Chapter 5
Comparison of Bridge and Modified Assessments

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.

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.

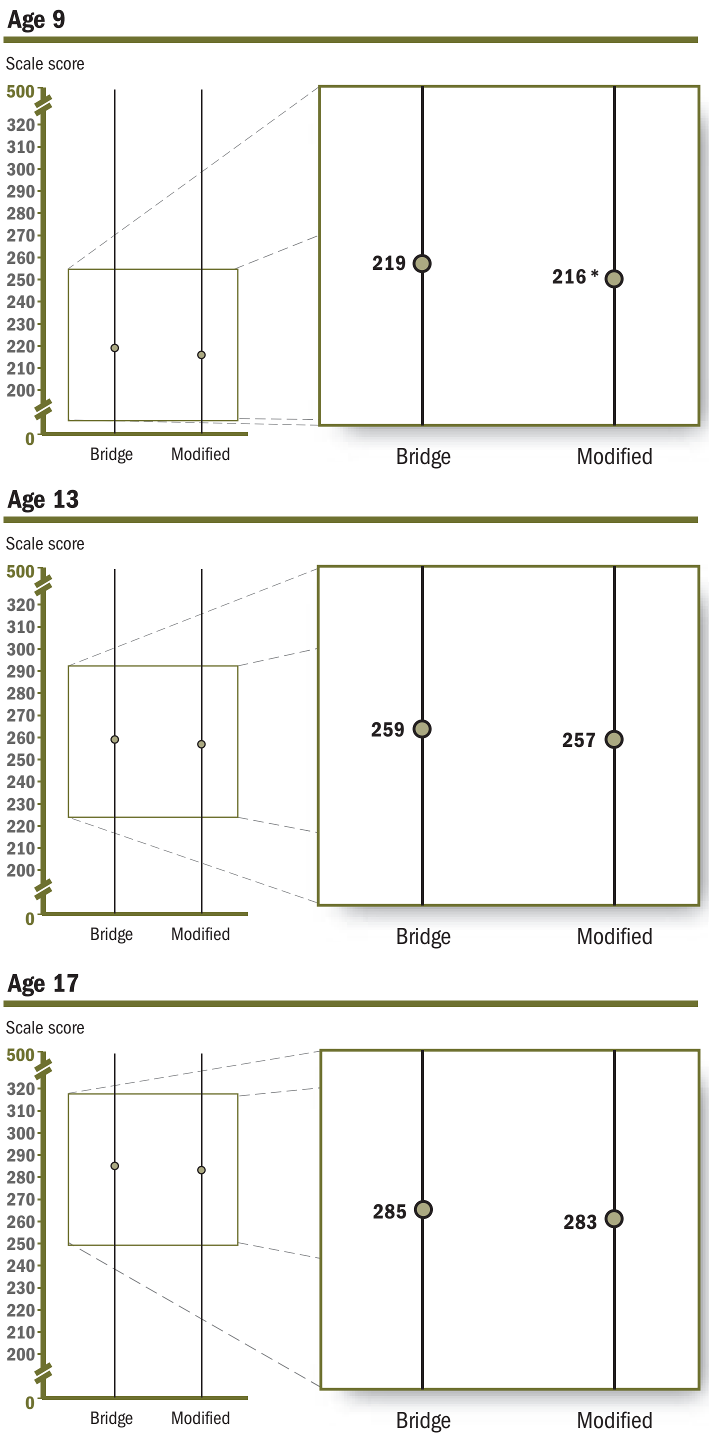
Comparison of Bridge and Modified Results for Reading

