statistic.

It is standard NAEP practice to treat all nonrespondents to the last question in a block as if they had not reached the question. For multiple-choice and short constructed-response questions, this practice produces a reasonable pattern of results in that the proportion reaching the last question is not dramatically smaller than the proportion reaching the next-to-last question. However, for science blocks that ended with extended constructed-response questions, the standard practice would result in extremely large drops in the proportion of students attempting the final question. Therefore, for blocks ending with an extended constructed-response question, students who answered the next-to-last question but did not respond to the extended constructed-response question were classified as having intentionally omitted the last question.

Item Response Theory (IRT) was used to estimate average science scale scores for the nation and for various subgroups of interest within the nation. IRT models the probability of answering a question in a certain way as a mathematical function of proficiency or skill. The main purpose of IRT analysis is to provide a common scale on which performance can be compared across groups such as those defined by characteristics including gender and race/ethnicity.

In producing the science scales, three distinct IRT models were used. Multiple-choice questions were scaled using the three-parameter logistic (3PL) model; short constructed-response questions rated as acceptable or unacceptable were scaled using the two-parameter logistic (2PL) model; and short constructed-response questions rated according to a three-level rubric, as well as extended constructed-response questions rated on a four- or five-level rubric, were scaled using a Generalized Partial-Credit (GPC) model.[15] Developed by ETS and first used in 1992, the GPC model permits the scaling of questions scored according to multipoint rating

[14] Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[15] Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement* (*16)*2, 159–176.

schemes. The model takes full advantage of the information available from each of the student response categories used for these more complex constructed-response questions.[16]

The science scale is composed of three types of questions: multiple-choice, short constructed-response (scored either dichotomously or allowing for partial credit), and extended constructed response (scored according to a partial-credit model). One question about the science scales concerns the amount of information contributed by each type of question. Unfortunately, this question has no simple answer for the NAEP science assessment, due to the procedures used to form the composite science scale. The information provided by a given question is determined by the IRT model used to scale the question. It is a function of the item parameters and varies by level of science proficiency.[17] Thus, the answer to the query "How much information do the different types of questions provide?" will differ for each level of science performance. When considering the composite science scale, the answer is even more complicated. The science data are scaled separately by the three fields of science (Earth, physical, and life), resulting in three separate subscales at each grade. The composite scale is a weighted combination of these subscales. IRT information functions are only strictly comparable when the item parameters are estimated together. Because the composite scale is based on three separate estimation runs, there is no direct way to compare the information provided by the questions on the composite scale.

Because of the BIB-spiraling design used by NAEP, students do not receive enough questions about a specific topic to provide reliable information about individual performance. (For more information on BIB-spiraling, see "The Assessment Design" section earlier in this document.) Traditional test scores for individual students, even those based on IRT, would lead to misleading estimates of population characteristics, such as subgroup means and percentages of students at or above a certain scale-score level. Consequently, NAEP constructs sets of plausible values designed to represent the distribution of performance in the population. A plausible value for an individual is not a scale score for that individual, but may be regarded as a representative value from the distribution of potential scale scores for all students in the population with similar characteristics and identical patterns of item response. Statistics describing performance on the NAEP science scale are based on the plausible values. 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.[18]

16 More detailed information regarding the IRT analyses used in NAEP assessments will be provided in a forthcoming technical report on the 2000 NAEP assessments. In addition, the reader may consult the *NAEP 1998 Technical Report* for a discussion of analysis procedures that are common to all NAEP assessments.

17 Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement* (*31)*4, 295–311.

18 For theoretical and empirical justification of the procedures employed, see Mislevy, R. J. (1988). Randomization-based inferences about latent variables from complex samples. *Psychometrika (56)*2, 177–196.

For computational details, see the forthcoming *NAEP 2000 Technical Report*.

Item Mapping Procedures

The science performance of fourth-, eighth-, and twelfth-graders can be illustrated by maps that position question or "item" descriptions along the NAEP science scale at each grade where questions are likely to be answered successfully by students. The descriptions used on these maps focus on the science knowledge or skill needed to answer the question. For multiple-choice questions, the description indicates the knowledge or skill demonstrated by selection of the correct option; for constructed-response questions, the description takes into account the knowledge or skill specified by the different levels of scoring criteria for that question.

