NCES
National Center for
Education Statistics

The Nation's Report Card NAEP

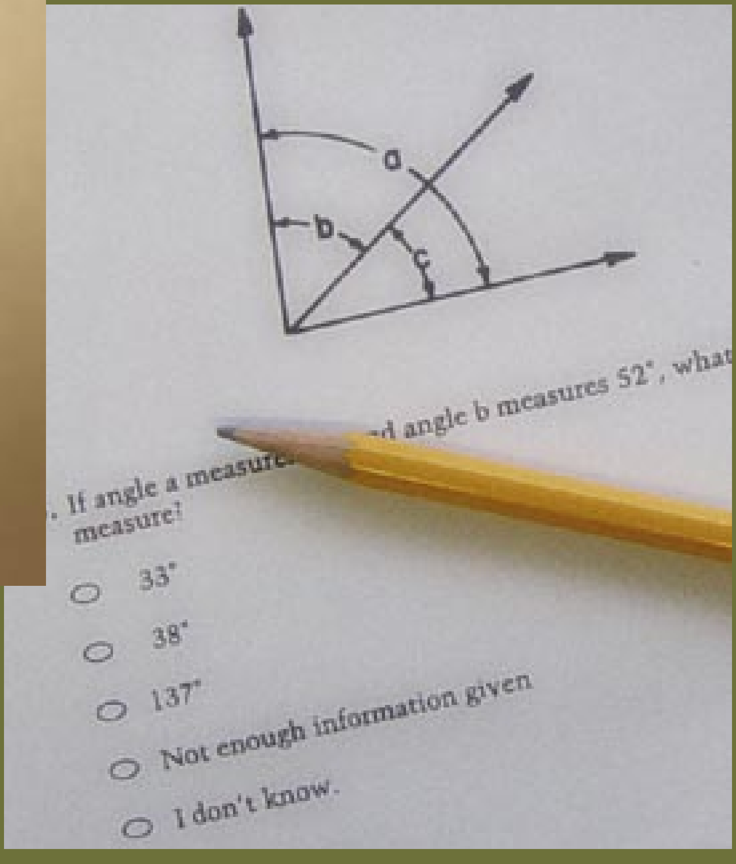
The Nation's Report Card™

NAEP 2004
Trends in Academic Progress

Three Decades of Student Performance in Reading and Mathematics

U.S. Department of Education
Institute of Education Sciences
NCES 2005–464

READING, 1971–2004



MATHEMATICS, 1973–2004

What is The Nation's Report Card™?

The Nation's Report Card™, the National Assessment of Educational Progress (NAEP), is a nationally representative and continuing assessment of what America's students know and can do in various subject areas. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history, geography, and other subjects.

NAEP is a congressionally mandated project of the National Center for Education Statistics within the Institute of Education Sciences of the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations.

By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement and relevant variables is collected under this program. The privacy of individual students and their families is protected, and the identities of participating schools are not released.

In 1988, Congress established the National Assessment Governing Board (NAGB) to oversee and set policy for NAEP. The Board is responsible for selecting the subject areas to be assessed; setting appropriate student achievement levels; developing assessment objectives and test specifications; developing a process for the review of the assessment; designing the assessment methodology; developing guidelines for reporting and disseminating NAEP results; developing standards and procedures for interstate, regional, and national comparisons; determining the appropriateness of all assessment items and ensuring the assessment items are free from bias and are secular, neutral, and nonideological; taking actions to improve the form, content, use, and reporting of results of the National Assessment; and planning and executing the initial public release of NAEP reports.

The National Assessment Governing Board

NCES
National Center for
Education Statistics

The Nation's Report Card
NAEP

NAEP 2004 Trends in Academic Progress

Three Decades of Student Performance in Reading and Mathematics

U.S. Department of Education
Institute of Education Sciences
NCES 2005–464

July 2005

Marianne Perie
Rebecca Moran
Anthony D. Lutkus
Educational Testing Service

William Tirre
Project Officer
National Center for Education Statistics

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Statistics
Grover J. Whitehurst
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in other countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

July 2005

The NCES World Wide Web Home Page address is http://nces.ed.gov
The NCES World Wide Web Electronic Catalog address is http://nces.ed.gov/pubsearch

SUGGESTED CITATION

Perie, M., Moran, R., and Lutkus, A.D. (2005). *NAEP 2004 Trends in Academic Progress: Three Decades of Student Performance in Reading and Mathematics* (NCES 2005–464). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington, DC: Government Printing Office.

FOR ORDERING INFORMATION ON THIS REPORT, WRITE:

U.S. Department of Education, ED Pubs, P.O. Box 1398, Jessup, MD 20794-1398, or call toll free 1-877-4ED-Pubs; or order online at http://www.edpubs.org

CONTENT CONTACT:

William Tirre
202-502-7361
William.Tirre@ed.gov

The work upon which this publication is based was performed for the National Center for Education Statistics by Educational Testing Service, the Education Statistics Services Institute, Pearson Educational Measurement, and Westat.

Executive Summary

Since its inception in 1969, NAEP has tracked trends in student performance over time.

The citizens and leaders of the United States have long valued education as a foundation for democracy, a resource for economic prosperity, and a means of realizing personal goals and individual potential. Throughout the nation's history, the commitment to educate children has grown stronger and more inclusive, and in recent decades, so has the expectation that our nation's schools and teachers be accountable (Ravitch 2002). In 2002, the reauthorization of the Elementary and Secondary Education Act—also known as the No Child Left Behind (NCLB) Act—further strengthened that commitment and expectation.

Since its inception in 1969, the National Assessment of Educational Progress (NAEP) has served the important function of measuring our nation's educational progress by regularly administering various subject-area assessments to nationally representative samples of students. One of the primary objectives of NAEP is to track trends in student performance over time. This report presents the results of NAEP long-term trend assessments in reading and mathematics, which were most recently administered in 2004 to students ages 9, 13, and 17. Because the assessments have been administered at different times in the 35-year history of NAEP, they make it possible to chart educational progress since 1971 in reading and 1973 in mathematics. Prior to 2004, the most recent long-term trend assessment was given in 1999, when results were reported for reading, mathematics, and science.

It should be noted that these long-term trend assessments are different from more recently developed assessments in the same subjects that make up the "main NAEP" assessment program. Because the instruments and methodologies of the two assessment programs are different, comparisons between the long-term trend results presented in this report and the main assessment results presented in other NAEP reports are not possible.

Approximately 38,000 students participated in the reading assessment, and 37,000 participated in the mathematics assessment. Appendix A provides technical information on this study, including sample sizes and a description of the significance tests done on each set of results. Only differences that have been determined to be statistically significant at the 0.05 level after controlling for multiple comparisons are included in this report.

National Results

National results, provided in chapter 2, are described in three ways: average score, score at selected percentiles, and percentage of students performing at or above each performance level. Student performance in each subject area is summarized as an average score on a 0–500 scale. The five long-term trend performance levels presented in this report were set at 50-point intervals on the two subject-area scales to provide a verbal description of student performance at different points on the scale. All national findings are reported from 1971–2004 for reading and 1973–2004 for mathematics. The primary findings include the following:

Average Scores

- Between 1999 and 2004, average reading scores increased at age 9 and average mathematics scores increased at ages 9 and 13. No measurable changes in average scores were found at age 17 in either subject between 1999 and 2004.
- In reading, 9-year-olds scored higher in 2004 than in any previous assessment year, with an increase of 7 points between 1999 and 2004. Average scores for age 13 showed no measurable differences between assessment years 1999 and 2004, but still were higher in 2004 than the scores in 1971 and 1975. For age 17, the average score in 2004 was not measurably different from the average score in the first assessment year, 1971.
- The average score in mathematics at age 9 was higher in 2004 than in any previous year—9 points higher than in 1999. The average score for 13-year-olds increased between 1999 and 2004 by 5 points. The average score at age 17 was not measurably different from 1973 or 1999.

Percentiles

- The reading score of 9-year-olds at the median (50th percentile) was higher in 2004 than the median score in every other year.
- Overall gains in reading scores for 13-year-olds were evident among higher performing students—those scoring at the 75th and 90th percentiles—between 1971 and 2004.
- Seventeen-year-olds showed no measurable improvements in reading scores at any of the selected percentiles between 1999 and 2004 or between 1971 and 2004.
- Mathematics scores for 9-year-olds at each of the selected percentiles showed gains between 1978 and 2004, increasing 26 points at the 10th percentile, 23 points at the 50th percentile, and 18 points at the 90th percentile.
- The mathematics score for 13-year-olds at each of the five percentile levels was higher in 2004 than in every previous assessment year, except at the 10th percentile.
- Mathematics scores for 17-year-olds in 2004 showed no measurable change since 1992 at any of the five percentiles.

Performance Levels

- The partially developed skills and understanding associated with reading at level 200 were demonstrated by 70 percent of 9-year-olds in 2004, more than in any other assessment year except 1980; by 94 percent of 13-year-olds; and by almost all 17-year-olds.
- The percentages of 13-year-olds and 17-year-olds who demonstrated the ability to interrelate ideas and make generalizations in reading (level 250) were 61 percent and 80 percent, respectively, in 2004, not measurably different from those in 1971 and 1999.
- Reading performance at or above level 300—understanding complicated information—was demonstrated by 38 percent of 17-year-olds in 2004, down from 41 percent a decade earlier in 1994.

- The beginning skills and understandings characteristic of level 200 in mathematics were demonstrated by 89 percent of 9-year-olds in 2004, more than in any other assessment year. Approximately 99 percent of 13-year-olds also demonstrated at least this level of performance in 2004.
- At age 13, the percentages of students at level 300 in mathematics increased from 17 percent in 1990 to 23 percent in 1999 and then to 29 percent in 2004. Students at this level could perform moderately complex procedures and use logical reasoning to solve problems. In 2004, 59 percent of 17-year-olds were at or above level 300 in mathematics, an increase of 7 percentage points from 1978.
- Across the assessment years in mathematics, between 5 and 8 percent of 17-year-olds performed at level 350, the highest performance level, in which students applied a range of reasoning skills to solve multistep problems.

Student Group Results

Chapter 3 describes the average scores for various groups of students, including male and female students; White, Black, and Hispanic students; and student-reported levels of parents' education, which included less than high school, graduated from high school, some education after high school and graduated from college. Some of the results were as follows:

Gender

- At all three ages in 2004, female students had higher average reading scores than their male counterparts.
- In 2004, there was no measurable difference between the average mathematics scores of male and female students at age 9, but at ages 13 and 17, male students scored higher on average than female students.
- The gender gap for 9-year-olds' reading scores in 2004 was smaller than the gaps in the first three assessment years and 1996. This gap did not change measurably between 2004 and any previous assessment year for 13-year-olds. This score gap in 2004 showed no measurable difference for 17-year-olds from the gap in 1999 or 1971.

Race/Ethnicity

- White students had higher average reading scores in 2004 than in 1971 at ages 9 and 13.
- For Black students at all three ages, average reading scores in 2004 were higher than in 1971.
- Although White students continue to outscore Black students, the White-Black score gap in reading narrowed from 1971 to 2004 at all three ages. The White-Black reading score gap for 9-year-olds decreased from 35 points in 1999 to 26 points in 2004.
- For Hispanic students, the average reading score at age 9 was higher in 2004 than in any other assessment year. Their average score at age 13 was higher in 2004 than in 1975, but not measurably different from that in 1999. No measurable difference was found between the average score for Hispanic students at age 17 in 2004 and that in 1999.
- Although White students continue to outscore Hispanic students, the White-Hispanic reading score gap for students at age 9 in 2004 was smaller than it was in 1994, 1984, 1980, and 1975. The White-Hispanic reading score gap for 13-year-olds showed no measurable difference between 2004 and 1999 or 1975. The score gap between White and Hispanic students at age 17 was measurably smaller in 2004 than in 1975.
- White students at all three ages scored higher, on average, in 2004 than in 1973 in mathematics.
- The average mathematics scores for Black students were higher in 2004 than in 1973 at all three ages. Average scores for Black students at ages 9 and 13 were higher in 2004 than in any previous assessment year.
- The differences in average scores for White and Black students at all ages decreased between the first (1973) and the most recent (2004) assessment in mathematics, although White students continued to outscore Black students in 2004. During this same period, the White-Black score gaps in mathematics narrowed by 12, 19, and 12 points for ages 9, 13, and 17, respectively.

- Hispanic students' performance in mathematics was higher at all three ages in 2004 than in any assessment year from 1973 through 1982. Average scores for Hispanic students at ages 9 and 13 were higher in 2004 than in any previous assessment year.
- White students scored higher on average than Hispanic students at all three age levels in 2004. For ages 13 and 17, the White-Hispanic score gap was smaller in 2004 than in 1973, but for age 9 there was no measurable difference in the size of the score gap between the first (1973) and most recent (2004) assessment year.

Parents' Education

- In 2004, the percentage of students reporting that at least one parent graduated from college has increased since 1980 for reading and 1978 for mathematics, while the percentage of students reporting that the highest level of education for their parents was a high school diploma or less has decreased.
- At age 13, there have been no measurable changes in average reading scores between 2004 and any previous assessment year regardless of the level of parents' education reported by the student.
- The average reading score for 17-year-olds who indicated that at least one parent had some education after high school was lower in 2004 than in any previous assessment year. For 17-year-olds who indicated that at least one parent graduated from college, the average score in 2004 (298) was lower than the average scores in 1990 (302) and 1984 (302).
- Students who reported that their parents had less than a high school education showed no measurable change in average mathematics score between 1999 and 2004 at either age 13 or 17, but their 2004 scores were higher than those in 1978.
- For students whose parents' highest education level was high school graduation or some education after high school, the average mathematics score at age 13 was higher in 2004 than in any other assessment year, while at age 17 there were no measurable changes between 1978 and 2004.
- For students with at least one parent who graduated from college, the average mathematics score in 2004 was higher than in any other assessment year at age 13; no measurable difference was seen at age 17 between 1978 and 2004.

Contextual Variables

As described in chapter 4, examining student scores in the context of their learning and home environments provides useful information. Learning and home factors for which trends are reported include students' reports of how often they read for fun, completed homework, used computers, and watched television, and the advanced mathematics courses they had taken. Some of the findings include the following:

Homework. Students who took the reading assessment were asked how many hours they had spent on homework the previous day.

- The percentage of students at age 9 indicating that no homework was assigned or that they did not do any homework decreased between 1984 and 2004. In 2004, a greater percentage of 9-year-olds indicated that they spent less than 1 hour on homework than in any other year in which the question was asked.
- In 2004, the average reading score of 9-year-olds who spent less than 1 hour on homework was higher than the average reading scores of students who did not do the homework that was assigned or who spent more than 2 hours on homework.
- At age 13, the percentage of students spending less than 1 hour on homework increased from 36 percent in 1984 to 40 percent in 2004. At the same time, the percentage of students spending 1 to 2 hours on homework decreased from 29 percent in 1984 to 26 percent in 2004.

- At age 13, students who spent 1 to 2 hours or 2 or more hours on homework had higher average reading scores than their peers who spent less than 1 hour on homework, did not do their homework, or did not have any homework to do.
- At age 17, the percentage of students reporting that they were not assigned homework increased from 22 to 26 percent. At the same time, the percentage of 17-year-olds indicating they had spent 1 to 2 hours on homework the previous day decreased from 27 to 22 percent between 1984 and 2004.
- At age 17, students who spent 2 or more hours on homework had higher average reading scores in 2004 than those who spent 1 to 2 hours, whose scores were higher than those who spent less than 1 hour, whose scores in turn were higher than those who did not do any homework.

Reading for Fun. Students who took the reading assessment were asked to estimate how often they read for fun.

- There were no measurable changes between 1984 and 2004 in the percentage of 9-year-olds indicating that they read for fun almost every day. At ages 13 and 17, the percentage saying they read for fun almost every day was lower in 2004 than in 1984. This trend was accompanied by an increase over the same 20-year time period in the percentage indicating that they never or hardly ever read for fun.
- At all three ages, students who indicated that they read for fun almost every day had higher average reading scores in 2004 than those who said that they never or hardly ever read for fun. Students at all three age levels who said that they read for fun once or twice a week had higher average scores than those who never or hardly ever read for fun.

Computer Access and Usage. Students at ages 13 and 17 who took the mathematics assessment were asked three questions about their access to computers and how they used them.

- The percentage of 13-year-olds with access to computers in schools increased from 12 percent in 1978 to 57 percent in 2004. The percentage of students receiving instruction in computers at age 13 also increased, from 14 percent in 1978 to 48 percent in 2004. In the 2004 assessment, 69 percent of 13-year-olds said that they had used a computer to solve a mathematical problem.
- Similar increases were also seen among 17-year-olds, where the percentage of students with access to a computer in school increased by 33 percentage points between 1978 and 2004. The percentage of 17-year-olds using a computer to solve mathematics problems increased from 46 percent in 1978 to 66 percent in 1999, then to 70 percent in 2004. In that year, 36 percent reported that they had studied mathematics using computers.
- There were no measurable differences in mathematics scores between 13-year-olds who responded positively and those who responded negatively to any of the computer access and usage questions in 2004. At age 17, students who indicated that they had access to a computer at school scored 5 points higher in 2004 than students who did not have such access.
- In 2004, students at age 17 who reported that they had used a computer to solve a mathematical problem scored 6 points higher on average than students who had not used a computer for that purpose. There was no measurable difference in average mathematics scores for 17-year-olds based on whether or not they had studied mathematics using computers.

Course-Taking Patterns in Mathematics. Students at age 17 who took the mathematics assessment were asked to check all the mathematics courses they had taken or were currently taking. The highest course checked was used for the analyses.

- A greater percentage of 17-year-olds indicated they were taking or had taken calculus in 2004 than in any previous assessment year. The percentage taking second-year algebra increased from 37 percent in 1978 to 53 percent in 2004, while the percentage of students who indicated that the highest level of mathematics they had taken by age 17 was pre-algebra or algebra was lower in 2004 than in 1978.
- The trend towards higher-level course-taking was seen across all three racial/ethnic groups shown. The percentage of White, Black, and Hispanic students who indicated that their highest course was second-year algebra was higher in 2004 than in 1978. In 2004, a higher percentage of White students took calculus (19 percent) compared to Black students at the same age (8 percent). At 14 percent, the percentage of Hispanic students taking calculus was not measurably different from the percentage of either White or Black students in 2004.

2004 Bridge Study

Several changes were made to the long-term trend assessment in 2004 to align it with current assessment practices and policies applicable to the NAEP main assessments. These changes, discussed in detail in chapter 5, included replacing items that had outdated material, eliminating blocks of items for subjects no longer reported, replacing background questions, and changing some administration procedures. In addition, the 2004 modified assessment provided for the inclusion of and accommodations for students with disabilities and English language learners.

A bridge study was conducted to ensure that the interpretation of the assessment results remains constant over time. A bridge study involves administering two assessments: one that replicates the assessment given in the previous assessment year (a bridge assessment), and one that represents the new design (a modified assessment). In 2003–2004, students were randomly assigned to take either the bridge assessment or the modified assessment. The bridge assessment replicated the instrument given in 1999 and used the same administration techniques. The modified assessment included the new items and features discussed above. This modified assessment will provide the basis of comparison for all future assessments, and the bridge study will link its results back to the results of the past 33 years. The results from the bridge study are presented in chapters 2 and 4, and comparisons between the two assessments are provided in chapter 5.

- Comparing the results of the modified and bridge assessments demonstrates that the link between the 2004 bridge and modified assessments was successful.

Acknowledgments

The authors wish to thank all of those who contributed to the design, writing, production, and review of this report for their thoughtful critique, careful fact checking, insightful suggestions, and creativity. In particular, the authors acknowledge the contributions of Steve Gorman, James Griffith, Lisa Hudson, Andrew Kolstad, Taslima Rahman, and Marilyn Seastrom of the National Center for Education Statistics (NCES); Mary Crovo, Larry Feinberg, Ray Fields, Susan Loomis, and Sharif Shakrani of the National Assessment Governing Board; and Tajuana Bates, Yang Chin, Kim Gettis, Dana Kelly, and Alan Vanneman of the Education Statistics Services Institute. Outside reviewers included Lisa Clarke and Nancy Horkay.

This report would not have been possible without the ETS team of researchers and data analysts who designed the administration and analyzed the results to produce the data included in this report. These ETS staff include Andreas Oranje, Brenda Tay-Lim, Amy Dresher, Lydia Gladkova, and Catherine McClellan from the research group; and Tatyana Petrovicheva, Yuxin Tang, David Freund, Laura Jerry, Matthew Kandathil, Edward Kulick, Youn-Hee Lim, Mei-Jang Lin, and Haiyang Liu from the data analysis group. Gloria Dion, David Garber, and Patricia Donahue prepared the items for release. Marilyn Binkley from NCES also worked on the assessment development and design of the long-term trend.

Many people at ETS were responsible for the careful edit and review of these chapters. In particular the authors wish to thank our advisors, Nancy Mead and Wendy Grigg; our editors, Arlene Weiner, Mary Daane, and Linda Myers; and our quality control staff, Ming Kuang, Carmen Payton, Janice Goodis, and Chrystal Murphy. Finally, we wish to thank the ETS production staff, Loretta Casalaina, Rick Hasney, and Susan Mills, for transforming our final chapters into a crisp and attractive report, and our special staff assistant, Karen Damiano, who helped ensure that the final details were completed.

Contents

List of Tables

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK.

The long-term trend assessment has been measuring student progress in reading for 33 years and in mathematics for 31 years.

Chapter 1 Introduction

The citizens and leaders of the United States have long valued education as a foundation for democracy, a resource for economic prosperity, and a means of realizing personal goals and individual potential. Throughout the nation's history, the commitment to educate children has grown stronger and more inclusive, and in recent decades, so has the expectation that our nation's schools and teachers be accountable (Ravitch 2002). In 2002 the reauthorization of the Elementary and Secondary Education Act—also known as the No Child Left Behind (NCLB) Act—further expanded that commitment and expectation.

As educators and policymakers turn their attention to student achievement as measured by assessments, examining trends—student performance now compared to in the past—can inform efforts to increase student performance in the future. The National Assessment of Educational Progress (NAEP) is one of the most important resources for monitoring the student achievement. Since its inception in 1969, NAEP has served the important function of measuring our nation's educational progress by regularly administering various subject-area assessments to nationally representative samples of students. One of the primary objectives of NAEP is to track trends in student performance over time. This report presents the results of NAEP long-term trend assessments in reading and mathematics, which were administered in the 2003–2004 school year (referred to hereafter as 2004) to students ages 9, 13, and 17. Because the same assessments have been administered at different times in the 35-year history of NAEP, they make it possible to chart educational progress since 1971 in reading and 1973 in mathematics.

The specific focus of this long-term trend report is to compare student performance in 2004 to past performance, measured by the most recent assessment in 1999 and previous assessments back to the early 1970s.

NAEP Assessments

NAEP is a project of the National Center for Education Statistics (NCES) within the Institute of Education Sciences of the U.S. Department of Education. The National Assessment Governing Board (NAGB), an independent group created by Congress in 1988, provides policy direction for NAEP. (Information about NAGB can be found on its website, http://www.nagb.org/.)

NAEP includes two components: the long-term trend assessments and the main assessments. The existence of the two national assessment programs—long-term trend and main—makes it possible for NAEP to meet two important objectives. The long-term trend program uses substantially the same assessments decade after decade, each time a subject is assessed, in order to measure student progress in that subject over time. In contrast, the main NAEP assessments are periodically adapted to reflect contemporary curriculum policies, content currently in use in the nation's schools, and improvements in techniques of educational measurement. In this way, main NAEP can provide valid data for those seeking evidence for contemporary questions, and long-term trend NAEP can provide data for evaluating change over long periods. For example, while the current main NAEP reading assessment, given in 2005, was first administered in 1992, the long-term trend reading assessment dates back to 1971.

This report presents the results from the long-term trend assessments only. Because the long-term trend assessments use different questions from those used in the main assessments, and because students are sampled by age for the long-term trend assessments, rather than by grade as in the main assessments, it is not possible to compare results from the two assessment programs.

Overview of the 2004 Long-Term Trend Assessments

The long-term trend assessment originally was given in four subjects: mathematics, science, reading, and writing. At the time of the last long-term trend report (1999), NAGB discontinued the assessment in writing for technical reasons. More recently, NAGB decided that changes were needed to the design of the science assessment and, given recent advances in the field of science, to its content. For instance, many science questions that were written in the late 1960s are no longer relevant, as they were first written before Neil Armstrong set foot on the moon, before computers could fit onto a desk, and without the knowledge of many medical and biotechnology breakthroughs of the late 20th century. NAGB decided that the long-term trend assessment in science required technical studies of the required changes, so that valid comparisons between the updated assessment and the original assessment could still be made. To allow time to update the assessment and study the changes, the decision was made not to assess science in 2004.

According to NAGB's new policy, reading and mathematics would continue to be assessed by the long-term trend and main NAEP instruments, but science and writing would be assessed only in main NAEP. As a result, changes were needed to separate out the sets of questions (blocks) for science and writing, which had been intermixed with the reading and mathematics blocks in the long-term assessment instruments. New booklets consist only of reading or only of mathematics blocks. The changes provided an opportunity to bring other aspects of the assessment up to date. Considerable progress in testing theory has been made since the late 1960s, and the 2004 administration provided a platform to bring these improvements to the long-term trend assessments, in areas such as scoring and scaling. In addition, main NAEP assessments had begun providing accommodations to allow students with disabilities and students who were not fluent in English to participate. In 2004, it was possible to implement the modifications to the long-term trend assessments resulting in the assessment of a greater proportion of students using accommodations.

Any time changes are made in a long-term trend assessment, studies are required to ensure that the results can continue to be reported on the same trend line—that is, that they are validly comparable to earlier results. So analyses were needed to ensure that the 2004 results under the new design were comparable to the results from 1971 through 1999, under the design that existed earlier. Therefore, two assessments were conducted in 2004. The modified assessment used the new design, and the "bridge" assessment replicated the former design. Comparisons of the results can then detect any shifts in results due to changes in test design. This bridge assessment links the old assessments to the new one.

2004 Bridge Study

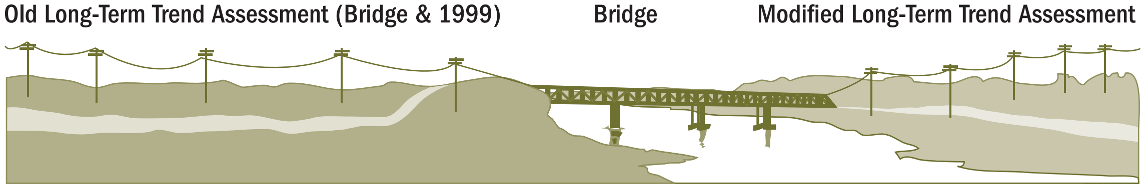
This section of the report presents a brief description of the 2004 bridge study, the modified assessment, and the long-term trend instruments. (More detailed information about the instruments and methodology is provided in appendix A.) The changes made for the modified 2004 assessment included replacing items containing outdated material, eliminating blocks of items for subjects no longer reported, replacing background questions, allowing accommodations for students who needed them, and changing some administrative procedures. For example, previous long-term trend assessments in mathematics included an audio portion that paced students, so they were always at the same place in the test booklet at the same time. The audiotape was eliminated in the modified design so that students could move at their own pace within each section. Another example is that students used to have the option of selecting “I don’t know” as a response to a multiple-choice item. That response was eliminated in the modified assessment. Also, in prior assessments, the student’s race/ethnicity was reported based on a test administrator’s classification of the student’s visual appearance. In 2004, both schools and students were asked to report each student’s race/ethnicity as part of the school and student questionnaires. Finally, the 2004 modified assessment provided for the inclusion of and accommodations for students with disabilities and English language learners.

The changes were intended to improve the validity of the results while continuing to maintain the integrity of the long-term trend. Thus, studies were needed to ensure that the modifications did not affect the interpretation of the results. In other words, it was important to assess whether any changes in scores were due to actual changes in student performance rather than changes in the assessments themselves that may have made them easier or harder.

The bridge study was conducted to ensure that the interpretation of the assessment results remains constant over time. A bridge study involves developing two assessments: one that replicates the assessment given in the previous assessment year using the same questions and administration procedures (a bridge assessment), and one that represents the new design (a modified assessment). In 2004, students were randomly assigned to take either the bridge assessment or the modified assessment. The bridge assessment replicated the instrument given in 1999 and used the same administration techniques. The modified assessment included the new items and features discussed previously. This modified assessment will provide the basis of comparison for all future assessments, and the bridge will link its results back to the results of the past 30 years (see figure 1-1). Further detail on this study is provided in appendix A.

This report will be the final report of new results acquired under the old design using the bridge assessment. The greater part of the report uses the results from the bridge assessment to maintain the trend lines from 1971 (in reading) and 1973 (in mathematics). Differences between the old and modified formats are discussed only in one chapter, chapter 5. Beginning in 2008, only the modified design will be used, and the results will be linked back to the previous assessments through the 2004 bridge study.

Figure 1-1. Comparison of the old and new long-term trend assessment



- “I don’t know” option
- Observed race/ethnicity
- Audio-paced portion
- No accommodations for SD/ELL students

- No “I don’t know” option
- School-reported race/ethnicity
- Self-paced throughout each section
- Accommodations permitted

Content of the Assessments

The content of the NAEP long-term trend reading and mathematics assessments has not changed since its beginning. The reading assessment contains a range of reading materials, from simple narrative passages to complex articles on specialized topics. The selections include stories, poems, essays, reports, and passages from textbooks, as well as a sample train schedule, telephone bill, and advertisements. Students' comprehension of these materials is assessed with both multiple-choice questions, for which students choose a response from a list, and constructed-response questions, for which students are asked to write a response.

The long-term trend mathematics assessment measures students' knowledge of basic facts, their ability to carry out numerical algorithms using paper and pencil, their knowledge of basic measurement formulas as they are applied in geometric settings, and their ability to apply mathematics to daily-living skills (such as those related to time and money). The computational focus of the long-term trend assessment provides a unique opportunity to measure how students perform in traditional procedural skills.

The Long-Term Trend Background Questionnaires

In addition to assessing students' progress in reading and mathematics, the NAEP long-term trend assessments include questions about students' home and school experiences that may be related to educational achievement. For example, students are asked about the courses they have taken, activities in their classrooms, the amount of time they spend on homework, and educationally relevant uses of their time out of school. Their responses to these questions provide an informative context for interpreting the assessment results.

In the previous long-term trend assessments, these background questions were intermixed with the assessment questions. For example, students would answer questions about a reading passage to assess their understanding of that passage, and then they would respond to background questions about their reading habits. In the modified design, these background questions were reduced in number and assembled together in a separate section that students completed after finishing the assessment.

The Student Sample

The NAEP long-term trend assessments measure the performance of students at three ages—9, 13, and 17. The NAEP assessments measure the achievement of students nationally and are not intended to provide a measure of individual student performance. A nationally representative sample of students is selected, and their results are generalized to the nation as a whole. Small percentages of students with disabilities (SD) and of English language learners (ELL) are excluded in each assessment year based on their schools' judgment that they cannot be meaningfully assessed. Formerly, NAEP did not permit students so identified to receive accommodations (such as extended time, assessment in small groups, or use of bilingual dictionaries). In 2004, accommodations were permitted on the modified assessment, and therefore fewer students were excluded. Specifically, approximately 14 to 19 percent of students across the three ages and two subjects were identified as SD/ELL in 2004, resulting in an exclusion rate of 7 to 8 percent, depending on the age and subject assessed, in the nonaccommodated format. When accommodations were permitted, the exclusion rates dropped to approximately 3 percent for mathematics and 4 to 5 percent for reading. (See appendix A for information regarding exclusion criteria and exclusion rates.)

This report contains results representing the performance of all in-school 9-, 13-, and 17-year-olds in the nation who are capable of being meaningfully assessed without accommodations, except for the results from the modified assessment shown in chapter 5. In addition, it describes the performance of groups of students, such as males and females, in each age group. In 2004, more than 11,000 students at each of the three ages were assessed in each subject area, including both public and private school students. To ensure that the sample was nationally representative, a sampling plan was created to randomly select schools and students to participate. This sampling plan targeted certain schools and students for participation in NAEP. The degree to which the students who actually participated in the assessment matched the target is a measure of the reliability of the results. In 2004, approximately 80 to 81 percent of the students originally selected for the assessment at age 9 were actually assessed, 76 to 77 percent of the students at age 13, and 55 to 57 percent at age 17. (See appendix A for more information on sampling procedures and appendix B for the percentages of students in various reporting groups who were assessed.)

Reporting the Trend Results

Students' performance on the long-term trend assessments is summarized on a 0–500 scale for each subject area. For each year in which the assessments were administered, achievement in a particular subject area is described for a group of students by their average scale score and the score at the selected percentiles. Trends in student achievement are determined by examining the average scale scores attained by students in the current assessment year or the score at the selected percentiles and comparing them to the same scores in other assessment years. While the score ranges in both subjects are identical, the scale was derived independently for each subject. Therefore, average scale scores between subjects cannot be compared.

In addition to reporting average scores, student performance is described in terms of the percentages of students attaining specific levels of performance. These performance levels correspond to five points on the reading and mathematics scales: 150, 200, 250, 300, and 350. For each subject area, the performance levels from lowest to highest are associated with increasingly advanced skills and knowledge (Allen, McClellan, and Stoeckel 2005, pp. 21–22). Examining the percentages of students in each year that attained each performance level provides additional insight into student achievement.

Because the results presented in this report are based on a nationally representative sample of students, they are considered estimates of all students' average performance (excluding students who cannot be meaningfully assessed). As such, the results are subject to a degree of uncertainty, which is reflected in the standard errors of the estimates. The standard errors for all of the scale scores and percentages presented in this report can be viewed using the NAEP Data Explorer found at http://nces.ed.gov/nationsreportcard/naepdata/. Statistical tests that take into account these standard errors were conducted to determine whether apparent changes or differences in the results are measurably different in a statistical sense. When the term "significant" is used, it does not imply a judgment about the absolute magnitude or educational relevance of changes and differences in student performance. Rather, it is used to indicate that the observed changes are not likely to be due to chance factors associated with sampling and measurement error. The differences described in this report have been determined to be statistically significant at the 0.05 level with appropriate adjustments for multiple comparisons. In the tables and charts in this report, the symbol (*) is used to indicate that a score or percentage is measurably different from another. (See appendix A for additional information on analysis procedures.)

The results presented here are meant to describe some aspects of the condition of education. They are best viewed as suggesting various ideas to be further examined in light of other data and in the context of the large research literature elaborating on the many factors contributing to educational achievement.

About This Report

This report describes trends in 9-, 13-, and 17-year-olds' achievement in reading and mathematics during the last three decades. Chapter 2 presents trends in terms of overall scale scores, percentiles, and percentages at selected performance levels for the nation. Chapter 3 examines trends in average scale scores for groups of students defined by gender, race/ethnicity, and the education level of the student's parents. Chapter 4 reports results from the NAEP long-term trend background questionnaires. In this chapter, students' school and home experiences, as shown in their responses to the background questions, are examined in relation to students' assessment scores. Chapter 5 explores the differences between the bridge assessment administered under the procedures used for earlier assessments and the modified assessment with the new design elements. The last chapter in this report provides sample items from the NAEP long-term trend assessments. For the first time, NCES is releasing items from the assessment, along with summary data that indicate how well students performed on these items. This report also contains three appendixes. Appendix A discusses technical procedures involved in collecting, analyzing, and reporting the assessment data, and appendix B is a data appendix showing the percentages of participating students in the bridge and modified samples by student groups. Appendix C provides a glossary of terms used in this report.

Additional information about the 2004 long-term trend assessments not included in this report, and other NAEP assessment reports and data, are available on the Internet at http://nces.ed.gov/nationsreportcard/. This site contains the data associated with all the figures in this report and further information on the technical features of the study. Additional data, such as the standard errors for each percentage, can also be found on this website.

Cautions in Interpreting the Long-Term Trend Results

The reader is cautioned against using the long-term trend results in this report to make simple causal inferences related to student performance, to the relative effectiveness of public and nonpublic schools, or to other educational variables discussed in this report. Simple cross-tabulations of a variable with measures of educational achievement, like the ones presented here, cannot constitute proof that differences in the variable cause differences in educational achievement. There are many possible reasons why the performance of one group of students will differ from that of another that are not discussed in this report. For example, group differences may be understood better by considering such factors as exposure to a rigorous curriculum, variations in course-taking patterns, and parental involvement.

A caution is also warranted for some small population group estimates. Smaller population groups may show increases or decreases across years in average scores; however, it is necessary to interpret such score changes with extreme caution. The effects of exclusion-rate changes for groups of students may be more marked for small groups than they are for the whole population. Another reason for caution is that the standard errors are often quite large around the score estimates for small groups, which in turn means the standard error around the gain is also large.

In addition, although in some figures trend lines for ages 9, 13, and 17 will appear in the same graphic, the reader is cautioned against making cohort comparisons. One cannot interpret the amount of growth between ages 9 and 13 from these figures by examining a 4-year time difference. Not all assessment years are four years apart, and the assessments were administered at different times of the year for the different ages. The relative merits of different types of comparisons are discussed in appendix A. Comparisons should be made within ages only.

National results are displayed using three reporting metrics: average scale scores, percentiles, and performance levels. Generally, all three metrics show improvements at age 9 in reading and mathematics.

Chapter 2
National Trends in Academic Achievement

For the past 35 years, NAEP's long-term trend assessments have documented trends in the academic achievement of America's students. Before the 2004 assessment, the last long-term trend assessment was conducted in 1999. This report examines the changes in students' performance in reading and mathematics over the past five years by comparing 2004 results to 1999 results and then provides a wider view of the overall trends in performance from the early 1970s through 2004.

This chapter presents the results by subject, first examining the trends in reading and then discussing mathematics results. There have been 11 administrations of the reading assessment since 1971 and 10 administrations of the mathematics assessment since 1973 for ages 9, 13, and 17. The next section describes the different ways of reporting results, and the remainder of this chapter describes the national trends in reading and mathematics.

How the Results Are Presented

Performance results in this chapter are reported in three ways: as average scale scores, as percentile scores, and as percentages of students reaching predetermined performance levels.

- ***Average scale scores.*** The average scale scores represent the performance of 9-, 13-, and 17-year-olds in reading or mathematics averaged across the nation. Student performance is summarized on a 0–500 scale for both reading and mathematics, where the different points on the scale represent what students know and can do at a given point in time. Although the results from both subjects are reported on the same scale, the results cannot be compared with one another, as they measure different content.

  Line graphs are provided to depict student performance on this scale across the years in both subject areas. The average scale score attained by students in each assessment year is indicated on the graph. The average scores for years prior to 2004 are highlighted with an asterisk (*) when the score is significantly higher or lower than the average score in 2004. (See appendix A for information on the statistical tests conducted.)

- ***Percentile scores.*** Going beyond average scores, useful information can be gained by examining trends of student scores falling at specified percentiles along the performance distribution. Percentiles indicate the percentage of students whose scores fell below a particular point on the NAEP scale. For example, 25 percent of assessed students' scores fell below the 25th percentile score; 75 percent fell below the 75th percentile score. This chapter provides such information by examining the scores of students at five distinct percentiles (10th, 25th, 50th, 75th, and 90th) of the score distribution in each year. Examining student performance at different percentiles on the 0–500 scale indicates whether or not the changes seen in the overall national average score results are reflected in the performance of lower-, middle-, and higher-performing students.

- ***Performance levels.*** More detailed information about what students know and can do in each subject area can be gained by examining their attainment of specific performance levels in each assessment year. For each of the subject area scales, performance levels were set at 50-point increments from 150 through 350. The five performance levels—150, 200, 250, 300, and 350—were then described in terms of the knowledge and skills likely to be demonstrated by students who reached each level. To develop these descriptions, assessment questions were identified that students at a particular performance level were more likely to answer successfully than students at lower levels. The descriptions of what students know and can do at each level are based on these sets of questions. This process of developing the performance-level descriptions is quite different from that used to develop achievement-level descriptions in the main NAEP reports as they are not set through a judgmental process. The levels for long-term trends were set arbitrarily and do not represent perfomance standards. Specific descriptions for each subject are presented later in this chapter along with the results. (The procedures for describing the performance levels are discussed in more detail in appendix A.)

National Trends in Reading Performance

National trends are shown through the average score, the percentile scores, and the percentage of students at or above each performance level. Although at first glance it may appear that this report provides the same results in three formats, these different reporting metrics actually provide different perspectives. The average score summarizes student performance in one measure. The percentiles examine performance at five different points, demonstrating whether any changes in average score are more likely due to changes in the scores of lower-performing students or higher-performing students. These percentiles are based on a normative measure, while the performance levels are based on a criterion measure. That is, the performance levels show trends in student performance at five benchmarks. These benchmarks are valid within all three age groups, permitting comparisons of the attainment of absolute performance levels over time. Cross-age comparisons can be supported, but readers are encouraged to focus more appropriately on within-grade comparisons.

Overall, the national trend in reading shows improvement across most reporting metrics at age 9 between 1999 and 2004 as well as between 1971 and 2004. Students at age 13 show no significant improvement in recent years, although most reporting metrics indicate that performance in 2004 was higher than in 1971. At age 17, no measurable differences in performance were found between 1971 and 2004 for any reporting metric.

Average Scores

This measure provides a summary account of student performance. Figure 2-1 displays the trend lines for each age, and further details are given below.

Nine-year-olds. The average reading score at age 9 was higher in 2004 than in any previous assessment year.

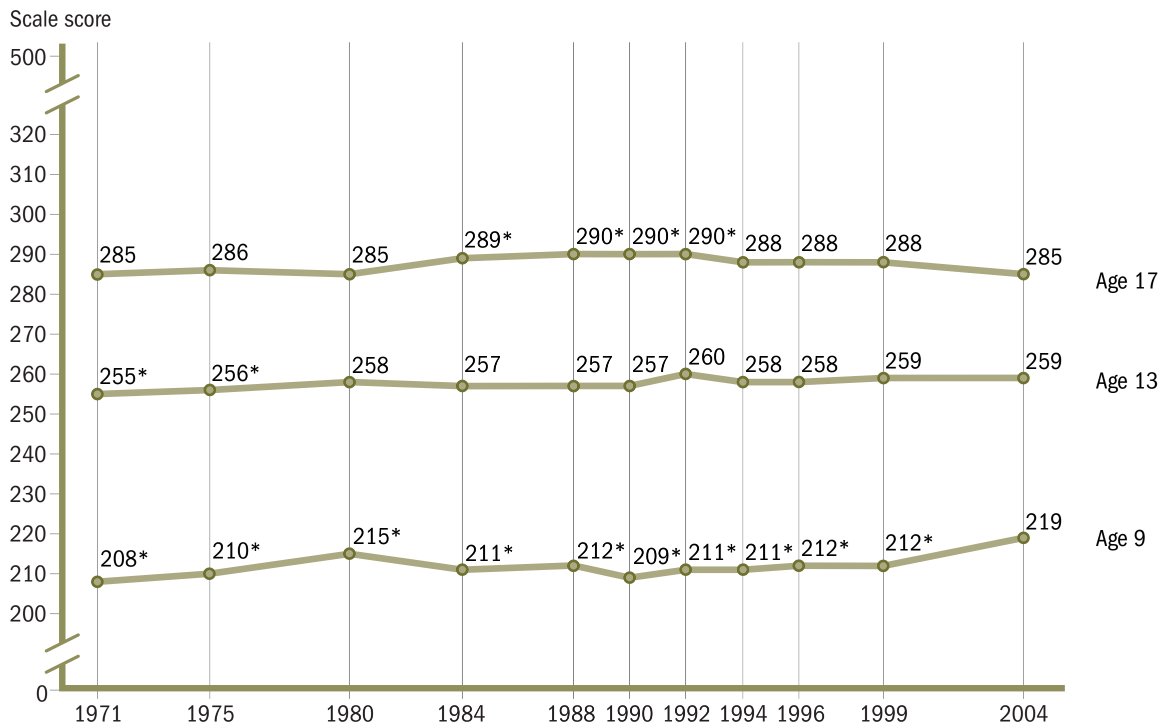
Thirteen-year-olds. The average score at age 13 was higher in 2004 than in 1971, but not measurably different from the average score in 1999.

Seventeen-year-olds. Between 1999 and 2004, average reading scores at age 17 showed no measurable changes. The average score in 2004 was similar to that in 1971.

How to interpret this graphic . . .

Graphics like these show the average scale score at each age for each year the assessment was given. Each score is plotted, and lines are drawn to connect the scores between the different years, creating trend lines. Examining the trend lines helps to determine whether scores appear to be increasing over time, or if there are any peaks or valleys in the 33-year trend. Statistically significant differences in scores between 2004 and previous years are marked with an asterisk. For example, figure 2-1 shows the trend lines of the average scores in reading for all three ages. The graphic shows that the average score at age 17 was about the same in 1971 as in 2004.

Figure 2-1. Trends in average reading scale scores for students ages 9, 13, and 17: 1971–2004



*Significantly different from 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading Assessments.

Percentile Scores

Examining the national trends at five percentiles shows whether changes seen in the national averages were sustained at every level of performance or were more likely to occur for students of specific ability levels. Figure 2-2 displays trends in reading scores for 9-, 13-, and 17-year-old students in the five percentile ranges. The results are discussed below for each age level.

Nine-year-olds. As seen in figure 2-2, only one significant increase was seen at the 90th percentile as compared to 2004. However, the score at the 50th percentile—the median—was higher in 2004 than in any other assessment year. The scores at the 10th, 25th, and 75th percentiles showed increases in performance between 1999 and 2004 and between 1971 and 2004.

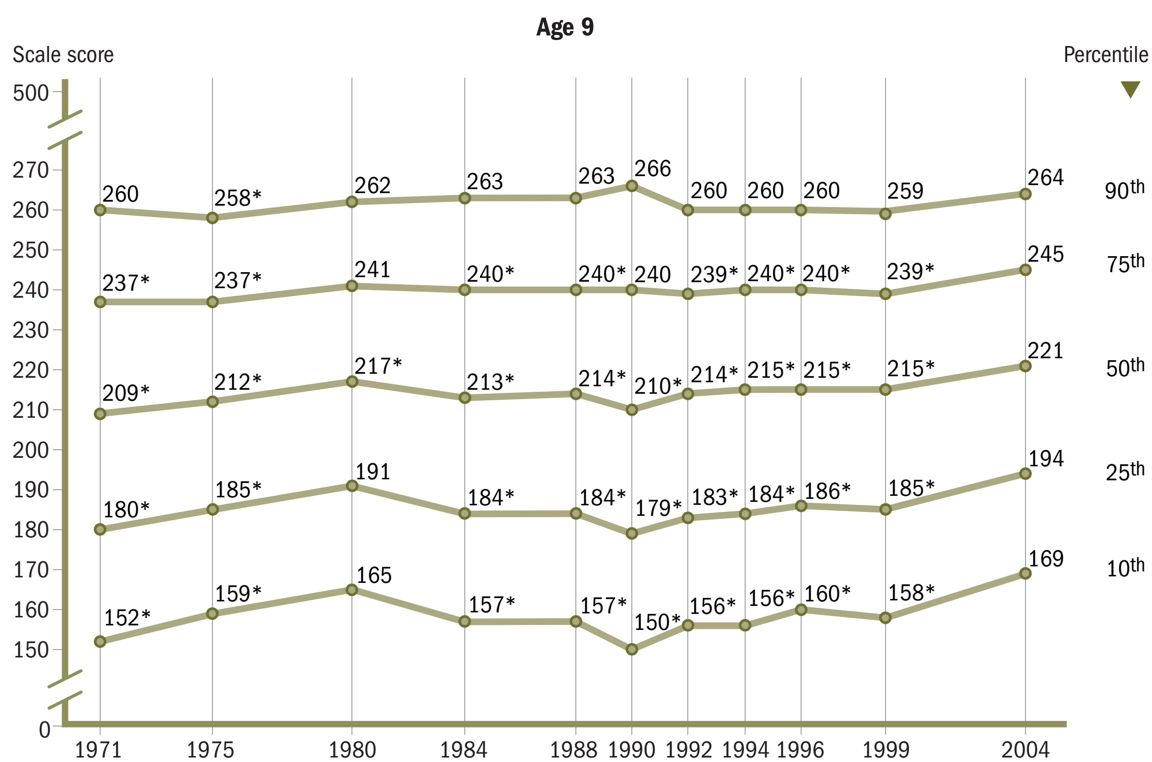
Thirteen-year-olds. The trends differ between upper and lower percentiles. The scores at the 10th, 25th, and 50th percentiles showed no measurable differences between 2004 and any previous assessment year. At the 75th and 90th percentiles, scores in 2004 were higher than in 1971, although no measurable differences were detected between the score in 2004 and that in 1999.

