

### Accuracy of Data

The accuracy of any statistic is determined by the joint effects of "sampling" and "nonsampling" errors. Estimates based on a sample will differ somewhat from the figures that would have been obtained if a complete census had been taken using the same survey instruments, instructions, and procedures. In addition to such sampling errors, all surveys, both universe and sample, are subject to design, reporting, and processing errors and errors due to nonresponse. To the extent possible, these nonsampling errors are kept to a minimum by methods built into the survey procedures. In general, however, the effects of nonsampling errors are more difficult to gauge than those produced by sampling variability.

### Sampling Errors

The samples used in surveys are selected from a large number of possible samples of the same size that could have been selected using the same sample design. Estimates derived from the different samples would differ from each other. The difference between a sample estimate and the average of all possible samples is called the sampling deviation. The sampling error of a survey estimate is a measure of the variation among the estimates from all possible samples and, thus, is a measure of the precision with which an estimate from a particular sample approximates the average result of all possible samples.

The sample estimate and an estimate of its standard error permit us to construct interval estimates with prescribed confidence that the interval includes the average result of all possible samples. If all possible samples were selected under essentially the same conditions and an estimate and its estimated standard error were calculated from each sample, then: 1) approximately 2/3 of the intervals from one standard error below the estimate to one standard error above the estimate would include the average value of the possible samples; and 2) approximately 19/20 of the intervals from two standard errors above the estimate to two standard errors below the estimate would include the average value of all possible samples. We call an interval from two standard errors below the

estimate to two standard errors above the estimate a 95 percent confidence interval.

Analysis of standard errors can help assess how valid a comparison between two estimates might be. The standard error of a difference between two independent sample estimates is equal to the square root of the sum of the squared standard errors of the estimates.

The standard error (se) of the difference between independent sample estimates "a" and "b" is:

$$se_{a,b} = \sqrt{se_a^2 + se_b^2}$$

To compare changes in between-group differences (groups "a" and "b") over time (years "1" and "2"), we approximate the standard error of the difference as:

$$se = \sqrt{se_{a1}^2 + se_{b1}^2 + se_{a2}^2 + se_{b2}^2}$$

This method overestimates the standard error because it does not account for covariance (the covariance figures were not available). Because of this overestimation, the approach is conservative; that is, one is less likely to obtain significant results.

### State and U.S. Comparisons

For the state-level indicators on student achievement, the state data include public school students only, while the U.S. data include public and nonpublic school students.

### Multiple State Comparisons

The procedure used in Part 1 of the state pages to determine whether the test scores in two years are significantly different is a statistical test based on the assumption that only one test of statistical significance is being performed. However, in Part 2 of the state pages, many different average test scores are being compared (one state must be compared to all other participating jurisdictions). In a case such as this where there are multiple comparisons, statistical theory indicates that the certainty associated with the entire data set is less than that attributable to each individual comparison. To hold the significance level for the entire

set of comparisons to 0.05, adjustments called multiple comparison procedures must be made. A powerful multiple comparison procedure designed by Benjamini and Hochberg was used in this case. This method controls the proportion of falsely rejected hypotheses from among all rejections. The Benjamini/Hochberg application of the False Discovery Rate (FDR) criterion can be described as follows. Let  $m$  be the number of significance tests made, and let  $P_1 \leq P_2 \leq \dots \leq P_m$  be the ordered significance levels of the  $m$  tests, from lowest to highest probability. Let  $\alpha$  be the combined significance level of 0.05. The procedure will compare  $P_m$  with  $\alpha$ ,  $P_{m-1}$  with  $\alpha(m-1)/m$ , ...,  $P_j$  with  $\alpha j/m$ , stopping the comparisons with the first  $j$  such that  $P_j \leq \alpha j/m$ . All tests associated with  $P_1, \dots, P_j$  are declared significant; all tests associated with  $P_{j+1}, \dots, P_m$  are declared not significant.

**Source:** Benjamini, Y., & Hochberg, Y. (1994). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57 (1): 289-300.

### National Assessment of Educational Progress (NAEP)

The National Assessment of Educational Progress, or NAEP, is the only nationally representative and ongoing assessment of what students in the United States know and are able to do in various academic subjects. Since 1969, NAEP has periodically assessed U.S. 4th, 8th, and 12th graders in reading, writing, mathematics, science, history, geography, the arts, and civics. NAEP is funded by Congress and is administered by the U.S. Department of Education's National Center for Education Statistics.

Congress expanded NAEP to allow the reporting of comparable state by state results, beginning with the 1990 mathematics assessment. Participation in state-level NAEP is voluntary, and has increased from 40 states and territories in the initial 1990 assessment, to 45 in the 1996 mathematics and science assessments. To date, state-level NAEP assessments have been administered in reading, mathematics, and science. During 1998, a new state-level assessment in writing was administered at Grade 8. Reading was assessed again at Grade

4 and, for the first time, at Grade 8. During 2000, state-level NAEP assessments will be administered once again in mathematics at Grades 4 and 8, and in science at Grade 8. Science will also be assessed at Grade 4 for the first time at the state level.