To map questions to particular points on the NAEP science scale, a response probability convention was adopted that would divide those who had a higher probability of success from those who had a lower probability. Establishing a response probability convention has an impact on the mapping of the test questions onto the science scale. A lower boundary convention maps the science questions at lower points along the scale, and a higher boundary convention maps the same questions at higher points on the scale. The underlying distribution of science skills in the population does not change, but the choice of a response probability convention does have an impact on the proportion of the student population that is reported as "able to do" the questions on the science scales.

There is no obvious choice of a point along the probability scale that is clearly superior to any other point. If the convention were set with a boundary at 50 percent, those above the boundary would be more likely to get a question right than get it wrong, while those below the boundary would be more likely to get the question wrong than right. Although this convention has some intuitive appeal, it was rejected on the grounds that having a 50/50 chance of getting the question right shows an insufficient degree of mastery. If the convention were set with a boundary at 80 percent, students above the criterion would have a high probability of success with an question. However, many students below this criterion show some level of science ability that would be ignored by such a stringent criterion. In particular, those in the range between 50 and 80 percent correct would be more likely to get the question right than wrong, yet would not be in the group described as "able to do" the question.

In a compromise between the 50 percent and the 80 percent conventions, NAEP has adopted two related response probability conventions: 74 percent for multiple-choice questions with four response options (to correct for the possibility of answering correctly by guessing), and 65 percent for constructed-response questions (where guessing is not a factor). These probability conventions were established, in part, based on an intuitive judgment that they would provide the best picture of students' science skills.

Some additional support for the dual conventions adopted by NAEP was provided by Huynh.[19] He examined the IRT information provided by items, according to the IRT model used in scaling NAEP questions. ("Information" is used here in a technical sense. See the forthcoming *NAEP 2000 Technical Report* for details.) Following Bock, Huynh decomposed the item information into that provided by a correct response [P(q) I(q)] and that provided by an incorrect response [(1- P(q)) I(q)].[20] Huynh showed that the item information provided by a correct response to a constructed-response item is maximized at the point along the science scale at which the probability of a correct response is two-thirds (for multiple-choice items, the information provided by a correct response is maximized at the point at which the probability of getting the item correct is 0.74). It should be noted, however, that maximizing the item information I(q), rather than the information provided by a correct response [P(q) I(q)], would imply an item mapping criterion closer to 50 percent.

Results are presented in terms of the composite science scale. However, the science assessment was scaled separately for the three fields of science at grades 4, 8, and 12. The composite scale is a weighted combination of the three subscales for the three fields of science. To obtain item map information, a procedure developed by Donoghue was used.[21] This method models the relationship between the item response function for the subscale and the subscale structure to derive the relationship between the item score and the composite scale (i.e., an item response function for the composite scale). This item response function is then used to derive the probability used in the mapping.

Weighting and Variance Estimation

A multistage sampling design was used to select the students who were assessed. The properties of a sample selected through such a design could be 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 data collection 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 assessment data were estimated using sampling weights. 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. Two components of uncertainty

19 Huynh, H. (1994, October). *Some technical aspects of standard setting*. Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessment, Washington, DC.

20 Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika*, 37, 29–51.

21 Donoghue, J. R. (1997, March). *Item mapping to a weighted composite scale*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

are accounted for in the variability of statistics based on student ability: 1) the uncertainty due to sampling only a relatively small number of students, and 2) the uncertainty due to sampling only a relatively small number of cognitive questions. The first component accounts for the variability associated with the estimated percentages of students who had certain background characteristics or who answered a certain cognitive question correctly.

Because NAEP uses multistage sampling procedures, conventional formulas for estimating sampling variability that assume simple random sampling are inappropriate. NAEP uses a jackknife replication procedure to estimate standard errors. The jackknife standard error provides a reasonable measure of uncertainty for any student information that can be observed without error. However, because each student typically responds to only a few questions within any field of science, the scale score for any single student would be imprecise. In this case, plausible values methodology can be used to describe the performance of groups and subgroups of students, but the underlying imprecision involved in this step adds another component of variability to statistics based on NAEP scale scores.[22]

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 estimation of standard errors may be quite large. Estimates of standard errors subject to a large degree of uncertainty are followed by the "!" symbol to indicate that the nature of the sample does not allow accurate determination of the variability of the statistic. In such cases, the standard errors—and any confidence intervals or significance tests involving these standard errors—should be interpreted cautiously. Additional details concerning procedures for identifying such standard errors are discussed in the forthcoming *NAEP 2000 Technical Report*.