Seventeen-year-olds. Examining the scores at the five selected percentiles shows no measurable difference in the scores in 2004 compared to either 1971 or 1999.

How to interpret this graphic . . .

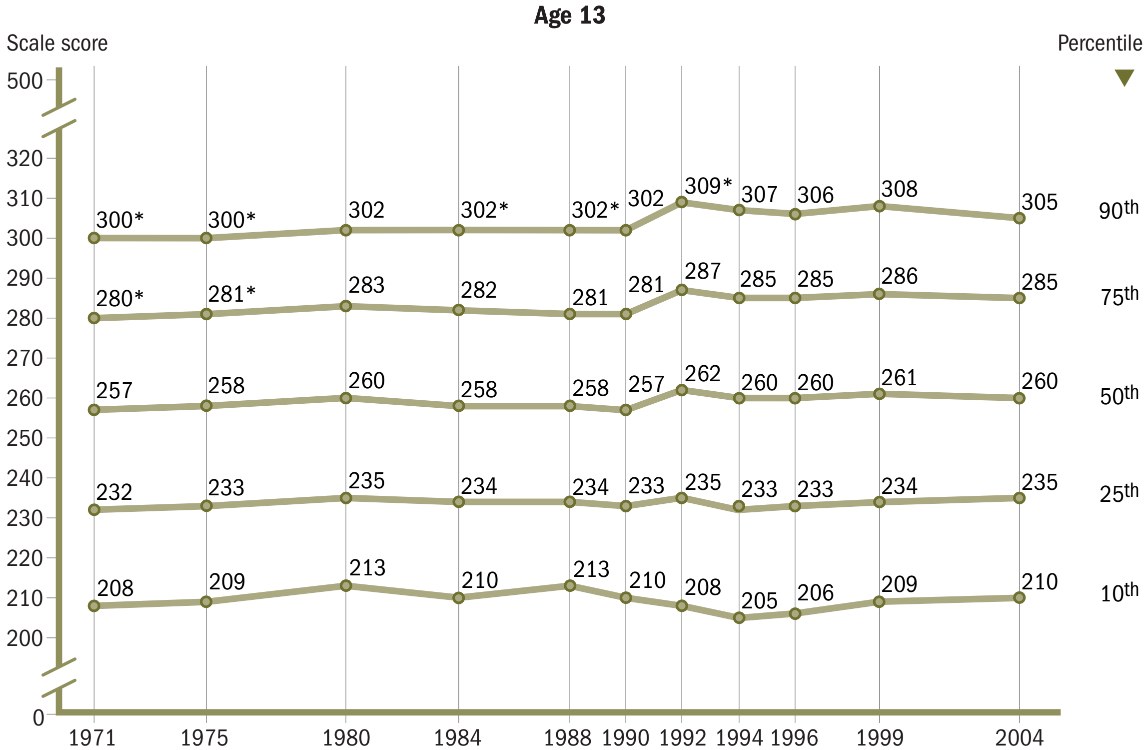
Graphics like figure 2-2 show the score at each percentile for five selected percentiles. For example, at age 9 in 2004, students at the 10th percentile scored 169 in reading, while students at the 90th percentile scored 264. Looking at the five trend lines together, it can be determined if more improvement took place at the upper end or at the lower end, or if the trend lines look the same at all five levels. For example, at age 9, the scores at the 10th, 25th, 50th, and 75th percentiles showed increases in performance between 1999 and 2004 and between 1971 and 2004.

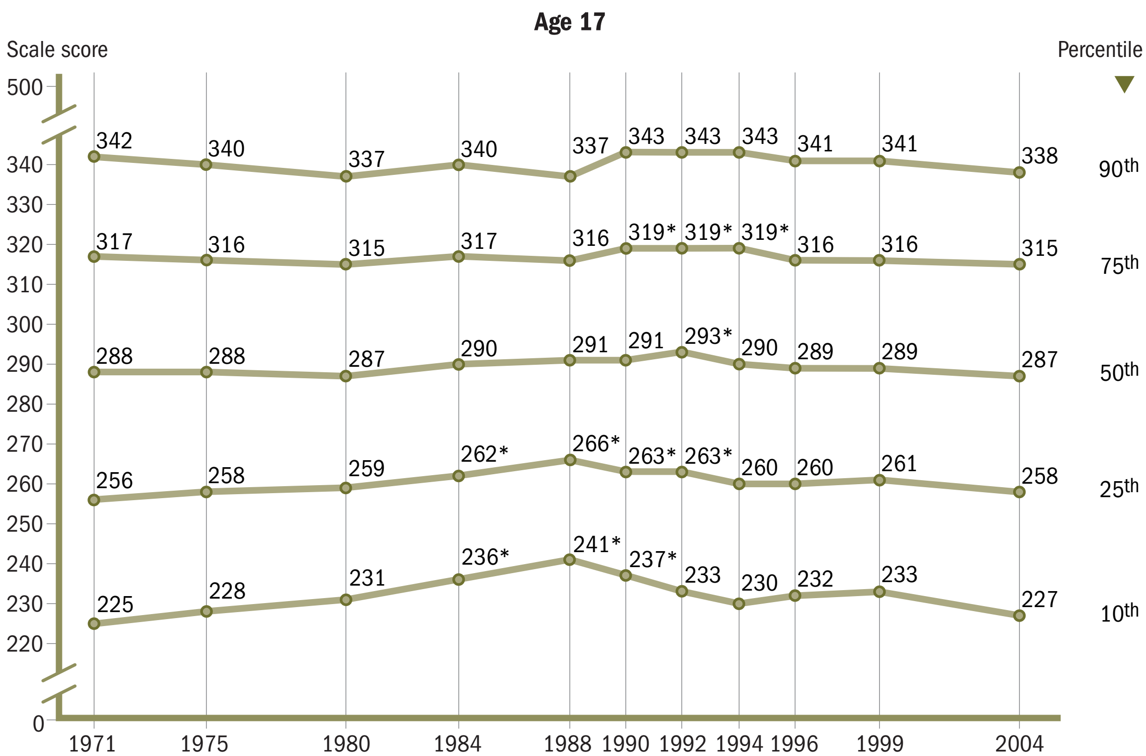
Figure 2-2. Trends in reading scale score at selected percentiles for students ages 9, 13, and 17: 1971–2004



See notes at end of figure.

Figure 2-2. Trends in reading scale score at selected percentiles for students ages 9, 13, and 17: 1971–2004—Continued





*Significantly different from 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading Assessments.

Performance Levels

This section reports trend results using the performance-level reporting metric, examining the percentage of students demonstrating particular levels of performance over the past three decades. Although one would expect these trends to follow closely the trends in average scores, it is instructive to examine changes in what students now seem to know and be able to do.

The skills and abilities demonstrated by students at each reading performance level are described below. The five performance levels are applicable at all three age groups, although the likelihood of attaining higher performance levels is directly related to a student's age, because older students have completed more education in both subject areas. For this reason, only three performance levels are discussed for each age: levels 150, 200, and 250 for age 9; levels 200, 250, and 300 for age 13; and levels 250, 300, and 350 for age 17. One might expect younger students to reach only the first performance levels, as they have not yet been taught the material in the higher performance levels, and it is expected that nearly 100 percent of older students will meet the lowest performance levels. Thus, the performance-level results displayed for each age are those that are most likely to show significant change across the assessment years. The levels not shown here are those that nearly all or almost no students attained at a particular age in each year.

Reading Performance-Level Descriptions

LEVEL 350: ***Learn from Specialized Reading Materials***
Readers at this level can extend and restructure the ideas presented in specialized and complex texts. Examples include scientific materials, literary essays, and historical documents. Readers are also able to understand the links between ideas, even when those links are not explicitly stated, and to make appropriate generalizations. Performance at this level suggests the ability to synthesize and learn from specialized reading materials.

LEVEL 300: ***Understand Complicated Information***
Readers at this level can understand complicated literary and informational passages, including material about topics they study at school. They can also analyze and integrate less familiar material about topics they study at school as well as provide reactions to and explanations of the text as a whole. Performance at this level suggests the ability to find, understand, summarize, and explain relatively complicated information.

LEVEL 250: ***Interrelate Ideas and Make Generalizations***
Readers at this level use intermediate skills and strategies to search for, locate, and organize the information they find in relatively lengthy passages and can recognize paraphrases of what they have read. They can also make inferences and reach generalizations about main ideas and author's purpose from passages dealing with literature, science, and social studies. Performance at this level suggests the ability to search for specific information, interrelate ideas, and make generalizations.

LEVEL 200: ***Demonstrate Partially Developed Skills and Understanding***
Readers at this level can locate and identify facts from simple informational paragraphs, stories, and news articles. In addition, they can combine ideas and make inferences based on short, uncomplicated passages. Performance at this level suggests the ability to understand specific or sequentially related information.

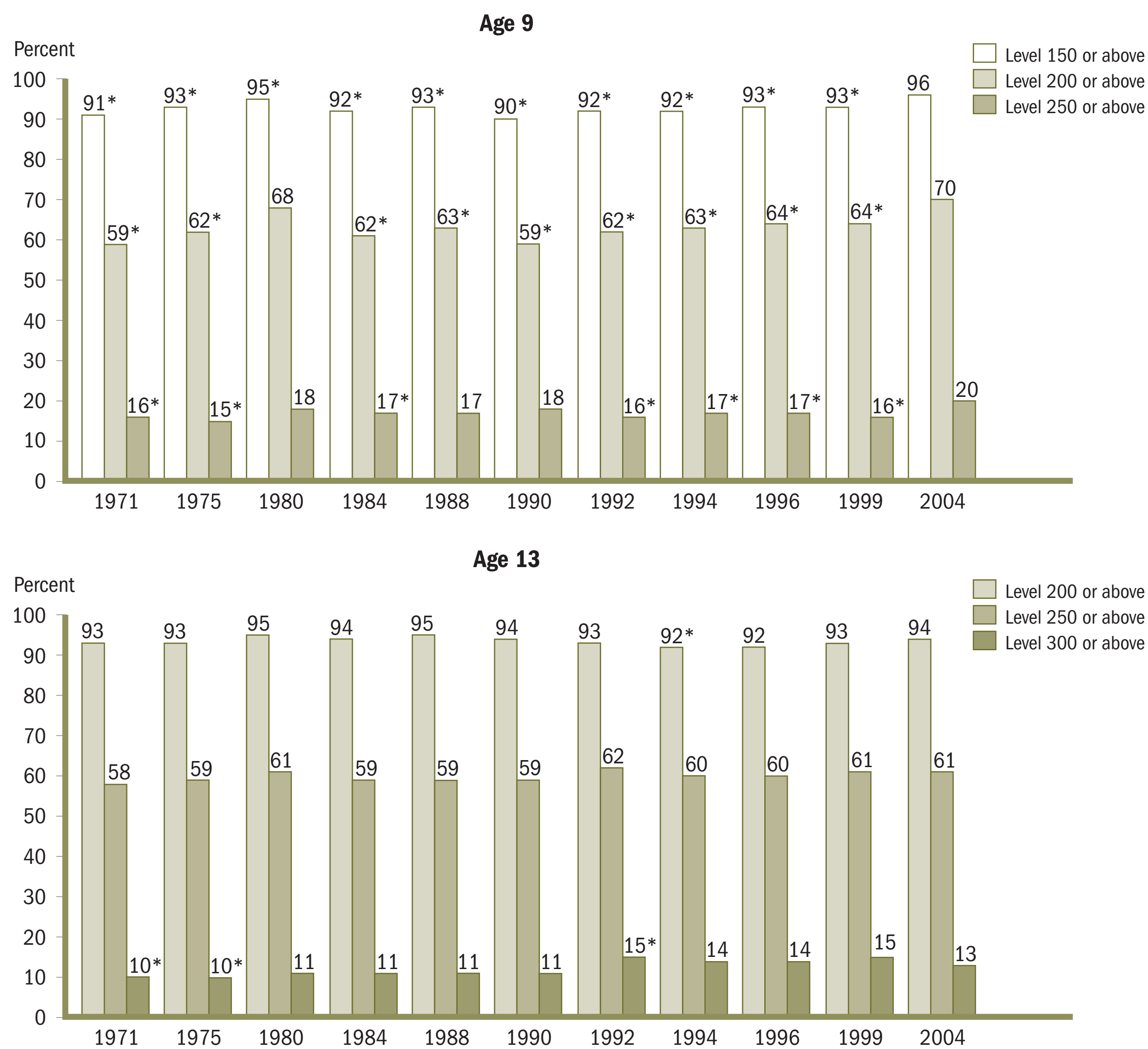
LEVEL 150: ***Carry Out Simple, Discrete Reading Tasks***
Readers at this level can follow brief written directions. They can also select words, phrases, or sentences to describe a simple picture and can interpret simple written clues to identify a common object. Performance at this level suggests the ability to carry out simple, discrete reading tasks.

Figure 2-3 shows the percentage of students reaching each performance level by age and assessment year. The following sections discuss the data for each age. It is important to keep in mind that the percentages reported for each level are cumulative. That is, the percentage shown for level 200 reflects the percentage of students who scored at 200 or above, so it also includes those who scored at 250, 300, or 350.

How to interpret this graphic . . .

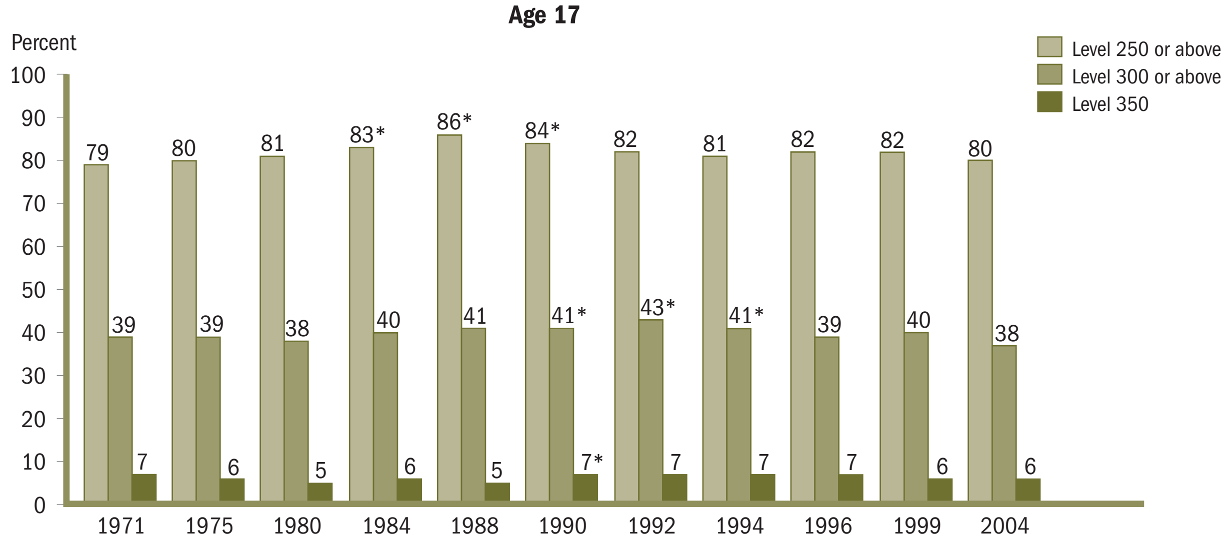
Bar charts are used to show the percentage of students who reach each performance level or above. For instance, figure 2-3 shows that 80 percent of 17-year-olds in 2004 reached level 250 or above, 38 percent reached level 300 or above, and 6 percent reached level 350. So, the 80 percent bar also includes those students in the 38 and 6 percent bars. Examining the height of the bars across years can help determine whether students are improving at the lower levels, higher levels, or both.

Figure 2-3. Trends in percentages at or above reading performance levels for students ages 9, 13, and 17: 1971–2004



See notes at end of figure.

Figure 2-3. Trends in percentages at or above reading performance levels for students ages 9, 13, and 17: 1971–2004—Continued



*Significantly different from 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading Assessments.

Nine-year-olds. Trends in the percentage of 9-year-olds scoring at or above reading performance levels 150, 200, and 250 are shown in the first panel of figure 2-3. In each assessment year, at least 90 percent of 9-year-olds performed the simple, discrete reading tasks described at level 150. In 2004, 96 percent of 9-year-olds reached level 150, a higher percentage than in any previous assessment year. The partially developed skills and understanding associated with level 200 were demonstrated by 70 percent of 9-year-olds in 2004. This number was higher than in every other assessment year with the exception of 1980, which showed no measurable difference from 2004. The ability to interrelate ideas and make generalizations (level 250) was demonstrated by 20 percent of 9-year-olds in 2004. This percentage was higher than both the more recent assessment year, 1999, and the first assessment year, 1971.

Thirteen-year-olds. The second panel of figure 2-3 displays trends in the percentage of 13-year-olds performing at or above reading performance levels 200, 250, and 300. In each assessment year, 92 percent or more of 13-year-old students performed at or above level 200, demonstrating at least partially developed skills and understanding. Ninety-four percent of students reached level 200 in 2004, which was not measurably different from the percentage in any other assessment year, except 1994, when the percentage fell to 92 percent. The ability to interrelate ideas and make generalizations (level 250) was demonstrated by 61 percent of 13-year-olds in 2004. Despite some apparent fluctuation, no measurable differences were found in the percentages of students at or above this level of performance across the assessment years. At level 300, students demonstrate the ability to understand complicated literary and informational passages. The percentage of students reaching level 300 in 2004 was higher than the percentage in 1971, mirroring the national trend for average score.

Seventeen-year-olds. Trends in the percentage of 17-year-olds scoring at or above reading performance levels 250 and 300 and at level 350 are shown in the last panel of figure 2-3. The ability to interrelate ideas and make generalizations (level 250) was demonstrated by 80 percent of 17-year-olds in 2004, which was not measurably different from 1999 or 1971. Performance at or above level 300—understanding complicated information—was demonstrated by 38 percent of 17-year-olds in 2004, which was not measurably different from the percentages in 1999 or 1971. Across all of the assessment years, only 5 to 7 percent of 17-year-olds demonstrated performance at level 350—the ability to learn from and synthesize specialized reading materials.

National Trends in Mathematics Performance

Overall, the national trend in mathematics shows improvement in performance at ages 9 and 13 in 2004 and few changes over the years at age 17. Note that the data from 1973 in figure 2-4 were extrapolated using a mean proportion correct, meaning that only average scores could be calculated. Results by percentile and performance levels are shown from 1978 through 2004. (See appendix A for further explanation of the extrapolated results.) The following sections examine the national results through the average score, the percentile scores, and the percentage of students at or above each performance level.

Average Scores

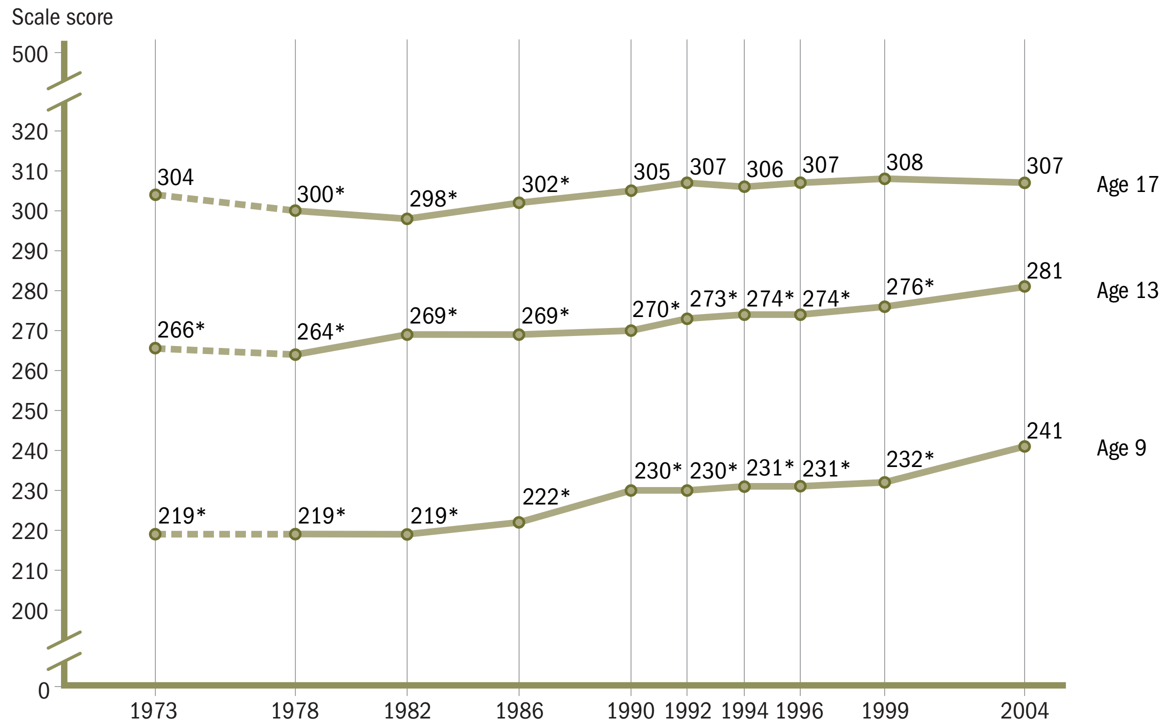
The first set of results shows trends in average scores in mathematics between 1973 and 2004. Figure 2-4 displays the trend lines for each age, and further details follow.

Nine-year-olds. At 241, the average score at age 9 was higher in 2004 than in any previous year—up 9 points from 1999 and 22 points from 1973.

Thirteen-year-olds. At age 13, the average score in 2004 was higher than in any other assessment year. The 5-point increase between 1999 and 2004 resulted in an average score in 2004 that was 15 points higher than the average score in 1973.

Seventeen-year-olds. The average score at age 17 was not measurably different from the average score in 1973 or 1999.

Figure 2-4. Trends in average mathematics scale scores for students ages 9, 13, and 17: 1973–2004



*Significantly different from 2004.
NOTE: Dashed lines represent extrapolated data.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1973–2004 Long-Term Trend Mathematics Assessments.

How to interpret this graphic . . .

Graphics like these show the average scale score at each age for each year the assessment was given. Each score is plotted, and lines are drawn to connect the scores between the subsequent assessment years, creating trend lines. Examining the trend lines helps to determine whether scores appear to be increasing over time, or if there are any peaks or valleys in the 31-year trend. Statistically significant differences in scores between 2004 and previous years are marked with an asterisk. For example, figure 2-4 shows that at age 17, the average score in 2004 was not measurably different from the scores shown in 1990 through 1999, but it was higher than the scores in 1978, 1982, and 1986.

Percentile Scores

This section examines the national trends at five percentiles to indicate whether changes seen in the national averages are sustained at every level of performance or occurred for students of specific ability levels. Figure 2-5 displays trends in mathematics scores for 9-, 13-, and 17-year-old students in the five percentile levels. Note that these trends are not available back to 1973 because only the overall average scores could be extrapolated for 1973.

Nine-year-olds. The trend lines shown in figure 2-5 appear very similar to one another at age 9. Nine-year-olds showed higher scores at each of the five selected percentiles in 2004 than in any other assessment year. Between the first year and the most recent assessment year—1978 and 2004—scores increased 26 points at the 10th percentile, 26 points[1] at the 25th percentile, 23 points at the 50th percentile, 21 points[1] at the 75th percentile, and 18 points at the 90th percentile.

Thirteen-year-olds. At age 13, the score at each of the five percentile levels was higher in 2004 than in every previous assessment year, with the exception of the 10th percentile. The score at the 10th percentile in 2004 was higher than in 1978, but showed no measurable gain between 1999 and 2004.

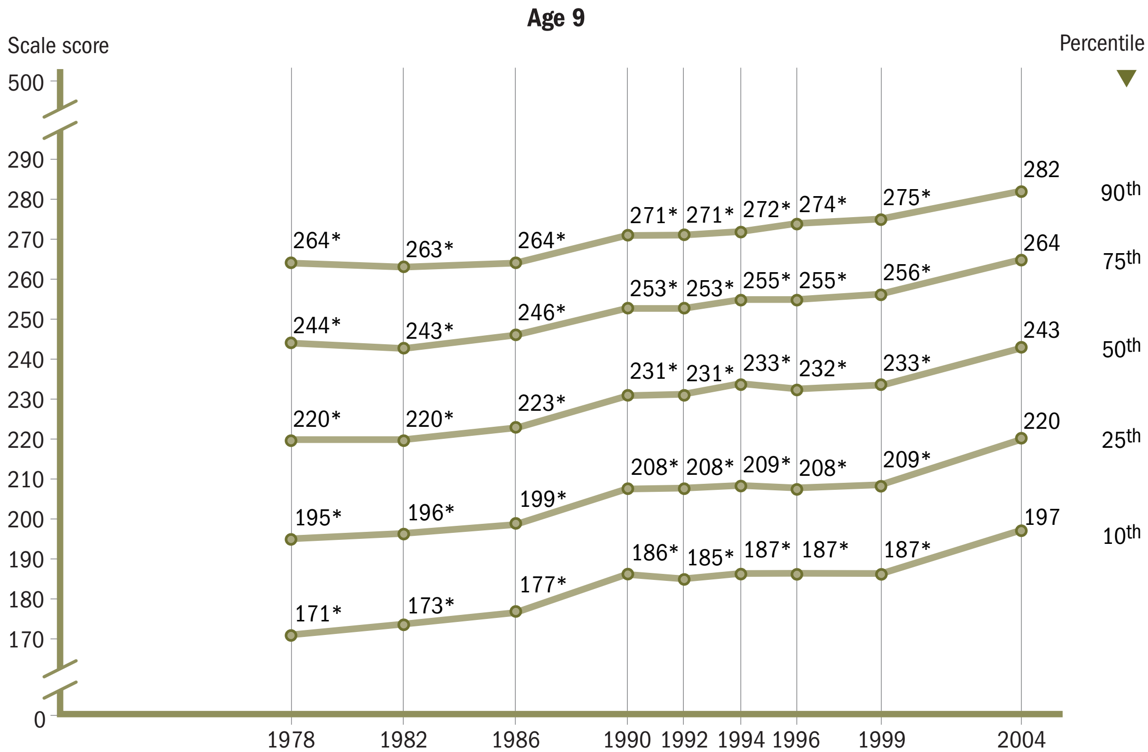
Seventeen-year-olds. Scores for 17-year-olds at the 10th, 25th, and 50th percentiles were higher in 2004 than in 1978. The scores at the 75th and 90th percentiles were not measurably different in 2004 compared to 1999 or 1978.

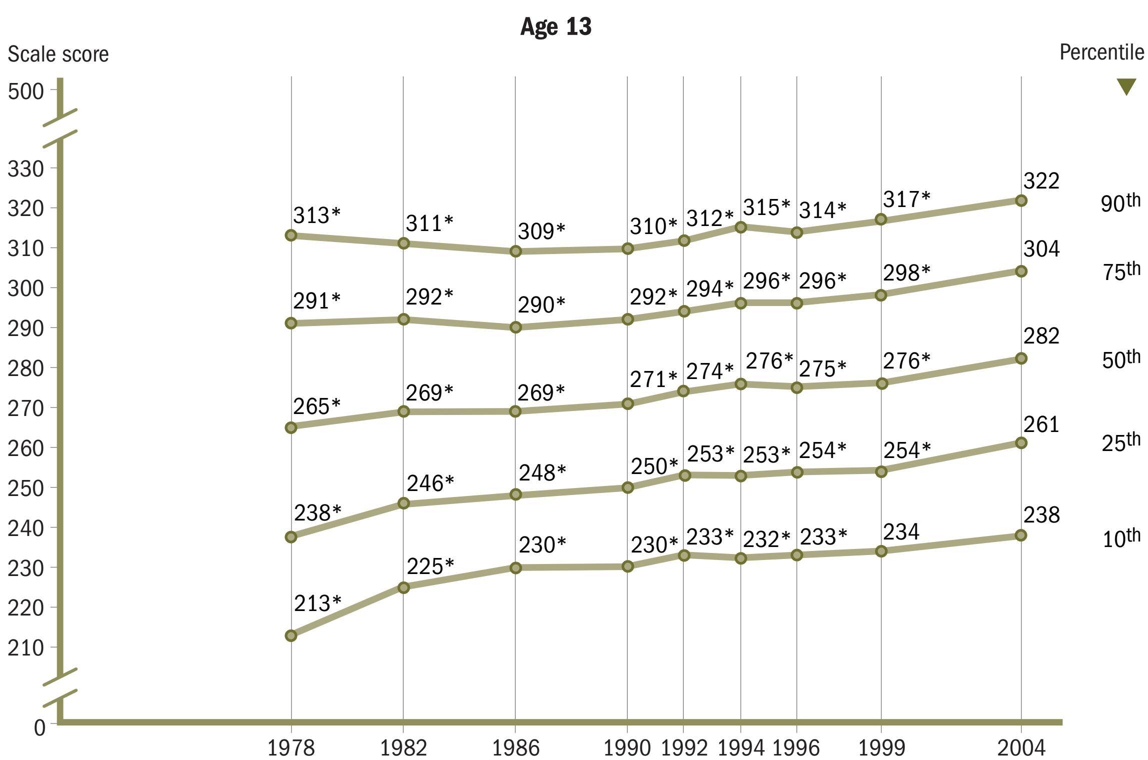
How to interpret this graphic . . .

Graphics like figure 2-5 show the score at each percentile for five selected percentiles. For example, at age 9 in 2004, students at the 10th percentile scored 197 in mathematics, while students at the 90th percentile scored 282. Both of these scores are higher than the scores in any previous assessment year. Looking at the five trend lines together, it can be determined if more improvement took place at the upper end or at the lower end, or if the trend lines look the same at all five levels.

[1] Detail may not sum to totals because of rounding. Differences between scores are calculated using unrounded values. In this instance, the result of the subtraction differs from what would be obtained by subtracting the rounded values shown in the accompanying figure.

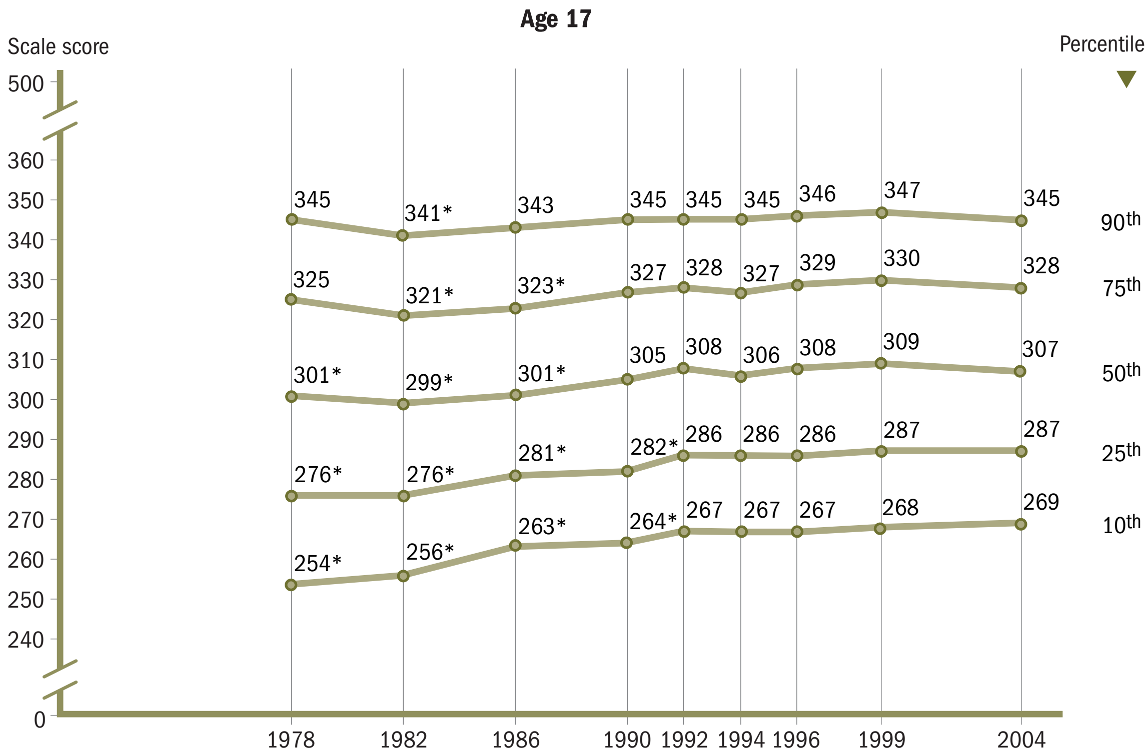
Figure 2-5. Trends in mathematics scale score at selected percentiles for students ages 9, 13, and 17: 1978–2004





See notes at end of figure.

Figure 2-5. Trends in mathematics scale score at selected percentiles for students ages 9, 13, and 17: 1978–2004—Continued



*Significantly different from 2004.

NOTE: Mathematics scores at selected percentiles are not available in 1973 because only the overall average scores were extrapolated for this year.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1978–2004 Long-Term Trend Mathematics Assessments.

Performance Levels

The skills and abilities demonstrated by students at each mathematics performance level are described below. As in reading, the five performance levels are applicable at all three ages, but only three performance levels are discussed for each age: levels 150, 200, and 250 for age 9; levels 200, 250, and 300 for age 13; and levels 250, 300, and 350 for age 17. These performance levels are the ones most likely to show significant change within an age across the assessment years and do not include the levels that nearly all or almost no students attained at a particular age in each year. Again, these trends are only available from 1978, because only the overall average scores could be extrapolated for 1973.

Mathematics Performance-Level Descriptions

LEVEL 350: *Multistep Problem Solving and Algebra*

Students at this level can apply a range of reasoning skills to solve multistep problems. They can solve routine problems involving fractions and percents, recognize properties of basic geometric figures, and work with exponents and square roots. They can solve a variety of two-step problems using variables, identify equivalent algebraic expressions, and solve linear equations and inequalities. They are developing an understanding of functions and coordinate systems.

LEVEL 300: *Moderately Complex Procedures and Reasoning*

Students at this level are developing an understanding of number systems. They can compute with decimals, simple fractions, and commonly encountered percents. They can identify geometric figures, measure lengths and angles, and calculate areas of rectangles. These students are also able to interpret simple inequalities, evaluate formulas, and solve simple linear equations. They can find averages, make decisions based on information drawn from graphs, and use logical reasoning to solve problems. They are developing the skills to operate with signed numbers, exponents, and square roots.

LEVEL 250: *Numerical Operations and Beginning Problem Solving*

Students at this level have an initial understanding of the four basic operations. They are able to apply whole number addition and subtraction skills to one-step word problems and money situations. In multiplication, they can find the product of a two-digit and a one-digit number. They can also compare information from graphs and charts and are developing an ability to analyze simple logical relations.

LEVEL 200: *Beginning Skills and Understandings*

Students at this level have considerable understanding of two-digit numbers. They can add two-digit numbers but are still developing an ability to regroup in subtraction. They know some basic multiplication and division facts, recognize relations among coins, can read information from charts and graphs, and use simple measurement instruments. They are developing some reasoning skills.

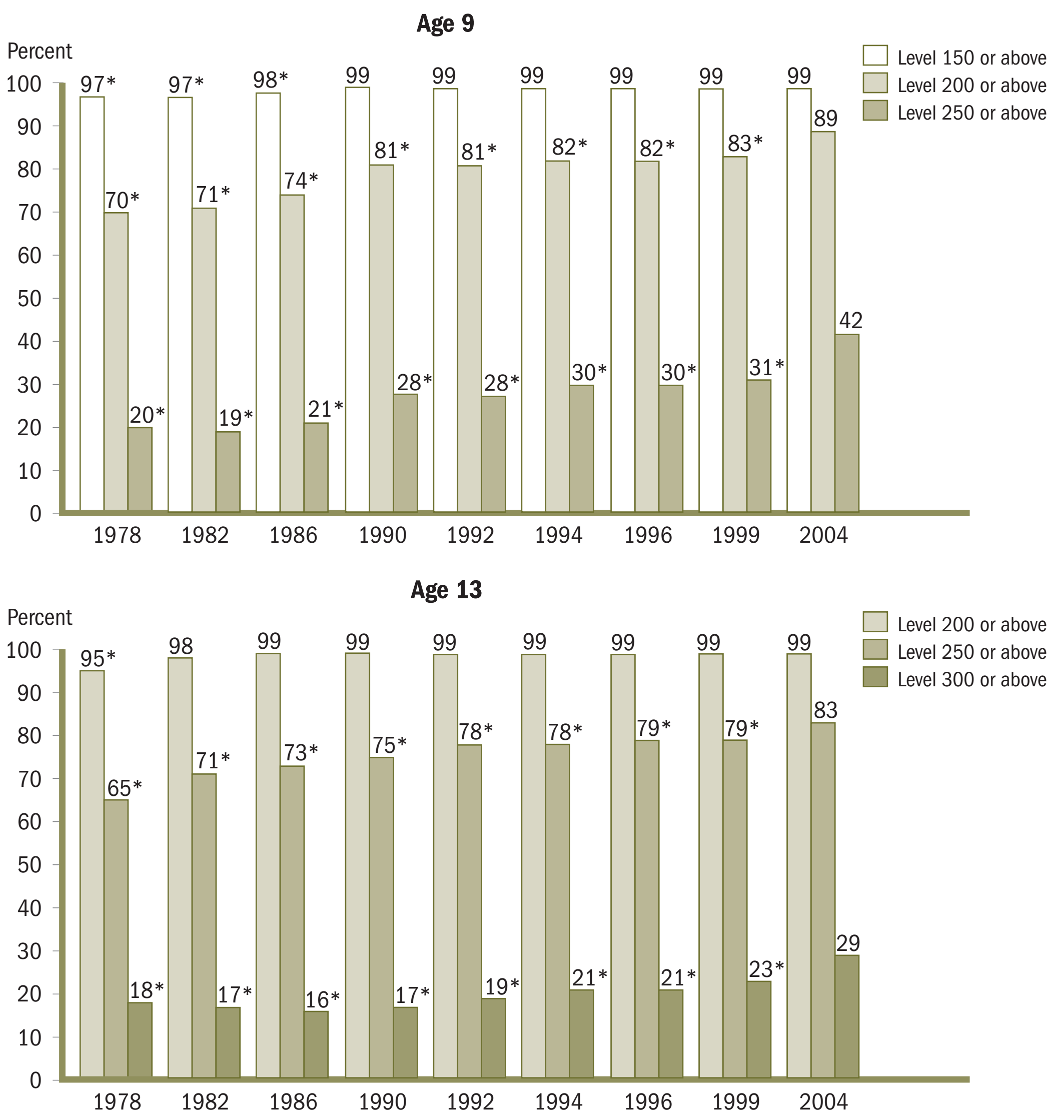
LEVEL 150: *Simple Arithmetic Facts*

Students at this level know some basic addition and subtraction facts, and most can add two-digit numbers without regrouping. They recognize simple situations in which addition and subtraction apply. They also are developing rudimentary classification skills.

Figure 2-6 shows the percentage of students reaching each performance level by age and assessment year. The following sections discuss the data for each age group.

Nine-year-olds. Trends in the percentage of 9-year-olds attaining mathematics performance levels 150, 200, and 250 are displayed in the upper panel of figure 2-6. In each assessment year, nearly all 9-year-olds (at least 97 percent) demonstrated understanding of simple arithmetic facts associated with level 150. In 2004, this percentage was 99, measurably higher by one percentage point than in 1986, and higher by three points[2] than in 1978, with no measurable change since 1990. The beginning skills and understandings characteristic of level 200 was demonstrated by 89 percent of 9-year-olds in 2004, higher than in any other assessment year. In the 2004 assessment, 42 percent of 9-year-olds performed the numerical operations and beginning problem solving associated with level 250, a higher percentage than in any other assessment year. There was an increase of 11 percentage points for 9-year-olds at this level between 1999 and 2004.

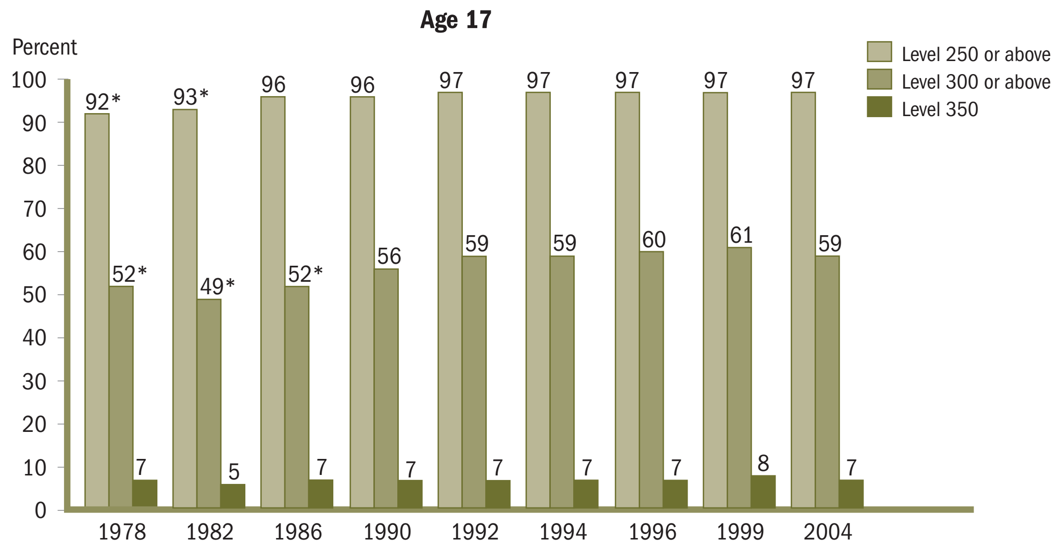
Figure 2-6. Trends in percentages at or above mathematics performance levels for students ages 9, 13, and 17: 1978–2004



See notes at end of figure.

[2] Detail may not sum to totals because of rounding. Differences between percentages are calculated using unrounded values. In this instance, the result of the subtraction differs from what would be obtained by subtracting the rounded values shown in the accompanying figure.

Figure 2-6. Trends in percentages at or above mathematics performance levels for students ages 9, 13, and 17: 1978–2004—Continued



*Significantly different from 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1978–2004 Long-Term Trend Mathematics Assessments.

Thirteen-year-olds. The percentage of 13-year-old students scoring at or above mathematics performance levels 200, 250, and 300 across the assessment years are displayed in the middle panel of figure 2-6. Since 1986, 99 percent of 13-year-olds demonstrated the beginning skills and understandings associated with level 200. In 2004, 83 percent scored at or above level 250, demonstrating the ability to perform numerical operations and beginning problem solving. Overall gains are also evident at level 300, where students performed moderately complex procedures and reasoning. The percentage of students who scored at or above this level increased from 18 percent in 1978, to 23 percent in 1999, and to 29 percent in 2004.

Seventeen-year-olds. Trends in the percentage of 17-year-olds scoring at or above mathematics performance levels 250, 300, and 350 are displayed in the last panel of figure 2-6. Since 1986, at least 96 percent of 17-year-olds have performed at or above level 250, demonstrating the ability to perform numerical operations and beginning problem solving. The percentage of 17-year-olds who performed moderately complex procedures and reasoning (level 300) showed no measurable change from 1990 to 2004, but has increased by 7 percentage points from 1978. No measurable change between 2004 and all the previous assessment years can be detected at 350, the highest performance level, in which students applied a range of reasoning skills to solve multistep problems. Across the assessment years, between 5 and 8 percent of students performed at this level.

Summary

The results presented in this chapter give an overall view of national trends in reading and mathematics achievement. Average scores for the nation, scores for students in five different ranges of the performance distribution, and attainment of specific performance levels were discussed. Looking across the 33 years, upward trends are most noticeable at age 9 in both reading and mathematics. Also of interest is the increase in performance at age 13 in mathematics.

The following figures provide an overview of the major findings presented in this chapter by comparing students' performance in 2004 to that of their counterparts in the first year data were collected. In addition, 2004 and 1999 results are compared, providing a summary of trends over the last five years.

Arrows pointing upward (↑) indicate improvement, and horizontal arrows (→) indicate no measurable change in performance. For example, the first line of the display in figure 2-7 indicates that the national average reading score for 9-year-olds was higher in 2004 than it was in 1971 or 1999.

Figure 2-7. Summary of trends in reading and mathematics average scale scores for students ages 9, 13, and 17: 1971–2004

Reading

↑ 9-year-olds' average scale scores since 1971 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1971 (→ since 1999)
→ 17-year-olds' average scale scores since 1971 (→ since 1999)

Mathematics

↑ 9-year-olds' average scale scores since 1973 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1973 (↑ since 1999)
→ 17-year-olds' average scale scores since 1973 (→ since 1999)

↑ Significantly higher in 2004.
→ Indicates no significant difference between earlier year and 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading and Mathematics Assessments.

Figure 2-8. Summary of trends in reading and mathematics scale score percentiles for students ages 9, 13, and 17: 1971–2004

Reading

9-year-olds

↑ 10th percentile since 1971 (↑ since 1999)
↑ 25th percentile since 1971 (↑ since 1999)
↑ 50th percentile since 1971 (↑ since 1999)
↑ 75th percentile since 1971 (↑ since 1999)
→ 90th percentile since 1971 (→ since 1999)

13-year-olds

→ 10th percentile since 1971 (→ since 1999)
→ 25th percentile since 1971 (→ since 1999)
→ 50th percentile since 1971 (→ since 1999)
↑ 75th percentile since 1971 (→ since 1999)
↑ 90th percentile since 1971 (→ since 1999)

17-year-olds

→ 10th percentile since 1971 (→ since 1999)
→ 25th percentile since 1971 (→ since 1999)
→ 50th percentile since 1971 (→ since 1999)
→ 75th percentile since 1971 (→ since 1999)
→ 90th percentile since 1971 (→ since 1999)

Mathematics

9-year-olds

↑ 10th percentile since 1978 (↑ since 1999)
↑ 25th percentile since 1978 (↑ since 1999)
↑ 50th percentile since 1978 (↑ since 1999)
↑ 75th percentile since 1978 (↑ since 1999)
↑ 90th percentile since 1978 (↑ since 1999)

13-year-olds

↑ 10th percentile since 1978 (→ since 1999)
↑ 25th percentile since 1978 (↑ since 1999)
↑ 50th percentile since 1978 (↑ since 1999)
↑ 75th percentile since 1978 (↑ since 1999)
↑ 90th percentile since 1978 (↑ since 1999)

17-year-olds

↑ 10th percentile since 1978 (→ since 1999)
↑ 25th percentile since 1978 (→ since 1999)
↑ 50th percentile since 1978 (→ since 1999)
→ 75th percentile since 1978 (→ since 1999)
→ 90th percentile since 1978 (→ since 1999)

↑ Significantly higher in 2004.
→ Indicates no significant difference between earlier year and 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading and Mathematics Assessments.