NAEP assessments include both multiple-choice and open-ended test items. NAEP also collects demographic, curricular, and instructional information through student, teacher, and school administrator surveys. Since NAEP is used for large-scale monitoring and is not designed to be an individual test, no participating student takes the entire NAEP examination. Instead, samples of students in Grades 4, 8, and 12 are selected to take different portions of the test.

This approach, called matrix sampling, minimizes the number of students and the amount of time needed for testing, yet still allows policymakers to draw valid conclusions about how all students would have performed if they had taken the entire test.

### National Assessment Governing Board (NAGB) Achievement Levels

The NAEP data shown in this report should be interpreted with caution. The Goals Panel's performance standard classifies student performance according to achievement levels adopted by the National Assessment Governing Board for the National Assessment of Educational Progress. This effort has resulted in three achievement levels: Basic, Proficient, and Advanced. The Goals Panel has set its performance standard at the Proficient or Advanced levels on NAEP.

The NAGB achievement levels are reasoned judgements of what students should know and be able to do. They are attempts to characterize overall student performance in particular subject matters. The NAGB achievement levels represent a useful way to categorize overall performance on NAEP. They are also consistent with the Panel's efforts to report such performance against a high-criterion standard.

Readers should exercise caution, however, in making particular inferences about what students at each level actually know and can do. A NAEP assessment is a complex picture of student achievement,

and applying external standards for performance is a difficult task. The process of setting achievement levels is still in transition and both NAGB and NCES regard the achievement levels as developmental. The Goals Panel acknowledges these limitations but believes that, used with caution, these levels convey important information about how American students are faring in reaching Goal 3.

**Basic:** *This level, below proficient, denotes partial mastery of knowledge and skills that are fundamental for proficient work at each grade – 4, 8, and 12.*

**Proficient:** *This central level represents solid academic performance for each grade tested – 4, 8, and 12. It reflects a consensus that students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling.*

**Advanced:** *This higher level signifies superior performance beyond proficient grade-level mastery at Grades 4, 8, and 12.*

Thus far, state-level assessments have been conducted in reading, mathematics, science, and writing. Student achievement levels have been established by NAGB in these subject areas, with the exception of writing.

### Mathematics Achievement

See general technical notes regarding NAEP and the NAGB achievement levels.

Forty jurisdictions (states and territories) participated in the 1990 trial mathematics assessment of 8th graders, and 44 jurisdictions participated in the 1992 state mathematics assessments of 4th and 8th graders.

In 1996, 45 jurisdictions participated in the voluntary assessment of 4th and 8th graders. However, three states (Nevada, New Hampshire, and New Jersey) failed to meet the minimum school participation guidelines for public schools at Grade 8 (i.e., an initial school participation rate of 70% for public schools); therefore, their results were not released. The following states did not satisfy one of the guidelines for school sample participation rates at Grade 4: Alaska, Arkansas, Iowa, Michigan, Montana, Nevada, New Jersey, New York,

Pennsylvania, South Carolina, and Vermont. The following states did not satisfy one of the guidelines for school sample participation rates at Grade 8: Alaska, Arkansas, Iowa, Maryland, Michigan, Montana, New York, South Carolina, Vermont, and Wisconsin.

**Sources:** Reese, C.M., Miller, K.E., Mazzeo, J., & Dossey, J.A. (1997, February). *NAEP 1996 mathematics report card for the nation and the states*. Washington, DC: National Center for Education Statistics.

National Center for Education Statistics, 1990 and 1992 NAEP mathematics data (revised), October 1996.

### Science Achievement

See general technical notes regarding NAEP and the NAGB achievement levels.

In 1996, 45 jurisdictions (states and territories) participated in the voluntary program. However, three states (Nevada, New Hampshire, and New Jersey) failed to meet the minimum school participation guidelines for public schools (i.e., an initial school participation rate of 70% for public schools); therefore, their results were not released. The following states did not satisfy one of the guidelines for school sample participation rates: Alaska, Arkansas, Iowa, Maryland, Michigan, Montana, New York, South Carolina, Vermont, and Wisconsin.

No school location data are reported for the 1996 NAEP science assessment. Although these data were collected via NAEP background questionnaires, the definitions used for school location have changed, and the National Assessment Governing Board has expressed reservations about the use of these data.

**Source:** Bourque, M.L., Champagne, A.B., & Crissman, S. (1997, October). *1996 science performance standards: Achievement results for the nation and the states*. Washington, DC: National Assessment Governing Board.

### NAEP Student Subgroups

NAEP results are reported for student subgroups only if they meet minimum requirements for student sample size and school representation. For public schools, the minimum number of

students per subgroup is 62, and students in the sample must be drawn from a minimum of 5 primary sampling units (PSUs). At the state level, a PSU is usually a single school. At the national level, a PSU is a region, such as a county, group of counties, or a metropolitan statistical area.

In this document, NAEP results are reported by five types of subgroups: sex, race/ethnicity, parents' highest level of education, school location, and student eligibility for free/reduced-price lunch, which is often used as a measure of poverty. Brief definitions and technical information about the five subgroups reported in this document follow.