The reader is reminded that, as with findings from all surveys, NAEP results are subject to other kinds of error, including the effects of imperfect adjustment for student and school nonresponse and unknowable effects associated with the particular instrumentation and data collection methods. Nonsampling errors can be attributed to a number of sources—inability to obtain complete information about all selected schools in the sample (some students or schools refused to participate, or students participated but answered only certain questions); ambiguous definitions; differences in interpreting questions; inability or unwillingness to give correct background information; mistakes in recording, coding, or scoring data; and other errors in collecting, processing, sampling, and estimating missing data. The extent of nonsampling error is difficult to estimate and, because of their nature, the impact of such errors cannot be reflected in the data-based estimates of uncertainty provided in NAEP reports.

[22] For further details, see Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics* (*17)*2, 175–190.

Drawing Inferences from the Results

The reported statistics are estimates and are therefore subject to a measure of uncertainty. There are two sources of such uncertainty. First, NAEP uses a sample of students rather than testing all students. Second, all assessments have some amount of uncertainty related to the fact that they cannot ask all questions that might be asked in a content area. The magnitude of this uncertainty is reflected in the standard error of each of the estimates. When the percentages or average scale scores of certain groups are compared, the standard error should be taken into account, and observed similarities or differences should not be relied on solely. Therefore, the comparisons are based on statistical tests that consider the standard errors of those statistics and the magnitude of the difference among the averages or percentages.

Using confidence intervals based on the standard errors provides a way to take into account the uncertainty associated with sample estimates and to make inferences about the population averages and percentages in a manner that reflects that uncertainty. An estimated sample average scale score plus or minus 1.96 standard errors approximates a 95 percent confidence interval for the corresponding population quantity. This statement means that one can conclude with approximately a 95 percent level of confidence that the average performance of the entire population of interest (e.g., all fourth-grade students in public and nonpublic schools) is within plus or minus 1.96 standard errors of the sample average.

As an example, suppose that the average science scale score of the students in a particular group was 156 with a standard error of 1.2. A 95 percent confidence interval for the population quantity would be as follows:

$$\text{Average} \pm 1.96 \text{ standard errors}$$

$$156 \pm 1.96 \times 1.2$$

$$156 \pm 2.35$$

$$(153.65, 158.35)$$

Thus, one can conclude with a 95 percent level of confidence that the average scale score for the entire population of students in that group is between 153.65 and 158.35. It should be noted that this example, and the examples in the following sections are illustrative. More precise estimates carried out to one or more decimal places are used in the actual analyses.

Similar confidence intervals can be constructed for percentages, if the percentages are not extremely large or extremely small. Extreme percentages should be interpreted with caution. Adding or subtracting the standard errors associated with extreme percentages could cause the confidence interval to exceed 100 percent or go below 0 percent, resulting in numbers that are not meaningful. The forthcoming *NAEP 2000 Technical Report* will contain a more complete discussion of extreme percentages.

Analyzing Group Differences in Averages and Percentages

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 percentages appear to be approximately the same.

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 the sample differences are likely to represent actual differences among the groups in the population.

To determine whether a real difference exists between the average scale scores (or percentages of a certain attribute) for two groups in the population, one needs to obtain an estimate of the degree of uncertainty associated with the difference between the averages (or percentages) of these groups 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 the squared standard errors, and taking the square root of that sum.

Standard Error of the Difference =

$$SE_{A\text{-}B} = \sqrt{(SE_A{}^2 + SE_B{}^2)}$$

Similar to how 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 among groups in the population are real. The difference between the averages or percentages of the two groups plus or minus 1.96 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 the groups in the population. If the interval does not contain zero, the difference between the groups is statistically significant (different) at the 0.05 level.