Figure 2-9. Summary of trends in reading and mathematics percentages at or above performance levels for students ages 9, 13, and 17: 1971–2004

Reading	
9-year-olds	
↑ Level 150 (simple, discrete reading tasks) since 1971	(↑ since 1999)
↑ Level 200 (partially developed skills and understanding) since 1971	(↑ since 1999)
↑ Level 250 (interrelate ideas and make generalizations) since 1971	(↑ since 1999)
13-year-olds	
→ Level 200 (partially developed skills and understanding) since 1971	(→ since 1999)
→ Level 250 (interrelate ideas and make generalizations) since 1971	(→ since 1999)
↑ Level 300 (understand complicated information) since 1971	(→ since 1999)
17-year-olds	
→ Level 250 (interrelate ideas and make generalizations) since 1971	(→ since 1999)
→ Level 300 (understand complicated information) since 1971	(→ since 1999)
→ Level 350 (learn from specialized reading materials) since 1971	(→ since 1999)
Mathematics	
9-year-olds	
↑ Level 150 (simple arithmetic facts) since 1978	(→ since 1999)
↑ Level 200 (beginning skills and understandings) since 1978	(↑ since 1999)
↑ Level 250 (numerical operations and beginning problem solving) since 1978	(↑ since 1999)
13-year-olds	
↑ Level 200 (beginning skills and understandings) since 1978	(→ since 1999)
↑ Level 250 (numerical operations and beginning problem solving) since 1978	(↑ since 1999)
↑ Level 300 (moderately complex procedures and reasoning) since 1978	(↑ since 1999)
17-year-olds	
↑ Level 250 (numerical operations and beginning problem solving) since 1978	(→ since 1999)
↑ Level 300 (moderately complex procedures and reasoning) since 1978	(→ since 1999)
→ Level 350 (multistep problem solving and algebra) since 1978	(→ since 1999)

↑ Significantly higher in 2004.
→ Indicates no significant difference between earlier year and 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading and Mathematics Assessments.

THIS PAGE INTENTIONALLY LEFT BLANK.

These long-term trend results show some progress since 1999 towards reducing the White-Black and White-Hispanic score gaps at age 9.

Chapter 3
Trends in Academic Achievement Among Student Groups

A key goal of the NAEP long-term trend assessment is to monitor the progress of various groups of students to determine whether any change in national scores is occurring across all student groups or is limited to a particular group. It is important to examine the performance gaps between student groups and any changes in these gaps over time as well as the overall achievement of all students. The assessment results presented in this chapter provide one source of information useful in monitoring progress of student achievement in this country.

Some of the student groups measured by this assessment are defined by gender, race/ethnicity, parental education level, and type of school (public or nonpublic). However, this report provides data only on those groups with sufficient sample size to produce reliable results. For instance, only White, Black, and Hispanic racial/ethnic groups are described here, as the sample sizes for Asian/Pacific Islander and American Indian/Alaska Native students were too small to provide reliable estimates. See tables B-1 and B-2 in the appendix for information on the percentage distribution of participating students by racial/ethnic group.

The NAEP long-term trend assessment has examined public and nonpublic school students' performance separately since 1980 in reading and 1978 in mathematics. However, in this report, results for nonpublic schools are neither displayed nor discussed because the participation rates for nonpublic schools were too low to produce valid and reliable results (see the School and the Student Sampling sections of appendix A for more detail). NAEP is preparing a report on the performance of nonpublic (private) school students with trend results from the main NAEP assessments (Perie, Vanneman, and Goldstein [forthcoming]).

The performance of students in each of these student groups is described in this chapter. First, descriptions of the student groups are given, and then the results for reading are displayed, followed by mathematics. Line graphs are used to display the average reading and mathematics scale scores attained by students in each group across the assessment years. Where appropriate, gaps between the student groups are also presented. For instance, the charts highlight any differences in scores of male and female students as well as the average score gaps between Black and White students and Hispanic and White students. The average score of each student group and age (9-, 13-, and 17-year-olds) is placed on a 0–500 scale in both subject areas to provide a numeric summary of students' performance.

Description of Student Groups

Results from the long-term trend assessment are presented in this chapter for gender, race/ethnicity, and highest level of parents' education. The following sections describe how the data were collected on each of the student groups discussed in this chapter, and give relevant background information about group membership and achievement.

Gender

In years past, gender differences have received considerable attention. Male students traditionally scored higher on average than female students in mathematics and science, while females scored higher on average than males in reading and writing (Baker and Jones 1993; Bauer, Park, and Sullivan 1998; Freeman 2004; Mullis et al. 1998). Now, gender differences are less pronounced in the United States than in other countries. For instance, in a recent international assessment of 15-year-olds, no differences were found in the United States between male and female students' scores in mathematics, but there were gender gaps in reading in which females scored higher than males in the United States (Lemke et al. 2002). So, although much of the nation's attention has shifted to the performance gaps between different racial/ethnic groups, it is important to continue to examine the trends in the male-female score gap.

The roster of sampled students from each participating school identifies the students as either male or female. These data are used to examine trends in male and female students' average reading and mathematics scores, which are presented in this chapter.

Race/Ethnicity

Previous main NAEP reports have shown a consistent finding of White and Asian students outperforming their Black and Hispanic peers. (See, for example, Braswell et al. 2005; Donahue, Daane, and Jin 2005.) Reducing the performance gaps between racial/ethnic groups is a primary goal of the recent federal legislation in education (NCLB 2002).

Although data are collected on five mutually exclusive racial/ethnic groups, the performance of only three groups is reported in this section—White, Black, and Hispanic students. The other racial/ethnic groups—Asian/Pacific Islander and American Indian/Alaska Native—are not reported, as the samples collected were of insufficient size to analyze and report separately. Data for Hispanic students were not available in 1971, so the trend in reading scores for this group runs from 1975 through 2004.

Relatively small numeric changes in scores are more likely to be statistically significant for White students than for Black or Hispanic students, because the weighted samples of White students tended to be larger than weighted samples for other racial/ethnic groups, with a corresponding lower margin of error. That is, the standard errors associated with larger groups, such as White students, are smaller than the standard errors associated with smaller groups, such as Hispanic students. Therefore, a similar difference between years in scale scores is more likely to be statistically significant for the larger group than for the smaller group.

Parents' Highest Level of Education

Parental education may influence student performance in school in a variety of ways. Earlier NAEP reports have shown that across all ages and subject areas, students who reported higher parental education levels tended to have higher assessment scores, on average. (See, for example, Braswell et al. 2005; Donahue, Daane, and Jin 2005.)

In the long-term trend assessment background questionnaires, students at all three ages are asked to identify the highest level of education attained by their parents. The student indicates how far each parent went in school, choosing from the following categories: did not finish high school, graduated from high school, went to another school after high school, graduated from college, and I don't know. The highest education level of either parent is used in these analyses. Data go back to 1978 in mathematics and 1980 in reading. In 1971 and 1975, students were asked to choose their parents' highest education level from among fewer categories. For purposes of this section, only the results from 1978 forward will be discussed so that "some education after high school" and "college graduate" can be analyzed separately. It should be noted that 9-year-olds' reports of their parents' education level may not be as reliable as those of older students and are therefore not reported.

Trends in Reading Scores by Student Groups

This section presents the results of the long-term trend reading assessment for each of the four types of groups. For gender and race/ethnicity, first the results are presented for each student group, and then the score gaps between the groups are examined.

Trends in Reading Scores by Gender

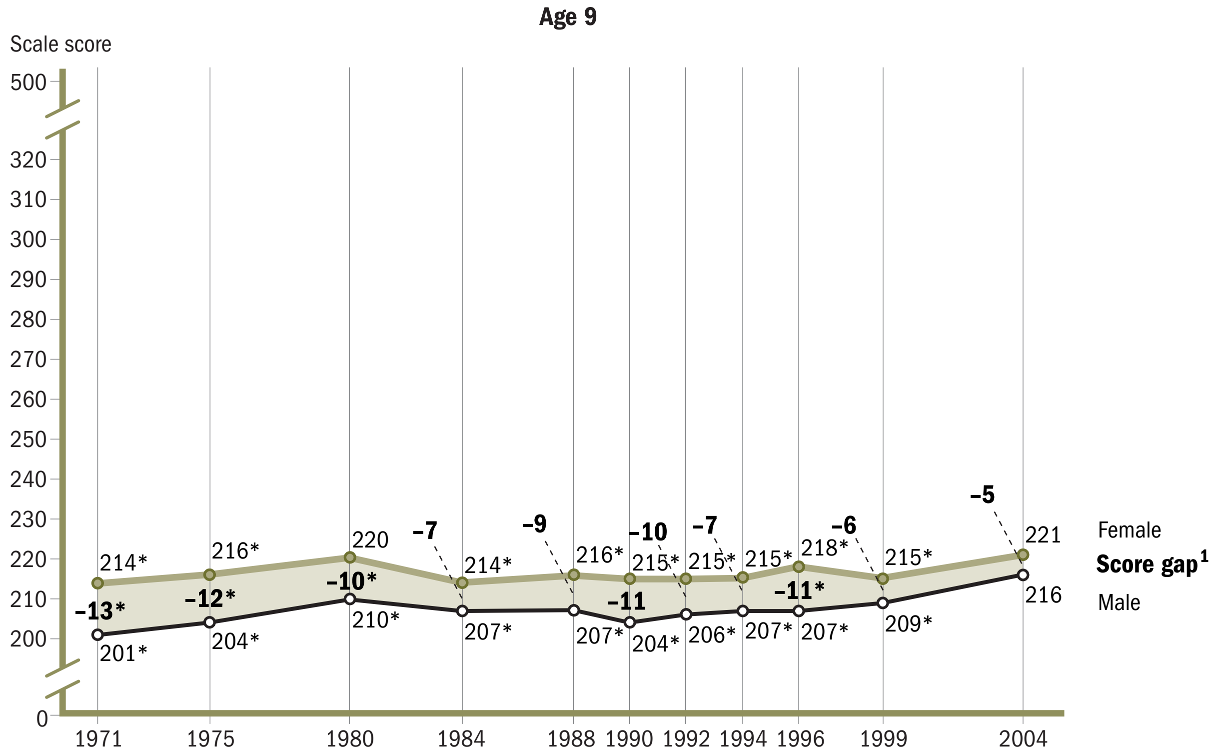
Trends in reading scores for both male and female students are shown in figure 3-1. Among male students, 9-year-olds had a higher average score in 2004 than in any previous assessment year. Thirteen-year-old males' average reading score in 2004 was higher than the scores in 1971 and 1975 but not measurably different from the scores in all other assessment years. In 2004, the average score of male 17-year-olds showed no measurable difference from 1971 or 1999.

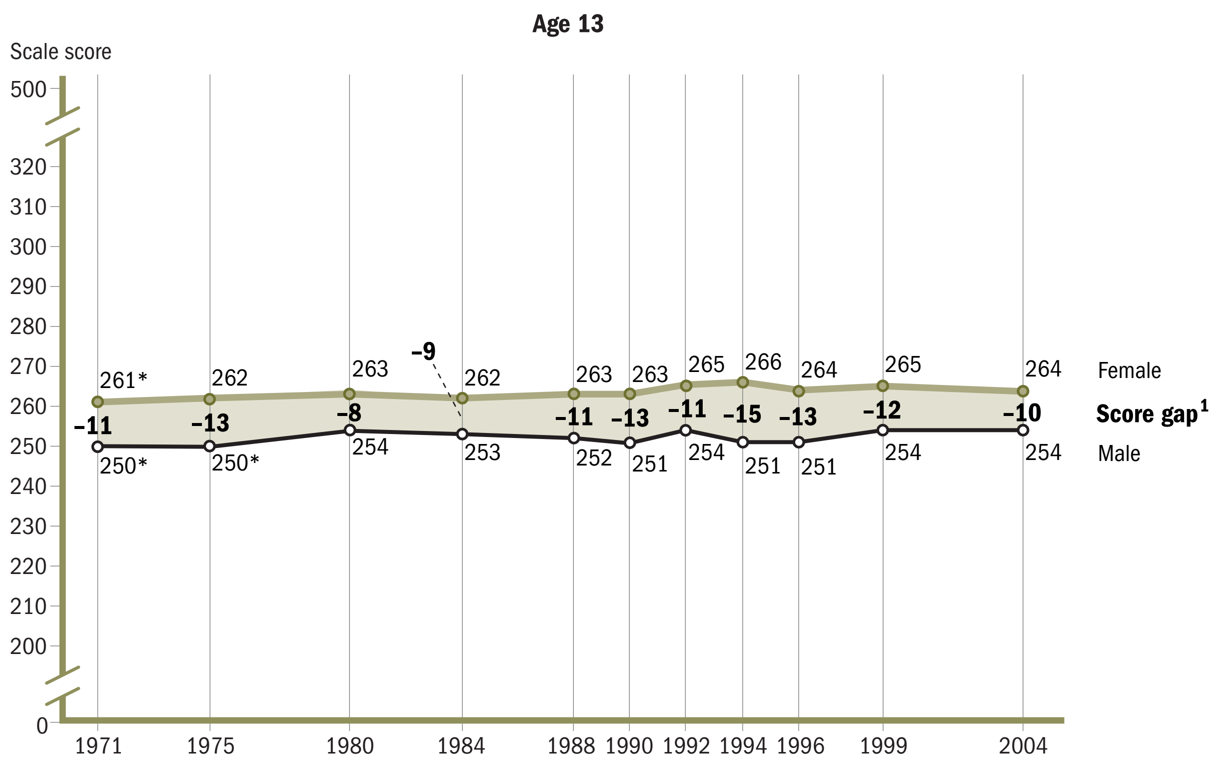
The reading trends of female students are similar to those of male students. At both ages 9 and 13 the female students' average reading scores were significantly higher in 2004 than in 1971. At age 9, the average score for female students was higher in 2004 than in any previous assessment year except 1980. There were no measurable differences in average scores for 13-year-old female students between 1975 and 2004. At age 17, female students' average score in 2004 was lower than those in 1990 and 1992 but not measurably different from that in 1971.

Score Differences Between Male and Female Students

Figure 3-1 also displays the gap between the male and female average scores. All reading score differences show female students scored higher on average than their male counterparts in 2004. The gender gap at age 9 decreased from 13 score points in 1971 to 5 score points in 2004. In contrast, there has been no measurable change in the score gap at age 13 between 2004 and any previous assessment year. For 17-year-olds, the score gap in 2004 was larger than the gaps in 1988 and 1980, but showed no measurable difference from the gaps in other assessment years.

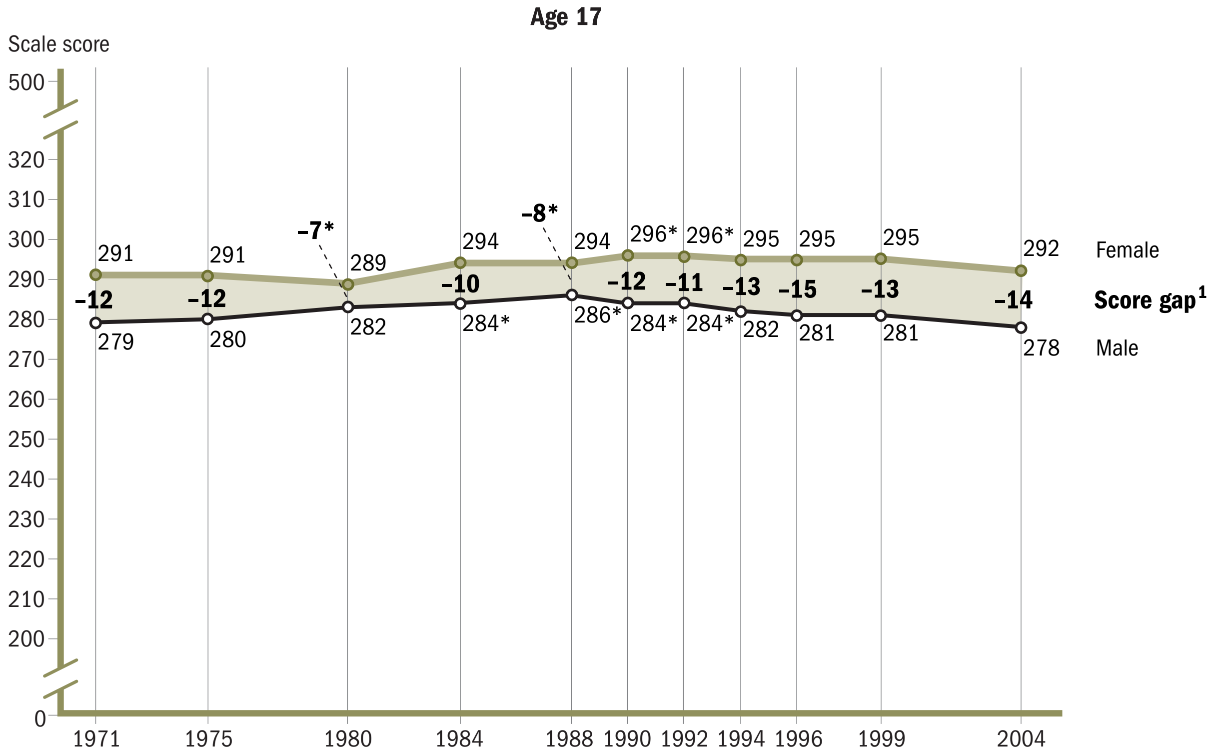
Figure 3-1. Trends in average reading scale scores and score gaps for students ages 9, 13, and 17, by gender: 1971–2004





See notes at end of figure.

Figure 3-1. Trends in average reading scale scores and score gaps for students ages 9, 13, and 17, by gender: 1971–2004—Continued



*Significantly different from 2004.

[1]Male average scale score minus female average scale score.

NOTE: Score gaps are calculated based on differences between unrounded average scale scores. Negative numbers indicate that the average scale score for male students was lower than the score for female students.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading Assessments.

How to interpret this graphic . . .

Graphics such as those in figures 3-1, 3-2, and 3-3 are called "gap charts." They are intended to show both the trend in performance of a single student group over time (such as female students) and the gap between two groups of students (such as males and females). In figure 3-1, the average reading scores of male and female students are graphed separately, and the difference between the two scores is shown. For example, in 2004, female 9-year-olds had an average score of 221, and male 9-year-olds had an average score of 216. When the average score for female students is subtracted from the average score of male students, the difference is –5 points. All differences are shaded.

Trends in Reading Scores by Race/Ethnicity

Figures 3-2 and 3-3 display the average reading scores across assessment years for White, Black, and Hispanic students as well as the score gaps between White and Black or White and Hispanic students.

Trends in Reading for White Students

For White students, the average scores for 9- and 13-year-olds were higher in 2004 than in 1971. As with the national sample, scores for White 9-year-olds were higher in 2004 than in any previous assessment year.

Trends in Reading for Black Students

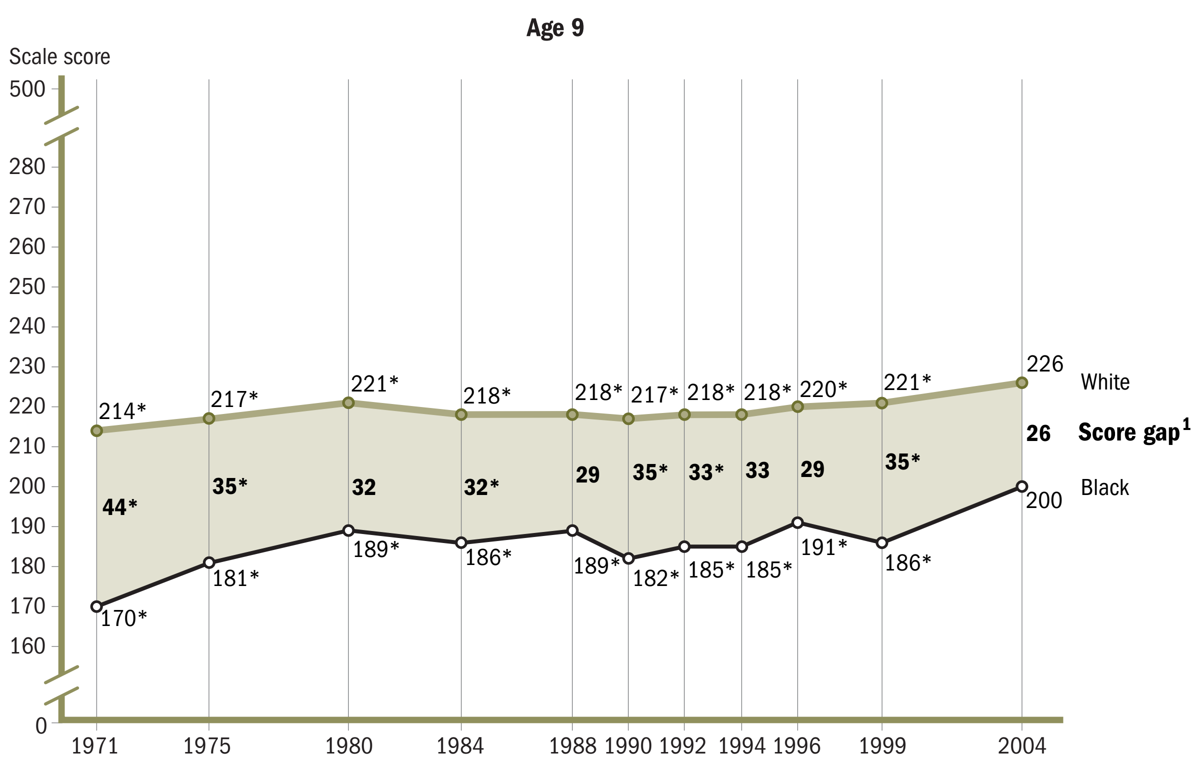
For Black students at all three ages, average reading scores in 2004 were higher than in 1971. At age 9, Black students scored higher on average in 2004 than in any previous administration year, up 30 points from 1971 and up 15 points[1] since 1999. For age 13, scores increased by 22 points between 1971 and 2004. Average scores for Black students at age 17 increased between 1971 and 2004 by 25 points.

Score Differences Between White and Black Students

As shown in figure 3-2, the differences in scores for White and Black students have decreased between the first (1971) and the most recent (2004) assessments across all three ages, although White students scored higher on average than Black students at each age level in 2004.

The score gap between Black and White students at age 9 decreased by 18 points between 1971 and 2004 and by 9 points between 1999 and 2004. At age 13, the gap decreased from 39 points in 1971 to 22 points in 2004. At age 17, the gap decreased by 24 points between 1971 and 2004.

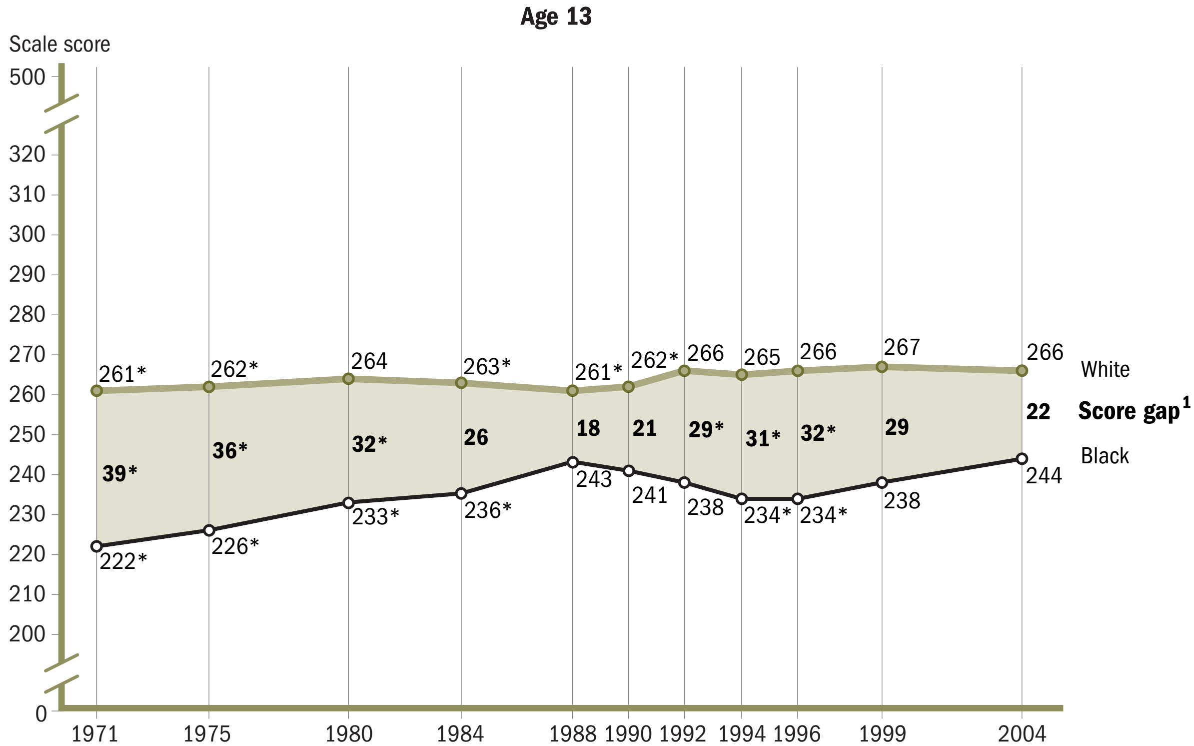
Figure 3-2. Trends in average reading scale scores and score gaps for White students and Black students ages 9, 13, and 17: 1971–2004

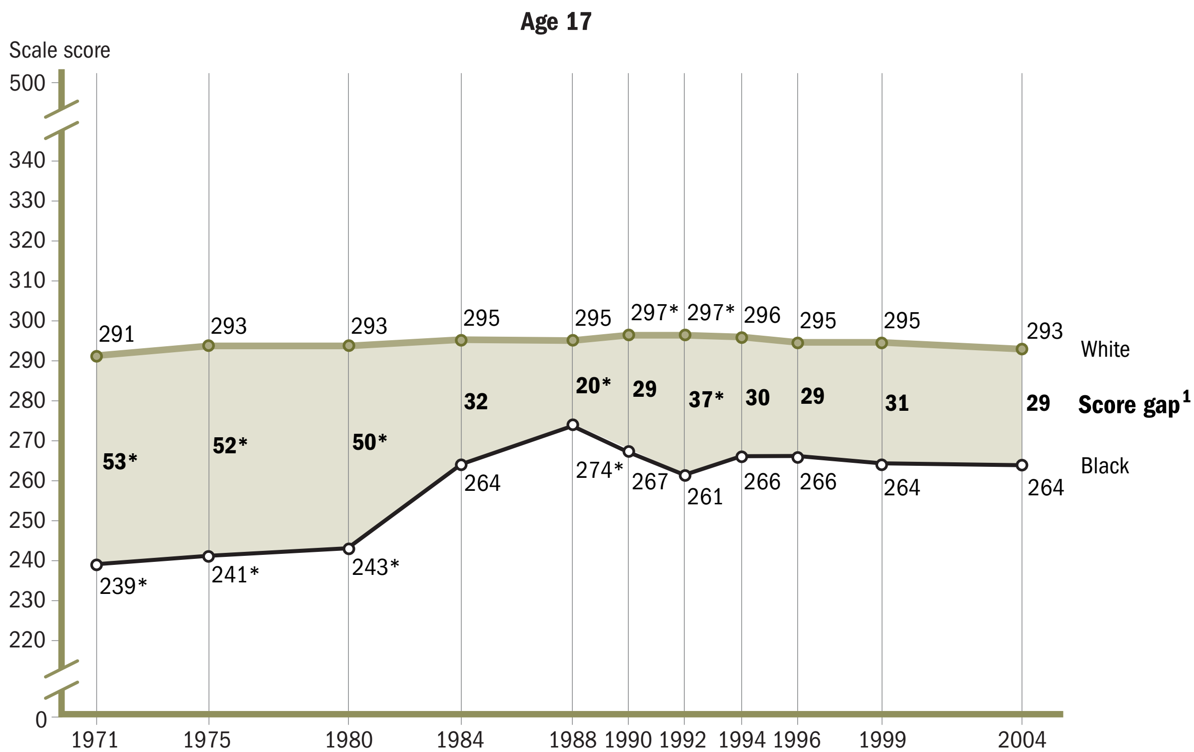


See notes at end of figure.

[1] Detail may not sum to totals because of rounding. Differences between scores are calculated using unrounded values. In this instance, the result of the subtraction differs from what would be obtained by subtracting the rounded values shown in the accompanying figure.

Figure 3-2. Trends in average reading scale scores and score gaps for White students and Black students ages 9, 13, and 17: 1971–2004—Continued





*Significantly different from 2004.

[1]White average scale score minus Black average scale score.

NOTE: Score gaps are calculated based on differences between unrounded average scale scores.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading Assessments.

Trends in Reading for Hispanic Students

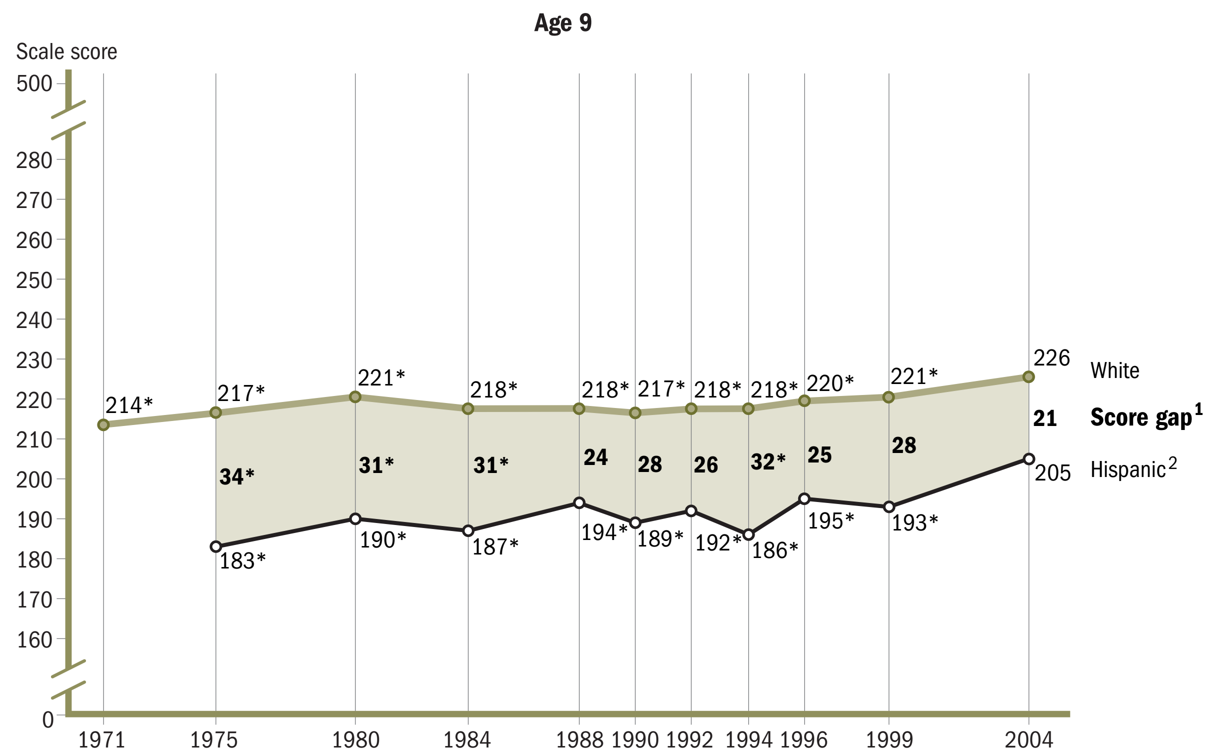
The average reading scores for Hispanic students show mixed results across the ages. As with the other racial/ethnic groups, the average reading score for Hispanic students at age 9 was higher in 2004 than in any other assessment year. The average score for Hispanic students at age 13 shows an increase of 10 points between 1975 and 2004. The scores for 17-year-old Hispanic students increased by 11 points[2] between 1975 and 2004, but no measurable changes were seen between 1999 and 2004. It is worth noting that due to smaller sample sizes, the standard errors associated with the scores of Hispanic students are relatively large, meaning that differences that look large may not be statistically significant.

Score Differences Between White and Hispanic Students

As shown in figure 3-3, White students scored higher on average than their Hispanic peers in reading at each age in 2004.

At age 9, the score gap between White and Hispanic students decreased from 34 points in 1975 to 21 points in 2004. At age 13, any apparent changes between 2004 and all previous assessment years in the size of the score gap were not statistically significant, except between 2004 and 1994, when the score gap narrowed by 6 points. At age 17, the score gap between White and Hispanic students was measurably smaller in 2004 than in 1975.

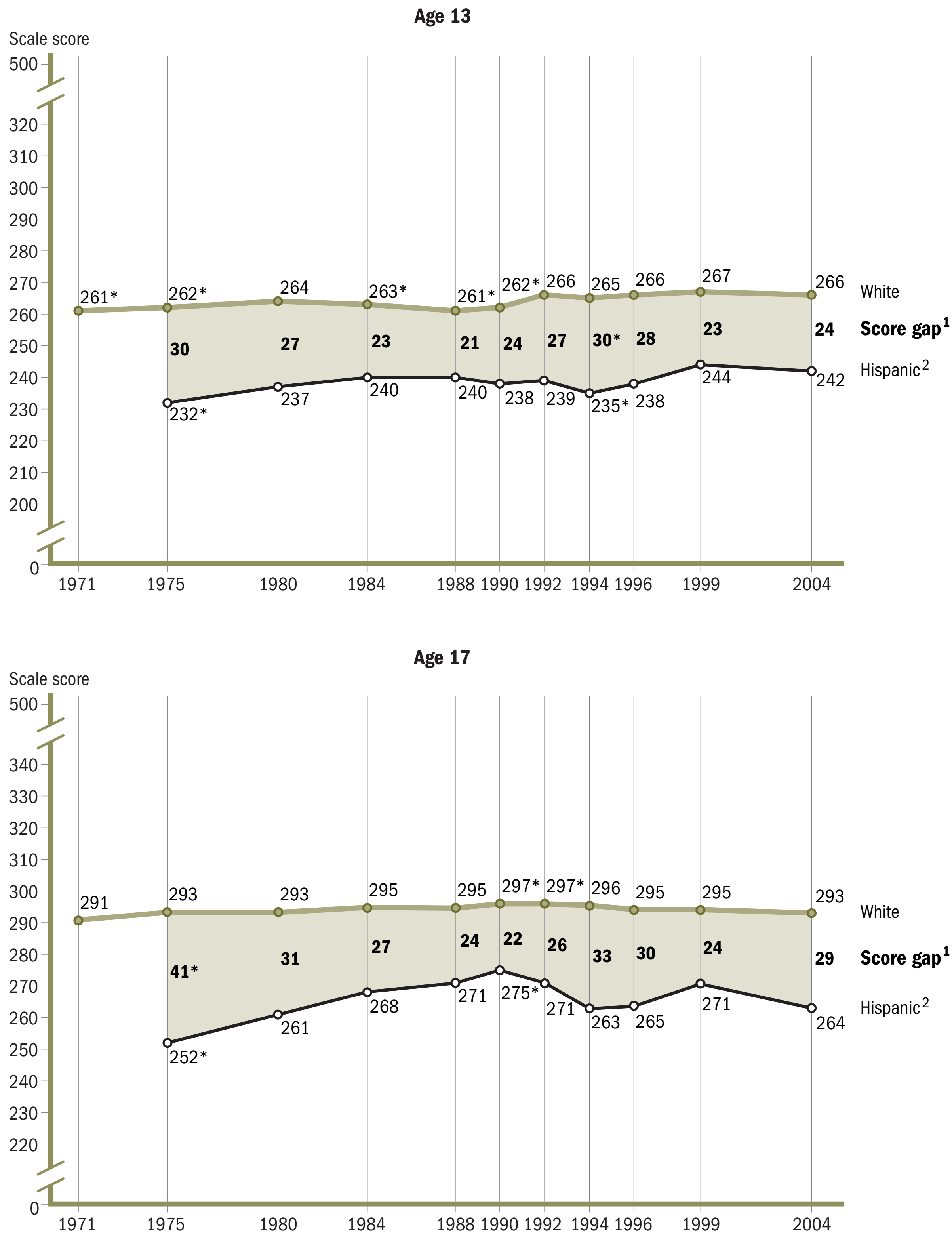
Figure 3-3. Trends in average reading scale scores and score gaps for White students and Hispanic students ages 9, 13, and 17: 1971–2004



See notes at end of figure.

[2] Detail may not sum to totals because of rounding. Differences between scores are calculated using unrounded values. In this instance, the result of the subtraction differs from what would be obtained by subtracting the rounded values shown in the accompanying figure.

Figure 3-3. Trends in average reading scale scores and score gaps for White students and Hispanic students ages 9, 13, and 17: 1971–2004—Continued



*Significantly different from 2004.

[1]White average scale score minus Hispanic average scale score.

[2]Data for Hispanic students are included in the overall national results but not reported as a separate racial/ethnic category in 1971.

NOTE: Score gaps are calculated based on differences between unrounded average scale scores.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading Assessments.

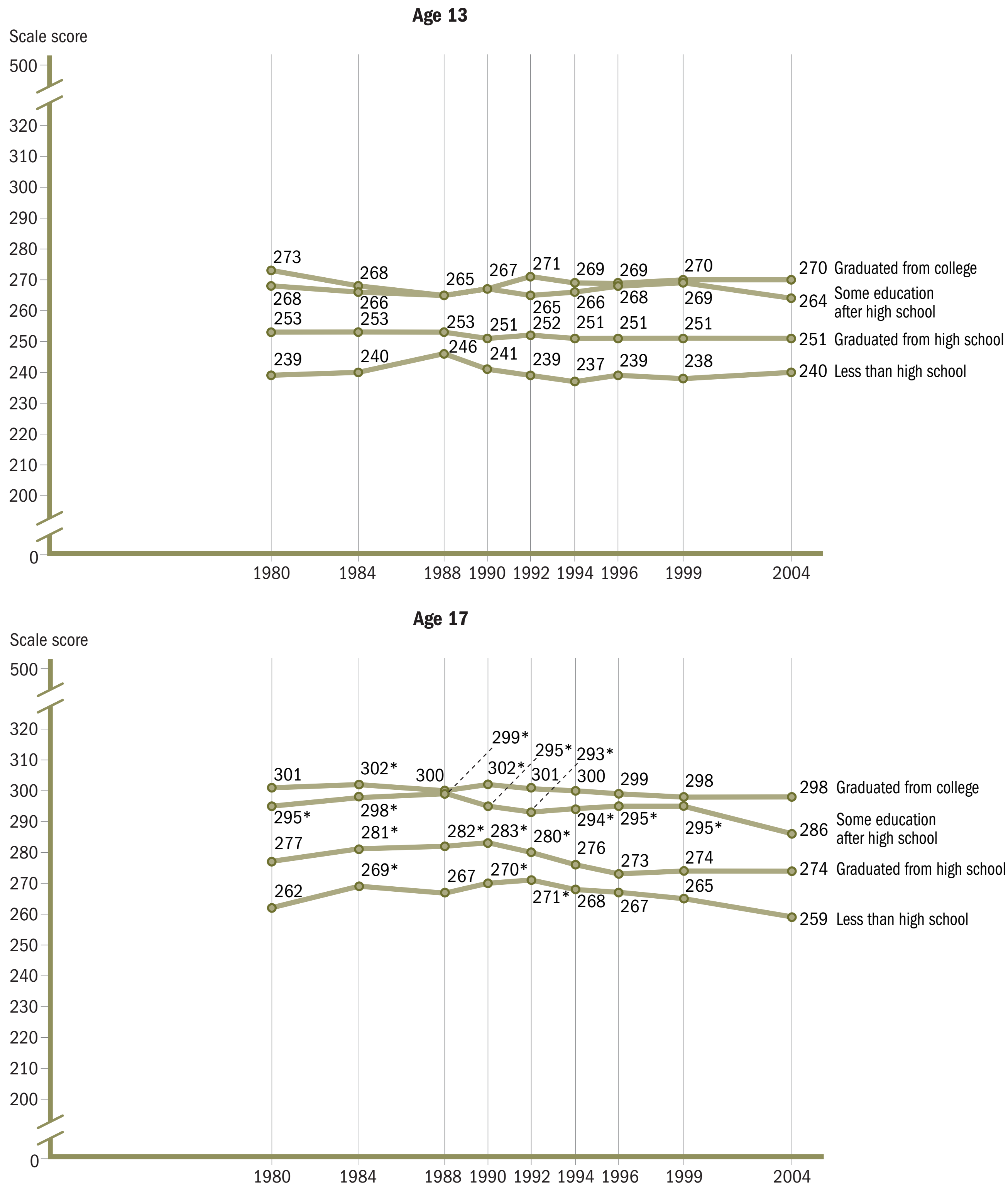
Trends in Reading Scores by Parents' Highest Level of Education

The average reading scores of students at ages 13 and 17 by students' reports of parents' highest education level across the assessment years are shown in figure 3-4. Results are not reported at age 9, because internal research shows that students' reports of their parents' education level are less reliable at this age. The percentage of students reporting that at least one parent had graduated from college has increased since 1980, while the percentages of students reporting that the highest level of education for their parents was a high school diploma or less has decreased (see table B-2).

Among 13-year-olds, there were no measurable differences in average scores between 2004 and all previous assessment years regardless of student-reported level of parental education. In 2004, scores averaged 251, 264, and 270, respectively, for students who reported that at least one parent graduated from high school, completed some education after high school, or graduated from college. None of these average scores was measurably different from the average scores in 1999 or 1980.

At age 17, there were no measurable differences in average scores in 2004 compared to average scores in 1980 and 1999 for three of the four student-reported levels of parents' education. The exception was for students who reported that at least one parent had some education after high school. At age 17, the average score for students who indicated their parents had some education after high school was lower in 2004 than in any previous assessment year, dropping from 295 to 286 between 1999 and 2004.

Figure 3-4. Trends in average reading scale scores for students ages 13 and 17, by student-reported parents' highest level of education: 1980–2004



*Significantly different from 2004.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1980–2004 Long-Term Trend Reading Assessments.

Trends in Mathematics Scores by Student Groups

This section presents the results of the long-term trend mathematics assessment for the various student groups. For gender and race/ethnicity, the results are presented first for each group separately and then the score gaps between the groups are examined.

Trends in Mathematics Scores by Gender

As discussed in chapter 2, the mathematics national trend showed higher average scores in 2004 than in previous assessment years for ages 9 and 13, while at age 17 there were no measurable changes in average scores between 2004 and 1973 or 1999. For the most part, the scores of male and female students paralleled that trend, as seen in figure 3-5.

For male students, the average mathematics scores at ages 9 and 13 were higher in 2004 than in any previous assessment year. Scores for males at age 9 increased by 25 points between 1973 and 2004 and by 10 points between 1999 and 2004. The average score for male students at age 13 was higher in 2004 than in 1999 by 5 points.[3] The average score for male students at age 17 was higher in 2004 than in 1978, but there was no measurable difference between the scores in 1999 and 2004.

The trends for female students were similar, as average scores in 2004 were higher than in any previous assessment year at ages 9 and 13. At age 13, there was a 5-point increase in the average scores of female students between 1999 and 2004. At age 17, female students scored higher in 2004 than in 1973 but showed no measurable difference between the scores in 1999 and 2004.

Score Differences Between Male and Female Students

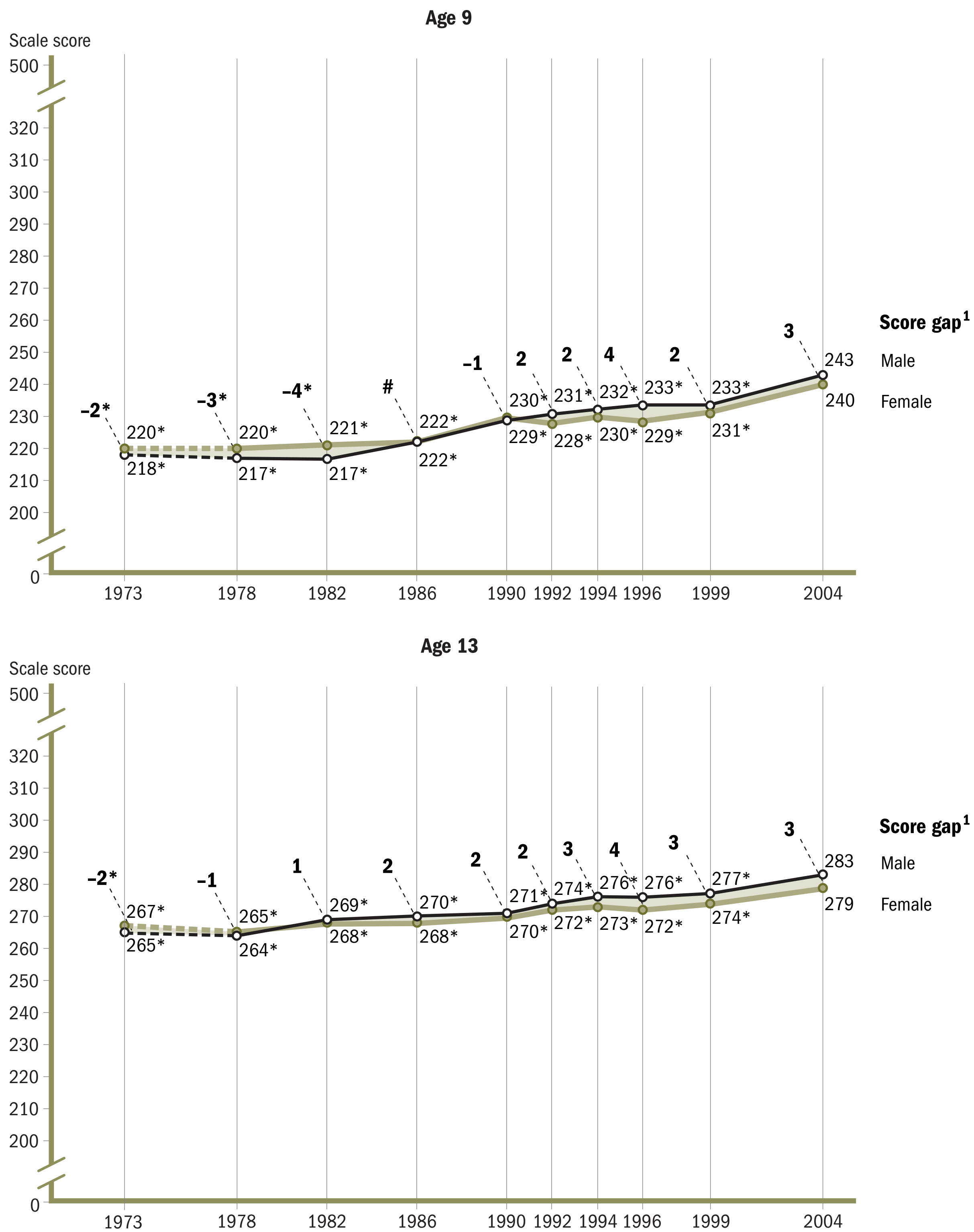
Figure 3-5 also shows the gap between the average mathematics scores of males and females. At age 9, the apparent difference between male and female students in 2004 was not statistically significant, while the change in the score gap between 1973 and 2004 was statistically significant. Males had higher average scores than females at ages 13 and 17. The gender score gaps for 13- and 17-year-olds were measurably different between 1973 and 2004.

How to interpret this graphic . . .