- **Sex.** Student results are reported separately for males and females. This information was collected on general student background questionnaires.
- **Race/ethnicity.** Student results are reported according to five federal reporting categories:
  - ◆ *American Indian/Alaskan Native;*
  - ◆ *Asian/Pacific Islander;*
  - ◆ *Black;*
  - ◆ *Hispanic;* and
  - ◆ *White.*

Classification was based on student self-reports to general background questions. A sixth response category, "Other," was also a response option.

**Parents' highest level of education.** Parents' highest level of education was based on student self-reports to general background questions. If a student indicated that his or her parents had completed different levels of education, the response was classified according to the higher of the two levels. In this document, student achievement data are reported by four levels of parental education:

- ◆ *less than high school;*
- ◆ *high school graduate;*
- ◆ *some education beyond high school;* and
- ◆ *college graduate.*

A fifth response category, "I don't know," was also a response option. The reader should note that nationally, 36% of 4th graders and 11% of 8th graders did not know the highest level of education completed by either parent.

- **School location.** Each student's school was assigned to one of three mutually exclusive categories of school location:
  - ◆ *central city;*
  - ◆ *urban fringe/large town;* or
  - ◆ *rural/small town.*

The definitions used by the National Center for Education Statistics for school location are as follows:

- ◆ **Central City:** The Central City category includes central cities of all Metropolitan Statistical Areas (MSAs). (Each Metropolitan Statistical Area (MSA) is defined by the Office of Management and Budget.) Central City is a geographic term and is not synonymous with "inner city."
- ◆ **Urban Fringe/Large Town:** An Urban Fringe includes all densely settled places and areas within MSAs that are classified as urban by the Bureau of the Census. A Large Town is defined as places outside MSAs with a population greater than or equal to 25,000.
- ◆ **Rural/Small Town:** Rural includes all places and areas with a population of less than 2,500 that are classified as rural by the Bureau of the Census. A Small Town is defined as places outside MSAs with a population of less than 25,000, but greater than or equal to 2,500.
- **Eligibility for free/reduced-price lunch program.** Student eligibility for the free/reduced-price lunch component of the U.S. Department of Agriculture's National School Lunch Program was based on school records. Eligibility referred only to the school year in which the NAEP assessment was administered.

### Third International Mathematics and Science Study (TIMSS)

The Third International Mathematics and Science Study, or TIMSS, is the most comprehensive international study of mathematics and science achievement conducted to date. TIMSS was administered in 1995, and tested half a million students in 30 different languages and in 41 countries, including the United States. In addition to the student assessments, TIMSS collected information through questionnaires administered to teachers, students, and school administrators; comparisons of mathematics and science curriculum guides and textbooks; videotapes of mathematics instruction in 8th grade classrooms in the United States, Japan, and Germany; and detailed case studies of education policies in the same three countries.

Three age groups were tested in the participating countries, corresponding roughly to Grades 4, 8, and 12 in the United States. Twenty-six nations took part in the mathematics and science assessments at Grade 4, 41 participated at Grade 8, and 23 participated at Grade 12. Both public and private schools participated, and the same students were tested in both mathematics and science. TIMSS drew random samples of virtually all students in the participating countries, not just those enrolled in mathematics and science courses. Nearly all countries in TIMSS accomplished high participation rates, and did not exempt large portions of their student bodies from testing. Exceptions among the countries that participated in the Grade 8 assessment follow.

The following countries did not meet international guidelines at Grade 8: Australia, Austria, Belgium (French), Bulgaria, Colombia, Denmark, Germany, Greece, Israel, Kuwait, Netherlands, Romania, Scotland, Slovenia, South Africa, and Thailand. In four countries, more than 10 percent of the population was excluded from testing at Grade 8: England, Germany, Israel, and Lithuania. In Belgium (Flemish), England, Germany, Latvia (LSS), Switzerland, and the United States, a participation rate of 75 percent of the

schools and students combined for Grade 8 was achieved only after replacements for refusals were substituted.

A 1998 research study linked state mathematics and science results from the 1996 National Assessment of Educational Progress (NAEP) and the 1995 country results from TIMSS. The linking study predicts TIMSS results for the states and jurisdictions that participated in the 1996 NAEP on the basis of their actual NAEP scores. Actual TIMSS results are also available for Minnesota, which tested a representative sample of 8th graders with the TIMSS instruments in 1995. Missouri and Oregon also tested representative samples of 8th graders with the TIMSS instruments in 1997, but their results have not yet been publicly released. For more detailed information about the statistical linking and validation procedures involved in this research and development effort, see the forthcoming technical report, *Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study at the eighth grade: A research report*.

**Sources:** U.S. Department of Education, National Center for Education Statistics. (1997). *Pursuing excellence: A study of U.S. eighth-grade mathematics and science teaching, learning, curriculum, and achievement in international context*, NCES 97-198, Washington, DC: U.S. Government Printing Office.

Johnson, E.G., & Siegendorf, A. (1998, May). *Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study: Eighth grade results*. Report prepared for the U.S. Department of Education, National Center for Education Statistics, NCES 98-500, Washington, DC: U.S. Government Printing Office.