As an example of comparing groups, consider the problem of determining whether the average science scale score of group A is higher than that of group B. Suppose that the sample estimates of the average scale scores and standard errors were as follows:

Group	Average Scale Score	Standard Error
A	218	0.9
B	216	1.1

The difference between the estimates of the average scale scores of groups A and B is two points (218 - 216). 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 plus or minus two standard errors of the difference

$$2 \pm 1.96 \times 1.4$$

$$2 \pm 2.74$$

$$(-0.74, 4.74)$$

The value zero is within the confidence interval; therefore, there is insufficient evidence to claim that group A outperformed group B.

Conducting Multiple Tests

The procedures in the previous 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, many different groups are being compared (i.e., multiple sets of confidence intervals are being analyzed). 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. To hold the significance level for the set of comparisons at a particular level (e.g., 0.05), adjustments (called "multiple comparison procedures"[23]) must be made to the methods described in the previous section. One such procedure, the False Discovery Rate (FDR) procedure[24] was used to control the certainty level.

Unlike the other multiple comparison procedures (e.g., the Bonferroni procedure) that control the familywise error rate (i.e., the probability of making even one false rejection in the set of comparisons), the FDR procedure controls the expected proportion of falsely rejected hypotheses. Furthermore, familywise procedures are considered conservative for large families of comparisons.[25] Therefore, the FDR procedure is more suitable for multiple comparisons in NAEP than other procedures. A detailed description of the FDR procedure appears in the forthcoming *NAEP 2000 Technical Report*.

To illustrate how the FDR procedure is used, consider the comparisons of current and previous years' average science scale scores for the five groups presented in table A.12. Note that the difference in average scale scores and the standard error of the difference are calculated in a way comparable with that of the example in the previous section. The test statistic shown is the difference in average scale scores divided by the standard error of the difference.

The difference in average scale scores and its standard error can be used to find an approximate 95 percent confidence interval as in the example in the previous section or they can be used to identify a confidence percentage. In the example in the previous section, because an approximate 95 percent confidence interval was desired, the number 1.96 was used to multiply the standard error of the difference to create the approximate confidence interval. In the current example, the confidence interval for the test statistics is identified from statistical tables. Instead of checking to see if zero is within the 95 percent confidence interval about the mean, the significance level from the statistical tables can be directly compared to 100-95 = 5 percent.

If the comparison of average scale scores across two years were made for only one of the five groups, there would be a significant difference between the average scale scores for the two years if the significance level were less than 5 percent. However, because

23 Miller, R. G. (1966). *Simultaneous statistical inference*. New York, NY: Wiley.

24 Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, No. 1.*, pp 289–300.

25 Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1994, December). *Controlling error in multiple comparisons with special attention to the National Assessment of Educational Progress.* Research Triangle Park, NC: National Institute of Statistical Sciences.

we are interested in the difference in average scale scores across the two years for all five of the groups, comparing each of the significance levels to 5 percent is not adequate. Groups of students defined by shared characteristics, such as race/ethnicity groups, are treated as sets or families when making comparisons. However, comparisons of average scale scores for each pair of years were treated separately. So the steps described in this example would be replicated for the comparison of other current and previous year average scale scores.

To use the FDR procedure to take into account that all comparisons are of interest to us, the percents of confidence in the example are ordered from largest to smallest: 62, 35, 20, 4, and 1. In the FDR procedure, 62 percent confidence for the Group 4 comparison would be compared to 5 percent, 35 percent for the Group 5 comparison would be compared to 0.05★(5-1)/5 = 0.04 = 4 percent,[26] 20 percent for the Group 1 comparison would be compared to 0.05★(5-2)/5 = 0.03 = 3 percent, 4 percent for the Group 3 comparison would be compared to 0.05★(5-3)/5 = 0.02 = 2 percent, and 1 percent for the Group 2 comparison (actually slightly smaller than 1 prior to rounding) would be compared to 0.05★(5-4)/5 = 0.01 = 1 percent. The last of these comparisons is the only one for which the percent confidence is smaller than the FDR procedure value. The difference in the current year and previous years' average scale scores for the Group 2 students is significant; for all of the other groups, average scale scores for current and previous year are not significantly different from one another. In practice, a very small number of counterintuitive results occur when using the FDR procedures to examine between-year differences in subgroup results by jurisdiction. In those cases, results were not included in this report. NCES is continuing to evaluate the use of FDR and multiple-comparison procedures for future reporting.