Graphics such as those in figures 3-5–3-7 are called "gap charts." They are intended to show both the trend in performance of a single student group over time (such as female students) and the gap between two groups of students (such as males and females). In figure 3-6, the average mathematics scores of male and female students are graphed separately, and the difference between the two scores is shown. For example, in 2004, female 9-year-olds had an average score of 240, and male 9-year-olds had an average score of 243. When the average score for female students is subtracted from the average score of male students, the difference is 3 points, shown with the dotted line. All differences are shaded.

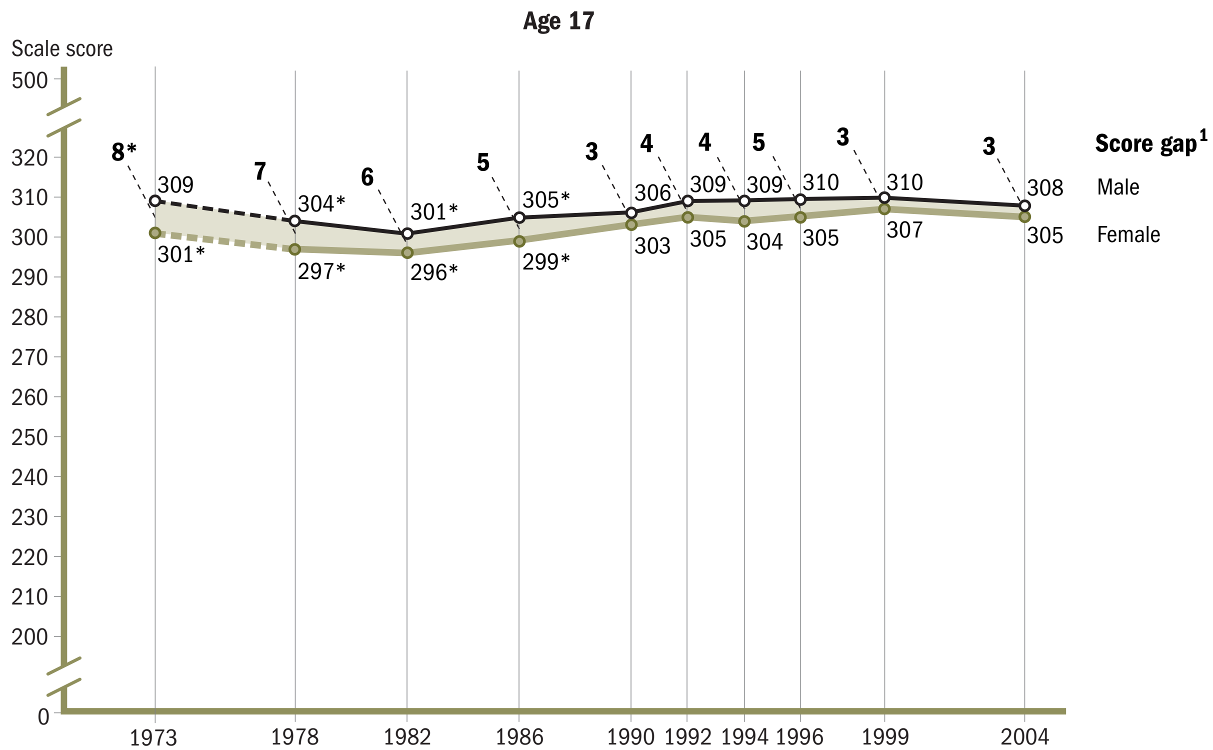
[3] Detail may not sum to totals because of rounding. Differences between scores are calculated using unrounded values. In this instance, the result of the subtraction differs from what would be obtained by subtracting the rounded values shown in the accompanying figure.

Figure 3-5. Trends in average mathematics scale scores and score gaps for students ages 9, 13, and 17, by gender: 1973–2004



See notes at end of figure.

Figure 3-5. Trends in average mathematics scale scores and score gaps for students ages 9, 13, and 17, by gender: 1973–2004—Continued



#The estimate rounds to zero.
*Significantly different from 2004.
[1]Male average scale score minus female average scale score. Negative numbers indicate that the average scale score for male students was lower than the score for female students.
NOTE: Dashed lines represent extrapolated data. Score gaps are calculated based on differences between unrounded average scale scores.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1973–2004 Long-Term Trend Mathematics Assessments.

Trends in Mathematics Scores by Race/Ethnicity

In 2004, the mathematics scores of the three largest racial/ethnic groups, as measured by the NAEP long-term trend assessment, show increases in performance at all ages. Oftentimes, these changes seem different from the overall trends. These differences are due to changes in the demographics in the population. Figure 3-6 displays the average scores and score gaps across assessment years in mathematics for White and Black 9-, 13-, and 17-year-old students.

Trends in Mathematics for White Students

The average score of 247 in 2004 for White students at age 9 was higher than in any previous assessment year. At age 13, White students had an average score of 288 in 2004, which was higher than in any previous assessment year. Average scores for White 17-year-olds showed no measurable difference between 1999 and 2004. However, their average score of 313 in 2004 was higher than the average score in 1973.

Trends in Mathematics for Black Students

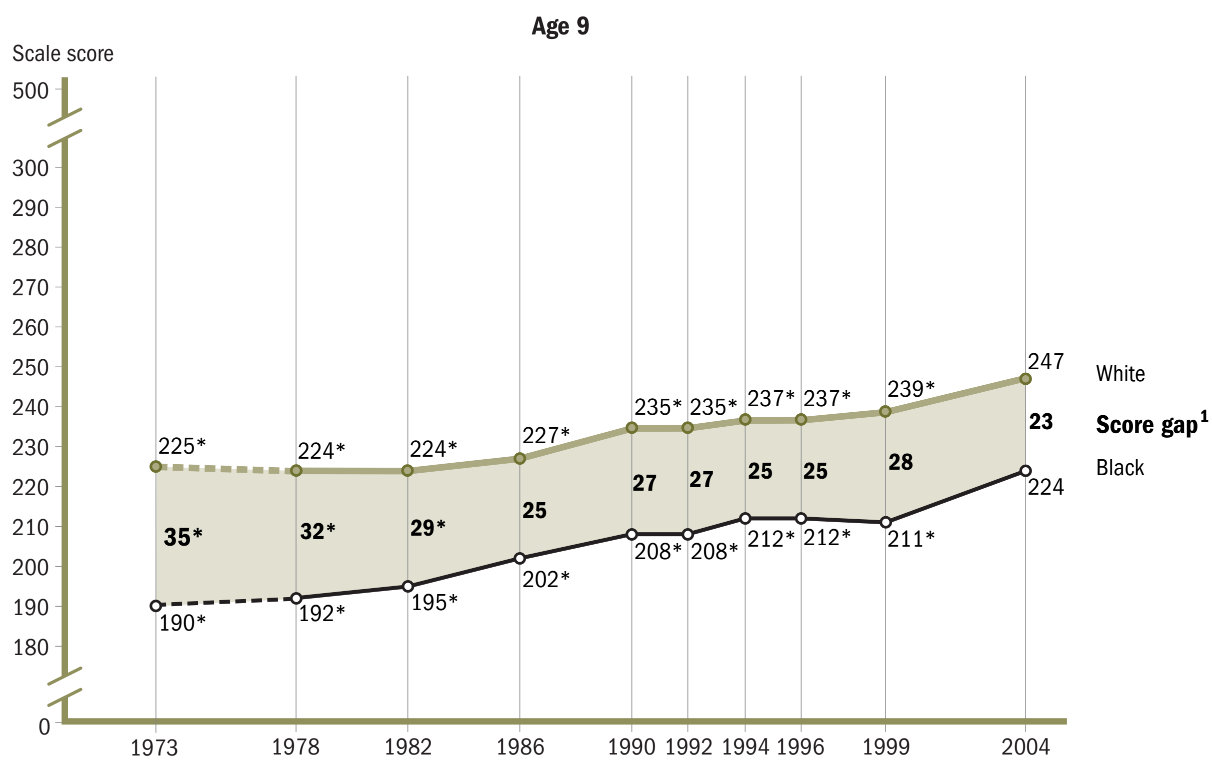
The average scores for Black students were higher in 2004 than in 1973 at all three ages. The scores for Black 9-year-olds showed measurable increases between 2004 and any previous assessment year. The score in 2004 was 34 points higher than the score in 1973 and 13 points higher than that in 1999. The 2004 mathematics score for Black 13-year-olds was higher than in any previous assessment year, and an 11-point increase in scores occurred between 1999 and 2004. The average score for Black 17-year-olds in 2004 was higher than the average score in 1973, but not measurably different from the average score in 1999.

Score Differences Between White and Black Students

As seen in figure 3-6, the differences in average scores for White and Black students at all ages decreased between the first (1973) and the most recent (2004) assessments in mathematics, although White students continued to outperform Black students in 2004.

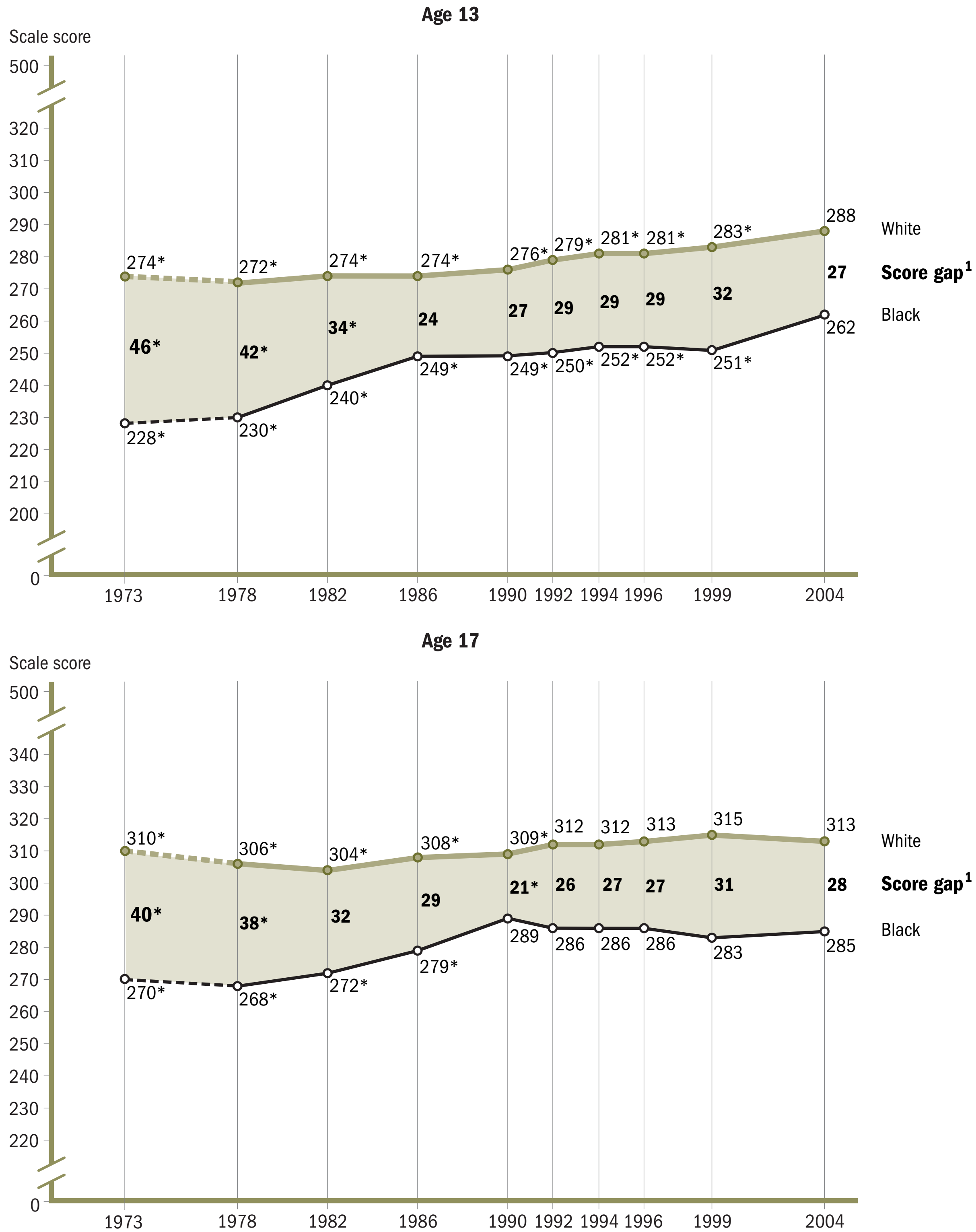
At age 9, the gap decreased from 35 points in 1973 to 23 points in 2004. At age 13, the gap decreased from 46 points in 1973 to 27 points in 2004, while the apparent difference in the gaps between 1999 and 2004 was not statistically significant. At age 17, the gap decreased from 40 points in 1973 to 28 points in 2004.

Figure 3-6. Trends in average mathematics scale scores and score gaps for White students and Black students ages 9, 13, and 17: 1973–2004



See notes at end of figure.

Figure 3-6. Trends in average mathematics scale scores and score gaps for White students and Black students ages 9, 13, and 17: 1973–2004—Continued



*Significantly different from 2004.

[1]White average scale score minus Black average scale score.

NOTE: Dashed lines represent extrapolated data. Score gaps are calculated based on differences between unrounded average scale scores.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1973–2004 Long-Term Trend Mathematics Assessments.

Trends in Mathematics for Hispanic Students

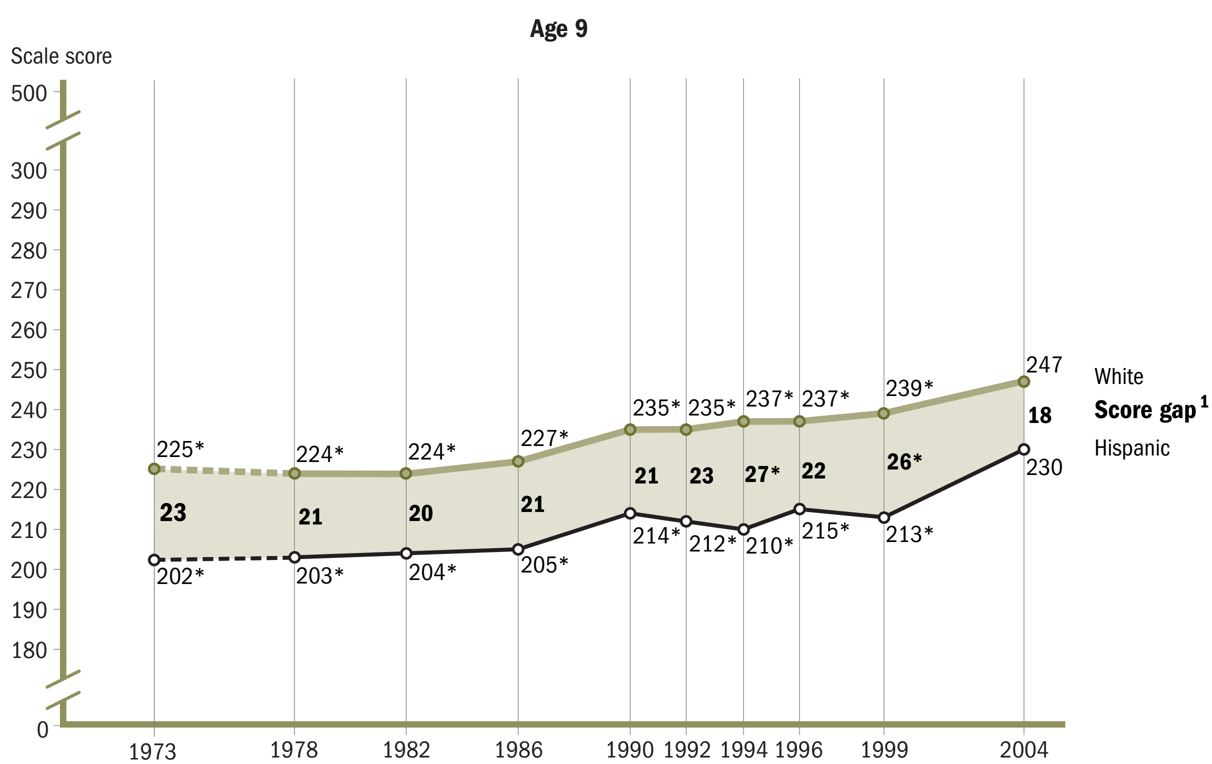
Figure 3-7 shows the trend lines for White and Hispanic students from 1973 to 2004. Hispanic students' average scores in mathematics were higher at all three ages in 2004 than in 1973. At age 9, the average score for Hispanics in 2004 was 28 points higher than the score in 1973 and higher than in any previous assessment year. At age 13, the average score in 2004 was higher than in any previous assessment year. At age 17, there was no measurable difference in average scores for Hispanic students between 1999 and 2004.

Score Differences Between White and Hispanic Students

As shown in figure 3-7, there were few changes in the score gap between White and Hispanic students. White students outscored Hispanic students at all three ages in 2004.

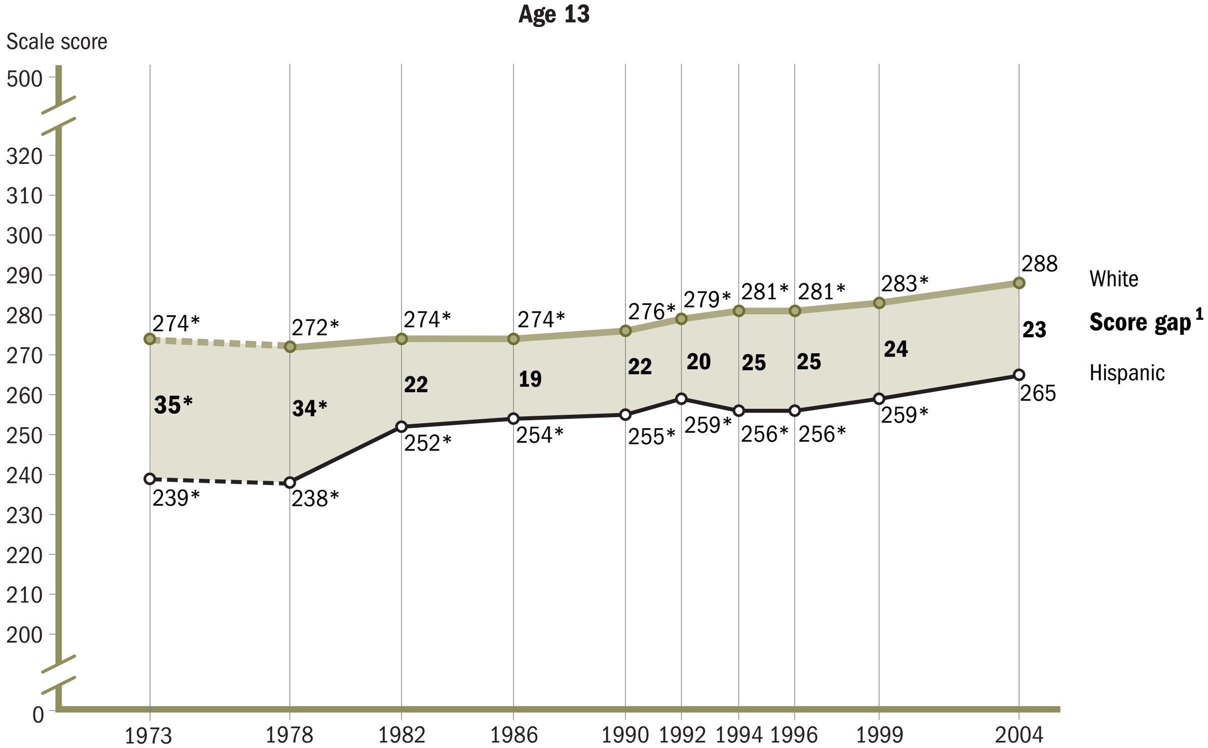
At age 9, the 2004 score gap between White and Hispanic students was measurably narrower than the gap in 1999, but showed no measurable difference from the gap in 1973. At age 13, the score gap in 2004 was narrower than the gaps in 1973 and 1978, but not measurably different from the gaps in any other assessment year. At age 17, the White-Hispanic score gap was smaller in 2004 than in 1973, but it was not measurably different from 1999 or any other assessment year after 1973.

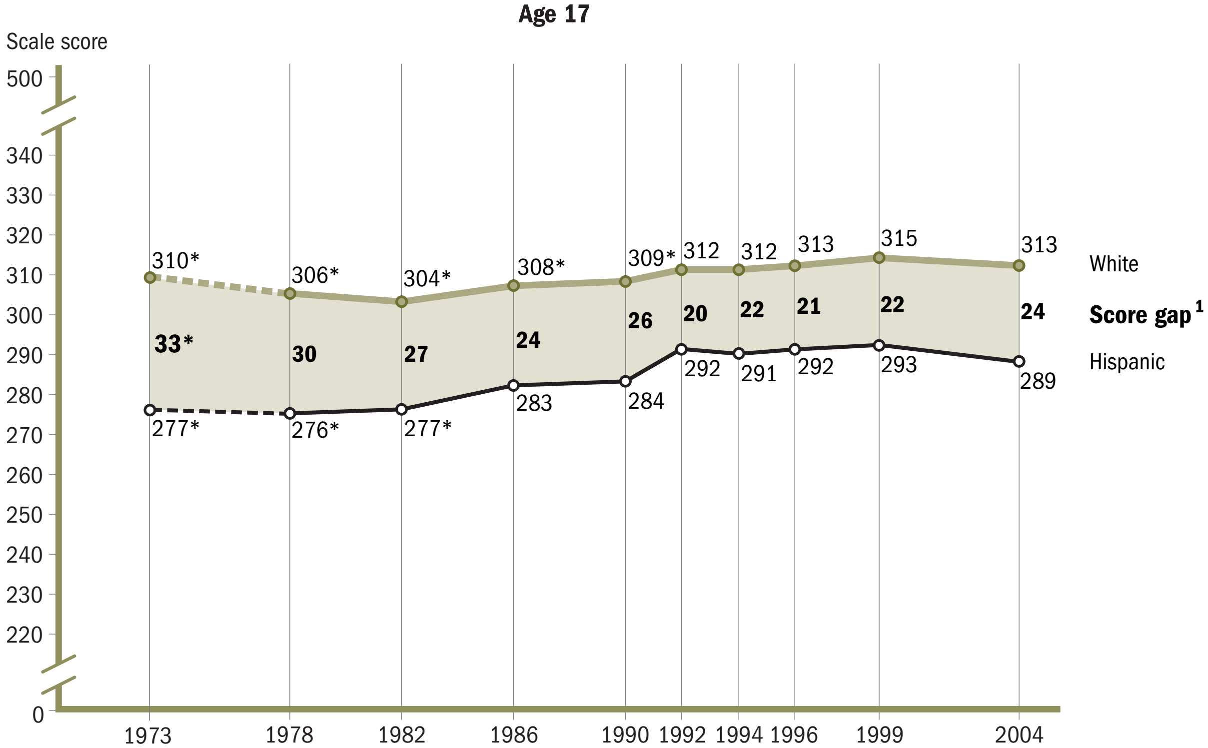
Figure 3-7. Trends in average mathematics scale scores and score gaps for White students and Hispanic students ages 9, 13, and 17: 1973–2004



See notes at end of figure.

Figure 3-7. Trends in average mathematics scale scores and score gaps for White students and Hispanic students ages 9, 13, and 17: 1973–2004—Continued





*Significantly different from 2004.

[1]White average scale score minus Hispanic average scale score.

NOTE: Dashed lines represent extrapolated data. Score gaps are calculated based on differences between unrounded average scale scores.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1973–2004 Long-Term Trend Mathematics Assessments.

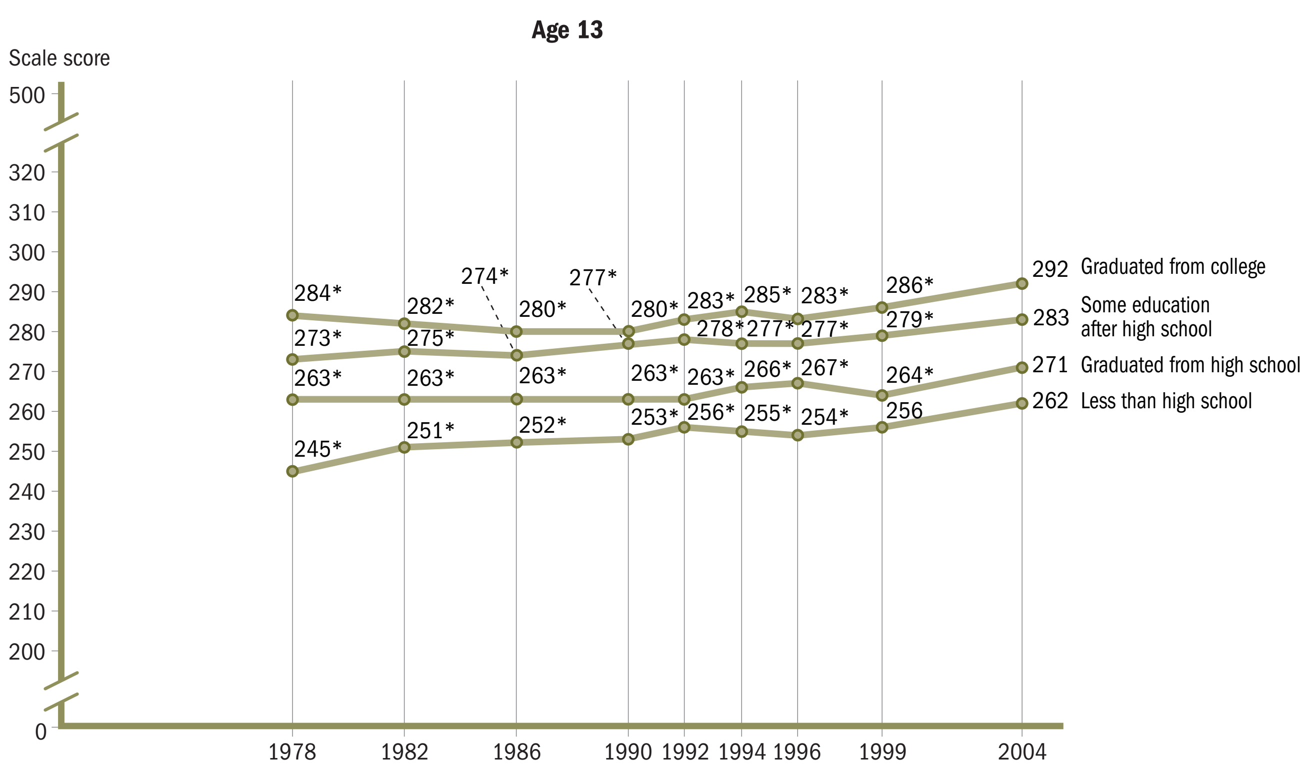
Trends in Mathematics Scores by Parents' Highest Level of Education

Average mathematics scores for students at ages 13 and 17 by highest level of parents' education as reported by the student are shown in figure 3-8. Results are not reported at age 9, because studies have shown that students' reports of their parents' education level are less reliable at this age.

At age 13, for students who reported that at least one parent had graduated from high school, had some education after high school, or had graduated from college, the average scores in 2004 were higher than in any other assessment year. Students who reported that their parents had less than a high school education had an average score in 2004 that was higher than the average score in 1978, but was not measurably different from the average score in 1999.

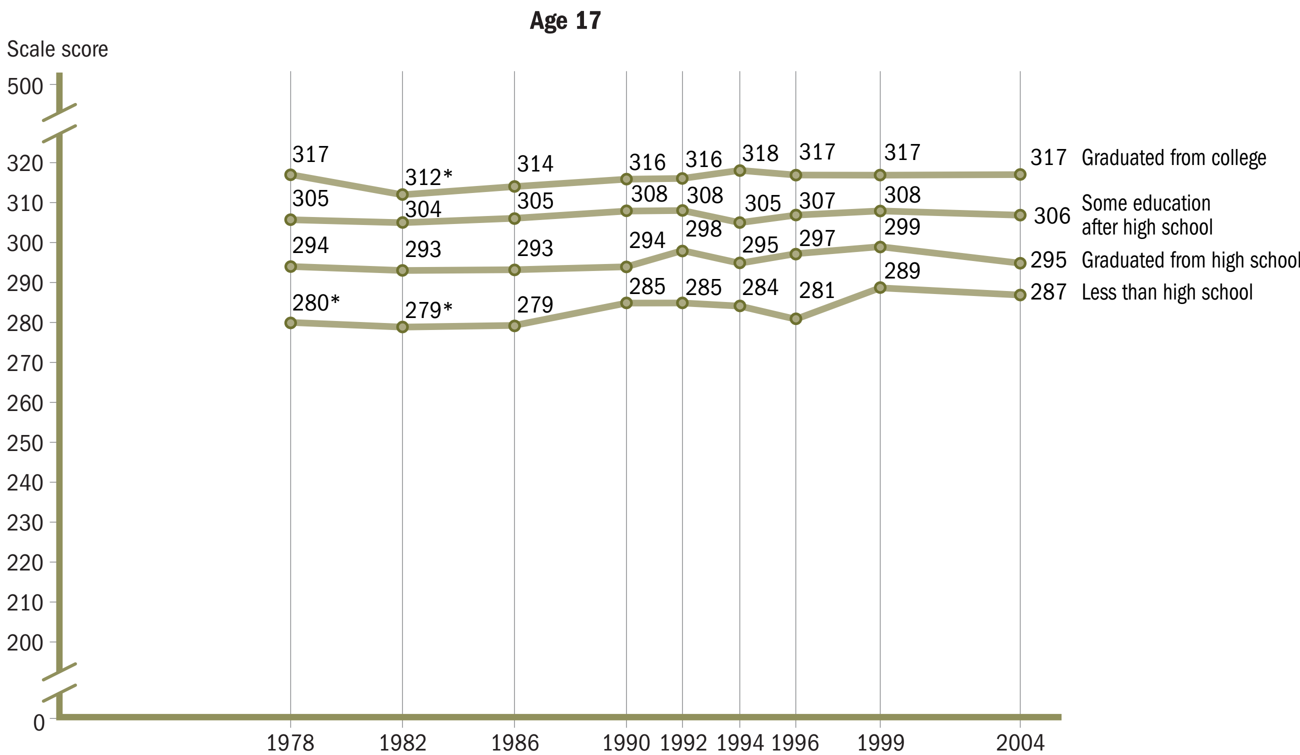
The average mathematics scores for 17-year-olds showed no measurable changes between 2004 and any previous assessment year for students who reported that at least one parent had graduated from high school or had some education after high school. For students with at least one parent who graduated from college, the average score of 17-year-olds was about the same in 2004 as in 1999 and in 1978 with an average score of 317. Students who reported that their parents had less than a high school education comprised the only group to show improvement between 1978 and 2004.

Figure 3-8. Trends in average mathematics scale scores for students ages 13 and 17, by student-reported parents' highest level of education: 1978–2004



See notes at end of figure.

Figure 3-8. Trends in average mathematics scale scores for students ages 13 and 17, by student-reported parents' highest level of education: 1978–2004—Continued



*Significantly different from 2004.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1978–2004 Long-Term Trend Mathematics Assessments.

Summary

This chapter presented results from the NAEP reading and mathematics long-term trend assessments for students in different reporting groups. The reporting groups examined were gender, race/ethnicity, and level of parental education.

The following figures, 3-9 through 3-11, provide an overview of the major findings presented in this chapter. In each line of the display, the average score for a particular group of students in 2004 is compared to that in the first assessment year in which data are available, and to that in 1999. Arrows pointing upward (↑) indicate increases, horizontal arrows (→) indicate no measurable change, and arrows pointing downward (↓) indicate decreases. For example, the first line of the display in figure 3-9 indicates that the average reading score for male 9-year-olds in 2004 was higher than in both 1999 and 1971.

Figure 3-9. Summary of trends in reading and mathematics average scale scores for students ages 9, 13, and 17, by gender: 1971–2004

Reading

Male

↑ 9-year-olds' average scale scores since 1971 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1971 (→ since 1999)
→ 17-year-olds' average scale scores since 1971 (→ since 1999)

Female

↑ 9-year-olds' average scale scores since 1971 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1971 (→ since 1999)
→ 17-year-olds' average scale scores since 1971 (→ since 1999)

Mathematics

Male

↑ 9-year-olds' average scale scores since 1973 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1973 (↑ since 1999)
→ 17-year-olds' average scale scores since 1973 (→ since 1999)

Female

↑ 9-year-olds' average scale scores since 1973 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1973 (↑ since 1999)
↑ 17-year-olds' average scale scores since 1973 (→ since 1999)

↑Significantly higher in 2004.
→Indicates no significant difference between earlier year and 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading and Mathematics Assessments.

Figure 3-10. Summary of trends in reading and mathematics average scale scores for students ages 9, 13, and 17, by race/ethnicity: 1971–2004

Reading

White

↑ 9-year-olds' average scale scores since 1971 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1971 (→ since 1999)
→ 17-year-olds' average scale scores since 1971 (→ since 1999)

Black

↑ 9-year-olds' average scale scores since 1971 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1971 (→ since 1999)
↑ 17-year-olds' average scale scores since 1971 (→ since 1999)

Hispanic

↑ 9-year-olds' average scale scores since 1975 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1975 (→ since 1999)
↑ 17-year-olds' average scale scores since 1975 (→ since 1999)

Mathematics

White

↑ 9-year-olds' average scale scores since 1973 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1973 (↑ since 1999)
↑ 17-year-olds' average scale scores since 1973 (→ since 1999)

Black

↑ 9-year-olds' average scale scores since 1973 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1973 (↑ since 1999)
↑ 17-year-olds' average scale scores since 1973 (→ since 1999)

Hispanic

↑ 9-year-olds' average scale scores since 1973 (↑ since 1999)
↑ 13-year-olds' average scale scores since 1973 (↑ since 1999)
↑ 17-year-olds' average scale scores since 1973 (→ since 1999)

↑Significantly higher in 2004.
→Indicates no significant difference between earlier year and 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading and Mathematics Assessments.

Figure 3-11. Summary of trends in reading and mathematics average scale scores for students ages 13 and 17, by student-reported parents' highest level of education: 1978-2004

Reading

Less than high school

→ 13-year-olds' average scale scores since 1980 (→ since 1999)
→ 17-year-olds' average scale scores since 1980 (→ since 1999)

Graduated from high school

→ 13-year-olds' average scale scores since 1980 (→ since 1999)
→ 17-year-olds' average scale scores since 1980 (→ since 1999)

Some education after high school

→ 13-year-olds' average scale scores since 1980 (→ since 1999)
↓ 17-year-olds' average scale scores since 1980 (↓ since 1999)

Graduated from college

→ 13-year-olds' average scale scores since 1980 (→ since 1999)
→ 17-year-olds' average scale scores since 1980 (→ since 1999)

Mathematics

Less than high school

↑ 13-year-olds' average scale scores since 1978 (→ since 1999)
↑ 17-year-olds' average scale scores since 1978 (→ since 1999)

Graduated from high school

↑ 13-year-olds' average scale scores since 1978 (↑ since 1999)
→ 17-year-olds' average scale scores since 1978 (→ since 1999)

Some education after high school

↑ 13-year-olds' average scale scores since 1978 (↑ since 1999)
→ 17-year-olds' average scale scores since 1978 (→ since 1999)

Graduated from college

↑ 13-year-olds' average scale scores since 1978 (↑ since 1999)
→ 17-year-olds' average scale scores since 1978 (→ since 1999)

↑Significantly higher in 2004.
→Indicates no significant difference between earlier year and 2004.
↓Significantly lower in 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1978–2004 Long-Term Trend Reading and Mathematics Assessments.

Students' home and learning environments have changed over time. Students have greater access to computers and are taking more upper-level mathematics classes. Students are also reading more in school and for homework.

Chapter 4
Trends in Students' School and Home Experiences

In examining trends in students' academic achievement, it is important also to consider the context of their learning. The context of learning today has changed since the assessment was first administered in the early 1970s. For example, computer technology plays a greater role in education as schools improve their infrastructure, use multimedia in their classrooms, and encourage students to explore research topics on the Internet. Calculators are used more often in the classroom, and algebra is being taught in earlier grades than it was three decades ago (Braswell et al. 2005).

Home environments have changed as well. Contextual variables such as availability of computers in the home or parental involvement may affect student learning (Cai, Moyer, and Wang 1997; Downes and Reddacliff 1997; Rathburn, West, and Hausken 2003). As part of NAEP's long-term trend assessments, students have responded to a variety of questions about their school and home experiences. The information gained from these responses provides insight into the activities and experiences that form the contexts in which students learn. This chapter highlights students' responses to NAEP background questions about several key factors associated with student achievement.

In the following sections, data are presented to show each variable's relationship to scores on the 2004 NAEP reading and mathematics long-term trend assessment. Different background questions were asked for reading and for mathematics, so the two sections highlight different variables. Trends associated with contextual factors are presented two ways. First, the relationship between the variable and the average NAEP score is examined. It should be noted, however, that a relationship between NAEP scores and students' responses to certain questions does not establish a causal relationship between a particular factor and student achievement. The relationship may be influenced by a number of other variables not accounted for in this report, such as family income or students' attitudes. In addition, the information examined here is based solely on student self-reports, which may vary in accuracy across ages and students.

Second, the contextual variable is shown on its own to clarify how students' responses to the background questions have changed over time. That is, the percentages of students selecting each response option in 2004 are compared with those from the first assessment year in which the question was asked. (The comparison year varies by question.) These comparisons, even without the associated performance scores, demonstrate how the context of education has changed over time.

Contextual Factors Associated With Reading

Students responded to several questions relating to reading as they took the long-term trend assessment. This chapter reports on three variables associated with reading: the amount of time spent on homework, the number of pages read per day for both school and homework, and the amount of time spent reading for fun.

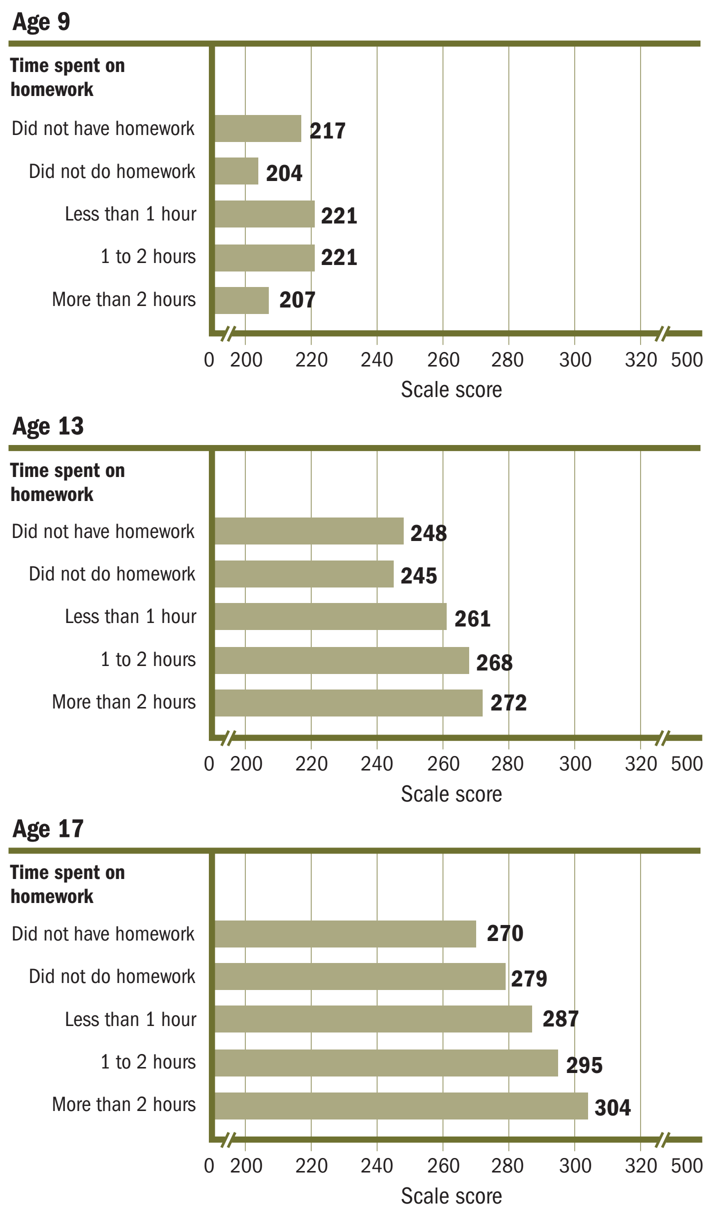
Amount of Homework

The first of two background questions pertaining to homework on the reading assessment is discussed in this section. Specifically, the question relating to time spent on homework asked, "How much time did you spend on homework yesterday?" The possible responses included the following:

- No homework was assigned.
- I had homework but didn't do it.
- Less than 1 hour
- 1 to 2 hours
- More than 2 hours

This question was asked at age 9 in assessment years 1984 through 2004 and at ages 13 and 17 in assessment years 1980 through 2004. Figure 4-1 shows the average reading scores in 2004 by the amount of time spent on homework for all three age groups, and figure 4-2 shows the trend in the percentages of students across the three age groups reporting they spent varying amounts of time on homework.

Figure 4-1. Average reading scale scores for students ages 9, 13, and 17, by amount of time spent on homework: 2004



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

How to interpret this graphic . . .

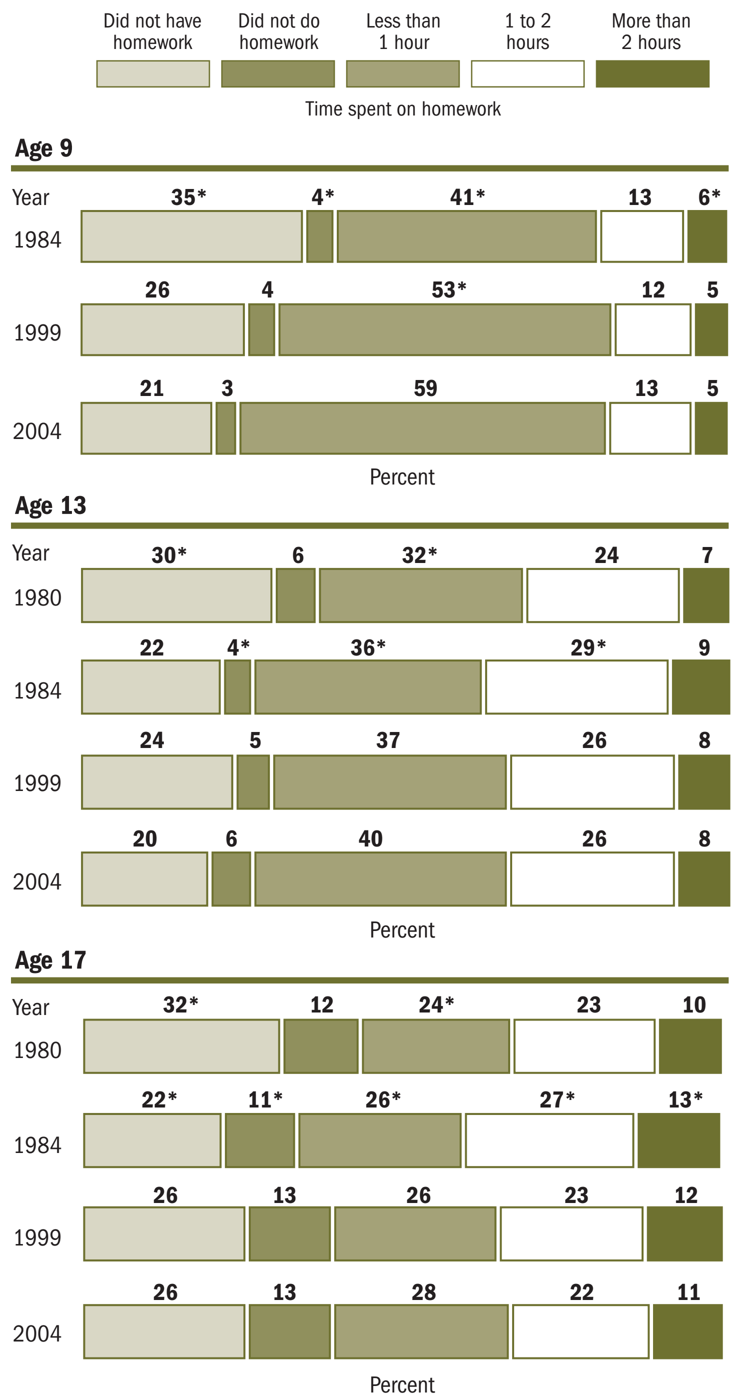
The graphics in this chapter differ from those in previous chapters in that the scale scores have been placed on the horizontal axis rather than on the vertical axis. The categories of the contextual variable analyzed are on the vertical axis. Thus, in figure 4-1, the five categories of "time spent on homework" are shown in order of amounts of time on the vertical axis, with the horizontal bar showing the average score for each category. For example, at age 17, students who did not have any homework had an average score of 270, and the average scores increased with each category of homework, up to 304 for the "more than two hours" category.

At all three ages less than one hour was the most commonly reported amount of time spent on homework the previous day (figure 4-2). However, the relationship between the amount of time spent on homework and average score on the NAEP reading assessment differed across the ages. In 2004, the average score of 9-year-olds who spent less than one hour on homework was higher than the average scores of students who did not do the assigned homework or who spent more than two hours on homework. The relationship between homework and achievement was more straightforward at age 13. In 2004, the average scores for 13-year-olds who spent either one to two hours or more than two hours on homework were higher than the average scores for their peers who spent less than one hour on homework, did not do their homework, or did not have any homework to do. At age 17, higher average scores on the long-term trend reading assessment were associated with more time spent on homework. That is, in 2004, those students who spent more than two hours on homework had higher average scores than those who spent one to two hours, whose scores were higher in turn than those of students who spent less than one hour, whose scores were higher than those of students who did not do any homework.

How to interpret this graphic . . .

The other type of graphic used in this chapter is a percentage distribution bar. Figure 4-2 shows the percentage of students who chose each category of a question, and the percentages add up to 100 percent of the assessed students. The years shown include the first years the question was asked (1980 and 1984), 1999, and 2004. So, figure 4-2 shows that at age 9 the percentage of students who reported that they spent less than one hour on homework was 41 percent in 1984 and 53 percent in 1999, both of which were lower than the 59 percent reported in 2004. At the same time, the percentage of students who reported they did not have any homework decreased from 35 percent in 1984 to 21 percent in 2004.

Figure 4-2. Percentages of students ages 9, 13, and 17, by amount of time spent on homework: 1980, 1984, 1999, and 2004



*Significantly different from 2004.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1980, 1984, 1999, and 2004 Long-Term Trend Reading Assessments.

In 2004, a greater percentage of 9-year-olds indicated that they spent less than one hour on homework than in any other year in which the question was asked. Simultaneously, the percentage of students indicating either that no homework was assigned or that they did not do any homework decreased between 1984 and 2004. The percentage of 13-year-old students spending less than one hour on homework has increased, from 32 percent in 1980 to 40 percent in 2004. At the same time, the percentage of students reporting that they did not have any homework decreased from 30 percent in 1980 to 20 percent in 2004. At age 17, the percentage of students indicating they spent less than one hour on homework the previous day increased from 24 to 28 percent between 1980 and 2004. At the same time, the percentage of 17-year-olds reporting that they were not assigned homework decreased from 32 to 26 percent.