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

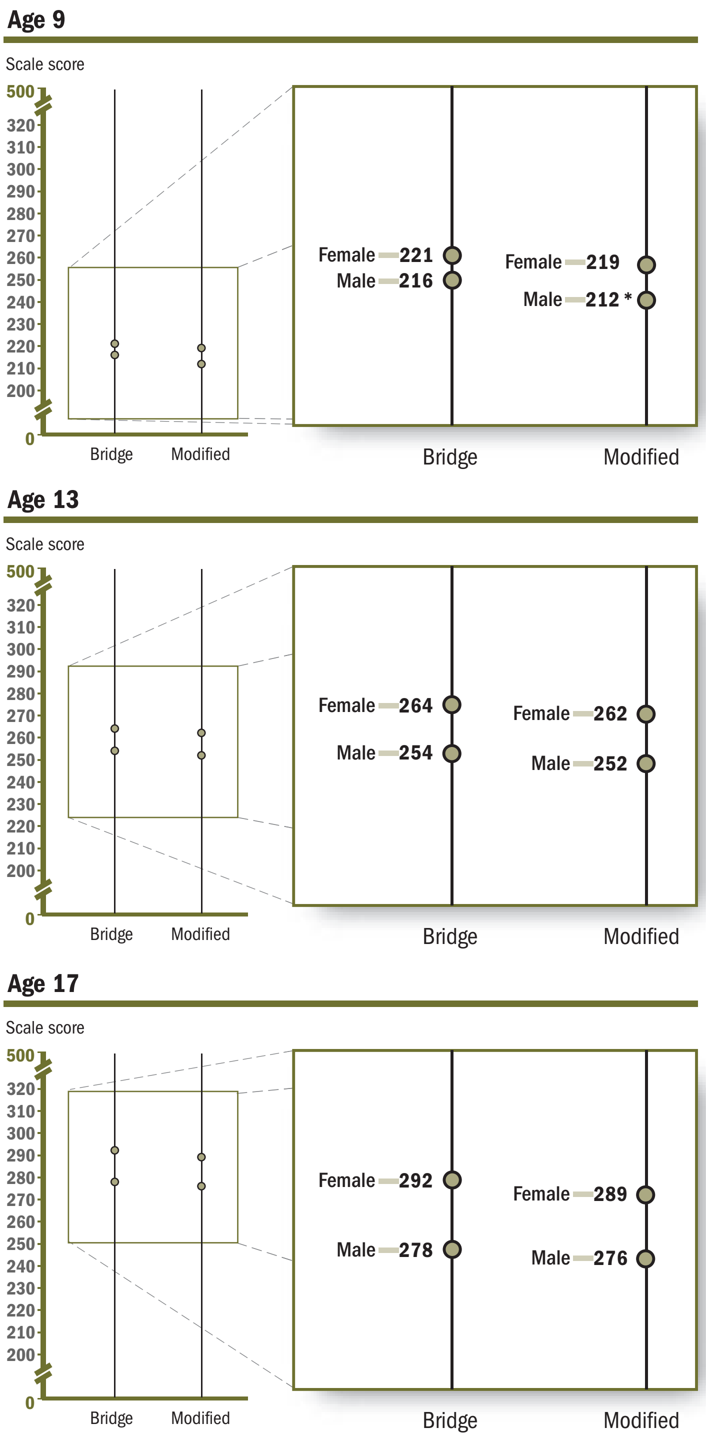


*Significantly different from 2004 bridge assessment.