Table A.12

FDR comparisons of average scale scores for different groups of students

	Previous year		Current year		Previous year and current year			
	Average scale score	Standard error	Average scale score	Standard error	Difference in averages	Standard error of difference	Test statistic	Percent confidence*
Group 1	224	1.3	226	1.0	2.08	1.62	1.29	20
Group 2	187	1.7	193	1.7	6.31	2.36	2.68	1
Group 3	191	2.6	197	1.7	6.63	3.08	2.15	4
Group 4	229	4.4	232	4.6	3.24	6.35	.51	62
Group 5	201	3.4	196	4.7	-5.51	5.81	-.95	35

* The percent confidence is 2(1−F(x)) where F(x) is the cumulative distribution of the t-distribution with the degrees of freedom adjusted to reflect the complexities of the sample design.

26 The level of confidence times the number of comparisons minus one divided by the number of comparisons is 0.05★(5–1)/5 = 0.04 = 4 percent.

NAEP Reporting Groups

Results are provided for groups of students defined by shared characteristics—region of the country, gender, race or ethnicity, school's type of location, eligibility for the free/reduced-price school lunch program, and type of school. Based on participation rate criteria, results are reported for subpopulations only when sufficient numbers of students and adequate school representation are present. The minimum requirement is at least 62 students in a particular subgroup from at least five primary sampling units (PSUs).[27] However, the data for all students, regardless of whether their subgroup was reported separately, were included in computing overall results. Definitions of the subpopulations are presented below.

Region

Results in NAEP are reported for four regions of the nation: Northeast, Southeast, Central, and West. Figure A.2 shows how states are subdivided into these NAEP regions. All 50 states and the District of Columbia are listed. Other jurisdictions, including territories and the two Department of Defense Educational Activities jurisdictions are not assigned to any region.

Figure A.2

States included in the four NAEP regions

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

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

[27] For the national assessment, a PSU is a selected geographic region (a county, group of counties, or metropolitan statistical area). For the state assessment program, a PSU is most often a single school. Further details about the procedure for determining minimum sample size appear in the *NAEP 1996 Technical Report* and the forthcoming *NAEP 2000 Technical Report*.

Gender

Results are reported separately for males and females.

Race/Ethnicity

The race/ethnicity variable is derived from two questions asked of students and from school records, and it is used for race/ethnicity subgroup comparisons. Two questions from the set of general student background questions were used to determine 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 following question were examined to determine their race/ethnicity.

Which best describes you?

- ❑ White (not Hispanic)
- ❑ Black (not Hispanic)
- ❑ Hispanic ("Hispanic" means someone who is Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or other Spanish or Hispanic background)
- ❑ Asian or Pacific Islander ("Asian or Pacific Islander" means someone who is from a Chinese, Japanese, Korean, Filipino, Vietnamese, Asian American or some other Asian or Pacific Islander 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 responses. For students who filled in the sixth oval ("Other"), provided illegible information or information that could not be classified, or did not respond at all, race/ethnicity was assigned as determined by school records.

Race/ethnicity could not be determined for students who did not respond to either of the demographic questions and whose schools did not provide information about race/ethnicity.

Also, some students indicated that they were from a Hispanic background (e.g., Puerto Rican or Cuban) and that a racial/ethnic category other than Hispanic best described them. These students were classified as Hispanic based on the rules described above.

Type of Location

Results from the 2000 assessment are reported for students attending schools in three mutually exclusive location types: central city, urban fringe/large town, and rural/small town:

Central City: This category includes central cities of all Standard Metropolitan Statistical Areas (SMSA) as defined by the Office of Management and Budget. Central City is a geographical term and is not synonymous with "inner city."

Urban Fringe/Large Town: The urban fringe category includes all densely settled places and areas within SMSA's that are classified as urban by the Bureau of the Census, but which do not qualify as Central City. A Large Town is defined as a place outside a SMSA with a population greater than or equal to 25,000.

Rural/Small Town: Rural includes all places and areas with populations of less than 2,500 that are classified as rural by the Bureau of the Census. A Small Town is defined as a place outside a SMSA with a population of less than 25,000, but greater than or equal to 2,500.

Results for each type of location are not compared across years. This was due to new methods used by NCES to identify the type of location assigned to each school in the Common Core of Data (CCD). The new methods were put into place by NCES in order to improve the quality of the assignments and they take into account more information about the exact physical location of the school.