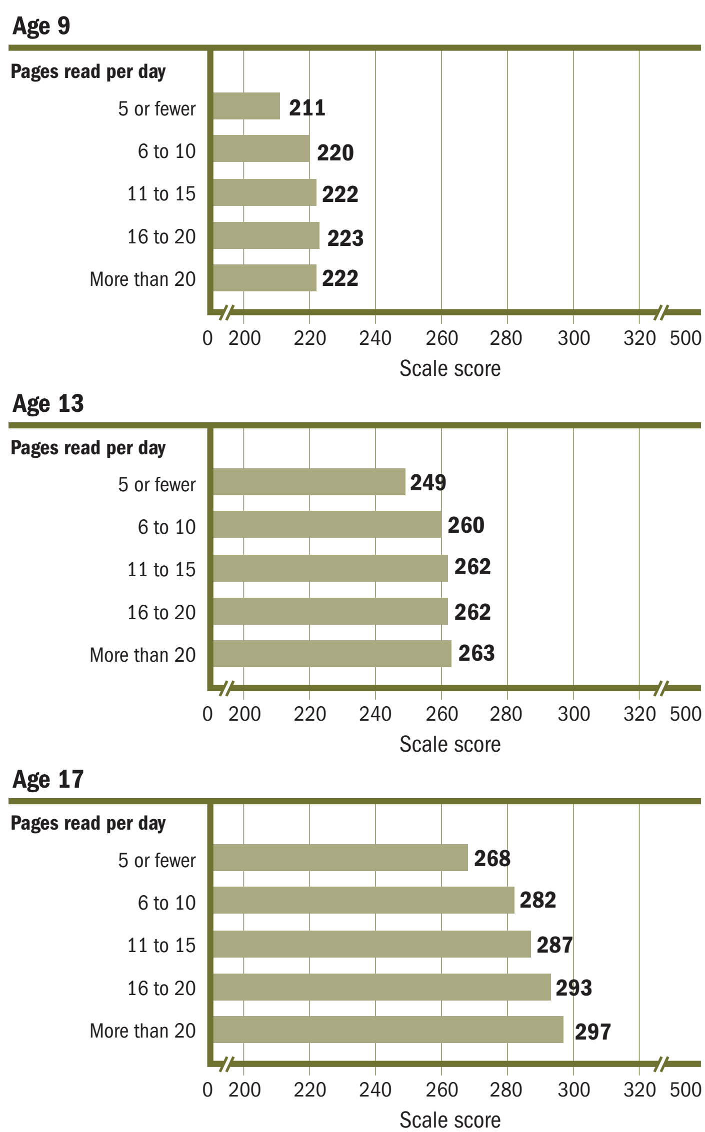
Pages Read Per Day

As part of the reading background questionnaire, students at all three ages were asked about the number of pages they read in school and for homework each day. The response options included the following:

- 5 or fewer
- 6 to 10
- 11 to 15
- 16 to 20
- More than 20

This question was first presented to students at ages 9, 13, and 17 in 1984. Figure 4-3 shows the average reading scores in 2004 by the number of pages read per day for all three ages, and figure 4-4 shows the trend in the percentage of students reporting reading various numbers of pages per day across the three ages.

Figure 4-3. Average reading scale scores for students ages 9, 13, and 17, by pages read per day in school and for homework: 2004

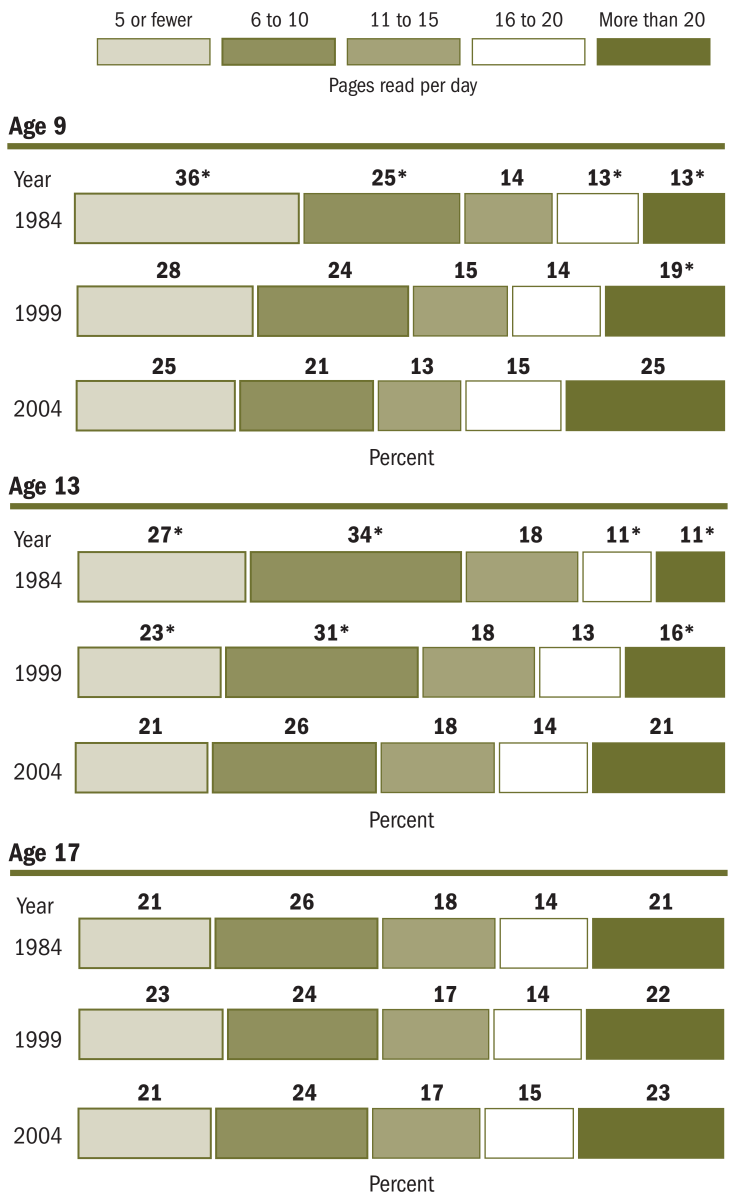


SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

In 2004, at ages 9, 13 and 17 students who indicated that they read 5 or fewer pages a day had lower reading scores than students in any other category; however, for students at ages 9 and 13, there were no differences in the average reading scores among students who read at least 6 pages a day. That is, students who indicated that they read more than 20 pages a day did not have reading scores that were measurably different from students who indicated they read 6–10, 11–15, or 16–20 pages per day. At age 17, there is a more linear relationship between the number of pages read per day and average reading scores. For example, students who read more than 20 pages a day had higher average reading scores than students who read 11–15, 6–10, or 5 or fewer pages a day. Students who selected any one of the four options indicating they read at least 6 pages a day had higher average scores than students who read 5 or fewer pages.

At age 9, the trend over the past 20 years has shown an increase in the number of pages students read for school and homework. Specifically, fewer students indicated that they read 5 or fewer pages in 2004 than in 1984. Likewise, the percentage of students indicating that they read more than 20 pages a day increased from 13 percent in 1984 to 25 percent in 2004. Similarly, a greater percentage of students at age 13 indicated that they read at least 16 pages per day in 2004 than in 1984. The percentage of 13-year-olds indicating they read either fewer than 5 pages or 6–10 pages decreased between 1984 and 2004. At age 17, there were no measurable changes in the percentage of students indicating various numbers of pages read per day over the 20-year period. In 1984, 1999, and 2004, between 21 and 23 percent of 17-year-olds indicated that they read more than 20 pages per day, and another 21 to 23 percent said they read 5 or fewer pages per day.

Figure 4-4. Percentages of students ages 9, 13, and 17, by pages read per day in school and for homework: 1984, 1999, and 2004



*Significantly different from 2004.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1984, 1999, and 2004 Long-Term Trend Reading Assessments.

Reading for Fun

Students at all three age levels were asked, “How often do you…read for fun on your own time?” The possible responses included the following:

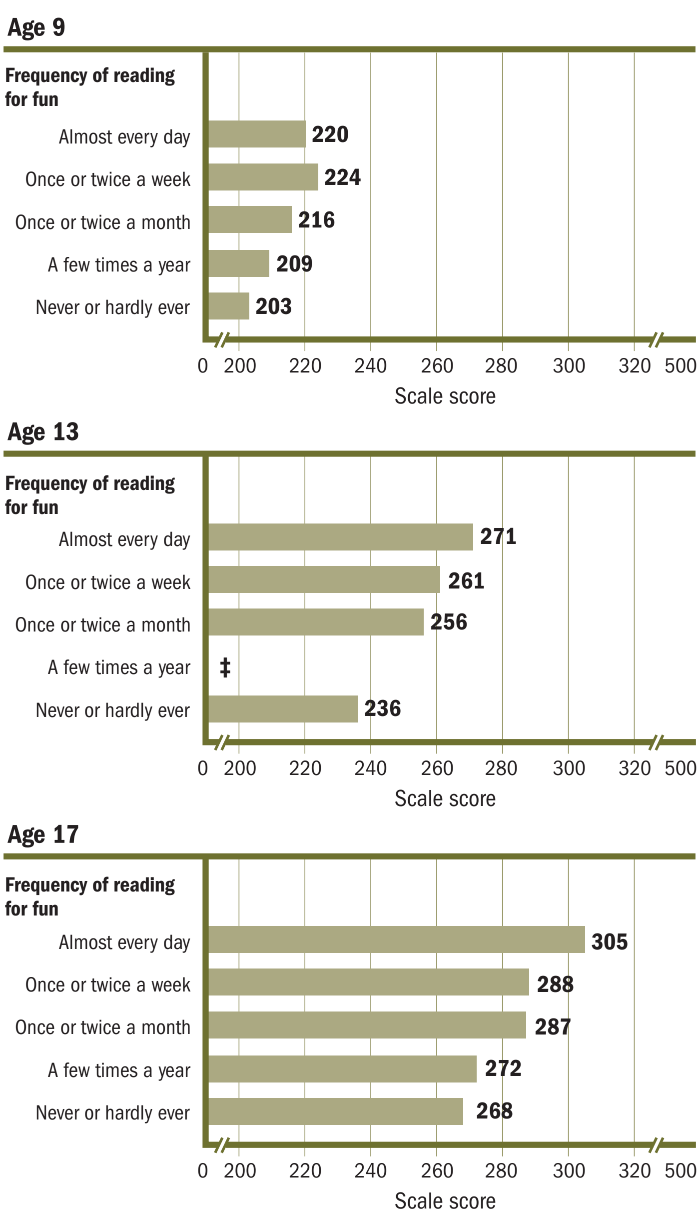
- Almost every day
- Once or twice a week
- Once or twice a month
- A few times a year
- Never or hardly ever

Responses are available for reporting from 1984 through 2004 at all three ages. Figure 4-5 shows the relationship between the amount of time spent reading for fun and average reading scores.

At all three ages, students who indicated that they read for fun almost every day had higher average scores in 2004 than those who said that they never or hardly ever read for fun. Students at all three ages who said that they read for fun once or twice a week also had higher average scores than those who never or hardly ever read for fun. At ages 13 and 17, those who read for fun almost every day had higher average scores than those who read for fun once or twice a week.

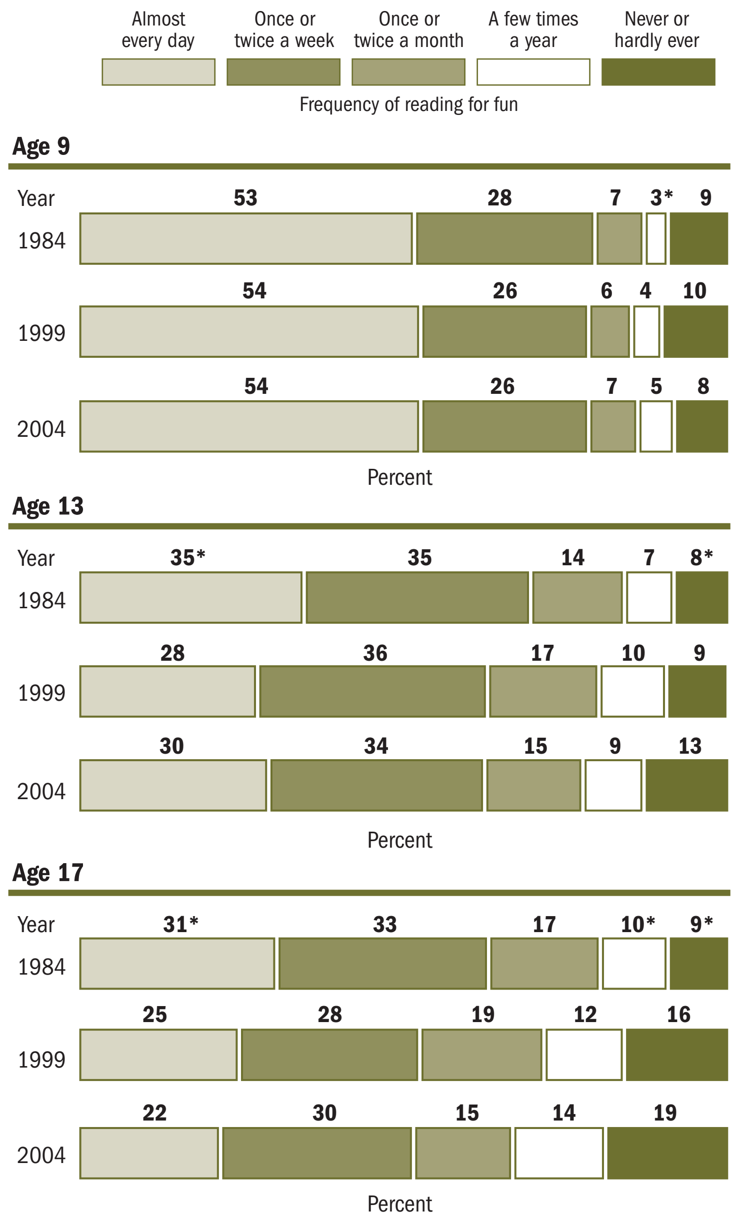
As seen in figure 4-6, at age 9 the only category showing a measurable change during this period was an increase in the percentage of students who indicated that they read a few times a year—up from 3 percent in 1984 to 5 percent in 2004. At ages 13 and 17, the percentage saying they read for fun almost every day was lower in 2004 than in 1984. This trend accompanied an increase over the same 20-year time period in the percentage indicating that they never or hardly ever read for fun.

Figure 4-5. Average reading scale scores for students ages 9, 13, and 17, by frequency of reading for fun: 2004



‡Reporting standards not met. Sample size is insufficient to permit a reliable estimate.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

Figure 4-6. Percentages of students ages 9, 13, and 17, by frequency of reading for fun: 1984, 1999, 2004



*Significantly different from 2004.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1984, 1999, and 2004 Long-Term Trend Reading Assessments.

Contextual Factors Associated With Mathematics

Students responded to several background questions relating to mathematics as they took the long-term trend assessment. This section reports on four types of factors associated with mathematics: course-taking patterns, availability of and amount of time spent on computers in mathematics studies, frequency of homework, and television-watching patterns. Each of these factors is analyzed to determine how it relates to performance in mathematics as measured by the long-term trend assessment and how the responses to these questions have changed over the past two to three decades.

Course-Taking Patterns

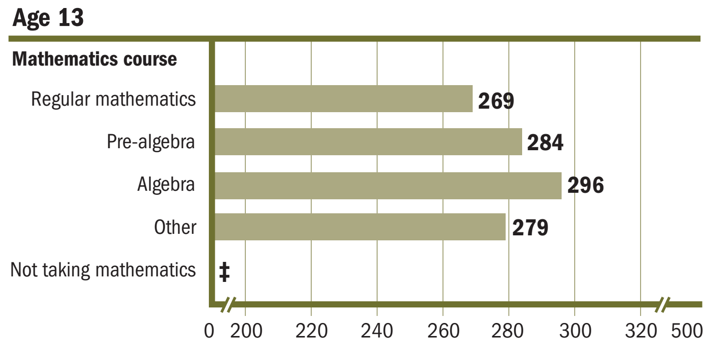
Questions on mathematics courses were given to students in the long-term trend background questionnaire at ages 13 and 17. At age 13, the question read: "What kind of mathematics class are you in this year?" The response options were the following:

- I am not taking mathematics this year.
- Regular mathematics
- Pre-algebra
- Algebra
- Other

In 2004, almost all 13-year-olds said that they were taking some mathematics course, and only 6 percent indicated that they were taking a mathematics class other than the ones listed (see figure 4-8). The remainder of the students at age 13 was split almost evenly among the choices of regular mathematics, pre-algebra, and algebra.

It was not possible to determine any variation in content or difficulty of mathematics classes across schools. As seen in figure 4-7, among those subjects, more advanced mathematics courses were associated with higher scores on the 2004 long-term trend mathematics assessment. That is, students who were in algebra scored higher than those in pre-algebra, who scored higher than those in regular mathematics classes.

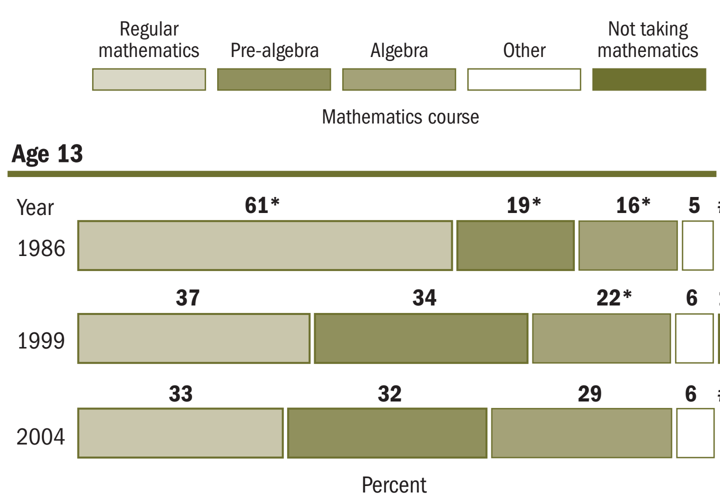
Figure 4-7. Average mathematics scale scores for students age 13, by type of mathematics course: 2004



‡ Reporting standard not met. Sample size is insufficient to permit a reliable estimate.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

Figure 4-8 shows the trends in mathematics course-taking patterns at age 13 from 1986 through 2004. Overall, more 13-year-olds are enrolled in algebra, up from 16 percent in 1986 to 29 percent in 2004—a higher percentage of students than in any previous assessment year. The percentage in pre-algebra has also increased from 19 percent in 1986 to 32 percent in 2004, while the percentage in regular mathematics decreased from 61 percent in 1986 to 33 percent in 2004.

Figure 4-8. Percentage of students age 13, by type of mathematics course: 1986, 1999, and 2004



The estimate rounds to zero.
*Significantly different from 2004.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1986, 1999, and 2004 Long-Term Trend Mathematics Assessments.

At age 17, the question was worded differently to focus on all mathematics classes taken. The question read: "Counting what you are taking now, have you ever taken any of the following mathematics courses?" Students indicated that they had or had not taken each of the following subjects:

- General, business, or consumer mathematics
- Pre-algebra or introduction to algebra
- First-year algebra
- Second-year algebra
- Geometry
- Trigonometry
- Pre-calculus or calculus

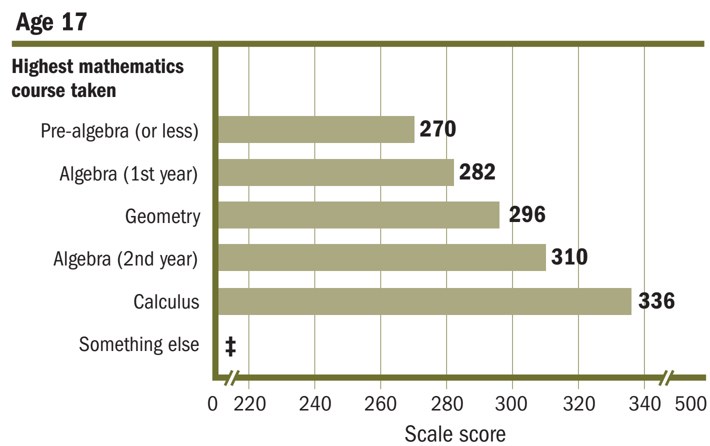
The most advanced mathematics class checked by the students was recorded as the highest level of mathematics taken.

The majority of students at age 17 (53 percent) indicated that the highest level of mathematics they had taken was second-year algebra (figure 4-10). Only 4 percent had not yet taken algebra, and 17 percent had taken calculus. As seen in figure 4-9, the highest level of mathematics taken was positively associated with average scores on the 2004 long-term trend assessment. That is, students who had taken calculus had a higher average score than those whose highest mathematics class was second-year algebra. Those who took algebra II had a higher average score than those whose highest class was geometry, and geometry students outperformed algebra I students. Pre-algebra students had a lower average score in mathematics than students who had taken any mathematics course beyond pre-algebra.

How to interpret this graphic . . .

Each variable in this section has two graphics. The first graphic, such as figure 4-7, shows the different categories of responses with horizontal bars showing the average score for each category. The second graphic, such as figure 4-8, shows the percentage of students selecting each response category in the first year the question was asked and in 1999 and 2004. The percentages should add up to 100 percent of assessed students but may not be exact due to rounding.

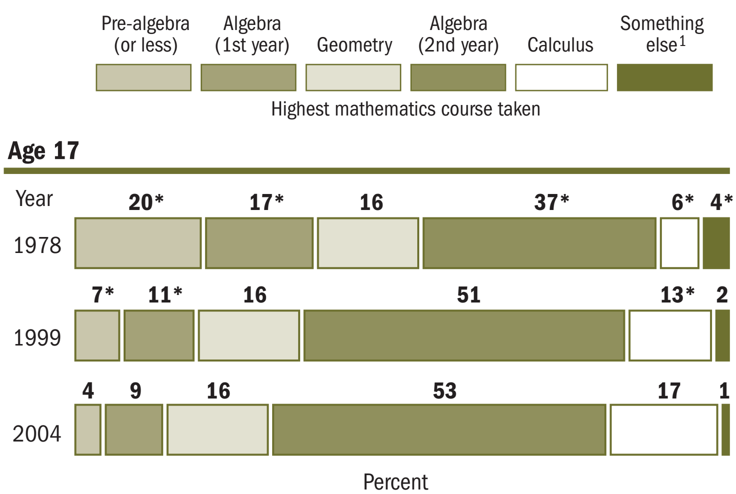
Figure 4-9. Average mathematics scale scores for students age 17, by highest mathematics course taken: 2004



‡ Reporting standards not met. Sample size is insufficient to permit a reliable estimate.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

Figure 4-10 shows the trend in course-taking patterns of 17-year-olds from 1978 through 2004. As with 13-year-olds, the trend at age 17 is for more advanced course-taking in mathematics. A greater percentage of 17-year-olds indicated they were taking or had taken calculus in 2004 than in any previous assessment year. The percentage taking second-year algebra as their highest class also increased from 37 percent in 1978 to 53 percent in 2004. Conversely, the percentage of students who indicated that the highest level of mathematics they had taken by age 17 was pre-algebra or algebra was lower in 2004 than in 1978.

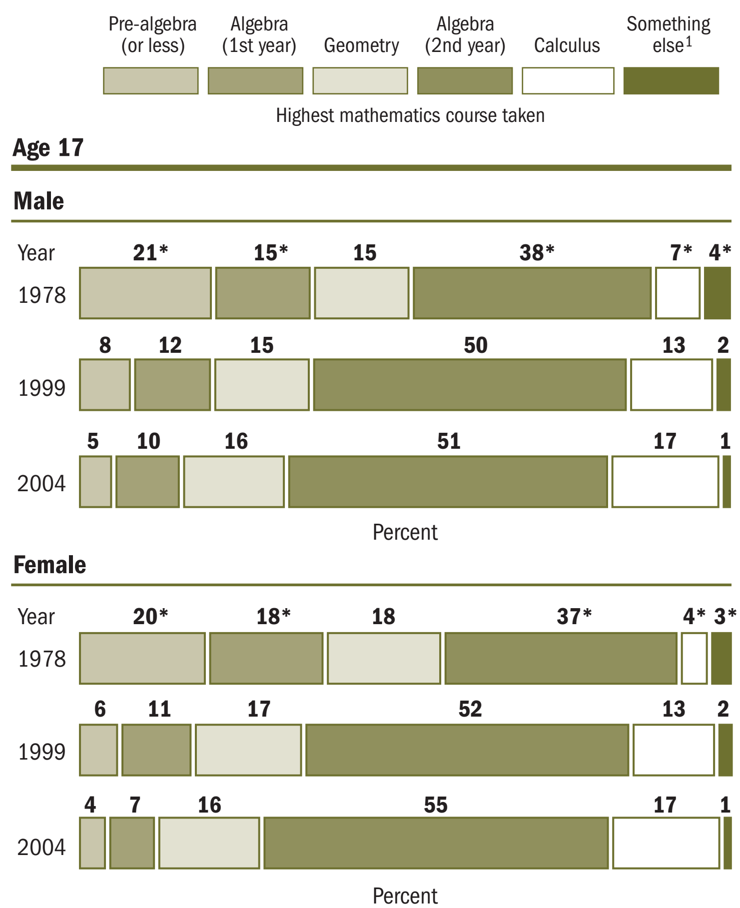
Figure 4-10. Percentage of students age 17, by highest mathematics course taken: 1978, 1999, and 2004



*Significantly different from 2004.
[1] "Something else" implies that students checked a series of courses that did not follow a logical course-taking pattern.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1978, 1999, and 2004 Long-Term Trend Mathematics Assessments.

Figure 4-11 shows students' course-taking patterns broken down by gender to analyze whether male students reported taking more advanced courses than female students. Almost no measurable differences by gender were evident in 2004. Similar percentages of males and females (17 percent each) took calculus. Although the percentages in 2004 did not differ measurably from those in 1999, more males and females took calculus in 2004 than in 1978, when 4 percent of female and 7 percent of male 17-year-olds said their highest mathematics class was calculus. In 2004, 55 percent of females and 51 percent of males at age 17 indicated the highest level of mathematics they had taken was second-year algebra, up from 37 and 38 percent, respectively, in 1978.

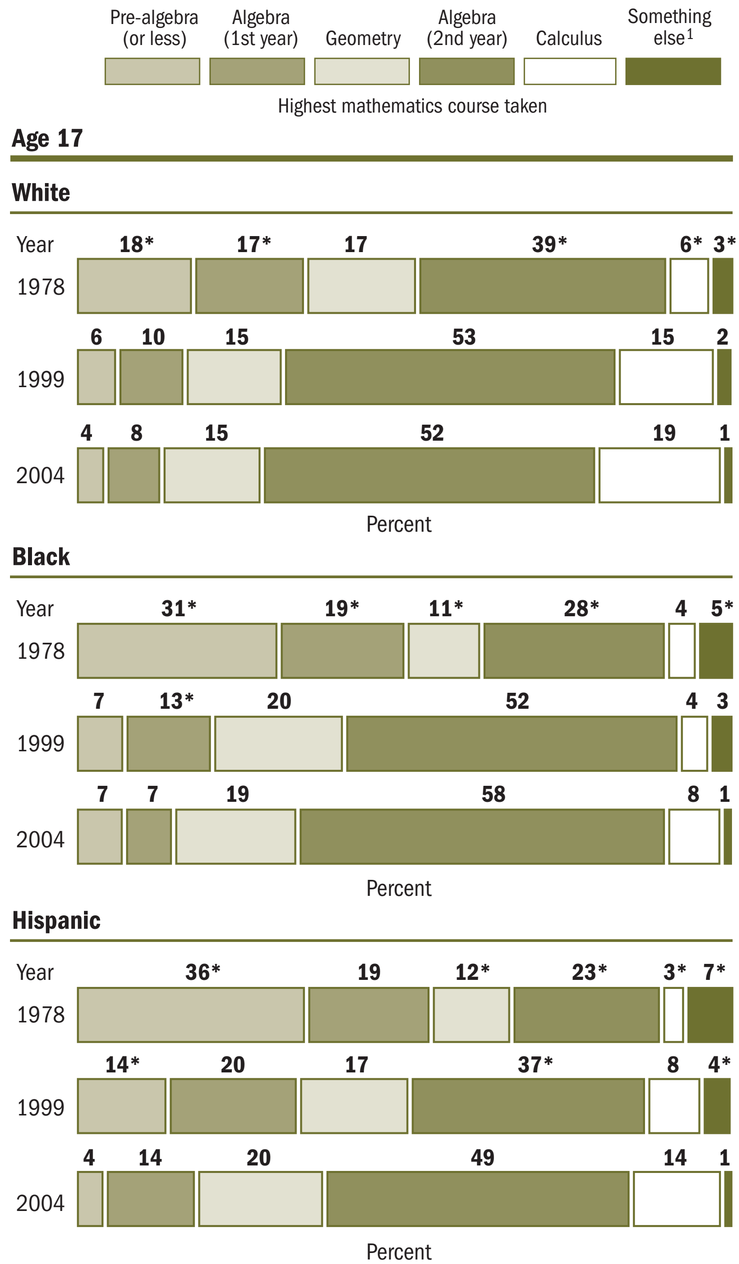
Figure 4-11. Percentage of students age 17, by gender and highest mathematics course taken: 1978, 1999, and 2004



*Significantly different from 2004.
[1] "Something else" implies that students checked a series of courses that did not follow a logical course-taking pattern.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1978, 1999, and 2004 Long-Term Trend Mathematics Assessments.

Figure 4-12 shows the highest mathematics course taken at age 17, by racial/ethnic group. In 2004, a higher percentage of White students took calculus (19 percent) compared to Black students at the same age (8 percent). At 14 percent, the percentage of Hispanic students taking calculus was not measurably different from either group. The pattern of higher-level course-taking was seen across all three racial/ethnic groups as a greater percentage of students in all three racial/ethnic groups took high-level courses in 2004 compared to 1999 or 1978. A greater percentage of Black, Hispanic, and White students indicated their highest course was second-year algebra in 2004 than in 1978. In each racial/ethnic group, a smaller percentage of students in 2004 compared to 1978 indicated that their highest mathematics course at age 17 was pre-algebra.

Figure 4-12. Percentage of students age 17, by race/ethnicity and highest mathematics course taken: 1978, 1999, and 2004



*Significantly different from 2004.

[1] "Something else" implies that students checked a series of courses that did not follow a logical course-taking pattern.

NOTE: Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1978, 1999, and 2004 Long-Term Trend Mathematics Assessments.

Availability and Use of Computers

Students at ages 13 and 17 were asked several questions regarding their access to and use of computers. From these questions, three factors relating to computer availability and usage discussed in this section were derived. The first question asked, "Have you ever studied mathematics through computer instruction?" and had the following response options:

- Often
- Sometimes
- Never
- I don't know.

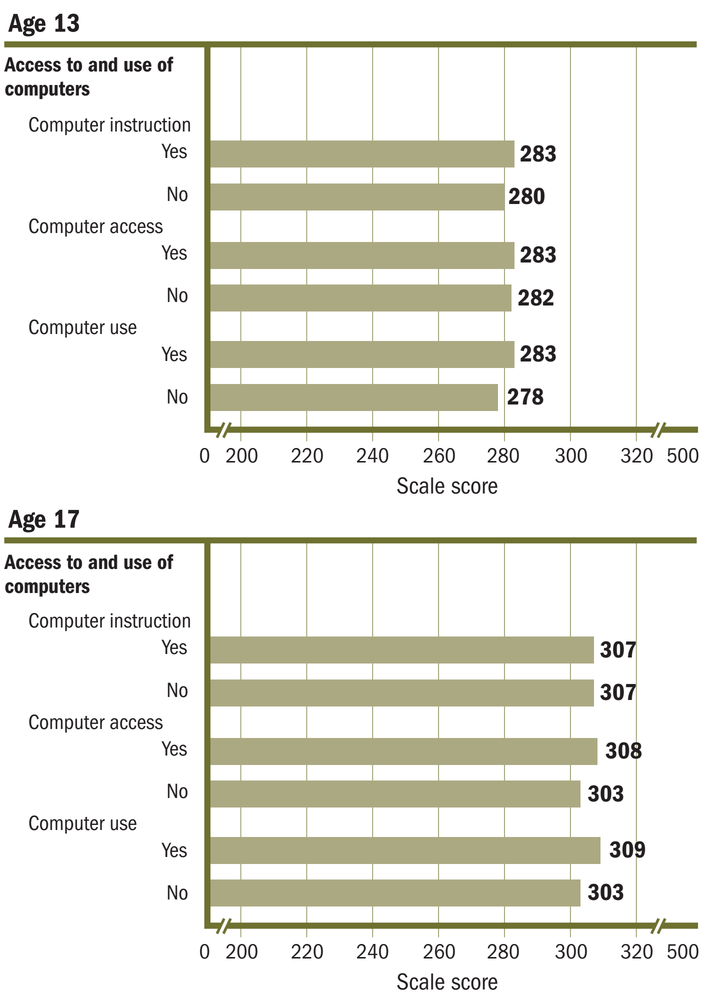
The first two categories—"often" and "sometimes"—were combined to indicate a positive response to the question. The second question asked, "Do you have access to a computer terminal in your school for learning mathematics?" and had the same response options as the previous question. The third question asked, "Have you ever used a computer to solve a mathematical problem?" and had the following response options:

- Yes
- No
- I don't know.

Figure 4-13 shows the relationship between these three questions and students' average scores on the long-term trend mathematics assessment at ages 13 and 17. Figure 4-14 shows the percentage of students at ages 13 and 17 responding positively to each question. In the 2004 assessment, 57 percent of 13-year-olds indicated that they had access to a computer at their school (either often or sometimes), and 69 percent said that they had used a computer to solve a mathematical problem (either often or sometimes). Just under one-half (48 percent) indicated that they had studied mathematics using computers. However, there were no measurable differences in mathematics scores between 13-year-olds who responded positively and those who responded negatively to any of these questions in 2004.

At age 17, the responses showed a similar pattern—57 percent said that they had access to a computer at their school, and 70 percent said they had used a computer to solve a mathematical problem. Because computer location was not specified in the question about using a computer to solve a mathematical problem but was specified in the question on access, it makes sense that more students indicated that they had used a computer than had access to a computer. Thirty-six percent responded that they had studied mathematics using computers. A relationship between computer access and use and long-term trend mathematics scores was seen at age 17. Students who indicated that they had access to a computer at school scored 5 points higher on average than students who did not have access. Likewise, students who responded that they had used a computer to solve a mathematical problem scored 6 points higher on average than students who had not used a computer for that purpose. There was no measurable difference in average mathematics scores for students based on whether or not they had studied mathematics through computer instruction.

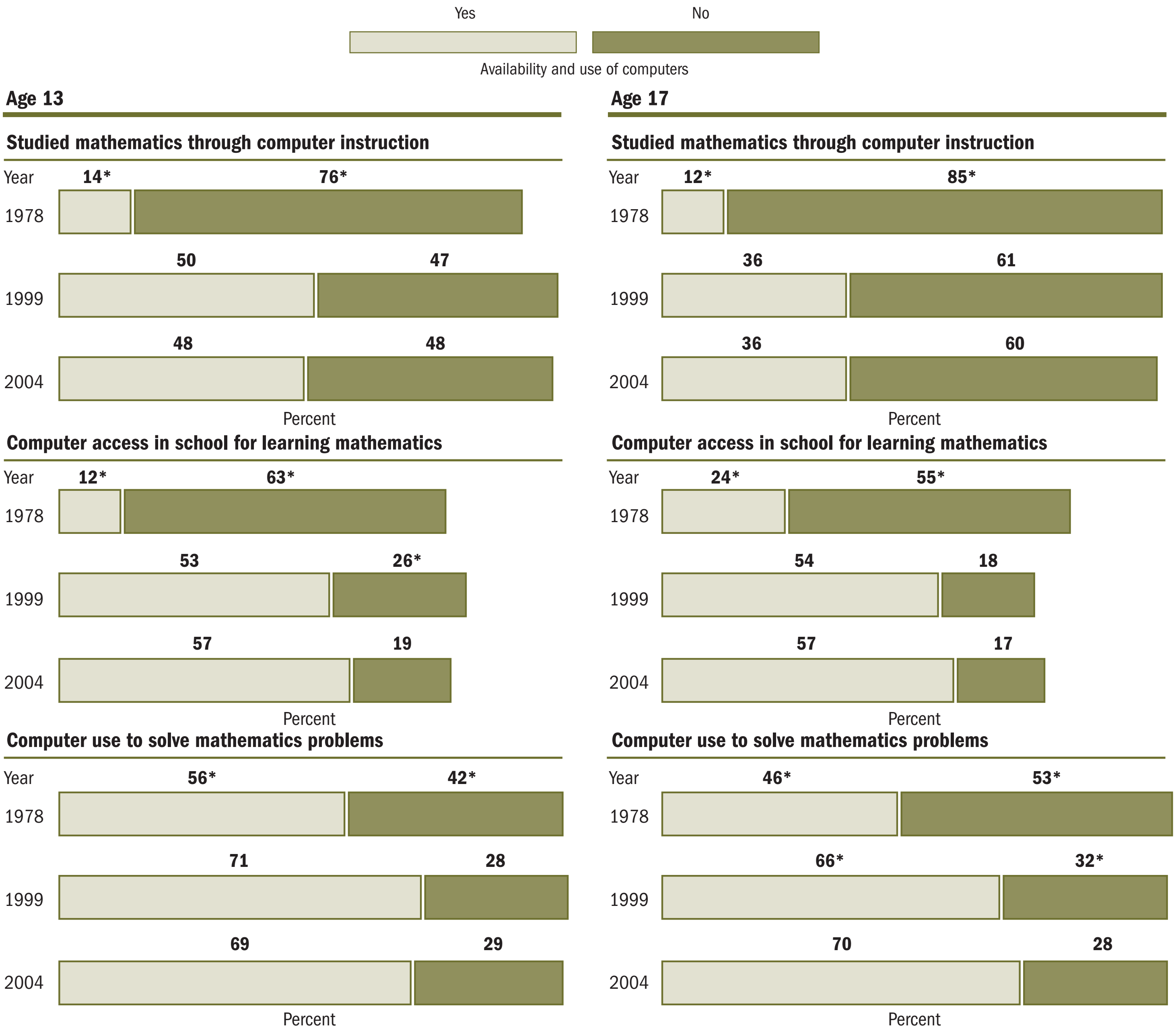
Figure 4-13. Average mathematics scale scores for students ages 13 and 17, by access to and use of computers for mathematics: 2004



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

Figure 4-14 shows the trends in computer access at school and usage in learning mathematics for both 13- and 17-year-olds. Although few differences were seen between 1999 and 2004 at age 13, measurable increases in the percentages of students with access to computers at school and of those who used computers for learning mathematics were seen between 1978 and 2004. The percentage of 13-year-olds with access to computers in schools increased from 12 percent in 1978 to 57 percent in 2004. The percentage of students receiving instruction in mathematics using computers at age 13 also showed a measurable increase, from 14 percent in 1978 to 48 percent in 2004. Similar increases were also seen at age 17, where the percentage of students with access to a computer in school increased by 33 percentage points between 1978 and 2004, from 24 to 57 percent. The percentage of 17-year-olds using a computer to solve mathematics problems increased from 46 percent in 1978 to 66 percent in 1999 to 70 percent in 2004. Small, but statistically significant, increases in the percentage of 17-year-olds studying mathematics through computer instruction occurred between 1978 and 2004 at both ages.

Figure 4-14. Percentages of students ages 13 and 17, by availability and use of computers: 1978, 1999, and 2004



*Significantly different from 2004.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1978, 1999, and 2004 Long-Term Trend Mathematics Assessments.

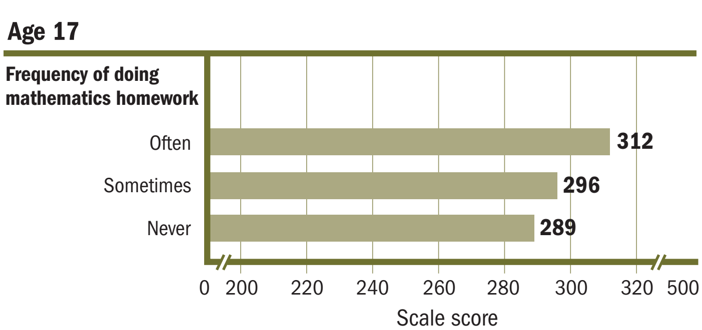
Homework

Students at age 17 were asked in the background questionnaire about the frequency with which they did homework. Specifically, the question asked, "How often did you do these activities in your high school mathematics courses?" Included in the list of activities was, "Do mathematics homework." The possible response options were the following:

- Often
- Sometimes
- Never

Figure 4-15 shows the average mathematics score as related to the frequency of doing mathematics homework at age 17. The majority (73 percent) of 17-year-olds indicated that they often did mathematics homework in 2004 (figure 4-16). The frequency of doing mathematics homework was associated with the average score on the 2004 long-term trend mathematics assessment. Those who often did mathematics homework had a higher average score in mathematics (312) than those who sometimes (296) or never (289) did mathematics homework. Likewise, those who indicated that they sometimes did mathematics homework had a higher average score than those who said they never did mathematics homework.

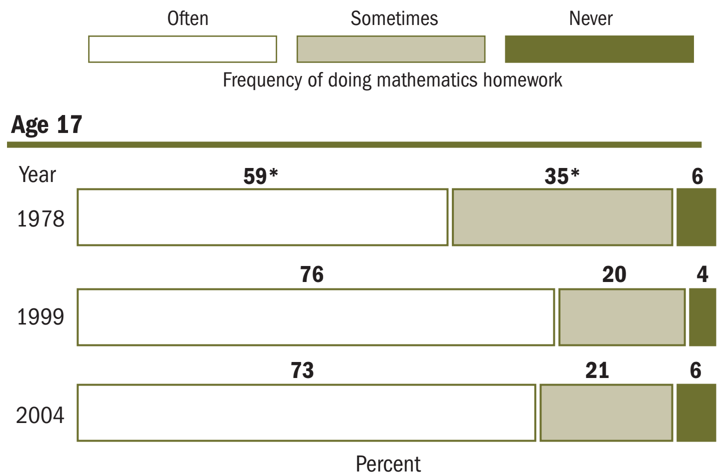
Figure 4-15. Average mathematics scale scores for students age 17, by frequency of doing mathematics homework: 2004



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

Figure 4-16 shows how the frequency of doing mathematics homework has changed from 1978 and 1999 to 2004 at age 17. There were no measurable differences in the percentage of students reporting various frequencies of doing mathematics homework between 1999 and 2004, but the percentage of students reporting that they often did mathematics homework increased by 14 percentage points between 1978 and 2004. The percentage of 17-year-olds indicating they sometimes did homework decreased by about the same amount. No measurable differences were found in the percentage of students who indicated they never did mathematics homework.

Figure 4-16. Percentage of students age 17, by frequency of doing mathematics homework: 1978, 1999, 2004



*Significantly different from 2004.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1978, 1999, and 2004 Long-Term Trend Mathematics Assessments.

Television Watching

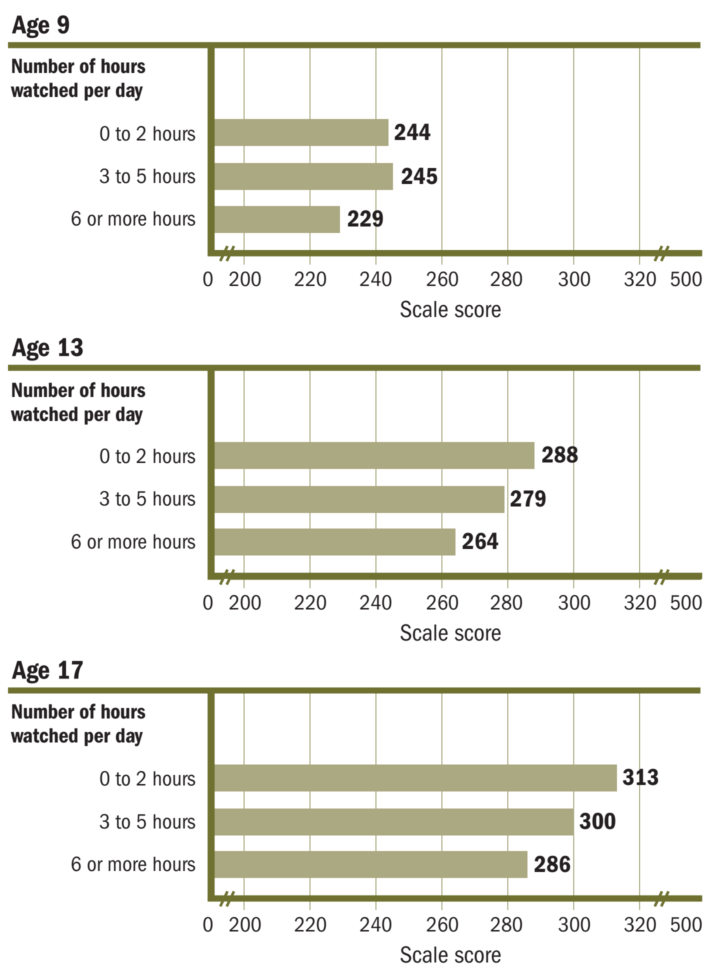
Examining television-watching habits provides information on the home environment, and specifically focuses on an activity that may compete with time spent on schoolwork. Students at all three ages were asked a question about their television-watching habits. Specifically, they were asked, "How much television do you usually watch each day?" The possible responses were the following:

- None
- 1 hour or less
- 2 hours
- 3 hours
- 4 hours
- 5 hours
- 6 hours or more

These options were then collapsed into three reporting categories: 0 to 2 hours, 3 to 5 hours, 6 or more hours. Information on television-watching habits is available for all assessment years from 1978 through 2004 for age 17 and from 1982 through 2004 for ages 9 and 13.

Figure 4-17 shows the average score on the 2004 long-term trend mathematics assessment by the amount of television watching for all three ages, and figure 4-18 shows the percentage of students watching varying amounts of television over time. In 2004, about half of 9-year-olds (51 percent) reported that they watched 0 to 2 hours of television each day. There were no measurable differences in average mathematics score at age 9 between students who watched 0 to 2 hours and those who watched 3 to 5 hours, but students in both these categories had higher average scores than students who watched 6 or more hours of television each day, 244 and 245 compared to 229, respectively. At age 13, students were about evenly split between those who watched 0 to 2 hours (45 percent) and those who watched 3 to 5 hours (44 percent), and 11 percent reported watching 6 or more hours of television each day. Thirteen-year-olds who reported watching 0 to 2 hours had higher average mathematics scores than those who watched 3 to 5 hours, and both groups had higher average scores than students who watched 6 or more hours of television each day. At age 17, the majority of students (58 percent) reported watching 0 to 2 hours of television each day, and 6 percent reported watching 6 or more hours per day. As with 13-year-olds, more television watching was associated with lower mathematics scores, as those watching 0 to 2 hours had higher average mathematics scores than those watching 3 to 5 hours, and both groups had higher average scores than students watching 6 or more hours of television each day.