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



Age 9

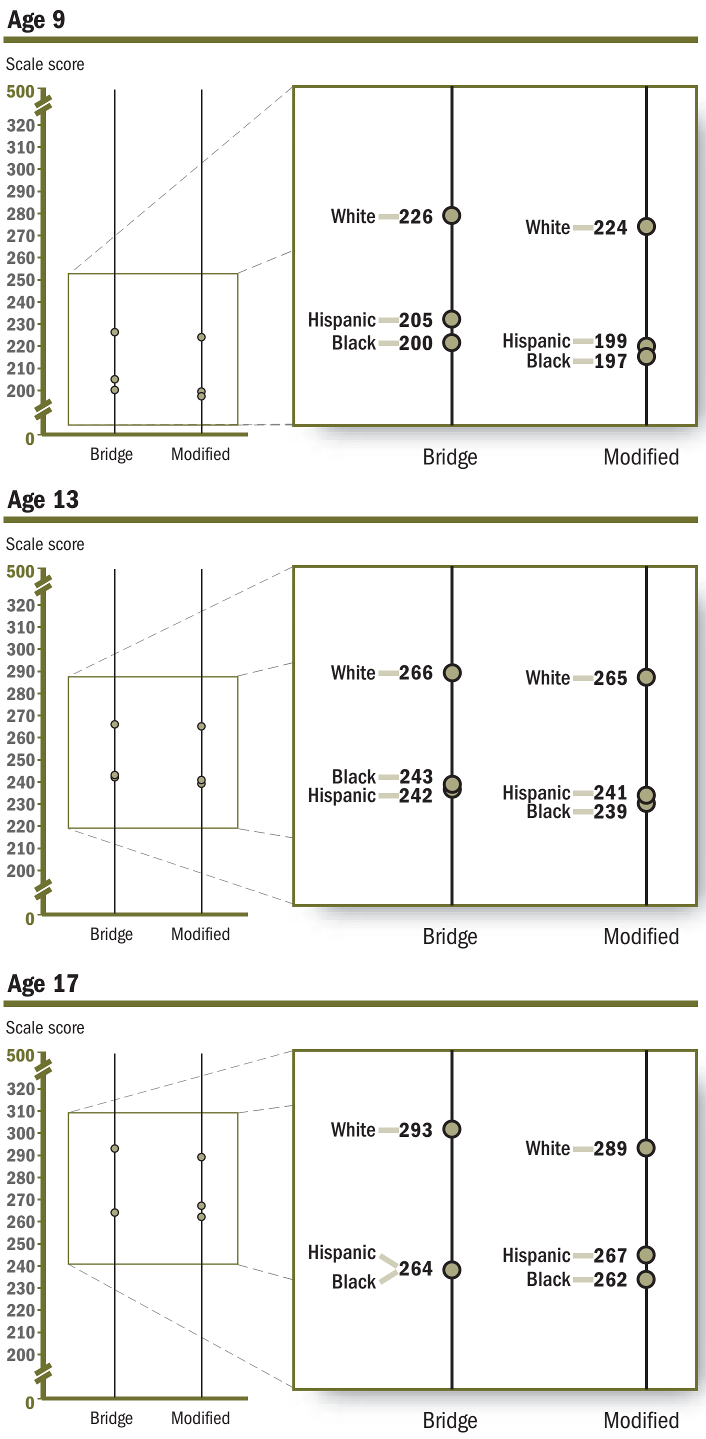
Age 13

Age 17

*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


Age 9

Age 13

Age 17

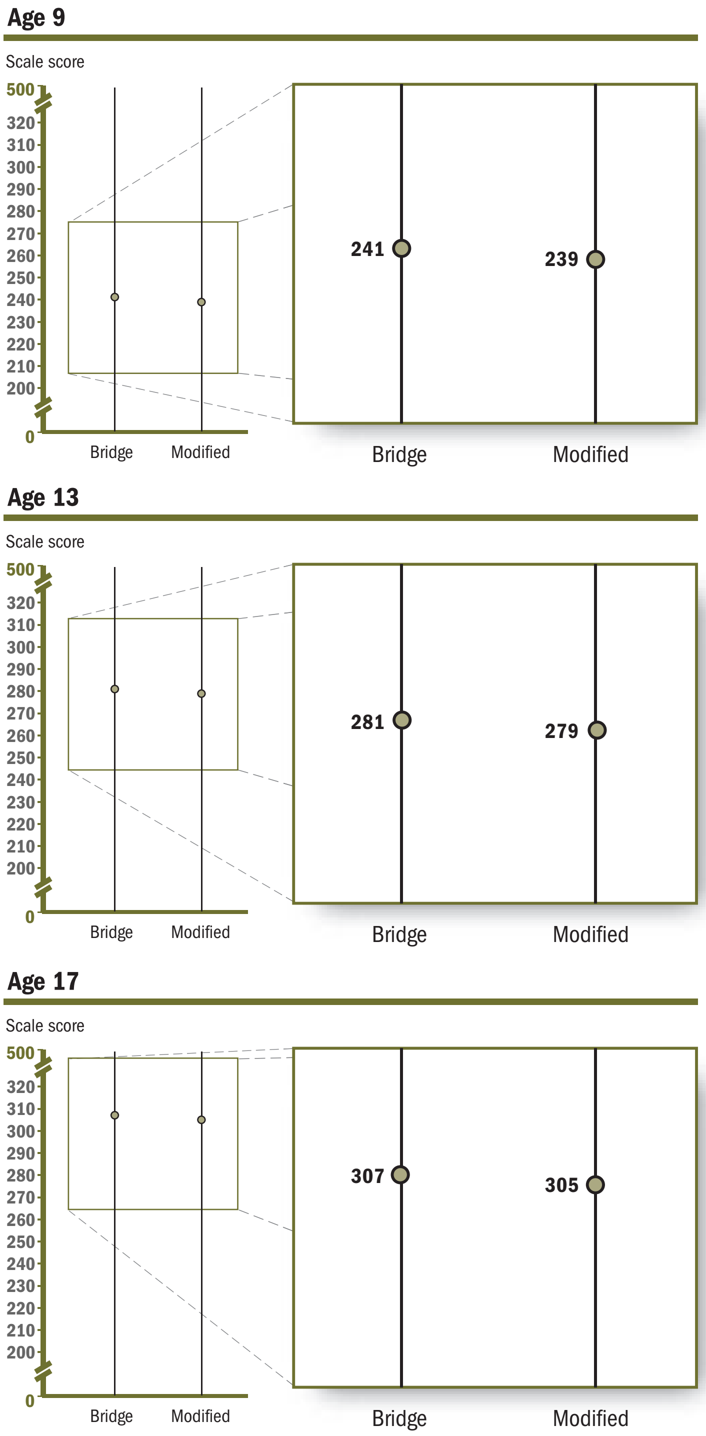
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Reading Assessment.