Eligibility for the Free/Reduced-Price School Lunch Program

Based on available school records, students were classified as either currently eligible for the free/reduced-price school lunch component of the Department of Agriculture's National School Lunch Program or not eligible. Eligibility for the program is determined by students' family income in relation to the federally established poverty level. Free lunch qualification is set at 130 percent of the poverty level, and reduced-price lunch qualification is set at 170 percent of the poverty level. The classification applies only to the school year when the assessment was administered (i.e., the 1999–2000 school year) and is not based on eligibility in previous years. If school records were not available, the student was classified as "Information not available." If the school did not participate in the program, all students in that school were classified as "Information not available."

Type of School

Results are reported by the type of school that the student attends—public or nonpublic. Nonpublic schools include Catholic and other private schools.[28] Because they are funded by federal authorities, not state/local governments, Bureau of Indian Affairs (BIA) schools and Department of Defense Domestic Dependent Elementary and Secondary Schools (DDESS) are not included in either the public or nonpublic categories; they are included in the overall national results.

Grade 12 Participation Rates and Motivation

NAEP has been described as a "low-stakes" assessment. That is, students receive no individual scores, and their NAEP performance has no effect on their grades, promotions, or graduation. There has been continued concern that this lack of consequences affects participation rates of students and schools, as well as the motivation of students to perform well on NAEP. Of particular concern has been the performance of twelfth-graders, who typically have lower student participation rates than fourth- and eighth-graders, and who are more likely to omit responses compared to the younger cohorts.

Participation Rates

In NAEP, there has been a consistent pattern of lower participation rates for older students. In the 2000 NAEP assessments, for example, the student participation rates were 96 percent and 92 percent at grades 4 and 8, respectively. At grade 12, however, the participation rate was 76 percent. School participation rates (the percentage of sampled schools that participated in the assessment) have also typically decreased with grade level. Again citing the 2000 assessments, the school participation rate was 88 percent for the fourth grade, 85 percent for the eighth grade, and 82 percent for the twelfth grade.

The effect of participation rates on student performance, however, is unclear. Students may choose not to participate in NAEP for many reasons, such as desire to attend regular classes so as not to miss important instruction or conflict with other school-based activities. Similarly, there are a variety of reasons for which various schools do not participate. The sampling weights and nonresponse adjustments, described earlier in this document, provide an approximate statistical adjustment for nonparticipation. However, the effect of some school and student nonparticipation may have some undetermined effect on results.

28 Through a pilot study, more detailed breakdowns of nonpublic school results are available on the NAEP web site (http://nces.ed.gov/nationsreportcard/science/results/index.asp).

Motivation

To the extent that students in the NAEP sample are not trying their hardest, NAEP results may underestimate student performance. The concern increases as students get older, and may be particularly pronounced for twelfth-graders. The students themselves furnish some evidence about their motivation. As part of the background questions, students were asked how important it was to do well on the NAEP science assessment. They were asked to indicate whether it was very important, important, somewhat important, or not very important to them. The percentage of students indicating they thought it was either important or very important to do well was 89 percent for fourth-graders, 58 percent for eighth-graders, and 31 percent for twelfth-graders.

It is also interesting to note that students who indicated it was very important for them to do well on NAEP did not have the highest average scores. In fact, at grades 8 and 12, students who reported it was not very important to do well also had higher average scores than those who reported it was very important to do well. These data further cloud the relationship between motivation and performance on NAEP.

Need for Future Research

More research is needed to delineate the factors that contribute to nonparticipation and lack of motivation. To that end, NCES is currently investigating how various types of incentives can be effectively used to increase participation in NAEP. One report that examines the impact of monetary incentives on student effort and performance is available on the NCES web site at http://nces.ed.gov/pubsearch/. Enter NCES#: 2001024.

Cautions in Interpretations

As described earlier, the NAEP science scale makes it possible to examine relationships between students' performance and various background factors measured by NAEP. However, a relationship that exists between achievement and another variable does not reveal its underlying cause, which may be influenced by a number of other variables. Similarly, the assessments do not capture the influence of unmeasured variables. The results are most useful when they are considered in combination with other knowledge about the student population and the educational system, such as trends in instruction, changes in the school-age population, and societal demands and expectations.