Figure 4-17. Average mathematics scale scores for students ages 9, 13, and 17, by amount of daily television watching: 2004



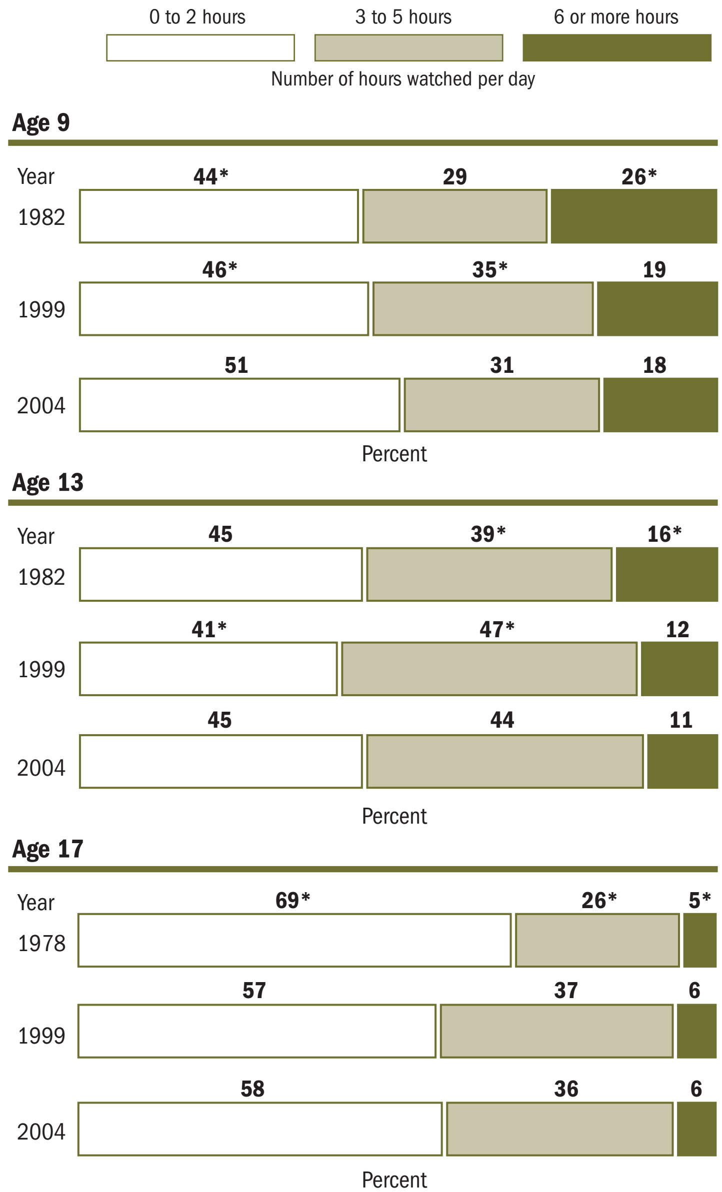
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

Examining trends in television watching over time shows that, overall, 9-year-olds are watching less television in 2004 than they were in 1982, while 17-year-olds appear to be watching more television in 2004 than they were in 1978. As seen in figure 4-18, more 9-year-olds reported that they watched 0 to 2 hours of television in 2004 compared to 1982, while fewer reported that they watched 6 or more hours. At age 13, the percentage of students watching 0 to 2 hours of television in 2004 was not measurably different from 1982, fewer students reported watching 6 or more hours of television, and more reported watching 3 to 5 hours in 2004 than in 1982. At age 17, fewer students reported watching 0 to 2 hours of television in 2004 than in 1978, and more students reported watching 3 to 5 hours and 6 or more hours. It is important to note, however, that, as the question is worded, students may not be reporting the time they spend watching movies or playing video games using the television. The question only asks about the amount of time spent watching television.

Summary

This chapter has provided a snapshot of how contextual variables may relate to performance in reading and mathematics. School variables, such as homework, pages read, mathematics course-taking, and access to and use of computers were explored, as were some home variables, including reading at home and watching television. In most cases there were relationships between these contextual factors and average scores, and trends over time were seen. However, readers again are cautioned against making causal inferences about a contextual factor producing a high or low score. Instead, these data should be used as a starting point to guide future research.

Figure 4-18. Percentages of students ages 9, 13, and 17, by amount of daily television watching: 1978, 1982, 1999, and 2004



*Significantly different from 2004.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1978, 1982, 1999, and 2004 Long-Term Trend Mathematics Assessments.

THIS PAGE INTENTIONALLY LEFT BLANK.

The 2004 long-term trend assessment marks the first time accommodations were provided for students with disabilities and English language learners in the long-term trend. The results from the modified assessment are now more inclusive of all students and reflect current assessment practices.

Chapter 5
Comparison of Bridge and Modified Assessments

Several changes were made to the long-term trend assessment in 2004 to align it with the best current assessment practices and with policies applicable to the NAEP main assessments. According to the new policy of the National Assessment Governing Board (NAGB), reading and mathematics are to be assessed by both the long-term trend instruments and the main NAEP instruments, but science and writing will be assessed only in main NAEP (http://www.nagb.org/release/policy_statement.doc). As a result, changes were needed to remove the sets, or blocks, of questions for science and writing, which had been intermixed with the reading and mathematics blocks in the long-term trend assessment instruments.

The changes provided an opportunity to bring other aspects of the assessment up to date. Considerable progress in testing theory has been made since the late 1960s, when these assessments were first designed, and the 2004 administration provided an opportunity to bring these improvements in scoring and scaling to the long-term trend assessments. In addition, since 1996, main NAEP assessments have been providing accommodations to allow more students with disabilities and students who are not fluent in English to participate. Traditionally, the long-term trend assessments have not provided such accommodations. However, in 2004, accommodations were provided, allowing NAEP to assess a greater proportion of students.

Thus, two assessments were given in 2004—a modified assessment that contained many changes from previous assessments, and a bridge assessment that was used to link the modified assessment to the 1999 assessment so the trend line could be continued. Approximately 14,000 students took the bridge assessment in each subject (28,000 total), while 24,000 took the modified reading assessment and 22,000 took the modified mathematics assessment. Results from the bridge assessment, which replicated the previous long-term trend assessment procedures, were reported in chapters 2–4. This chapter discusses the changes made in the modified assessment, the specifications of the bridge study, and the results of the two assessments.

Specific Changes Made for the 2004 Long-Term Trend Assessment

In addition to removing science and writing items and providing accommodations for students with disabilities and English language learners (ELL), the NAEP assessment instruments for the 2004 modified assessment were changed in the following ways:

- *Replacing items.* Many of the items in the long-term trend assessment were written in the late 1960s. Given changes in context over the past four decades, several items needed to be restructured or replaced with items more in line with current contexts. For example, reading passages that discussed outdated issues were replaced with more current passages.
- *Restructuring background questions.* Many of the background questions were eliminated because they appeared intrusive or outdated or were no longer supplying useful data for the analyses. In addition, methods for collecting student demographic data were also updated. For instance, race/ethnicity previously was determined by the test administrator, by observation. In 2004, the student's race/ethnicity was assigned based on student records supplied by schools. Although the observed data were still used to maintain the trend line in the bridge assessment (reported in chapter 3), comparisons between the two 2004 assessments were based on the school-reported race/ethnicity data. Students were also asked to indicate their racial/ethnic background on the background questionnaire as a second source of information.
- *Moving all background questions to the end of the administration time.* Previously, background questions were intermixed with the assessment questions. That is, a student would read a passage, answer the questions associated with that passage, and then answer questions about the student's own reading habits. The same was true for mathematics. In 2004, all questions pertaining to student demographics, reading habits, frequency of homework, and other contextual situations were gathered into a single section and given to the students after they had answered all assessment questions in either subject.
- *Eliminating "I don't know" as a response option for multiple-choice items.* In mathematics, the multiple-choice items had four possible answers to the question and a fifth option, "I don't know." Because this fifth option provided no useful information and represented an antiquated assessment technique, it was eliminated in the modified version of the 2004 assessment.
- *Eliminating audio paced tapes.* Use of an audio tape, which paced students during the assessment session so that they were at the same place in the test booklet at the same time, was discontinued for mathematics in 2004. That is, in the 1999 and previous assessments, students taking the mathematics portion would listen to an audio tape that spoke each question aloud, paused to allow the student time to respond, then spoke the next question aloud. The reading trend assessment eliminated the use of a paced tape in 1984.
- *Using assessment booklets that pertain only to a single subject area.* In the past, a single assessment booklet may have contained both reading and mathematics items. Science and writing items were also intermingled with the reading and mathematics items. In the 2004 modified assessment, students received a booklet that either contained only reading questions or only mathematics questions.

Other changes to the reading and mathematics trend assessments for 2004 included changing the number of items and the number of booklets used in the assessment. These changes are discussed in more detail in the following sections and in appendix A.

Changes to the Reading Assessment

Changes were made to the number of items and to the organization of the assessment booklets for reading. The 2004 modified trend instrument contained blocks with items used in the 1999 trend assessment (although reconfigured from the 1999 design), blocks with new pilot-tested items, and blocks with new items that had not been pilot tested. Items in this latter category were pilot tested for future assessments, since one aim of the modified assessment is to maintain the trend line while releasing items to the public. Pilot-tested items were not included in the score calculations for 2004, but student performance on these items will be used to determine which items should be incorporated into future assessments.

The 2004 bridge assessment maintained the same format as the 1999 assessment, with one exception. Each student received three blocks of questions. If the first or second block of questions in the 1999 design contained science or writing items, it was left as it was to preserve the context. However, if the last block contained science or writing items, it was replaced with a block of new reading or mathematics items.

In contrast, the 2004 modified assessment contained only blocks of reading or mathematics items. Because each assessment booklet in the modified design contained questions from only one subject area, the design is called a "focused" design. In previous administration years, more than one subject was assessed in each booklet, so the 2004 modified assessment marked the first focused design for the long-term trend assessment.

Overall, there were 10 blocks of questions arranged in 6 different booklets for each age sample of the bridge study. In the modified trend assessment, 20 different assessment booklets in all were administered to each age sample.

Table 5-1 shows the changes in the number of items from the bridge assessment (and previous years' assessments) to the modified assessment. Both the bridge and modified assessments used the same number of passages: 39, 42, and 36 at ages 9, 13, and 17, respectively (data not shown), but there are fewer items in the modified assessment. Because the modified assessments only assess one subject, fewer blocks were needed.

Table 5-1. Total number of questions of each format administered in the bridge and modified reading assessments, by age: 2004

	Bridge assessment			Modified assessment		
Age	Total	Multiple-choice	Constructed-response	Total	Multiple-choice	Constructed-response
9	110	105	5	91	87	4
13	137	130	7	106	99	7
17	125	117	8	104	96	8

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

Changes to the Mathematics Assessment

As with the reading assessment, the booklets used in the 2004 modified mathematics trend assessment consisted of blocks that had been administered in 1999 and previous trend assessments, blocks with new pilot-tested questions, and blocks with new questions that had not been pilot tested. Blocks that required calculators were eliminated. Pilot-tested items were not included in the score calculations for 2004, but student performance on these items will be used to determine which items should be incorporated into future assessments. For the bridge assessment, two booklets were constructed for age 9, two for age 13, and one for age 17. The modified assessment included six booklets at each age level. Changes were made to the number of items, as shown in table 5-2.

Table 5-2. Total number of questions of each format administered in the bridge and modified mathematics assessments, by age: 2004

	Bridge assessment			Modified assessment		
Age	Total	Multiple-choice	Constructed-response	Total	Multiple-choice	Constructed-response
9	119	91	28	140	103	37
13	172	144	28	166	129	37
17	121	102	19	162	128	34

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

Bridge Study

The changes to the long-term trend instruments in 2004 were intended to increase the validity of the results obtained while maintaining the integrity of the long-term trend assessments. It was important to ensure that any changes in assessment results could be attributed to actual changes in student performance rather than to changes in the assessments. A special bridge study was conducted in 2004 to evaluate how changes to the assessment design and administration procedures would affect assessment results. The bridge study involved the administration of the two assessments to two randomly assigned groups of students. One assessment, the bridge assessment, used the same assessment questions in reading and mathematics given under the same conditions as in previous years. The other assessment, the modified assessment, represented the new design with the changes discussed earlier.

The remainder of this chapter examines the results of the bridge and modified assessments after they have been linked together. Briefly, the two assessments were linked by first removing all accommodated students' data from the results and then setting the average scale scores to be equal for both assessments. Then the data for accommodated students were reintroduced, and the average scale scores were recalculated. (See appendix A for a complete explanation of how the assessments were linked.) Comparing the results from the two assessments, given in the same year to equivalent groups of students, provides an indication of whether there were any significant changes in results caused by the changes in the assessment. Although one might expect the results of the modified assessment to be lower than the results of the bridge assessment because greater percentages of ELL students and students with disabilities were assessed, the differences should be small. It is important to examine the magnitude of these differences and to determine whether the results for different groups of students are affected differentially.

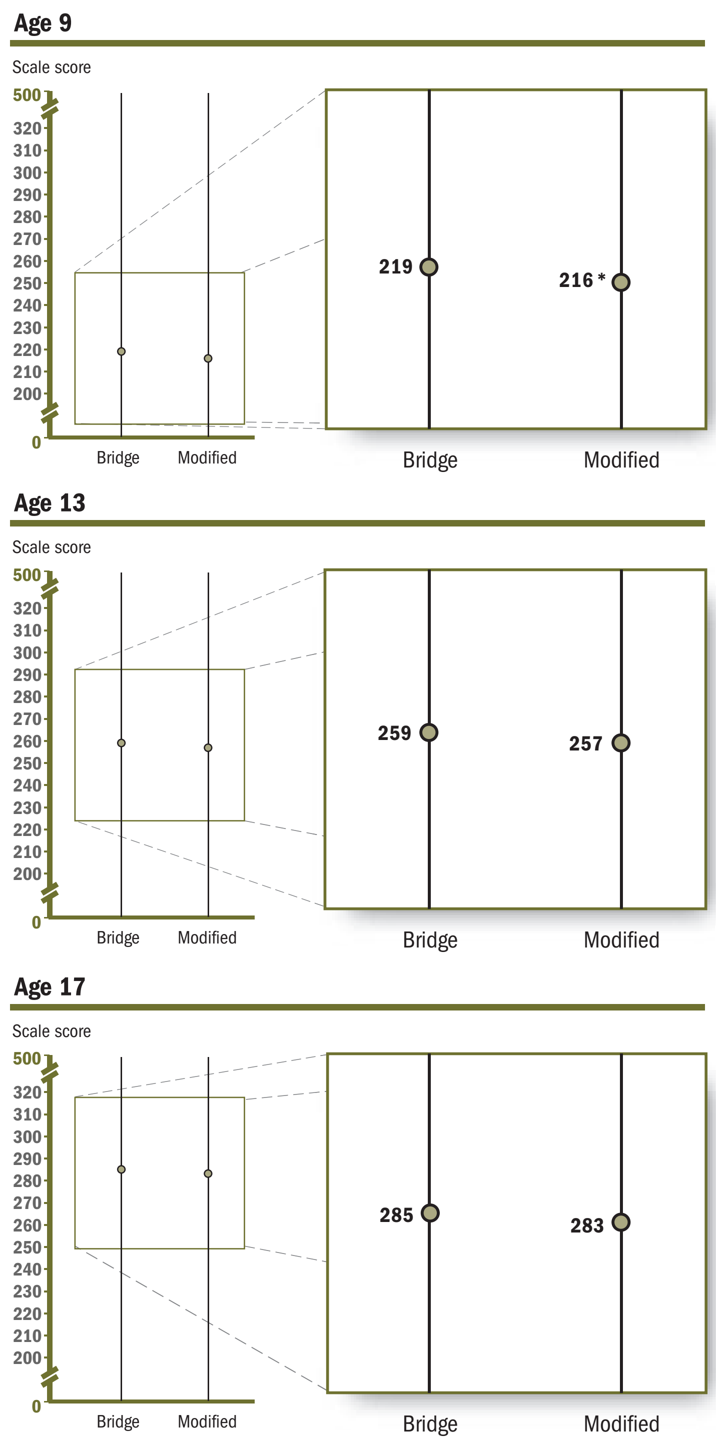
Comparison of Bridge and Modified Results for Reading

Almost no measurable differences were found between the average reading scores of students who took the bridge assessment and the average scores of those who took the modified assessment at ages 13 and 17. As seen in figure 5-1, at age 13 students who took the bridge assessment appear to score 2 points higher, on average, than students who took the modified assessment, but this difference was not statistically significant. However, at age 9 the average score of the students taking the bridge assessment was 219, 3 points higher than the average score of students taking the modified assessment, and this difference was statistically significant. Again, this difference is not unexpected, considering that the group of students taking the modified assessment was more inclusive, since accommodations were allowed on the modified assessment.

How to interpret this graphic . . .

The figures in this chapter show the average score for the 2004 bridge assessment compared to the average score for the 2004 modified assessment. Only one point is presented for each assessment. The full scale is shown in the smaller graphic, and the area of focus is enhanced to make it easier to read any group differences such as those between males and females in figures 5-2 and 5-5 or between different racial/ethnic groups in figures 5-3 and 5-6.

Figure 5-1. Average reading scale scores for students ages 9, 13, and 17 for bridge and modified assessments: 2004

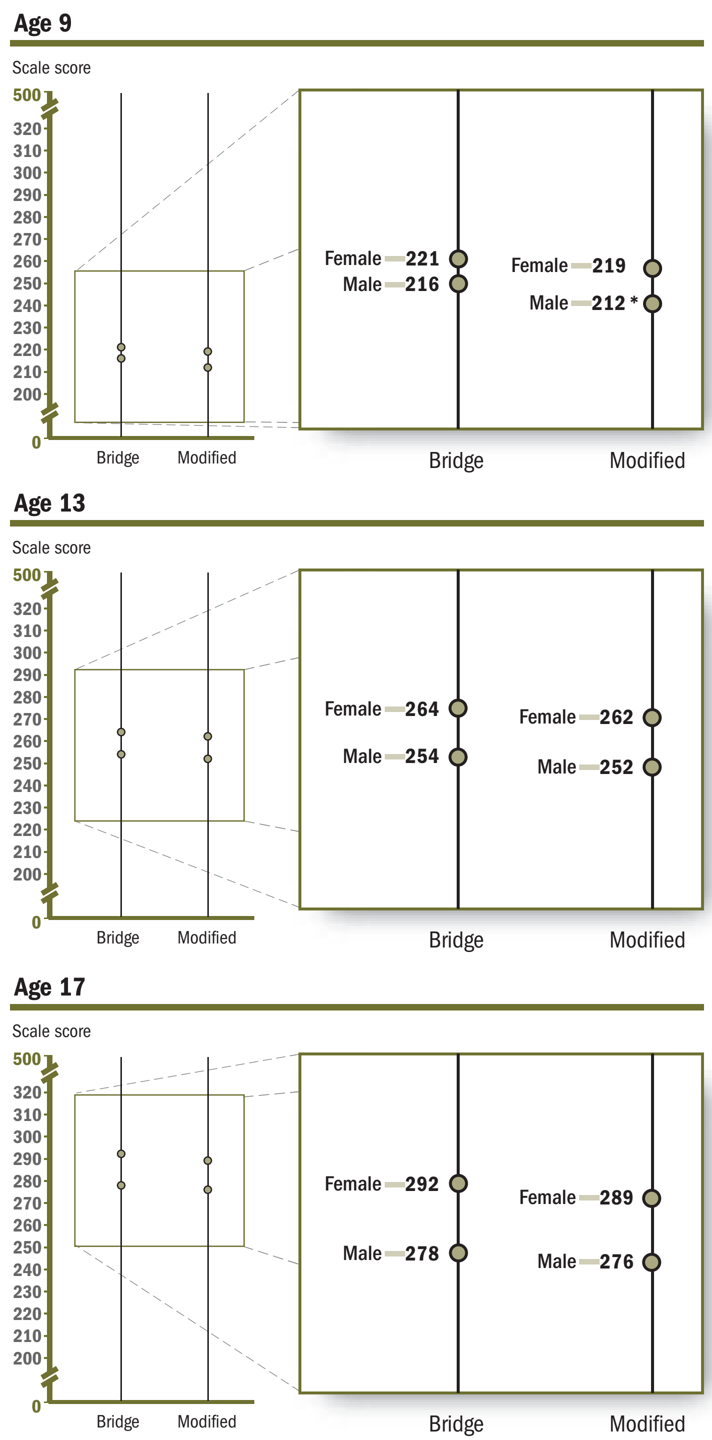


*Significantly different from 2004 bridge assessment.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

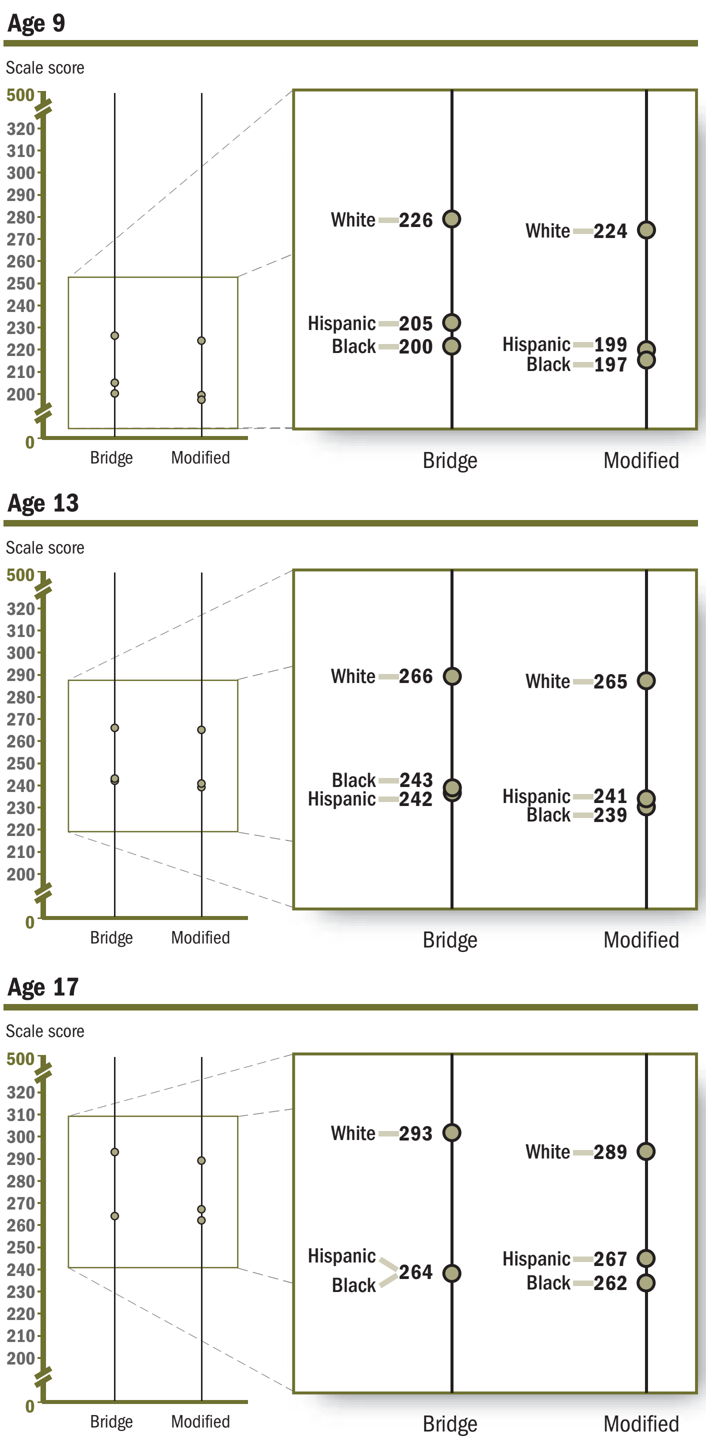
Even if the modifications did not affect students' results overall, they might affect some specific group of students. In order to examine whether the modification affected the results for any specific student group, comparisons were made between the results for each group on the modified assessment and the results for the corresponding group on the bridge assessment—for example, the results for female students on the two assessments were compared (figure 5-2). No measurable differences were found for any of the groups at any age with one exception. At age 9, the average score of male students taking the bridge assessment was 4 points higher than the score of students taking the modified assessment. Figure 5-3 shows the results of the two assessments by race/ethnicity. No measurable difference was found between the bridge and modified reports for any racial/ethnic group. As discussed earlier in this chapter, because the long-term trend assessment first used observational data to assign students to different race/ethnicity categories, that practice was maintained through 1999 and in the 2004 bridge study. However, for the 2004 modified assessment, those data have been supplemented with school records. For the analysis of differences by race/ethnicity, the information from the school roster was used to assign each student to a category. For the purposes of the comparisons in this chapter, school reports of race/ethnicity were used to report results for both the bridge and modified assessments. Future long-term trend assessments also will use school records to assign students to the various race/ethnicity categories.

Figure 5-2. Average reading scale scores for students ages 9, 13, and 17 for bridge and modified assessments, by gender: 2004



*Significantly different from 2004 bridge assessment.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

Figure 5-3. Average reading scale scores for students ages 9, 13, and 17 for bridge and modified assessments, by race/ethnicity: 2004



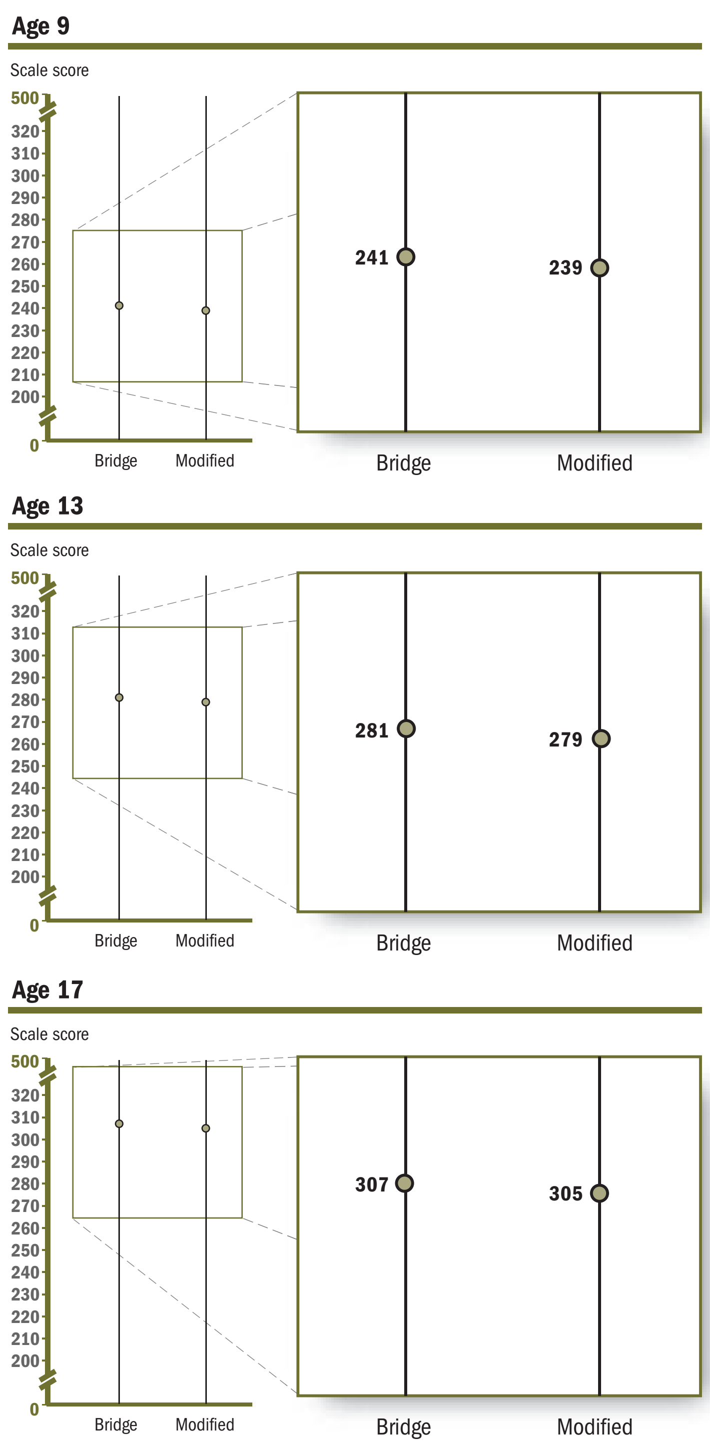
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

In addition to examining gender and race/ethnicity, other analyses compared scores across geographic region, community type, and school type. No measurable differences were found between the results of the bridge and modified assessments for any of these groups. These results validate the link between the bridge and modified assessments and imply that the trend line can continue, using the results of the modified assessment as the point connecting the former trend line to the new trend line. (A similar linkage was made to maintain the shorter trend line in main NAEP, which started with a single line when no accommodations were permitted, had an overlapping line in the years that allowed accommodations for a portion of the sample assessed, and then continued with a single line when accommodations were permitted for the students who required them.)

Comparison of Bridge and Modified Results for Mathematics

In mathematics, no differences were found between the average scores of students who took the bridge assessment and those who took the modified assessment at any age. Again, as seen in figure 5-4, it appears that students who took the bridge assessment scored slightly higher, on average, but these differences were not statistically significant.

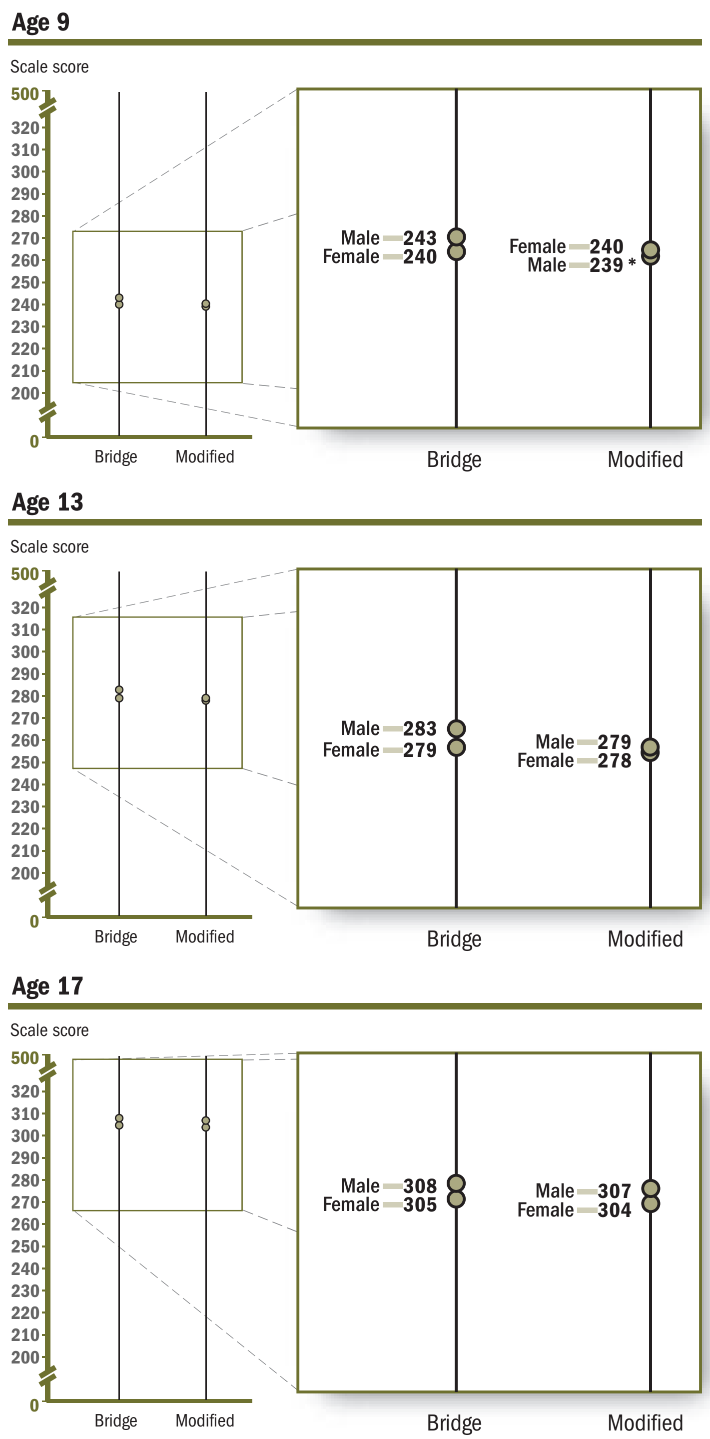
Figure 5-4. Average mathematics scale scores for students ages 9, 13, and 17 for bridge and modified assessments: 2004



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

Turning now to comparisons across various student groups, only one measurable difference was found. At age 9, male students scored, on average, 4 points higher on the bridge assessment in mathematics than on the modified assessment, scores of 243 and 239, respectively. However, figure 5-5 shows that the average mathematics scores for female 9-year-olds were not measurably different, and there were no measurable differences at the other age levels.

Figure 5-5. Average mathematics scale scores for students ages 9, 13, and 17 for bridge and modified assessments, by gender: 2004



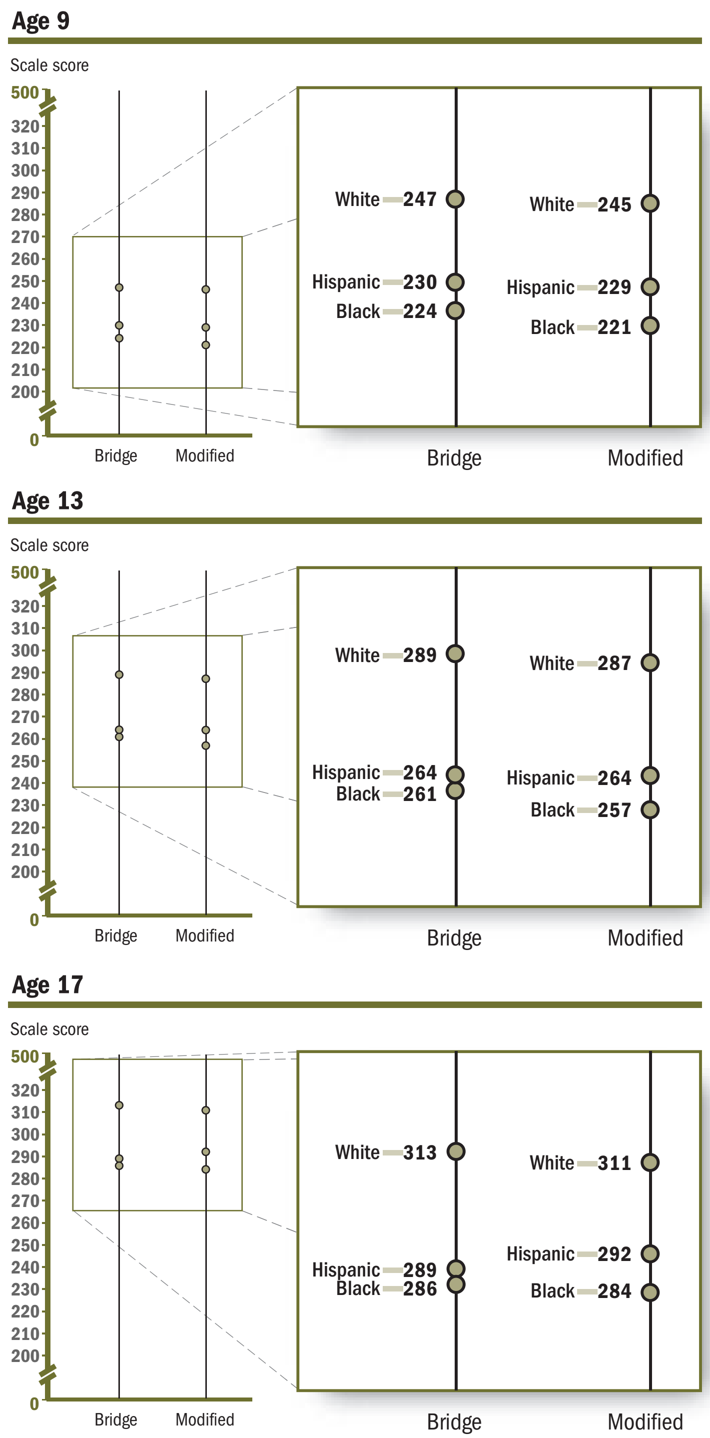
*Significantly different from 2004 bridge assessment.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

Comparisons of average scores for different racial/ethnic groups show no measurable differences between the two assessments (see figure 5-6). The average scores were also examined by other student and school demographic factors, and, again, showed almost no measurable differences. The one exception, which can be examined using the NAEP Data Explorer at http://nces.ed.gov/nationsreportcard/naepdata/, was found at age 17 when the comparison was made by type of community. Students in rural schools who took the bridge assessment had higher average scores than students in rural schools who took the modified assessment, with scores of 306 and 302, respectively. The fact that no other differences were statistically significant across any group for any age level implies that the link between the old and new assessments is valid and the trend line for mathematics can continue.

Summary

Overall, many changes were made to how the assessments were constructed and administered, but these changes resulted in minimal differences in student scores. The long-term trend assessment now uses more up-to-date assessment techniques, and obsolete items and methodologies from the late 1960s have been eliminated. Future assessments of NAEP long-term trend will use the modified format, and the link will allow for the comparisons back to 1971 for reading and 1973 for mathematics, preserving the more-than-30-year trend line.

Figure 5-6. Average mathematics scale scores for students ages 9, 13, and 17 for bridge and modified assessments, by race/ethnicity: 2004



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

As a result of the modifications made to the long-term trend assessment, it is now possible to share some questions with the public. For the first time, NCES is releasing items that have been administered to students since the early 1970s.

Chapter 6
Sample Questions

This chapter provides sample questions at all three ages for both reading and mathematics. These questions were administered to students in previous assessments but will no longer be used in NAEP assessments. They provide a glimpse of the types of skills and knowledge the long-term trend assessment measures.

For reading, two to three questions are provided for each age. The reading passage is followed by the questions. For multiple-choice items, all possible response options are given, with the oval corresponding to the correct answer filled in. Then, the percentage of students answering that item correctly in 2004 is shown. One constructed-response question is shown with sample student responses, a summary of the scoring criteria used to determine their score, and their actual assigned scores. The percentage of students receiving each possible score point is also given for the constructed-response question.

For mathematics, three questions are provided for each age. The response options are provided as the students saw them, and the correct answer is filled in. The constructed-response items are scored as correct or incorrect. The correct response is shown in the answer box. The percentage of students answering each item correctly in 2004 is stated below and to the right of each item.

Additional questions, as well as student performance data and scoring guides, are available through the NAEP Question Tool, located on the NAEP website at http://nces.ed.gov/nationsreportcard/itmrls/.

Reading: 9-Year-Olds

Questions 1–3. Read the passage below and answer the questions based on it.

New Folks

All the hill was boiling with excitement. On every side there rose a continual chattering and squeaking, whispering and whistling, as the animals discussed the great news. Through it all could be heard again and again the words, "New Folks are here!"

"Real sensible, knowledgeable Folks they seem to be," the Gray Fox said. "Quiet-like and friendly. Why just yesterday afternoon late I was prospecting around—sort of smelled chicken frying—I guess, and I came to that little walled-in garden where the benches are. I wasn't paying much attention and he, the Man, wasn't smoking his pipe or I'd have known he was around, when first thing I knew there I was right in front of him, face to face you might say. He was reading a book and he looked up and what do you suppose he did? Nothing, that's what. He just sat there and looked at me and I stood there and looked at him and then he said, "Oh, hello," and went back to reading his book, and I went on about my business. Now that's the sort of Folks is Folks."

Sample Reading Question 1

Sample question 1 asked students to make an inference based on the dialogue from the passage.

1. Who are the "Folks" in this story?
   - Ⓐ Larger foxes who live in a bigger den
   - Ⓑ Other animals who live on the hill
   - Ⓒ Foxes who live in a big, square cage
   - ● Humans who live nearby

55 percent of 9-year-olds answered this question correctly

Sample Reading Question 2

Sample question 2 asked students to identify a description made explicit in the passage.

2. What did the Gray Fox think about the Folks?

Ⓐ That they were strange

● That they were sensible

Ⓒ That they were frightening

Ⓓ That they were foolish

53 percent of 9-year-olds answered this question correctly

Sample Reading Question 3

Sample question 3 asked students to identify a character's action.

3. What was the man doing when the Gray Fox saw him?

Ⓐ Reading in his living room

● Sitting on a bench in the garden

Ⓒ Smoking on the front porch

Ⓓ Sitting in a chair smoking his pipe

38 percent of 9-year-olds answered this question correctly

Reading: 13-Year-Olds

Questions 4–6. Read the newspaper advertisement below and answer the questions based on it.

Wanted

Persons interested in earning between $35 and $45 per month delivering the *Post* newspaper. Help needed in most areas. Papers delivered to your home between 5 and 6 a.m.

Requirements for News Carrier

1. Must be at least nine years old
2. Must be reliable
3. Must deliver all papers by 7 a.m., 7 days a week
4. Must make collections during the last three days of every month

If you can meet these requirements, call 584-3640 Monday–Friday, 8 a.m. through 4 p.m. Ask for the Circulation Department.

Sample Reading Question 4

Sample question 4 asked students to connect text details to make an inference.

4. According to the advertisement, what should you do if you are interested in the job and meet the requirements?

   Ⓐ Apply in person at the *Post.*

   Ⓑ Write the *Post* for a job application form.

   Ⓒ Wait for the openings to be published in the *Post.*

   ● Call the *Post* Circulation Department.

84 percent of 13-year-olds answered this question correctly

Sample Reading Question 5

Sample question 5 asked students to make an inference based on details from the advertisement.

5. David and Mary are both reliable eight year olds and have applied for the job. What will probably happen?

- ● They will not get the job because they are too young.
- Ⓑ They will get the job since they are reliable.
- Ⓒ They will not get the job unless they have bicycles.
- Ⓓ They might get the job if they can work at the right times.

83 percent of 13-year-olds answered this question correctly

Sample Reading Question 6

Sample question 6 asked students to identify specific text details.

6. By what time must the news carrier deliver all the papers?

- Ⓐ By 6 every morning
- ● By 7 every morning
- Ⓐ By 8 every morning, except weekends
- Ⓐ By 7 every evening

85 percent of 13-year-olds answered this question correctly

Reading: 17-Year-Olds

Questions 7–8. Read the passage below and answer the questions based on it.

Throwing the Javelin

The scent of honeysuckle seemed to linger in the air and joined itself with the sweet odor of freshly cut grass. I slipped out of my bright red sweats and flung them to the base of the tree. I picked up the javelin, stuck point down in the turf. I stretched my arms with the javelin behind my neck. Out of habit, I stood and held the javelin in my left hand, and with the thumb of my right forced small clumps of dirt from the tip. I searched for a target. Picking a spot in a cloud moving towards me I cocked the javelin above my shoulder and regulated my breathing. My right foot was placed on the first mark and my left foot rested behind. My eyes were focused on one abstract point in the sky. Pierce it. I built up energy. Slowly, my legs flowed in motion, like pistons waiting for full power and speed. I could feel my legs churning faster, the muscles rippling momentarily, only to be solidified when foot and turf met like gears. Hitting the second mark, I escaped from the shadow of the tree and was bathed in sunlight . . . . Left foot forward . . . javelin back, straight back, . . . turn now, five steps . . . three, four . . . stretch, the clouds, the point . . . turn back, throw the hips . . . chest out . . . explode through the javelin . . . terminate forward motion, release.

The muscles of my right leg divided in thirds just above my knee, as the full weight of my body in motion was left to its support. Skipping, I followed through and watched the quivering javelin climb as it floated in the oncoming wind. For a moment, it reflected the sunlight and I lost sight of the javelin. The javelin landed quickly, piercing the ground. I heaved in exhaustion, and perspiration flowed from my face and hands. Before me the field stretched and I attempted to evaluate my throw. I was pleased. The smell of honeysuckle again drifted into my senses and somehow, I had a feeling of accomplishment I could just as easily have experienced had I thrown poorly.

Sample Reading Question 7

Sample question 7 asked students to identify the overall stylistic purpose of the author.

7. What is the main reason the writer wrote this story?

   Ⓐ To express an athlete's feeling of failure

   Ⓑ To provide information about javelin throwing

   ● To describe how it feels to throw the javelin

   Ⓒ To encourage people to take up javelin throwing

80 percent of 17-year-olds answered this question correctly

Sample Reading Question 8

Sample question 8 is a constructed-response question, which asked students to explain how the author of the passage created an effect. Responses to this task were rated according to a four-level scoring guide in one of the following score categories: "Elaborated Interpretation," "Satisfactory Interpretation," "Minimal Interpretation," or "Unsatisfactory Interpretation."

8. Here is one student's impression of the story:

> *When I watch throwing javelins on television, everything seems to happen in a split second. First, the javelin is in the thrower's hand and the next thing you know the official is out there measuring how far the javelin was thrown. In this story, though, throwing the javelin seems to take a long time.*

Think about the story. Think about the way in which the writer created the impression that this javelin throw took a long time. Write your explanation on the lines provided.

Sample "Elaborated Interpretation" response. The following sample response is rated "Elaborated Interpretation" because it explains ways the writer made the javelin throw seem to take a long time with multiple references to the passage.

In this story the writer created the impression that the javelin throw took a long time by explaining every motion and feeling of the thrower vividly. Every single movement of his body is maticulously explained. Everything the thrower feels, sees, and smells is stated to the reader. He is concentrating with all of his might to pierce the cloud with the javelin, to make every step of the throw fluidly, and to please himself with his throw. The reader uses similes to compare the throw and to make it more clear. The writer appeals to the senses; he explains the smell of honeysuckle and the grass.

1 percent of 17-year-olds wrote responses rated as "Elaborate."

Sample Reading Question 8

Sample "Satisfactory Interpretation" response. The following sample response is rated "Satisfactory Interpretation" because it explains how the javelin throw seems to take a long time with a specific example from the passage.

The writer wrote everything that was going through their mind as they were throwing. They expressed all their feelings and emotions. This story shows the thought process of a javelin thrower.

18 percent of 17-year-olds wrote responses rated as "Satisfactory."

Sample "Minimal Interpretation" response. The following sample response is rated "Minimal Interpretation" because it provides a passage-based generalization to explain why the javelin throw seems to take a long time.

I think the boy thought it took a long time because the author was so detailed and said every single thing step-by-step and it really all happened with in seconds.

44 percent of 17-year-olds wrote responses rated as "Minimal."

Mathematics: 9-Year-Olds

Sample Mathematics Question 1

Sample Question 1 was a multiple-choice question that asked students to demonstrate an understanding of place value by identifying the number that represents "nine tens." The incorrect choices in this question represent other place values (9 and 900) and a literal reading of "nine tens" (910).

1. Which one of the following represents "nine tens"?

   Ⓐ 9

   ● 90

   Ⓒ 900

   Ⓓ 910

80 percent of 9-year-olds answered this question correctly

Sample Mathematics Question 2

Sample Question 2 was a constructed-response question that asked students to show an understanding of the concept of "more than" and its relationship to the operation of addition.

2. What number is 10 MORE than 95?

   Answer: 105

77 percent of 9-year-olds answered this question correctly

Sample Mathematics Question 3

Sample Question 3 was a multiple-choice question that asked students to demonstrate knowledge of metric measurement by identifying a reasonable weight for a bicycle from among the choices given. The incorrect choices in this question represented misjudgments related to order of magnitude of an appropriate number of kilograms.

3. About how many kilograms does a bicycle weigh?

   Ⓐ 1.5 kilograms

   ● 15 kilograms

   Ⓒ 150 kilograms

   Ⓓ 1500 kilograms

40 percent of 9-year-olds answered this question correctly

Mathematics: 13-Year-Olds

Sample Mathematics Question 4

Sample Question 4 asked students to translate a situation in a real context into a number sentence.

4. Kathleen is packing baseballs into boxes. Each box holds 6 baseballs. She has 24 balls. Which number sentence will help her find out how many boxes she will need?

Ⓐ $24 - 6 = \square$

● $24 \div 6 = \square$

Ⓒ $24 + 6 = \square$

Ⓓ $24 \times 6 = \square$

80 percent of 13-year-olds answered this question correctly

Sample Mathematics Question 5

Sample Question 5 asked students to show an understanding of equivalent fractions by writing a mixed number as an improper fraction.

5. Write the following mixed numeral as an improper fraction.

$1\frac{1}{4} = \underline{\quad \frac{5}{4} \quad}$

70 percent of 13-year-olds answered this question correctly

Sample Mathematics Question 6

Sample Question 6 asked students to show an understanding of metric prefixes.

6. One liter is how many milliliters?

Ⓐ 10

Ⓑ 100

● 1000

37 percent of 13-year-olds answered this question correctly

Mathematics: 17-Year-Olds

Sample Mathematics Question 7

Sample Question 7 asked students to read data from a table and perform a computation with selected values.

POPULATIONS OF DETROIT AND LOS ANGELES
1920–1970

YEAR	CITY	
	DETROIT	LOS ANGELES
1920	950,000	500,000
1930	1,500,000	1,050,000
1940	1,800,000	1,500,000
1950	1,900,000	2,000,000
1960	1,700,000	2,500,000
1970	1,500,000	2,800,000

7. How many more people were living in Los Angeles in 1960 than 1940?

Ⓐ 100,000

Ⓑ 500,000

Ⓒ 800,000

● 1,000,000

Ⓔ 2,500,000

82 percent of 17-year-olds answered this question correctly

Sample Mathematics Question 8

Sample Question 8 asked students to show an understanding of percentages by identifying the “whole” when given the “part” and the percentage it represents.

8. 9 is 12% of what number?

Ⓐ .75

Ⓑ 1.08

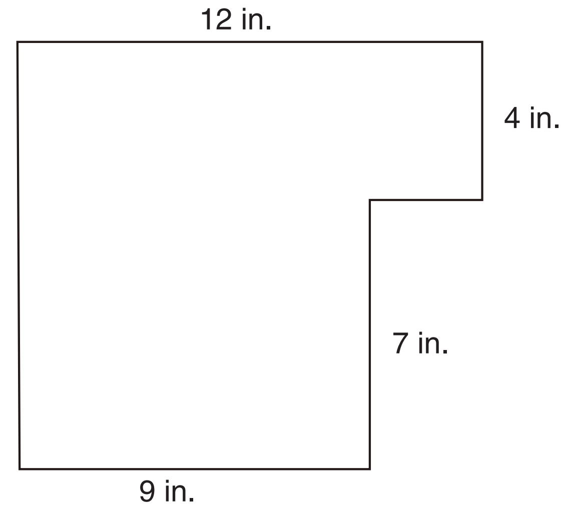
● 75

Ⓓ 108

56 percent of 17-year-olds answered this question correctly

Sample Mathematics Question 9

Sample Question 9 was a constructed-response question that asked students to determine the area of an irregularly shaped region. For this problem the student could partition the region into smaller rectangles and find the sum of the areas. For example, one way to do this is with rectangles that are 9 inches by 11 inches and 3 inches by 4 inches. The area is then given by $(9 \times 11) + (3 \times 4) = 99 + 12 = 111$ inches. An alternative approach is to find the area of a large rectangle (12 inches by 11 inches) and subtract from it the area of a smaller rectangle (3 inches by 7 inches) that has been removed. The area is then given by $(12 \times 11) - (3 \times 7) = 132 - 21 = 111$ inches.



9. What is the area of this figure?

ANSWER: 111 square in.

32 percent of 17-year-olds answered this question correctly

References

Allen, N.L., McClellan, C.A., and Stoeckel, J. (2005). *NAEP 1999 Long-Term Trend Technical Analysis Report: Three Decades of Student Performance* (NCES 2005–484).U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Baker, D.P., and Jones, D.P. (1993). Creating Gender Equality: A Cross-National Gender Stratification and Mathematical Performance. *Sociology of Education, 66*(2): 91–103.

Bauer, S.C., Park, H.S., and Sullivan, L.M. (1998). Gender Differences Among Top Performing Elementary School Students in Mathematical Ability. *Journal of Research and Development in Education, 31*(3): 133–141.

Benjamini, Y., and 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, 289–300.

Berliner, D., and Biddle, B. (1996). In Defense of Schools. *Vocational Education Journal, 71*(3): 36–38.

Braswell, J.S., Dion, G.S., Daane, M.C., and Jin, Y. (2005). *The Nations' Report Card: Mathematics 2003* (NCES 2005–451). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Cai, J., Moyer, J.C., and Wang, N. (1997, March). *Parental Roles in Students' Learning of Mathematics*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC ED412087)

Donahue, P.D., Daane, M.C., and Jin, Y. (2005). *The Nation's Report Card: Reading 2003* (NCES 2005–453). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Downes, T., and Reddacliff, C. (1997). *Children's Use of Electronic Technologies in the Home (Based on Structured Interviews with 14 Children, Their Parents, and Their Teachers From a Wide Cross Section of Urban Sydney).* Western Sydney Univeristy: Macarthur, Australia. (ERIC ED416843)

Freeman, C.E. (2004). *Trends in Educational Equity of Girls and Women: 2004* (NCES 2005–016). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Johnson, E.G. (1989). Considerations and Techniques for the Analysis of NAEP Data. *Journal of Educational Statistics, 14*(4): 303–334.

Lemke, M., Calsyn, C., Lippman, L., Jocelyn, L., Kastberg, D., Liu, Y., Roey, S., Williams, T., Kruger, T., and Bairu, G. (2002). *Outcomes of Learning: Results from the 2000 Program for International Student Assessment of 15-Year-Olds in Reading, Mathematics, and Science Literacy* (NCES 2002–115). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Miller, R.G. (1981). *Simultaneous Statistical Inference* (2nd ed.). New York: Springer-Verlag.

Mislevy, R.J. (1988). Randomization-Based Inferences About Latent Variables From Complex Samples. *Psychometrika, 56*(2): 177–196.

Mislevy, R.J., and Sheehan, K.M. (1987). Marginal Estimation Procedures. In A.E. Beaton (Ed.), *Implementing the New Design: The NAEP 1983–1984 Technical Report* (Technical Rep. No. 15-TR-20), pp. 293–260. Princeton, NJ: Educational Testing Service.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Kelly, D., and Smith, T.A. (1998). *Mathematics and Science Achievement in the Final Year of Secondary School.* Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*(2): 159–176.

National Assessment of Educational Progress. (1986). *Math Objectives: 1985–86 Assessment.* Princeton, NJ: Educational Testing Service. Report No. 17-M-10.

National Assessment of Educational Progress. (1984). *Reading Objectives: 1983–84 Assessment.* Princeton, NJ: Educational Testing Service. Report No. 17/19-R-10.

No Child Left Behind Act of 2001, P.L. 107–110, 115 Stat. 1425 (2002).

Perie, M., Vanneman, A., and Goldstein, A. (2005). *Private Schools in America: Results From the 2000–2003 NAEP* (NCES 2005–xxx). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Rathbun, A.H., West, J., and Hausken, E.G. (2003). *Young Children's Access to Computers in the Home and at School in 1999 and 2000* (NCES 2003–036). U.S. Department of Education, National Center for Education Statistics. (ERIC ED473848)

Ravitch, D. (2002). A Brief History of Testing and Accountability. *Hoover Digest, 4.* Retrieved November 19, 2004, from http://www.hooverdigest.org/024/ravitch.html.

Williams, V.S.L., Jones, L.V., and Tukey, J.W. (1999). Controlling Error in Multiple Comparisons with Examples From State-to-State Differences in Educational Achievement. *Journal of Educational and Behavioral Statistics, 24*(1): 42–69.

This procedural appendix discusses the development and administration of the long-term trend assessment. It includes information on technical aspects such as sampling, scoring, weighting, and scaling, as well as a glossary of terms.

Appendix A
Overview of Procedures Used in the 2004 NAEP Long-Term Trend Assessments

This appendix provides information about the methods and procedures used in the 2004 NAEP reading and mathematics long-term trend (LTT) assessments. More extensive information about these procedures will be available in the form of technical documentation on the NAEP website (http://nces.ed.gov/nationsreportcard/). Additional data from the 2004 long-term trend assessments, as well as data from other NAEP assessments, can be obtained from the NAEP Data Explorer at http://nces.ed.gov/nationsreportcard/naepdata/.

NAEP long-term trend assessments are designed to give information about changes in the basic achievement of America's youth. They have measured students' performance in mathematics, science, reading, and writing, and have monitored trend lines first established 35 years ago. Over the past three decades, results have been reported for students at ages 9, 13, and 17 in mathematics, reading, and science and in grades 4, 8, and 11 in writing. In 1999, the National Assessment Governing Board (NAGB) discontinued the writing trend assessment, and in 2002 NAGB decided that additional technical studies were required to update the design of the science trend assessment (for more information see http://www.nagb.org/release/policy_statement.doc). Therefore, this NAEP long-term trend report is based on results from 10 assessments of the mathematics performance and 11 assessments of the reading performance of 9-, 13-, and 17-year-old students, with the most recent assessment in each curriculum area conducted during the 2003–04 school year.

Measuring trends of student achievement, or change over time, requires the precise replication of past procedures. Since their inception, the design and methodology of the NAEP long-term trend assessments have remained constant, to the extent feasible, thereby enabling the continuous monitoring of a fixed set of curriculum topics. The long-term trend instruments do not evolve based on changes in curricula or in educational practices; in this way, the long-term trend assessments differ from the main national and state NAEP assessments. The results presented in this report are based solely on the most recent and past administrations of the NAEP long-term trend reading and mathematics assessments and not on the main NAEP assessments. Because the long-term trend assessments use different instruments from those used in the main assessments, and because students are sampled by age for the long-term trend assessments rather than by grade as in the main assessments, it is not possible to compare results from the two assessment programs.

The NAGB decision to discontinue the writing and science trend assessments provided an opportunity to modify the NAEP long-term trend assessments to reflect current assessment designs and practices. Consequently, a number of changes were implemented in 2004 to revitalize the long-term trend assessments. In order to align the trend assessment procedures with the policies applicable to the NAEP main assessments, assessment accommodations were provided for students with disabilities and English language learners. Traditionally, such accommodations were not available to students participating in the long-term trend assessments. Changes to the assessment instruments included replacing items that used outdated contexts, replacing background questions, eliminating "I don't know" as a response option for multiple-choice items, and using assessment booklets that pertain only to a single subject area (whereas, in the past, a single assessment booklet may have contained both reading and mathematics items). In addition, use of an audio paced tape, which paced students during the assessment session so that they were at the same place in the test booklet at the same time, was discontinued for mathematics. The reading trend assessment eliminated the use of a paced tape in 1984. Specific changes to the reading and mathematics trend assessments for 2004 are discussed in more detail later in this appendix.

The changes implemented in 2004 were intended to maintain the integrity of the long-term trend assessments and increase the validity of the results obtained. It was important to know that any changes in assessment results could be attributed to actual changes in student performance rather than to changes in the assessments. A special bridge study was conducted in 2004 to evaluate how the changes to the assessment design and administration procedures would affect assessment results. A bridge study involves the administration of two assessments to two randomly assigned groups of students. One assessment, the bridge assessment, is exactly the same as previous years' assessments, and the other assessment, the modified assessment, represents the modified design. Data from a bridge study are used to link the scale of the revised assessment to the scale established by the previous version of the assessment, so that trend reporting can be continued. Results from the 2004 bridge assessments are presented in chapters 2 through 4 of this report. These bridge study results maintain the trend lines established in 1971 for reading and in 1973 for mathematics. Results from the 2004 modified assessments are presented in chapter 5, and will serve as the base year of results for future long-term trend assessments.

NAEP assessments are designed to best support certain types of inferences. In the case of long-term trend, the items on the assessment have remained unchanged for a long period of time in an effort to provide a solid foundation for the measurement of trend. Subsequent to the baseline IRT scaling, when the cross-age scale was established, the assessment has been scaled within age. These within-age scalings involve jointly analyzing the data from the current and most recent NAEP long-term trend assessments. These separate within-age scalings are then linked to the cross-age scale that was originally established. This approach strengthens the evidence that the assessment provides to support within-age comparisons across time. Because the assessment was explicitly scaled in a cross-age manner only in the base year, cross-age comparisons are most strongly supported in that year rather than in later years. However, the items did not change between the initial cross-age scaling in the 1970s and the bridge assessment of 2004. Moreover, within-age scales from subsequent years have been aligned to the initial cross-age scale. Therefore, cross-age comparisons should be reasonably well supported, although the focus continues to be on within-age comparisons. It should be borne in mind, however, that NAEP is not a cohort or longitudinal design, and the LTT assessments have not been given at intervals that coincide with the age span (4 years apart) in the assessment and have been given at different times of the year for the three ages. As a result, inferences about the performance of cohorts of students over time should not be made based on NAEP LTT results.

The long-term trend comparisons described in this report are based on content specifications for reading and mathematics that have remained substantially constant over the assessments. More information about the composition of each of the trend assessments is presented in the following pages.

The Reading Assessment

NAEP has assessed student reading achievement at age 9, age 13, and age 17 in 11 reading assessments, conducted during the school years ending in 1971, 1975, 1980, 1984, 1988, 1990, 1992, 1994, 1996, 1999, and 2004. For historical reasons, the writing assessment results were based on a sample of students in grades 4, 8, and 11, and the reading assessment results were based on a sample of students aged 9, 13, and 17. Because the two subjects were administered together, NAEP long-term trend assessments in reading and writing were administered to the same sample of both age- and grade-eligible students (i.e., the sample included students who were either in grade 4 or at age 9, either in grade 8 or at age 13, and either in grade 11 or at age 17), and the results for the two subject areas were based on different subsamples of these students. Since the writing assessment is no longer administered, however, it was not necessary to sample students by grade in 2004. Consequently, only an age-eligible sample of students participated in the 2004 reading trend assessment.

The long-term trend reading tasks required students to read and answer questions based on a variety of materials, including informational passages, literary texts, and documents. The set of reading passages and questions included in the trend assessments has been kept essentially the same since 1984, and most closely reflects the objectives developed for that assessment (National Assessment of Educational Progress [NAEP] 1984). The selections include brief stories, passages from textbooks, and other age-appropriate reading material. Although some tasks required students to provide written responses, most questions were multiple choice. The assessment was designed to evaluate students' ability to locate specific information, to make inferences based on information in two or more parts of a passage, and to identify the main idea in a passage. Demonstration booklets from the 2004 NAEP long-term trend assessments, along with booklets from other NAEP assessments, are available on the NAEP website at http://nces.ed.gov/nationsreportcard/about/booklets.asp.

The assessment booklets used in the 2004 reading trend assessment contained segments or "blocks" of reading and background questions that were used in the 1984, 1988, 1990, 1992, 1994, 1996, and 1999 trend assessments. In addition, some questions that were common to several trend assessments before the mid-1980s were included in the 2004 assessment booklets. Each block contained reading passages and questions and a short set of background questions that pertained to students' reading habits and experiences. The blocks were assembled three to a booklet, together with a general background questionnaire that was common to all booklets. This background questionnaire included questions about demographic information and home environment. Overall, there were ten blocks of questions arranged in groups of three blocks in six different booklets for each age sample of the bridge study.

The reading assessment administered in the 2004 bridge study at age 9 was composed of 45 passages and 102 questions. Most questions were multiple choice; 5 questions required students to construct responses. At age 13, the bridge assessment was composed of 43 passages and 107 questions, 7 of which required constructed responses. The age 17 bridge assessment contained 36 passages and 95 questions, 8 of which required constructed responses. The assessment booklets for the 2004 bridge study are identical to those used in the 1999 assessment, except for booklets used in 1999 that had a block of writing items in the third (last) position. In those cases, the writing block was replaced with a new block of reading items developed for the 2004 modified trend assessment. (New blocks of items, developed for use in the 2004 modified trend assessment, were included in the bridge assessment to strengthen the link between the two assessments. As is described in later sections of this appendix, in order to report trends, it is necessary to place the results of the bridge and modified assessments on the same scale.) One booklet at age 9 was reconfigured in this way, as were three booklets for the age 13 assessment and three booklets for the age 17 assessment. Figure A-1 depicts these changes to the reading bridge assessment booklets.

Figure A-1. Changes to the 1999 reading long-term trend assessment booklets implemented in the 2004 reading bridge assessment

		1999 reading trend assessment				2004 reading trend bridge assessment		
		Position				Position		
		1	2	3		1	2	3
Age 9	Book 1	Writing	Reading block 4	Reading block 8	Book 1	Writing	Reading block 4	Reading block 8
	Book 2	Reading block 1	Writing	Reading block 9	Book 2	Reading block 1	Writing	Reading block 9
	Book 3	Writing	Reading block 3	Reading block 2	Book 3	Writing	Reading block 3	Reading block 2
	Book 4	Writing	Reading block 7	Writing	Book 4	Writing	Reading block 7	New reading block 1
	Book 5	Reading block 5	Writing	Reading block 6	Book 5	Reading block 5	Writing	Reading block 6
	Book 6	Reading block 10		Reading block 9	Book 6	Reading block 10		Reading block 9
Ages 13/17	Book 1	Reading block 5	Reading block 3	Writing	Book 1	Reading block 5	Reading block 3	New reading block 1
	Book 2	Writing	Reading block 4	Reading block 9	Book 2	Writing	Reading block 4	Reading block 9
	Book 3	Reading block 1	Writing	Reading block 10	Book 3	Reading block 1	Writing	Reading block 10
	Book 4	Reading block 6	Writing	Writing	Book 4	Reading block 6	Writing	New reading block 2
	Book 5	Writing	Reading block 7	Writing	Book 5	Writing	Reading block 7	New reading block 3
	Book 6	Writing	Reading block 2	Reading block 8	Book 6	Writing	Reading block 2	Reading block 8

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 and 2004 Long-Term Trend Reading Assessments.

Although writing results would not be reported, not all writing blocks were removed from the 2004 bridge study assessment booklets. The writing blocks that appeared in either the first or second booklet positions continued to be administered in order to preserve the context of the reading blocks.

The assessment booklets used in the modified trend assessment in 2004 were different from those used in the 2004 bridge study and all previous years' trend assessments. The 2004 modified trend instrument contained blocks with items used in the 1999 trend assessment (although reconfigured from the 1999 design), blocks with new pilot-tested items, and blocks with new items that had not been pilot tested. All background questions were presented together in a single block at the end of each booklet, rather than interspersed among the cognitive item blocks as in previous versions, and none of the multiple-choice questions included an "I don't know" response option. The modified instrument followed a focused, balanced, incomplete block (BIB) design, which ensures that each block, and therefore each question, is presented to a nationally representative sample of students, and that each question is presented in various positions with respect to other questions. Specifically, the 2004 NAEP long-term trend assessments utilize a partially balanced, incomplete block (pBIB) booklet design.

In a pBIB design, blocks may not appear an equal number of times in each booklet position, or may not be paired with every other block an equal number of times. Because each assessment booklet contained questions from only one subject area, the design of the modified assessment is called a “focused” design. In previous administration years, more than one subject was assessed in each booklet, so the 2004 modified assessment marks the first focused pBIB design for the long-term trend assessment.

The 2004 modified reading trend assessment administered at age 9 included 39 passages, or reading tasks, and 91 questions, including 4 questions that required students to construct written responses. At age 13, the modified assessment included 42 passages and 106 questions, 7 of which required constructed responses. At age 17, the modified assessment contained 36 passages and 104 questions, 8 of which required constructed responses. All told, 20 different assessment booklets were administered to each age sample in the modified trend assessment.

The Mathematics Assessment

NAEP has assessed the mathematics achievement of 9-, 13-, and 17-year-olds ten times: in the school years ending in 1973, 1978, 1982, 1986, 1990, 1992, 1994, 1996, 1999, and 2004. The mathematics trend assessments contained a range of constructed-response and multiple-choice questions designed to measure performance on sets of objectives developed by nationally representative panels of mathematics specialists, educators, and other interested parties. The 1986, 1990, 1992, 1994, 1996, 1999, and 2004 assessments shared common objectives (NAEP 1986). The objectives for each assessment prior to 1990 were based on the framework used for the previous assessment, with some revisions that reflected changes in the content of mathematics education. Although changes were made from assessment to assessment before 1990, some questions were retained from one assessment to the next in order to measure trends in achievement across time. This continuity allows comparisons to be made across all of the available assessments, other than the 1973 assessment, using item response theory (IRT). Results from the 1973 assessment were placed on the same scale using mean-proportion-correct extrapolation. (For further explanation of IRT and mean-proportion-correct extrapolation, see the section later in this appendix on Data Analysis and IRT Scaling.)

As for the reading assessment, changes to the design and administration of the 2004 mathematics trend assessment made a special bridge study necessary. The 2004 mathematics bridge study used procedures established in 1973. For all three age samples assessed in the bridge study, the mathematics questions were administered using an audio paced tape that accompanied the booklets. It standardized the timing of the administration, and was intended to help students with any difficulty they might have in reading the questions. Thus, in an administration session, all students were being paced through the same booklet.

The instrument used in the 2004 mathematics bridge study contained a number of questions that were also administered in the 1986, 1990, 1992, 1994, 1996, and 1999 mathematics long-term trend assessments. These common questions numbered 52, including 18 constructed-response questions at age 9; 74, including 17 constructed-response questions at age 13; and 70, including 15 constructed-response questions at age 17. The questions covered a range of content, including numbers and operations, measurement, geometry, and algebra. The process areas included knowledge, understanding, skills, applications, and problem solving. In the 2004 bridge study, two different assessment booklets were constructed for use at age 9, and two for use at age 13; one assessment booklet was constructed for the age 17 bridge study sample.

The booklets used in the 1999 mathematics long-term trend assessment and in all previous long-term trend assessments consisted of one block of mathematics questions, one block of science questions, and one block of reading questions at ages 9 and 13. For age 17, the booklets consisted of either two blocks of mathematics questions and one block of science questions, or one block of mathematics questions and two blocks of science questions. Neither the reading nor science blocks in these booklets were analyzed in the 2004 NAEP long-term trend assessment, but the blocks were

Figure A-2. Changes to the 1999 mathematics long-term trend assessment booklets implemented in the 2004 mathematics bridge assessment

	1999 mathematics trend assessment			
		Position		
		1	2	3
Ages 9/13	Book 1	Reading	Mathematics block 1	Science
	Book 2	Mathematics block 2	Science	Reading
	Book 3	Science	Reading	Mathematics (Calculator)
Age 17	Book 1	Mathematics block 1	Mathematics block 2	Science
	Book 2	Science	Science	Mathematics (calculator)

2004 mathematics trend bridge assessment			
	Position		
	1	2	3
Book 1	Reading	Mathematics block 1	New mathematics block 1
Book 2	Mathematics block 2	New mathematics block 1	New mathematics block 2
Book 1	Mathematics block 1	Mathematics block 2	New mathematics block 1

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 and 2004 Long-Term Trend Mathematics Assessments.

included in the bridge study assessment booklets in order to preserve the context of the mathematics questions. Thus, the bridge assessment used a partial BIB design, and the modified assessment used a focused partial BIB design. Reading or science blocks that appeared in trailing positions of the booklets (i.e., did not precede a mathematics block) did not affect the context of mathematics questions and therefore could be replaced. Newly developed blocks of mathematics questions replaced trailing reading and science blocks in both age 9 booklets, both age 13 booklets, and the single age 17 booklet. A reading block remained in the first position in one age 9 booklet and one age 13 booklet. A comparison of booklets from the 1999 mathematics trend assessment and the 2004 bridge assessment appears in figure A-2.

The booklets used in the 2004 modified trend assessment consisted of blocks that had been administered in 1999 and previous trend assessments, blocks with new pilot-tested questions, and blocks with new questions that had not been pilot tested. The modified assessment contained 140 questions at age 9, including 37 constructed-response questions; 166 questions at age 13, including 37 constructed-response questions; and 162 questions, including 34 constructed-response questions at age 17. These questions formed blocks that were assembled into six different assessment booklets for each age sample.

Sampling and Data Collection

Sampling and data collection activities for the 2004 NAEP trend assessments in reading and mathematics were conducted by Westat, Inc. The target population for the 2004 NAEP long-term trend assessments consisted of 9-, 13-, and 17-year-old students enrolled in public and nonpublic elementary and secondary schools nationwide. Accordingly, a sample of students in each of these age groups was assessed. Eligibility for the age 9 and age 13 samples was based on calendar year: students in the age 9 sample were 9 years old on January 1, 2004, with birth months January 1994 through December 1994, and students in the age 13 sample were 13 years old on January 1, 2004, with birth months January 1990 through December 1990. Students eligible for the age 17 sample had to be 17 years old on October 1, 2004, with birth months October 1986 through September 1987.

The sampling plan was determined by the target number of students to be assessed in each subject, each assessment type—modified or bridge—and each age level. A total of nine different assessment sessions, corresponding to nine different samples, was required. As

Table A-1. Target student sample size in reading and mathematics, by type of school and type of assessment: 2004

Type of assessment	Total	Public schools	Private schools
AGE 9			
Reading			
Modified	8,000	6,400	1,600
Bridge	5,000	4,000	1,000
Mathematics			
Modified	8,000	6,400	1,600
Bridge	6,000	4,800	1,200
AGE 13			
Reading			
Modified	8,000	6,400	1,600
Bridge	5,000	4,000	1,000
Mathematics			
Modified	8,000	6,400	1,600
Bridge	6,000	4,800	1,200
AGE 17			
Reading			
Modified	8,000	6,400	1,600
Bridge	5,000	4,000	1,000
Mathematics			
Modified	8,000	6,400	1,600
Bridge	4,000	3,200	800

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading and Mathematics Assessments.

described earlier, changes implemented in the 2004 modified assessment allowed reading and mathematics to be assessed together in a single assessment session. In the bridge assessment, however, each subject required a separate assessment session at each grade level. Therefore, the sampling plan called for a total of six samples for the bridge assessment and three samples for the modified assessment. The plan took into consideration the necessary sample size to permit accurate estimation of performance for certain student groups. Sample size requirements were determined separately for public and nonpublic schools. The target sample sizes for the age 9, age 13, and age 17 samples are shown in table A-1. Note that these targets are for completed assessments. In assigning student sample sizes and developing a sampling plan, it was necessary to account for losses from absent, refusing, and ineligible students.

Consistent with past national long-term trend assessments, students were selected for participation based on a stratified three-stage sampling plan. In the first stage, geographic primary sampling units (PSUs) were defined and selected. In the second stage, schools, both public and nonpublic, were selected within PSUs. In the third stage, eligible students were selected within schools. Stratification occurred at both the school level and the PSU level. A full description of the sampling plan is beyond the scope of this appendix; for additional details regarding the design and structure of the 2004 trend assessment samples, the reader should refer to the technical documentation section of the NAEP website (http://nces.ed.gov/nationsreportcard/ltt).

Primary Sampling Units

The first-stage sampling units, PSUs, were drawn from a list—a sampling frame—developed by Westat using the metropolitan area designations of the U.S. Census Bureau. Each NAEP PSU in the frame was intended to encompass one county or contiguous multiple counties, generally not crossing state boundaries, and contained a minimum number of school-aged children. (In previous NAEP long-term trend assessments, PSUs were permitted to cross state boundaries. However, NAEP field personnel reported that contacts with state officials are very important in the process of recruiting schools, making single-state PSUs easier to manage. In 2004, therefore, PSUs were defined within single states, to the extent possible.) The minimum size constraint was 15,000 students (aged 9 to 17, based on data from the 2000 decennial census) for the Northeast and Southeast regions of the nation, and 10,000 students for the Central and West regions. Census-defined Metropolitan Statistical Areas (MSAs), Consolidated Metropolitan Statistical Areas (CMSAs), and New England County Metropolitan Statistical Areas[1] (NECMAs) were split

[1]The MSAs in New England are defined in terms of townships and sometimes split across counties. The NECMAs are close approximations to the MSAs, which are defined in terms of counties (i.e., do not split across counties).

according to state boundaries to form metropolitan PSUs. For example, the New York-Northern New Jersey-Long Island NY-NJ-CT-PA CMSA was partitioned into four separate metropolitan PSUs by state. In some cases, the partitioned PSUs violated the minimum size constraint. There were 11 such PSUs, corresponding to 12 counties. In four of these cases, the youth populations within the partitioned PSUs were not far from the minimum, so they were allowed to stand as metropolitan PSUs.

Then the pool of remaining nonmetropolitan counties, together with the seven counties that were partitioned from a metropolitan area but did not meet the minimum size constraint of a metropolitan PSU, were grouped into nonmetropolitan PSUs. These nonmetropolitan PSUs were formed by fitting together counties within a single state that covered a minimum geographic area and met the minimum size constraints. In many instances, counties could not be combined into PSUs that satisfied the minimum size constraints while still remaining within a single state. In some cases, counties were combined across state lines to form a PSU that met the minimum size requirement. In other cases, PSUs that were below the minimum size requirement were allowed to stand, if satisfying the minimum size requirement was not possible.

The overall frame of metropolitan and nonmetropolitan PSUs was divided into certainty and noncertainty PSUs. Certainty PSUs were those with target populations so large that it was efficient to include them in the sample with certainty (i.e., they had a probability of selection equal to 1). In general, a PSU was included with certainty if it contained more than 800,000 students. Seventeen metropolitan PSUs met this criterion and were considered certainty PSUs. The remaining metropolitan PSUs and all nonmetropolitan PSUs were considered noncertainty PSUs.

Once the PSUs were determined, the certainty PSUs were set aside, and the remaining noncertainty PSUs were stratified. Hard strata consisted of NAEP region (Northeast, Southeast, Central, and West) and metropolitan status (metropolitan or nonmetropolitan). Within each of the eight hard strata, the total number of youths within the stratum was computed as a measure of size; this measure of size determined its relative share of the 60 PSU strata. Next, PSUs were assigned to implicit strata using four stratification variables: percentages of racial/ethnic groups, income levels, education levels in the population, and percentage of renters (as opposed to homeowners). It was desirable that the PSU strata be as equal in size as possible and homogeneous across variables.

After stratification was completed, measures of size and probabilities of selection were defined, and PSUs were sampled from the 60 strata. All together, 77 PSUs were drawn in this first stage of sampling.

School Sampling

In the second stage of sampling, schools were sampled from within the selected PSUs. Schools were selected with probability proportional to a measure of size based on the estimated number of age-eligible students in the school. This in turn was estimated by applying population-level percentages of age-eligible students within each grade to estimated grade enrollments for each grade, and aggregating to an age-eligible total for the school.

The sampling frame (i.e., list of eligible schools) for public schools was the National Center for Education Statistics (NCES) Common Core of Data (CCD) Public Elementary and Secondary School Universe file corresponding to the 2001–2002 school year. (The CCD is a program of NCES that annually compiles information about the nation's public schools and school districts, and makes this information available through a public database. For more information, see http://nces.ed.gov/ccd.) The CCD file lists all public schools that were open in the 2001–2002 school year. This frame was pared down, or subsetted, to include only the sampled PSUs, and schools with no grade higher than first were also excluded. Table A-2 presents tabulations of the number of public schools on the subsetted frame, within the eight strata defined by NAEP region and metropolitan status.

The sampling frame for private schools was developed from the 2002 Private School Survey (PSS), which was carried out by the U.S. Census Bureau for NCES. PSS is a biennial mail survey of all private schools in the 50 states and the District of Columbia. This frame was

Table A-2. Number of schools and estimated number of students within the sampled primary sampling units (PSUs) for public schools, by NAEP region and metropolitan status: 2004

NAEP region and metropolitan status	Number of schools in sampled PSUs	Estimated number of students		
		Age 9	Age 13	Age 17
Total	34,873	3,696,519	3,691,174	3,316,088
Northeast				
Metropolitan	9,369	635,170	632,780	566,142
Nonmetropolitan	86	75,905	82,040	66,444
Southeast				
Metropolitan	4,557	629,426	636,754	520,962
Nonmetropolitan	181	252,517	263,131	222,149
Central				
Metropolitan	6,258	605,103	605,099	584,520
Nonmetropolitan	198	230,639	248,327	239,233
West				
Metropolitan	13,954	1,093,435	1,041,868	934,229
Nonmetropolitan	270	174,324	181,175	182,409

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Assessment.

Table A-3. Number of schools and estimated number of students within the sampled primary sampling units (PSUs), by private school affiliation: 2004

School affiliation	Number of schools in sampled PSUs	Estimated number of students		
		Age 9	Age 13	Age 17
Total	17,128	416,030	393,478	305,595
Roman Catholic	4,078	196,072	194,655	149,881
Lutheran	856	19,048	16,628	4,363
Conservative Christian	2,120	65,745	58,389	38,681
Other religious	3,431	79,534	71,909	55,718
Other nonreligious private	5,840	55,631	51,897	56,952
Unknown affiliation	—	—	—	—

— Not available.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Assessment.

also subsetted to the sampled PSUs. Tabulations of the number of private schools on the subsetted frame are presented in table A-3.

The 2004 NAEP school frame was derived from the 2001–2002 CCD and the 2002 PSS, whereas the 2003 NAEP school frame was derived from the 2000–2001 CCD and the 2001 PSS. The 2003 NAEP school frame was a grade-based school frame, consisting of schools that included a fourth grade, eighth grade, or twelfth grade. The frame totals were of estimated grade enrollment of the schools in the frame. The NAEP 2004 school frame was an age-based school frame in a PSU sample. The 2004 frame totals were of estimated student enrollment at the target ages in each school, and were divided by the PSU probabilities of selection. The 2003 and 2004 frames were compared with respect to percentages of students who were Black, Hispanic, Asian/Pacific Islander, and American Indian/Alaska Native; median family income; and type of location. Any differences between the frames were small and reflected minor student population shifts over the one-year period.

Stratification of the school frame was an implicit stratification, using systematic sampling through a sorted file. Implicit stratification gains some of the benefits of stratification by considerably reducing the variability in the sample size between targeted student groups (so that the percentage for these student groups in the sample is close to the percentage in the population). The highest levels of school stratification were public/private status and certainty/noncertainty PSU status. Within the certainty PSU strata, the next highest level was NAEP region. Within noncertainty PSU strata, the next level was PSU stratum. This difference between certainty and noncertainty strata reflects the very different sample designs within these two types of PSUs: in the first case, schools are the first stage of selection; in the second case, PSUs are the first stage of selection, and schools are the second stage of selection. The sort order for the remaining stratifiers varied for public and private schools and for certainty and noncertainty PSUs. These stratifiers included type of location, racial/ethnic stratum, age-eligible students, and school type (for private schools).

To account for the possibility of a sampled school refusing to participate in NAEP, a set of replacement schools was identified. Any unsampled school that was neighbor to a sampled school in the implicit-stratification sort order was identified as a potential replacement, respecting such "hard boundaries" as NAEP region; PSU stratum; type of location stratum and race/ethnicity stratum for public schools; and school type for private schools. If no unsampled school satisfied these criteria for a particular sampled school, then that sampled school had no replacement. A replacement school was recruited only after the originally sampled school gave a firm and final refusal.

Student Sampling

In the third stage of sampling, students were sampled from within schools. Sampled schools were asked to list all students with the appropriate birth dates for each specified age sample. All eligible students up to a prespecified maximum were then selected for the assessment. The maxima were 128 students for ages 9 and 13 and 121 students for age 17. For instance, if a school selected for the age 9 or age 13 samples had 128 or fewer students, all age-eligible students were selected into the sample for that school. Otherwise, a sample of 128 age-eligible students was taken.

In the 2004 NAEP long-term trend assessments, there were multiple session types, corresponding to the modified trend assessments in mathematics and reading and the bridge assessments in mathematics and reading. The target sample sizes varied according to assessment session type (as shown in table A-1). Within schools, sampled students were randomly assigned to the various types of assessment sessions in such a way that the correct proportions of students were assigned to each type of session. For the age 9 and age 13 samples, roughly 60 percent of sampled students within each school were assigned to the modified assessment sessions, and 40 percent were assigned to bridge assessment sessions. The proportions were slightly different for the age 17 samples, in which roughly 67 percent of sampled students within each school were assigned to a modified assessment session, and 33 percent were assigned to a bridge assessment session. These within-school sampling procedures helped to ensure that the target sample sizes for each session type were met.

The actual student sample sizes obtained in the NAEP long-term trend assessments, as well as the school and student participation rates, are presented in tables A-4 through A-7. Student sample sizes appear in tables A-4 and A-6. School and student participation rates are shown in tables A-5 and A-7. For assessments conducted before 1984, the school and student participation rates were obtained from the NCES public use data tape user guides. Rates for more recent assessments were obtained from reports of NAEP field operation and data collection activities. Although sampled schools that refused to participate were replaced, school participation rates were computed based on the schools originally selected for participation in the assessments. The student participation rates represent the percentage of students assessed of those invited to be assessed, including those assessed in follow-up sessions when necessary.

Table A-4. Student sample sizes for the reading long-term trend scaling: 1971–2004

Age	1971	1975	1980	1984	1988	1990	1992	1994	1996	1999	2004 Bridge	2004 Modified
Total	72,400	62,700	61,600	70,200	11,400	13,300	13,400	15,700	15,200	17,000	13,900	24,100
Age 9	23,200	21,700	21,200	22,300	3,800	4,300	4,900	5,300	5,000	5,800	4,600	7,500
Age 13	25,500	21,400	22,300	22,700	4,000	4,600	4,000	5,500	5,500	5,900	4,700	8,300
Age 17	23,700	19,600	18,100	25,200	3,700	4,400	4,400	4,800	4,700	5,300	4,600	8,300

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading Assessments.

Table A-5. School and student participation rates for the reading long-term trend assessments: 1971–2004

Year	Age	Weighted percent of schools participating before substitution	Weighted percent of students participating	Overall participation
1971	9	92.5	90.9	84.1
	13	92.0	84.2	77.5
	17	90.5	73.5	66.5
1975	9	93.9	87.2	81.9
	13	92.8	85.2	79.1
	17	91.0	73.2	66.6
1980	9	94.5	90.5	85.5
	13	93.2	85.5	79.7
	17	90.5	74.2	67.2
1984	9	88.6	92.9	82.3
	13	90.3	89.2	80.5
	17	83.9	78.9	66.2
1988	9	87.2	92.5	80.7
	13	92.7	90.2	83.6
	17	78.1	82.1	64.1
1990	9	87.0	92.5	80.5
	13	89.0	90.2	80.3
	17	79.0	82.1	64.9
1992	9	87.0	93.8	81.6
	13	85.3	90.8	77.5
	17	80.9	83.3	67.4
1994	9	86.7	94.1	81.6
	13	79.7	91.8	73.2
	17	80.1	84.2	67.4
1996	9	83.5	95.6	79.9
	13	82.0	92.2	75.6
	17	81.7	83.8	68.5
1999	9	84.9	94.4	80.2
	13	80.8	92.1	74.4
	17	74.0	80.2	59.4
2004 bridge	9	85.4	94.5	80.7
	13	83.2	92.4	76.9
	17	73.4	75.5	55.4
2004 modified	9	85.1	94.3	80.2
	13	82.5	92.4	76.2
	17	74.1	76.0	56.3

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading Assessments.

Table A-6. Student sample sizes for the mathematics long-term trend scaling: 1978–2004

Age	1978	1982	1986	1990	1992	1994	1996	1999	2004 Bridge	2004 Modified
Total	65,700	44,100	17,000	17,300	17,600	15,500	14,600	15,800	14,700	22,400
Age 9	14,800	12,000	6,900	6,200	7,300	5,700	5,400	6,000	5,200	7,300
Age 13	24,200	15,800	6,200	6,600	5,900	6,100	5,700	5,900	5,700	7,500
Age 17	26,800	16,300	3,900	4,400	4,400	3,800	3,500	3,800	3,800	7,600

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1978–2004 Long-Term Trend Mathematics Assessments.

Table A-7. School and student participation rates for the mathematics long-term trend assessments: 1973–2004

Year	Age	Weighted percent of schools participating before substitution	Weighted percent of students participating	Overall participation
1973	9	93.9	90.9	85.4
	13	93.8	84.2	79.0
	17	92.4	73.5	67.9
1978	9	91.5	87.2	79.8
	13	91.5	85.2	78.0
	17	89.5	73.2	65.5
1982	9	88.3	90.5	79.9
	13	89.2	85.5	76.3
	17	86.5	74.2	64.2
1986	9	88.7	92.9	82.4
	13	88.1	89.2	78.6
	17	82.7	78.9	65.3
1990	9	87.0	92.5	80.5
	13	89.0	90.2	80.3
	17	79.0	82.1	64.9
1992	9	87.8	94.4	82.9
	13	85.6	90.9	77.8
	17	81.0	82.3	66.7
1994	9	87.1	94.4	82.2
	13	80.4	92.3	74.2
	17	79.5	84.8	67.4
1996	9	82.6	95.4	78.8
	13	80.8	92.6	74.8
	17	75.6	84.1	63.6
1999	9	83.5	93.7	78.3
	13	79.3	92.5	73.4
	17	72.1	81.3	58.6
2004 bridge	9	85.3	94.3	80.4
	13	82.5	92.1	76.0
	17	74.4	76.4	56.8
2004 modified	9	85.1	93.9	79.9
	13	82.5	91.8	75.7
	17	74.1	75.4	55.9

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1973–2004 Long-Term Trend Mathematics Assessments.

The overall response rate (the product of the weighted school participation rate before substitution and the weighted student participation rate) for age 17 fell below the NCES reporting target of 85 percent for ages 13 and 17 at the school level and for age 17 at the student level. At age 13, a bias was found for private schools, as a greater proportion of nonresponses were from other private schools as compared to Catholic schools, which were more likely to respond. In addition, nonrespondent schools had a lower percentage of Black students than schools that participated in the long-term trend assessment. Likewise, at age 17, private schools were disproportionately less likely to participate, and within private schools, Catholics and Conservative Christian schools had higher participation rates than other private schools. Nonrespondent schools also had a slightly higher percentage of Asian students compared to participating schools at age 17. At the student level at age 17, some bias was shown for race/ethnicity, free lunch eligibility, and disabled students.

Student Exclusion Rates

Some students selected for participation in the NAEP long-term trend assessments were identified as English language learners (ELL) or students with disabilities (SD). In previous long-term trend assessments, if it was decided that a student classified as SD or ELL could not meaningfully participate in the NAEP assessment for which he or she was selected, the student was, according to NAEP guidelines, excluded from the assessment.

For each student selected to participate in NAEP who was identified as either SD or ELL, a member of the school staff most knowledgeable about the student completed an SD/ELL questionnaire. Students with disabilities were excluded from the assessment if an IEP (individualized education program) team or equivalent group determined that the student could not participate in assessments such as NAEP; if the student's cognitive functioning was so severely impaired that the student could not participate; or if the student's IEP required that the student be tested with an accommodation or adaptation not permitted or available in NAEP, and the student could not demonstrate his/her knowledge of the assessment subject area without that accommodation or adaptation. A student who was identified as ELL and who was a native speaker of a language other than English was excluded if the student had received instruction in the assessment's subject area (e.g., reading or mathematics) primarily in English for less than three school years, including the current year, or if the student could not demonstrate his or her knowledge of reading or mathematics in English without an accommodation or adaptation.

In recent years, changes in policy and legislation pertaining to civil rights have resulted in assessment accommodations being permitted for SD and ELL students selected to participate in NAEP. Such accommodations enable students needing accommodations to participate in the NAEP assessments under modified conditions whereas, before, they were excluded. Future NAEP long-term trend assessments, beginning with the 2004 modified trend assessments, will offer accommodations for these students. For consistency with trend assessments in past years, however, accommodations were not offered to students in the 2004 bridge assessment samples. The exclusion rates percentage of sampled students who were excluded from the assessment for NAEP long-term trend assessments administered since 1990 are presented in table A-8.

Table A-8. Student exclusion rates for the reading and mathematics long-term trend assessments: 1990–2004

Assessment and age	1990	1992	1994	1996	1999	2004 Bridge	2004 Modified
Reading							
Age 9	5.5 (0.45)*	6.6 (0.37)	7.4 (0.56)	8.1 (0.88)	7.9 (0.73)	8.1 (0.68)	5.2 (0.48)*
Age 13	5.3 (0.47)*	5.7 (0.40)*	6.1 (0.53)*	6.9 (0.53)	6.5 (0.64)*	8.3 (0.51)	4.9 (0.39)*
Age 17	4.5 (0.28)*	5.3 (0.33)*	5.2 (0.45)*	7.3 (0.53)	6.0 (0.58)	6.7 (0.37)	3.7 (0.34)*
Mathematics							
Age 9	5.3 (0.44)*	6.7 (0.38)	7.8 (0.57)	7.8 (0.88)	7.4 (0.66)	7.0 (0.59)	2.9 (0.28)*
Age 13	5.3 (0.47)*	6.0 (0.43)*	6.2 (0.54)*	6.5 (0.52)	6.1 (0.64)	7.9 (0.58)	3.2 (0.29)*
Age 17	4.5 (0.27)*	5.4 (0.34)*	5.3 (0.45)*	7.4 (0.53)	6.1 (0.59)	7.3 (0.47)	3.2 (0.31)*

* Significantly different from 2004 Bridge.

NOTE: Standard errors of the exclusion rates appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1990–2004 Long-Term Trend Reading and Mathematics Assessments.

Data Collection and Scoring

Scoring the Booklets

Materials from the NAEP 2004 trend assessments were shipped to Pearson Educational Measurement in Iowa City, Iowa, for processing. Receipt and quality control were managed through a sophisticated bar coding and tracking system. After all appropriate materials were received from a school, the assessment booklets were scored. The reading and mathematics trend assessments included multiple-choice questions, which were machine-scored by optical-mark reflex scanning, and constructed-response questions, which were scored by professional scoring personnel using an image-based scoring system that routes student responses directly to each scorer. Each constructed-response question had a unique scoring guide that defined the criteria to be used in evaluating students' responses. Scorer consistency was monitored throughout the process through ongoing reliability checks and frequent backreading of scored papers by scoring supervisors. After the professional scoring, the booklets were scanned, and all information was transcribed to the NAEP database at Educational Testing Service (ETS). Each processing activity was conducted with rigorous quality control. An overview of the professional scoring for reading and mathematics constructed-response questions follows.

Scoring the Reading Constructed-Response Questions

The 2004 reading bridge assessment included five questions at age 9 for which students were required to construct written responses, seven such questions at age 13, and eight such questions at age 17. The 2004 modified trend assessment included four constructed-response questions at age 9, seven such questions at age 13, and eight such questions at age 17. Some of the questions were administered to more than one age group of students.

The scoring guides for the constructed-response reading questions focused on the students' ability to perform various reading tasks—for example, identifying the author's message or mood, making predictions based on given details, supporting an interpretation, and comparing and contrasting information. The scoring guides for the reading questions varied somewhat, but typically included a distribution of five rating categories. Some of the scoring guides included secondary scores, which typically involved categorizing the kind of evidence or details the student used as support for an interpretation.

The training program for scoring the constructed-response questions in reading was carried out on each assessment question separately for each age group and covered the range of student responses. Because the purpose of the scoring was to measure trends since the 1984 assessment, preparation for training included rereading hundreds of 1984 responses and compiling training sets. In order to ensure continuity with the past scoring of the trend questions, at least half of the sample papers in the training sets were taken from the 1984 training sets, and previously scored 1984 booklets were masked to ensure that scoring for training and the subsequent trend reliability scoring would be done without knowledge of the previous scores given.

The training was conducted by ETS staff assisted by Pearson's scoring director and team leaders. Training began with each reader receiving a photocopied packet of materials consisting of a scoring guide, a set of 15 to 20 scored samples, and an additional 20 to 40 response samples to be scored. The trainers reviewed the scoring guide, explained all the applicable score points, and elaborated on the rationale used to arrive at a particular score. The readers then reviewed the 15 to 20 scored samples as the trainers clarified and elaborated on the scoring guide. After this explanation, the additional samples were scored and discussed until the readers were in agreement. If necessary, additional packets of 1984 responses were used for practice scoring.

As a further step to achieve reliability with 1984, a 25 percent sample of the 1984 responses was scored on separate scoring sheets following the formal training session. These sheets were key-entered and a computerized report was generated comparing the new scores with those assigned in 1984. After some further dis-

Table A-9. Percentage exact agreement between readers for the reading long-term trend assessment scoring: 2004

	1996 responses rescored in 2004		1999 responses rescored in 2004		2004 bridge responses scored twice		2004 modified responses scored twice	
Age	Mean percent agreement	Range of agreement	Mean percent agreement	Range of agreement	Mean percent agreement	Range of agreement	Mean percent agreement	Range of agreement
Age 9	79.5	76.0-83.1	79.2	79.2-79.2	90.5	76.8-97.4	86.1	78.1-92.8
Age 13	72.3	60.6-83.6	60.5	60.5-60.5	84.2	74.2-91.5	81.4	75.6-88.7
Age 17	72.7	60.8-84.8	73.7	73.7-73.7	90.2	86.3-94.1	88.2	84.1-93.6

NOTE: The reading scoring was generally based on 5 scoring categories.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

cussion, scoring of the 2004 responses began. Three reliability studies were conducted as part of this scoring. For the 2004 material, 25 percent of the constructed responses from the modified assessment and 33 percent of the constructed responses from the bridge assessment were scored by a second reader to produce interreader reliability statistics. In addition, a trend reliability study was conducted by rereading 20 percent of the 1996 responses. Finally, a trend reliability study was conducted by rereading 20 percent of the 1999 responses. The reliability information from these studies for reading is shown in table A-9.

Scoring the Mathematics Constructed-Response Questions

The 2004 mathematics bridge study included 28 constructed-response questions at ages 9, 27 at age 13, and 19 at age 17. The modified trend assessment included 37 constructed-response questions at ages 9 and 13 and 34 at age 17.

Most of the constructed-response questions in the mathematics trend assessment were scored dichotomously, as either correct or incorrect. The scoring guides identified the correct or acceptable answers for each question in each block. The scores for these questions included 0 for no response; 1 for an incorrect response or, for the bridge assessment only, an "I don't know" response; and 2 for a correct response. The values of 1 (incorrect) and 2 (correct) were subsequently rescaled to 0 and 1, respectively, for the estimation of scale scores using item response theory (IRT). The IRT scaling procedures are described later in this appendix. Because of the straightforward nature of the scoring, lengthy training was not required. In an orientation period, the readers were trained to follow the procedures for scoring the mathematics questions and given an opportunity to become familiar with the scoring guides, which listed the correct answers for the questions in each of the blocks.

During the scoring, 25 percent of constructed responses from the modified assessment and 33 percent of constructed responses from the bridge assessment were scored by a second reader to provide a quality check. These quality checks were recorded on a separate sheet with the few discrepancies noted, and the scores were corrected. For the most part, the discrepancies were due to a score not being coded for a response to a question. Percent agreement rates between readers for mathematics constructed-response questions are shown in table A-10. Note that only within-year reliability information was obtained for mathematics; mathematics trend papers from previous assessment years were not available.

Table A-10. Percentage exact agreement between readers for the mathematics long-term trend assessment scoring: 2004

	2004 bridge responses scored twice		2004 modified responses scored twice	
Age	Mean percent agreement	Range of agreement	Mean percent agreement	Range of agreement
Age 9	99.3	98.3-100	99.7	99.1-100
Age 13	99.2	97.4-100	99.5	97.9-100
Age 17	98.5	96.2-99.9	99.3	98.0-100

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

Weighting

A complex sampling design was used to select the students who were assessed. The properties of a sample selected through such a design can 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. The weights permit valid inferences to be drawn between the student samples and the respective populations from which they were drawn and, most importantly, ensure that the results of the assessments are fully representative of the populations under study. This procedure also permits the preparation of unbiased estimates of standard error. All population and subpopulation characteristics based on the assessment data were estimated using sampling weights. These weights included adjustments for school and student nonresponse.

The final weights assigned to each school and student as a result of the estimation procedures are the product of the following steps: assignment of a base weight reflecting the reciprocal of the initial probabilities of school and student selection; adjustment of the school base weights to reduce variability; adjustments for school and student nonresponse; adjustment (if needed) to reflect assignment to a specified assessment subject; and poststratification (if applicable), which adjusts the student weights to reduce variability, by benchmarking to known student counts obtained from independent sources, such as the U.S. Census Bureau. Detailed descriptions of the weighting procedures applied to the trend assessment sample design and population structure are lengthy and complex; only a general overview of the trend assessment weighting is provided in this appendix. Further detail is available on the NAEP website at http://nces.ed.gov/nationsreportcard.

School base weights are assigned separately by age level and, as noted above, are the reciprocal of the school's probability of selection.

Each sampled student received a student base weight, whether or not the student participated in the assessment process. The base weight represents the number of students in the population of interest that the sampled students represent. Summing the student base weights for a given student group provides an estimate of the total number of students in that group.

Since nonresponse is unavoidable in any survey of a human population, a weighting adjustment is introduced to compensate for the loss of sample data and to improve the precision of the assessment estimates. Nonresponse adjustment is applied at both the school and the student levels: the weights of responding schools are adjusted to reflect the nonresponding schools, and the weights of responding students, in turn, receive an adjustment to reflect nonresponding students.

Students are assigned in a random fashion to assessment booklets. Any nonresponse bias resulting from unequal nonresponse is adjusted for across different kinds of schools and students by ensuring homogeneity either in response propensity or in characteristics associated with achievement level.

The complexity of the sample selection process as well as the variations in school enrollment can result in extremely large weights for both some schools and some students. Since unusually large weights are likely to produce large sampling variances for statistics of interest, and especially so when the large weights are associated with sample cases reflective of rare or atypical characteristics, such weights usually undergo an adjustment procedure that "trims" or reduces extreme weights. Again, the motivation is to improve the precision of the survey estimates.

Prior to 2004, NAEP long-term trend samples used weights that had been poststratified to the census or Current Population Survey (CPS) totals for the populations being assessed. Due to concerns about the availability of appropriate targets for poststratification as a result of changes in the reporting of race in the 2000 census, nonpoststratified weights have been used in the analysis of main NAEP national samples since 2002. The 2004 NAEP trend assessment samples for both assessment types were analyzed using nonpoststratified weights.

Estimates of the sampling variance of statistics derived through the assessment effort are developed through a replication method known as "jackknife." This process of replication involves the repeated selection of portions of the sample (replicates). A separate set of weights is produced for each replicate, using the same weighting procedures as for the full sample. The replicate weights, in turn, are used to produce estimates for each replicate (replicate estimates). The variability among the calculated replicate estimates is then used to obtain the variance of the full-sample estimate.

Data Analysis and IRT Scaling

After the assessment information in the NAEP database was compiled and the sampling weights applied, a variety of analyses were performed to check the accuracy of results in the database. Analyses were first 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 after the last question the student answered) and missing responses before the last observed response. Missing responses before the last observed response were considered intentional omissions. In analysis, omitted responses to multiple-choice questions were scored as fractionally correct (Lord 1980, p. 229). Omitted responses for constructed-response questions were placed into the lowest score category. Missing responses after the last observed response were considered "not reached" and treated as if the questions had not been presented to the student. Average percent missing rates were calculated by first averaging across items. 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 mathematics blocks that end with extended constructed-response questions, there may be extremely large drops in the proportion of students attempting some of the final questions. 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 rates for the reading trend assessments are presented in table A-11. Similar information for the mathematics trend assessments appears in table A-12.

Item response theory (IRT) was used to estimate average proficiency for the nation and various student groups of interest within the nation. IRT scaling was performed separately within each age level for each of the two trend assessments (reading and mathematics). Each assessment employs slightly different steps in data analysis and IRT scaling. Because these descriptions are rather lengthy, they are not repeated in this appendix but can be found online at http://nces.ed.gov/nationsreportcard. IRT models the probability of answering a question correctly 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 age, assessment year, or other characteristics (e.g., race/ethnicity or gender), even when students receive different blocks of items. One desirable feature of IRT is that it locates items and students on this common scale. In contrast to classical test

Table A-11. Summary item response rates for the reading long-term trend assessment, by different types of response: 2004

Item response types	Bridge assessment: Multiple-choice items	Bridge assessment: Constructed-response items	Modified assessment: Multiple-choice items	Modified assessment: Constructed-response items
Age 9				
Average percent missing[1]	6.33	21.27	2.79	8.76
Minimum	0.57	8.45	0.41	4.85
Maximum	24.04	36.87	13.71	13.67
Average percent off-task[2]	†	1.32	†	1.09
Minimum	†	0.77	†	0.38
Maximum	†	2.27	†	1.69
Average weighted proportion correct	53.52	19.54	62.02	33.15
Age 13				
Average percent missing[1]	3.73	9.88	2.30	4.82
Minimum	0.15	3.32	0.22	2.52
Maximum	26.56	26.66	9.99	7.54
Average percent off-task[2]	†	0.52	†	0.32
Minimum	†	0.00	†	0.20
Maximum	†	1.21	†	0.39
Average weighted proportion correct	66.43	45.96	74.49	48.38
Age 17				
Average percent missing[1]	3.87	11.69	2.14	6.10
Minimum	0.13	2.58	0.28	3.53
Maximum	23.35	32.97	13.40	9.18
Average percent off-task[2]	†	1.17	†	1.04
Minimum	†	0.00	†	0.34
Maximum	†	2.52	†	1.62
Average weighted proportion correct	72.59	52.35	78.23	52.87

† Not applicable.
[1]"Missing" includes the categories "omitted" and "not reached." The percentages are calculated first across students within an item and then averaged across all items.
[2]"Off-task" is only relevant for constructed-response items and refers to responses that are unrelated to the question and are considered inappropriate.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

Table A-12. Summary item response rates for the mathematics long-term trend assessment, by different types of response: 2004

Item response types	Bridge assessment: Multiple-choice items	Bridge assessment: Constructed-response items	Modified assessment: Multiple-choice items	Modified assessment: Constructed-response items
Age 9				
Average percent missing[1]	2.10	2.51	5.26	5.35
Minimum	0.08	0.12	0.44	0.67
Maximum	9.39	6.69	19.10	16.37
Average percent off-task[2]	†	0.03	†	0.10
Minimum	†	0.00	†	0.00
Maximum	†	0.19	†	2.11
Average weighted proportion correct	63.57	69.01	63.55	66.52
Age 13				
Average percent missing[1]	0.74	2.53	3.03	3.70
Minimum	0.12	0.16	0.18	0.11
Maximum	2.95	11.52	21.88	21.45
Average percent off-task[2]	†	0.05	†	0.04
Minimum	†	0.00	†	0.00
Maximum	†	0.19	†	0.28
Average weighted proportion correct	67.15	70.94	67.05	67.69
Age 17				
Average percent missing[1]	1.18	5.84	2.34	7.50
Minimum	0.27	0.29	0.48	0.94
Maximum	3.16	12.76	15.90	26.73
Average percent off-task[2]	†	0.27	†	0.32
Minimum	†	0.00	†	0.02
Maximum	†	0.53	†	1.13
Average weighted proportion correct	69.52	48.33	70.64	55.95

† Not applicable.
[1]"Missing" includes the categories "omitted" and "not reached." The percentages are calculated first across students within an item and then averaged across all items.
[2]"Off-task" is only relevant for constructed-response items and refers to responses that are unrelated to the question and are considered inappropriate.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

theory, IRT does not rely solely on the total number of correct item responses, but uses the particular patterns of student responses to items in determining the student location on the scale. As a result, adding items that function at a particular point on the scale to the assessment does not change the location of the students on the scale, even though students may respond correctly to more items. It does increase the precision with which students are measured, particularly for those students whose scale locations are close to the additional items.

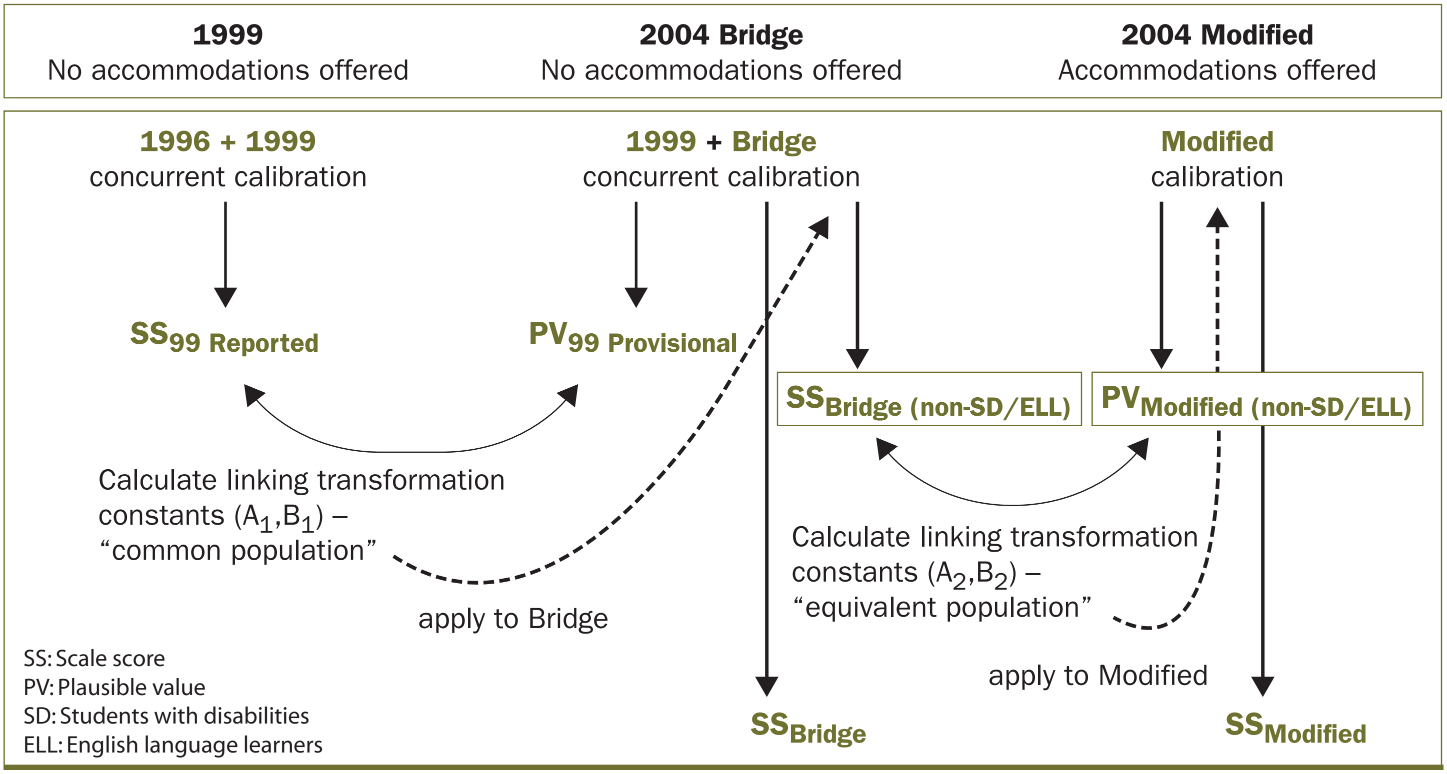
The reading and mathematics trend assessments are composed of three types of questions: multiple-choice, short constructed-response (scored either dichotomously or allowing for partial credit), and extended constructed-response (scored according to a partial-credit model). Prior to 2004, all constructed-response items were dichotomized for analysis. In all assessments, multiple-choice questions were scaled using the three-parameter logistic (3PL) IRT model; short constructed-response questions rated as acceptable or unacceptable were scaled using a two-parameter (2PL) IRT model. In the 2004 modified assessment, the constructed-response items scored in three or more categories were not dichotomized for analysis; an additional IRT model was introduced. Short constructed-response questions, rated according to a three-level guide, as well as extended constructed-response questions, rated on a four- or five-level guide, were scaled using a generalized partial-credit (GPC) model (Muraki 1992). Developed by ETS and first used in 1992, the GPC model permits the scaling of questions scored according to multipoint rating 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.

In NAEP assessments, students do not receive enough questions about a specific topic to permit reliable estimates of individual performance. Traditional test scores for individual students, even those based on IRT, would result in misleading estimates of population characteristics, such as student group means and percentages of students at or above a certain scale-score level. However, it is NAEP's goal to estimate these population characteristics. NAEP's objectives can be achieved with methodologies that produce estimates of the population-level parameters directly, without the intermediary computation of estimates for individuals. This is accomplished using marginal estimation scaling model techniques for latent variables (Mislevy and Sheehan 1987). 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. This would not be the case for population estimates obtained by aggregating optimal estimates of individual performance. (For theoretical and empirical justification of the procedures employed, see Mislevy 1988.)

Linking the Bridge and Modified Assessments

For the 2004 reading and mathematics trend assessments, separate IRT scales were constructed within each age level; results are reported on a scale ranging from 0 to 500. These scales were linked to the previously established scales within each subject area using common-population linking procedures. Specifically, the bridge assessment results were linked to the trend scales established in 1984 (for reading) or 1986 (for mathematics) and extending to the most recent trend point in 1999. Results for the modified trend assessments were subsequently linked to the trend scale using equivalent populations. The linking mechanism used is shown in figure A-3.

Figure A-3. Linking design for the long-term trend assessment: 2004



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading and Mathematics Assessments.

In 1999, scale scores on the trend line were obtained (as noted by the SS_{99} Reported). The 2004 bridge assessment was identical to the 1999 assessment, and accommodations were not offered in either assessment. Therefore, standard NAEP linking procedures, known as common population design, were used: a concurrent calibration was performed using the 1999 data and the 2004 bridge data. The item parameters from this concurrent calibration were then used to obtain plausible values for both 1999 (as noted by the $PV_{99\ Provisional}$) and the 2004 bridge. Linking transformation constants (A_1 and B_1) were calculated to place the two sets of 1999 results onto the same scale (i.e., set the mean and standard deviation equal). These same transformation constants were then applied to the 2004 bridge plausible values, creating 2004 bridge scale scores that are on the trend line (as noted by the SS_{Bridge} at the bottom of the diagram).

To link the 2004 modified assessment to the 2004 bridge assessment, the equivalent population design was used. Recall that accommodations were offered in the modified assessment, but not in the bridge assessment. Therefore, the non-SD/ELL portions (i.e., students who were not identified as being SD and/or ELL) of both the bridge and modified samples are theoretically randomly equivalent samples. Plausible values were obtained for the modified sample using a single-sample calibration. The subset of non-SD/ELL students of the modified sample (from these plausible values) was used to calculate the linking transformation constants (A_2 and B_2) to the non-SD/ELL subset of scale scores obtained for the bridge sample. These linking transformation constants were then applied to all students in the modified sample, thus creating 2004 modified scale scores that are linked to the existing trend line (as noted by the $SS_{Modified}$ at the bottom of the diagram).

Creating the Trend Lines

The reading trend scale was constructed based on the 1984 assessment and included all previous reading assessments. The mathematics trend scale was developed based on the 1986 assessment and also included previous mathematics trend assessments. The initial trend scaling, however, did not include the 1973 mathematics assessment because it had too few questions in common with subsequent mathematics assessments. To provide a link to the early assessment results for the nation and for student groups defined by race/ethnicity and gender at each of three age levels, estimates of average scale scores were extrapolated from previous analyses. The extrapolated estimates were obtained by assuming that, within a given age level, the relationship between the logit transformation of a student group's average p-value (i.e., average proportion correct) for common questions and its scale score average was linear, and that the same line held for all assessment years and for all student groups within the age level. More details about how these estimates were extrapolated appear in the 1986 NAEP technical report (Beaton and Barone 1988). Because of the need to use extrapolation of the average scale scores for these early assessments, caution should be used in interpreting the patterns of trends across those assessment years.

As described earlier, the NAEP scales for the trend assessment subjects make it possible to examine relationships between students' performance and a variety of background factors measured by NAEP. The fact that a relationship exists between achievement and another variable, however, does not reveal the underlying cause of the relationship, 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 information about the student population and the educational system, such as trends in instruction, changes in the school-age population, and societal demands and expectations.

Setting the Performance Levels

To facilitate interpretation of the NAEP results, the scales were divided into successive levels of performance, and a "scale anchoring" process was used to define what it means to score in each of these levels. NAEP's scale anchoring follows an empirical procedure whereby the scaled assessment results are analyzed to delineate sets of questions that discriminate between adjacent performance levels on the scales. For the reading and mathematics trend scales, these levels are 150, 200, 250, 300, and 350. For these five levels, questions were identified that were likely to be answered correctly by students performing at a particular level on the scale and much less likely to be answered correctly by students performing at the next lower level. The guidelines used to select such questions were as follows: students at a given level must have at least a specified probability of success with the questions (65 percent for mathematics, 80 percent for reading), while students at the next lower level must have a much lower probability of success (that is, the difference in probabilities between adjacent levels must exceed 30 percent). For each curriculum area, subject-matter specialists examined these empirically selected question sets and used their professional judgment to characterize each level. The reading scale anchoring was conducted on the basis of the 1984 assessment, and the scale anchoring for mathematics trend reporting was based on the 1986 assessment.

NAEP Reporting Groups

This report contains results for the nation and for groups of students within the nation defined by shared characteristics. The student groups defined by gender, race/ethnicity, parents' education level, and type of school are discussed below.

Gender. Results are reported separately for males and females. Gender was reported by the student.

Race/Ethnicity. Results are presented for students in different racial/ethnic groups according to the following mutually exclusive categories: White, Black, and Hispanic. Results for Asian/Pacific Islander and American Indian (including Alaska Native) students are not reported separately because there were too few students in the groups for statistical reliability. The data for all students, regardless of whether their racial/ethnic group was reported separately, were included in computing the overall national results.

In NAEP long-term trend assessments, data about student race/ethnicity have been collected in three ways: through observation, school records, and student self-reports.

Observed Race/Ethnicity. Students were assigned to a racial/ethnic category based on the assessment administrator's observation. Reports of NAEP long-term trend assessment results have been based upon this method of identifying students' race/ethnicity since 1971. A category for Hispanic students did not exist in 1971, but was included in subsequent years. The 2004 bridge assessment and all the previous assessments results presented in this report are based on observed race/ethnicity.

Student-Reported Race/Ethnicity. Although students participating in NAEP assessments since 1984 have been asked to self-report race/ethnicity, long-term trend assessment results have not been reported based on this method. As in previous long-term trend assessments, data on students' self-reports of ethnicity were collected in 2004.

School-Reported Race/Ethnicity. Data about students' race/ethnicity from school records were collected in 2004, but were not collected in any previous NAEP long-term trend assessment. The 2004 modified assessment results presented in this report are based on school-reported race/ethnicity.

Parents' Education Level. Students were asked to indicate the extent of schooling for each of their parents, choosing among the following options: did not finish high school, graduated from high school, had some education after high school, or graduated from college. The response indicating the higher level of education for either parent was selected for reporting. In the 2004 bridge study, the questions were presented to students in the age 9, age 13, and age 17 samples. In the 2004 modified trend assessment, however, the questions were presented only to the students in the age 13 and age 17 samples.

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. 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 category; they are included in the overall national results. Response rates for nonpublic schools selected for participation in the 2004 trend assessments failed to reach the necessary threshold for reporting; therefore, only results for the total sample and public schools are reported.

Estimating Variability

The statistics presented in this report are estimates of group performance based on samples of students, rather than the values that could be calculated if every student in the nation answered every assessment question. It is therefore important to have measures of the degree of uncertainty of the estimates. Accordingly, in addition to providing estimates of percentages of students and their average scale score, this report provides information about the uncertainty of each statistic. The corresponding standard errors for the statistics presented in this report are available from the NAEP Data Explorer at http://nces.ed.gov/nationsreportcard/naepdata/.

Two components of uncertainty are accounted for in the variability of statistics based on scale scores: the uncertainty due to sampling only a small number of students relative to the whole population, and the uncertainty due to sampling only a relatively small number of questions from the content domain. The variability of estimates of percentages of students having certain background characteristics or answering a certain cognitive question correctly is accounted for by the first component alone. Because NAEP uses complex sampling procedures, conventional formulas for estimating sampling variability that assume simple random sampling are inappropriate. For this reason, NAEP uses a jackknife replication procedure to estimate standard errors. The jackknife standard error provides a reasonable measure of uncertainty for any information about students that can be observed without error, but each student typically responds to so few questions within any content area that the scale score for any single student would be imprecise. In this case, using the plausible values methodology makes it possible to describe the performance of groups of students, but the underlying imprecision that makes this step necessary adds an additional component of variability to statistics based on NAEP scale scores (for further details see Johnson 1989).

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 on the tables in the NAEP Data Explorer 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.

The reader is reminded that NAEP results, like those from all surveys, are also subject to other kinds of errors, including the effects of necessarily imperfect adjustments for student and school nonresponse and other largely unknowable effects associated with the particular instrumentation and data collection methods used. Nonsampling errors can be attributed to a number of sources: inability to obtain complete information about all selected students in 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; respondents' inability or unwillingness to give correct information; mistakes in recording, coding, or scoring data; and other errors of collecting, processing, and estimating missing data. The extent of nonsampling errors is difficult to estimate. By their nature, the impact of such errors cannot be reflected in the data-based estimates of uncertainty provided in NAEP reports.

Drawing Inferences from the Results

The use of confidence intervals, based on the standard errors, provides a way to make inferences about the population averages and percentages in a manner that reflects the uncertainty associated with the sample estimates. An estimated sample scale score average ±2 standard errors represents about a 95 percent confidence interval for the corresponding population quantity. This means that, with 95 percent certainty, the average performance of the entire population of interest is within about ± 2 standard errors of the sample average.

For the data in this report, all the estimates have corresponding estimated standard errors of the estimate. For example, table A-13 shows the average national scale score for 2004 in reading and mathematics at all three age levels. The estimated standard errors appear in parentheses next to each estimated scale score. The estimated standard errors corresponding to other data in this report can be found in the NAEP Data Explorer on the NCES website at http://nces.ed.gov/nationsreportcard/naepdata.

Table A-13. Trends in reading and mathematics average scale scores for students ages 9, 13, and 17: 1971–2004

Assessment and age	1971 (Reading)/ 1973 (Mathematics)	1999	2004
Reading			
Age 9	208 (1.0) *	212 (1.3) *	219 (1.1)
Age 13	255 (0.9) *	259 (1.0)	259 (1.0)
Age 17	285 (1.2)	288 (1.3)	285 (1.2)
Mathematics			
Age 9	219 (0.8) *	232 (0.8) *	241 (0.9)
Age 13	266 (1.1) *	276 (0.8) *	281 (1.0)
Age 17	304 (1.1)	308 (1.0)	307 (0.8)

* Significantly different from 2004.
NOTE: Standard errors of the average scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading and Mathematics Assessments.

As an example, suppose that the average mathematics scale score of students in a particular group was 256, with a standard error of 1.2. (The calculations were completed with compounded numbers.) A 95 percent confidence interval for the population quantity would be as follows:

$$\text{Average} \pm 2 \text{ standard errors} =$$
$$256 \pm 2\,(1.2) = 256 \pm 2.4 =$$
$$256 - 2.4 \text{ and } 256 + 2.4 =$$
$$(253.6,\ 258.4)$$

Thus, one can conclude with close to 95 percent certainty that the average scale score for the entire population of students in that group is between 253.6 and 258.4.

Similar confidence intervals can be constructed for percentages, provided that the percentages are not extremely large or extremely small. For percentages, confidence intervals constructed in the above manner work best when sample sizes are large and the percentages being tested have magnitudes relatively close to 50 percent. Statements about group differences should be interpreted with caution if at least one of the groups being compared is small in size or if “extreme” percentages are being compared. Percentages, P, were treated as “extreme” if:

$$P < P_{\lim} = \frac{200}{N_{EFF} + 2}$$

where the effective sample size is

$$N_{EFF} = \frac{P(100 - P)}{(SE)^2}$$

and SE is the jackknife standard error of P.

Similarly, at the other end of the 0 to100 scale, a percentage is deemed extreme if $100 - P < P_{\lim}$. This “rule of thumb” cutoff leads to flagging a large proportion of confidence intervals that would otherwise include values less than zero or greater than one. In either extreme case, the confidence intervals described above are not appropriate, and procedures for obtaining accurate confidence intervals are quite complicated. In this case, the value of *P* was reported, but no standard error was estimated and hence no tests were conducted.

As for percentages, confidence intervals for average scale scores are most accurate when sample sizes are large. For some of the groups of students for which average scale scores or percentages were reported, student sample sizes could be quite small. For results to be reported for any group of students, a minimum sample size of 62 was required.

If students in a particular group were clustered within a small number of geographic primary sampling units (PSUs), the estimates of the standard errors might also be inaccurate. So, data for student groups were required to come from a minimum of five PSUs.

Analyzing Group Differences in Averages and Percentages

To determine whether there is a real difference between the average scale score (or percentage 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 average scale scores 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 squaring each group's standard error, summing these squared standard errors, and then taking the square root of this sum ($SE_{A-B} = \sqrt{(SE_A{}^2 + SE_B{}^2)}$). This procedure produces a conservative estimate of the standard error of the difference, since the estimates of the group averages or percentages will be positively correlated to an unknown extent due to the sampling plan. Direct estimation of the standard errors of all reported differences would involve a heavy computational burden. As with group averages or percentages, the standard error of the difference can be used to help determine whether differences between assessment years are likely to be real. If zero is within the confidence interval for the differences, there is no statistically significant difference between the groups.

To be more specific about the way in which differences between average scale scores for two groups were shown to be statistically significant with 95 percent certainty, whenever comparisons were made with the students assessed in an assessment year for which average scale scores were extrapolated (1973 for mathematics), the confidence interval was constructed using ± 2 standard errors (from a normal distribution). However, when the two groups that were being compared were from other assessments (those with scale scores estimated without extrapolation), the number multiplied by the standard error varied. This multiplier is the .975 (= 1 – .025) percentile from a *t* distribution with the degrees of freedom that vary by the values of the average scale scores, their standard errors, and the number of PSUs that contribute to the average scale scores. It is possible that scale scores that appear equal when rounded for two assessment years or two groups of students may not have the same significance test results when compared to another year or group of students. This may be due to the actual nonrounded value of the data and/or the standard error of the differences.

Conducting Multiple Tests

The procedures used to determine whether group differences in the samples represent actual differences among the groups in the population 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, there are times when many different groups are being compared (i.e., multiple sets of confidence intervals are being analyzed). To hold the significance level for the set of comparisons at a particular level (e.g., .05), the standard methods must be adjusted by multiple comparison procedures (Miller 1981). One such procedure, the Benjamini-Hochberg False Discovery Rate (FDR) procedure, was used to control the certainty level (Benjamini and Hochberg 1995).

Unlike other multiple comparison procedures, such as Bonferroni, 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, the FDR procedure used in NAEP is considered appropriately less conservative than familywise procedures for large families of comparisons (Williams, Jones, and Tukey 1999). Therefore, the FDR procedure is more suitable for multiple comparisons in NAEP than other procedures.

To illustrate how the FDR procedure is used, consider the comparisons of current and previous years' average scale scores for the five groups presented in table A-14. The test statistic shown is the difference in average scale scores divided by the estimated standard error of the difference. (Rounding of the data occurs after the test is done.)

The difference in average scale scores and its estimated standard error can be used to find an approximately 95 percent confidence interval or to identify a confidence percentage. The confidence percentage for the test statistics is identified from statistical tables. 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 was made for only one of the five groups, there would be a significant difference between the average scale scores for the two years at a significance level of less than 5 percent. However, because of interest 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 racial/ethnic 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.

Table A-14. Example of False Discovery Rate comparisons of average scale scores for different groups of students

	Previous year		Current year		Previous year and current year			
Students	Average scale score	Standard error	Average scale score	Standard error	Differences in averages	Standard error of differences	Test statistic	Percent confidence[1]
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	0.51	62
Group 5	201	3.4	196	4.7	-5.51	5.81	-0.95	35

[1]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.

Using the FDR procedure to take into account that all comparisons are of interest, the confidence percentages in the example are ordered from largest to smallest: 62, 35, 20, 4, and 1. In the FDR procedure, the adjusted level of confidence percentage is determined by the level of confidence desired times the number of comparisons minus one divided by the number of comparisons. So, 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 $0.05 \times (5 - 1)/5 = 0.04$ = 4 percent, 20 percent for the group 1 comparison would be compared to $0.05 \times (5 - 2)/5 = 0.03$ = 3 percent, 4 percent for the group 3 comparison would be compared to $0.05 \times (5 - 3)/5 = 0.02$ = 2 percent, and 1 percent for the group 2 comparison (actually slightly smaller than 1 prior to rounding) would be compared to $0.05 \times (5 - 4)/5 = 0.01$ = 1 percent. The procedure stops with the first contrast found to be significant. The last of these comparisons is the only one for which the confidence percentage is smaller than the FDR procedure value. Therefore, the difference between the current year's and previous year's average scale scores for the group 2 students is statistically significant; for all of the other groups, average scale scores for the current and the earlier year are not significantly different from one another. In practice, a very small number of counterintuitive results occur when the FDR procedures are used to examine between-year differences in results for student groups.

Cautions in Interpretations

As previously stated, the NAEP reading and mathematics trend scales make it possible to examine relationships between students' performance and various background factors measured by NAEP. However, a relationship 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 reflect 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. Additional data collected during the 2004 trend assessments in reading and mathematics not presented in this report are available from the NAEP Data Explorer at http://nces.ed.gov/nationsreportcard/naepdata/.

A caution is also warranted for some small population group estimates. Smaller population groups may show increases or decreases across years in average scores; however, it is necessary to interpret such score changes with extreme caution. The effects of exclusion-rate changes for small groups of students may be more marked for small groups than they are for the whole population. Another reason for caution is that the standard errors are often quite large around the score estimates for small groups, which in turn means the standard error around the gain is also large.

THIS PAGE INTENTIONALLY LEFT BLANK.

Appendix B
Percentage Distribution of Students Taking Each Assessment in 2004 Across Various Student Groups

The tables in appendix B show the percentages of students who took the reading and mathematics bridge and modified assessments. These percentages are broken out by various student groups. For reading, table B-1 shows the percentage distribution of students taking the bridge and modified assessments by gender, race/ethnicity, and school type. For mathematics, table B-2 shows the percentage distribution of students taking the bridge and modified assessments by gender, race/ethnicity, highest level of parental education, and school type. As discussed in appendix A, the sampling plan was designed to make the percentages as similar as possible between the bridge and modified assessments.

Table B-1. Percentage of students assessed in reading at ages 9, 13, and 17, by type of assessment and student and school characteristics: 2004

Characteristic	Age 9 Bridge	Age 9 Modified	Age 13 Bridge	Age 13 Modified	Age 17 Bridge	Age 17 Modified
Gender						
Male	50	51	49	51	49	50
Female	50	49	51	49	51	50
Race/ethnicity						
White	59	59	63	63	67	70
Black	17	16	15	16	12	12
Hispanic	17	17	17	16	15	13
Other[1]	8	7	5	5	6	6
Public school	89	89	92	90	91	90

[1]Other includes Asian/Pacific Islander students and American Indian/Alaska Native students, and students categorized in school records as another race or ethnicity.
NOTE: Detail may not sum to totals because of rounding. Data by parents' level of education are not shown because the questions used to gather these data changed in the modified reading assessments, resulting in noncomparable response percentages.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

Table B-2. Percentage of students assessed in mathematics at ages 9, 13, and 17, by type of assessment and student and school characteristics: 2004

Characteristic	Age 9 Bridge	Age 9 Modified	Age 13 Bridge	Age 13 Modified	Age 17 Bridge	Age 17 Modified
Gender						
Male	49	50	48	49	48	50
Female	51	50	52	51	52	50
Race/ethnicity						
White	59	59	64	62	68	69
Black	14	16	14	16	12	12
Hispanic	19	18	16	17	14	14
Other[1]	8	7	6	5	6	5
Parents' highest education level						
Less than high school	†	†	7	7	9	8
Graduated from high school	†	†	19	18	19	19
Some education after high school	†	†	15	15	22	22
Graduated from college	†	†	47	47	47	47
Unknown	†	†	12	13	3	4
Public school	88	90	91	92	91	91

†Not applicable.
[1]Other includes Asian/Pacific Islander students and American Indian/Alaska Native students, and students categorized in school records as another race or ethnicity.
NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

Table B-3. Percentage of students assessed in reading at ages 9, 13, and 17, by student and school characteristics: 1971, 1999 and 2004

	Age 9			Age 13			Age 17		
Characteristic	1971[1]	1999	2004	1971[1]	1999	2004	1971[1]	1999	2004
Gender									
Male	50	49	50	50	49	49	49	52	49
Female	50	51	50	50	51	51	51	48	51
Race/ethnicity									
White	84 *	69 *	59	84 *	70 *	64	87 *	72	68
Black	14 *	18 *	17	15	16	15	11	14	12
Hispanic	5 *	9 *	17	5 *	10 *	16	3 *	9 *	14
Parents' highest education level									
Less than high school	†	†	†	10 *	6	7	13 *	7	9
Graduated from high school	†	†	†	31 *	25	25	32 *	25	24
Some education after high school	†	†	†	16 *	10	8	18	17	18
Graduated from college	†	†	†	33 *	48	48	34 *	48	46
Unknown	†	†	†	10	11	12	4	3	4
Public school	89	88	89	88	87	92	93	90	91

†Not applicable.
*Significantly different from 2004.
[1]Data for Hispanic students were first available in 1975, and data for parents' education level and public schools were first available in 1980. Therefore, the data shown in the 1971 column in the table for these categories are from the 1975 and 1980 assessments, respectively.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1971–2004 Long-Term Trend Reading Assessments.

Table B-4. Percentage of students assessed in mathematics at ages 9, 13, and 17, by student and school characteristics: 1978, 1999, and 2004

	Age 9			Age 13			Age 17		
Characteristic	1978	1999	2004	1978	1999	2004	1978	1999	2004
Gender									
Male	50	49	49	50	50	48	49	48	48
Female	50	51	51	50	50	52	51	52	52
Race/ethnicity									
White	79 *	70 *	60	80 *	71 *	66	83 *	72	69
Black	14 *	18 *	15	13	15	15	12	15	13
Hispanic	5 *	8 *	18	6 *	10 *	15	4 *	10 *	14
Parents' highest education level									
Less than high school	†	†	†	12 *	6	7	13 *	7	9
Graduated from high school	†	†	†	33 *	21	19	33 *	20	19
Some education after high school	†	†	†	14	17	15	16 *	23	22
Graduated from college	†	†	†	26 *	48	47	32 *	48	47
Unknown	†	†	†	15 *	9 *	12	5	3	3
Public school	89	88	88	91	88	91	94	89	91

†Not applicable.
*Significantly different from 2004.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1978–2004 Long-Term Trend Mathematics Assessments.

THIS PAGE INTENTIONALLY LEFT BLANK.

Appendix C
Glossary of Terms

Accommodation. A change in how an assessment is presented or administered or in how a test taker is allowed to respond. In NAEP accommodations are provided to students with disabilities (SD) as specified in the student's Individualized Education Program (IEP), and to English language learners (ELL).

Assessment administrator. A trained proctor who administers an assessment.

Background questions, background questionnaires. The instruments used to collect information about student demographics and educational experiences.

Backreading. A monitoring function conducted by scoring supervisors during the scoring of NAEP constructed-response questions. These supervisors read a subset (typically 10 percent) of all scores assigned by each team of scorers daily to ensure scoring accuracy and inform additional training (group or individual) as needed.

BIB (Balanced Incomplete Block) design. A design used to pair blocks of assessment questions together in order to form NAEP assessment booklets. In a BIB design, blocks of questions are balanced; each block appears an equal number of times in each booklet position. Each block is also paired with every other block in an assessment booklet exactly the same number of times. The 2004 NAEP long-term trend assessments use a partially balanced incomplete block (pBIB) booklet design. In a pBIB design, blocks may not appear an equal number of times in each booklet position, or may not be paired with every other block an equal number of times.

Block. A group of assessment items created by dividing the item pool for an age or grade into subsets.

Booklet. The assessment instrument presented to an individual student, which is created by combining blocks of assessment items.

Bridge study. A special study that involves administering two assessments to randomly assigned samples of students. The purpose of the 2004 trend bridge study was to determine the impact (if any) on assessment results of changes to the design and administration of the NAEP long-term trend assessments. A bridge assessment (which replicated the long-term trend assessment instrument used in 1999 and all previous trend assessments) and a modified assessment (which reflected the design changes) were administered.

Constructed-response item. A non-multiple-choice item that requires the student to produce some type of written or oral response.

English language learners (ELL). NAEP relies on state and local school districts to identify which students are English language learners (ELL). States and school districts use a variety of methods to identify ELLs and to monitor the progress of the students in special programs in which they may be placed. These methods include registration and enrollment records, home language surveys, interviews, observations, referrals, classroom grades and performance, and test results.

Individualized Education Program (IEP). A program created for an individual public school student, generally for each student who receives special education and related services that is developed, reviewed, and revised in accordance with Title 42 U.S.C. Section 1414(d). It specifies any accommodations needed in order for the student to participate in standardized tests such as NAEP.

Item. The basic scorable part of an assessment; a test question.

Item response theory (IRT). Test analysis procedures that determine a mathematical model for the probability that a given examinee will respond correctly to a given assessment item.

Mean Proportion Correct. The average percentage of students answering each question correctly.

Measurable difference. A difference between statistics that has been tested by a statistical procedure and found to be unlikely to be due to sampling or measurement error. See Statistically significant.

Metropolitan statistical area (MSA). An area defined by the U.S. Census Bureau for the purposes of presenting general-purpose statistics for metropolitan areas. Typically, an MSA contains a city with a population of at least 50,000 and includes its adjacent areas.

Multiple-choice item. An item that consists of one or more introductory sentences or prompts and a question, followed by a list of response options that include the correct answer and several incorrect alternatives.

Multistage sample design. A sampling design that consists of two or more stages of sampling. The following is an example of three-stage sampling: (1) sample of counties (primary sampling units or PSUs), (2) sample of schools within each sampled PSU, and (3) sample of students within each sample school.

National Assessment Governing Board (NAGB). Independent organization whose members are appointed by the U.S. Secretary of Education. NAGB provides overall policy direction to the NAEP program. It is an independent, bipartisan group whose members include governors, state legislators, local and state school officials, educators, business representatives, and members of the general public.

National School Lunch Program (NSLP). A federally assisted meal program that provides low-cost or free lunches to eligible students. It is sometimes referred to as the free/reduced-price lunch program. Free lunches are offered to those students whose family incomes are at or below 130 percent of the poverty level; reduced-price lunches are offered to those students whose family incomes are between 130 percent and 185 percent of the poverty level.

Nonresponse. The failure to obtain responses or measurements for all of the elements in a sample.

Observed race/ethnicity. Race or ethnicity of an assessed student as perceived by the assessment administrator.

Parental education. The level of education of the mother or father of an assessed student, whichever is higher, as derived from the student's response to two background questions. It defines a NAEP reporting group.

Pilot test. A pretest of assessment questions done to obtain information regarding clarity, difficulty levels, timing, feasibility, and special administrative situations. The pilot test is done before revising and selecting the items to be used in the operational NAEP assessment.

Population. In the case of the NAEP long-term trend assessments, the population of interest is the entire collection of America's students in public or private schools at ages 9, 13, and 17 years. The small samples of students that NAEP selects for the assessment permit inferences about academic performance to be made for all school students at the three age levels.

Poststratification. Classification and weighting to correspond to external values of selected sampling units by a set of strata definitions after the sample has been selected.

Primary sampling unit (PSU). The basic geographic sampling unit for NAEP. It can be either a single county or a set of contiguous counties.

Probability sample. A sample in which every element of the population has a known, nonzero probability of being selected.

Reporting group. Groups within the national population for which NAEP data are reported; for example, those defined by gender, by race/ethnicity, by grade, by age, by level of parental education, by region, and by type of location.

Response options. In a multiple-choice question, alternatives that can be selected by a respondent.

Sample. A portion of a population, or a subset from a set of units, that is selected by some probability mechanism for the purpose of investigating the properties of the population. NAEP does not assess an entire population but rather selects a representative sample from the group to answer assessment items.

Sampling error. The error in survey estimates that occurs because only a sample of the population is observed. It is the error associated with the variation in samples drawn from the score frame population.

Sampling frame. The list of sampling units from which the sample is selected.

Sampling weight. A multiplicative factor equal to the reciprocal of the probability of a respondent being selected for assessment with adjustment for nonresponse and, perhaps, poststratification. The sum of the weights provides an estimate of the number of persons in the population represented by a respondent in the sample.

Scale score. Derived from overall level of performance of groups of students on NAEP assessment items. NAEP subject-area scale scores for the long-term trend assessments are typically expressed on 0–500 scales. When used in conjunction with interpretive aids, such as performance levels, average scale scores provide information about what a particular aggregate of students in the population knows and can do.

NOTE: In other testing programs, the scale score is derived from individual student responses to assessment items and summarizes the overall level of performance attained by that student. In NAEP, no individual scale scores are available.

Scaling. The process of assigning numbers to reflect students' performance on an assessment based on a pattern of responses. In NAEP, scaling is based on item response theory (IRT) and results in a scale score for each subject area that can be used to summarize levels of performance attained by particular groups of students.

Scoring guide. Criteria for scoring an assessment item at each score category (also referred to as a scoring rubric).

Second-scoring. During the scoring of NAEP constructed-response questions, a subset of student responses is scored by a second reader in order to obtain within-year reliability data.

Standard deviation. A measure of the dispersion of a set of scores. Specifically, it is the square root of the average squared deviation of scores about their arithmetic mean.

Standard error. A measure of sampling variability and measurement error for a statistic. Because of NAEP's complex sample design, sampling standard errors are estimated by jackknifing the samples from first-stage sample estimates. Standard errors may also include a component due to the error of measurement of individual scores estimated using plausible values.

Statistically significant. Statistical tests are conducted to determine whether the changes or differences between two resulting numbers are statistically significant. The term "significant" does not imply a judgment about the absolute magnitude or educational relevance of changes in student performance. Rather, it is used to indicate that the observed changes are not likely to be associated with sampling and measurement error, but are statistically dependable population differences.

All differences reported are significant at the .05 level with appropriate adjustments for multiple comparisons.

Stratification. The division of a population into parts, or strata.

Stratified sample. A sample selected from a population that has been stratified, with a sample selected independently in each stratum. The strata are defined for the purpose of reducing sampling error.

Stratum. A collection of sampled units defined by a characteristic. All sampling units belong to a stratum, and the strata are mutually exclusive.

Students with disabilities (SD). A student with a disability, who may need specially designed instruction to meet his or her learning goals. A student with a disability will usually have an Individualized Education Program (IEP), which guides his or her special education instruction. Students with disabilities are often referred to as special education students and may be classified by their school as learning disabled (LD), physically disabled (PD), or emotionally disturbed (ED).

Subject area. One of the areas assessed by NAEP, for example, reading or mathematics.

Student groups. Groups of the student population identified in terms of certain demographic or background characteristics. Some of the major reporting groups used for reporting NAEP results are based on students' gender, their race or ethnicity, the highest level of education they report for either parent, whether they are eligible for free or reduced-price school lunch, and the type of school (public or nonpublic) they attend. Information gathered from NAEP background questionnaires also makes it possible to report results based on variables such as course-taking, home discussions of schoolwork, and television-viewing habits.

Transformation. An equation used to convert values on one score scale to values on another score scale.

Weighted percentage. A percentage that has been calculated by differentially weighting observations to account for complex sampling procedures. It differs from a simple percentage, in which all cases are equally weighted. In NAEP, each sampled student is assigned a weight that makes proper allowances for NAEP's sampling design and reflects adjustments for school and student nonparticipation. Weighted percentages are estimates of the percentages of the total population, or population subgroup, that have a specified characteristic. For example, the weighted percentage of 9-year-old students in the NAEP sample who correctly answered a particular NAEP test item is an estimate of the percentage of age 9 students in the nation who can correctly answer that question.

THIS PAGE INTENTIONALLY LEFT BLANK.

U.S. Department of Education
ED Pubs
8242-B Sandy Court
Jessup, MD 20794-1398