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


Age 9

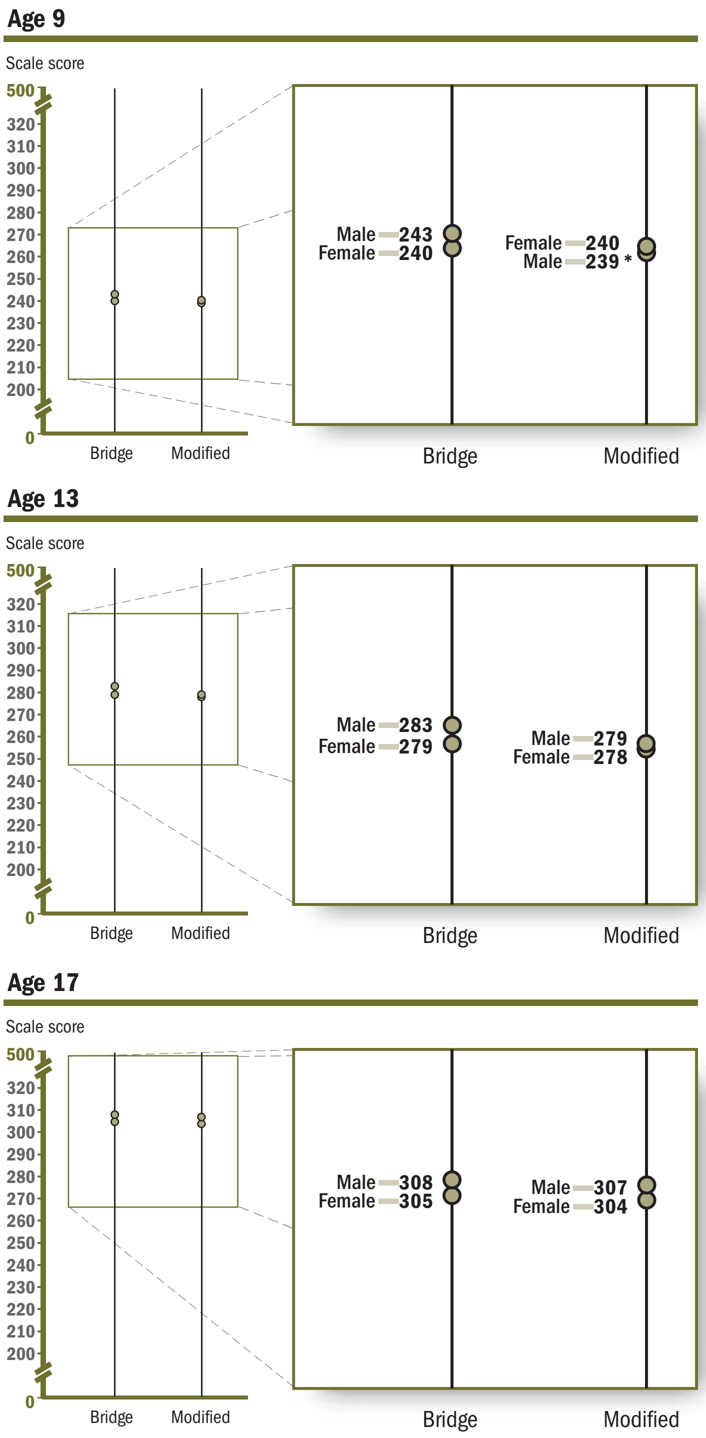
Age 13

Age 17

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



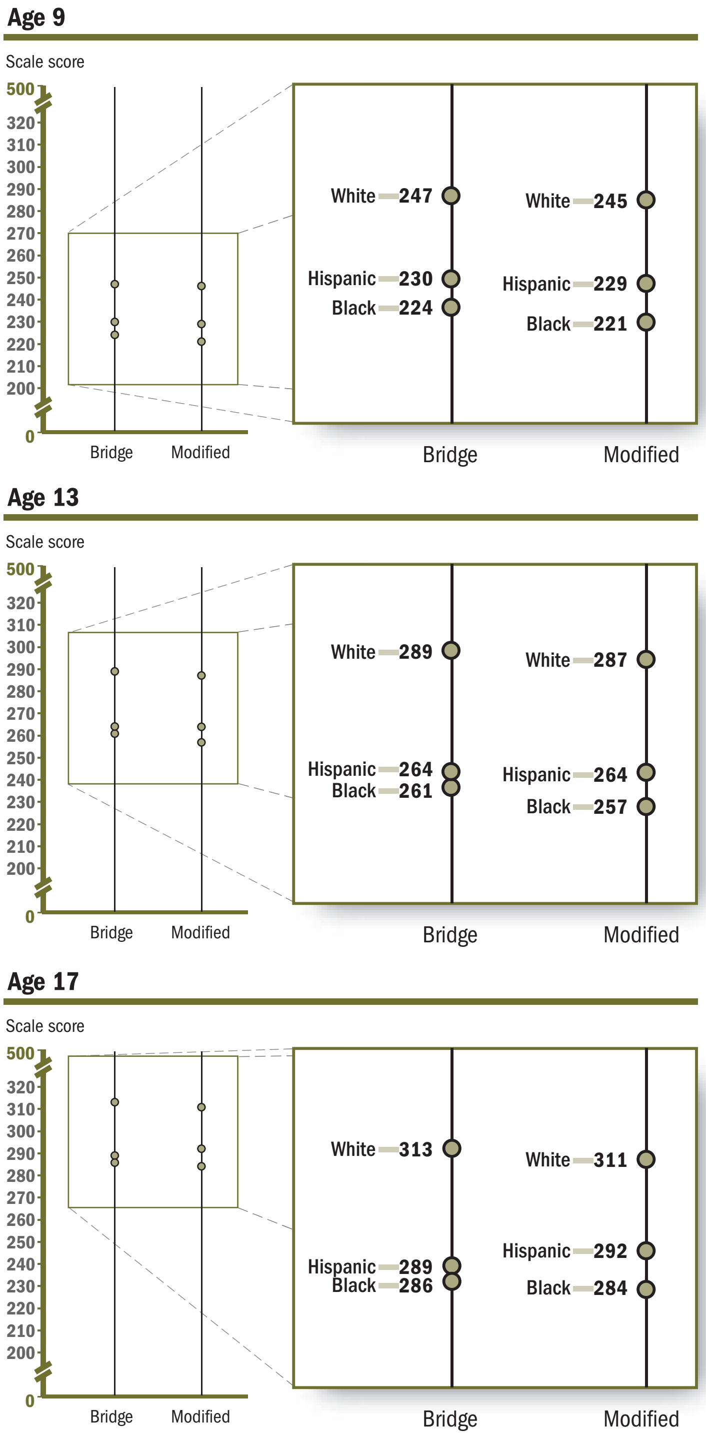
*Significantly different from 2004 bridge assessment.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2004 Long-Term Trend Mathematics Assessment.

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

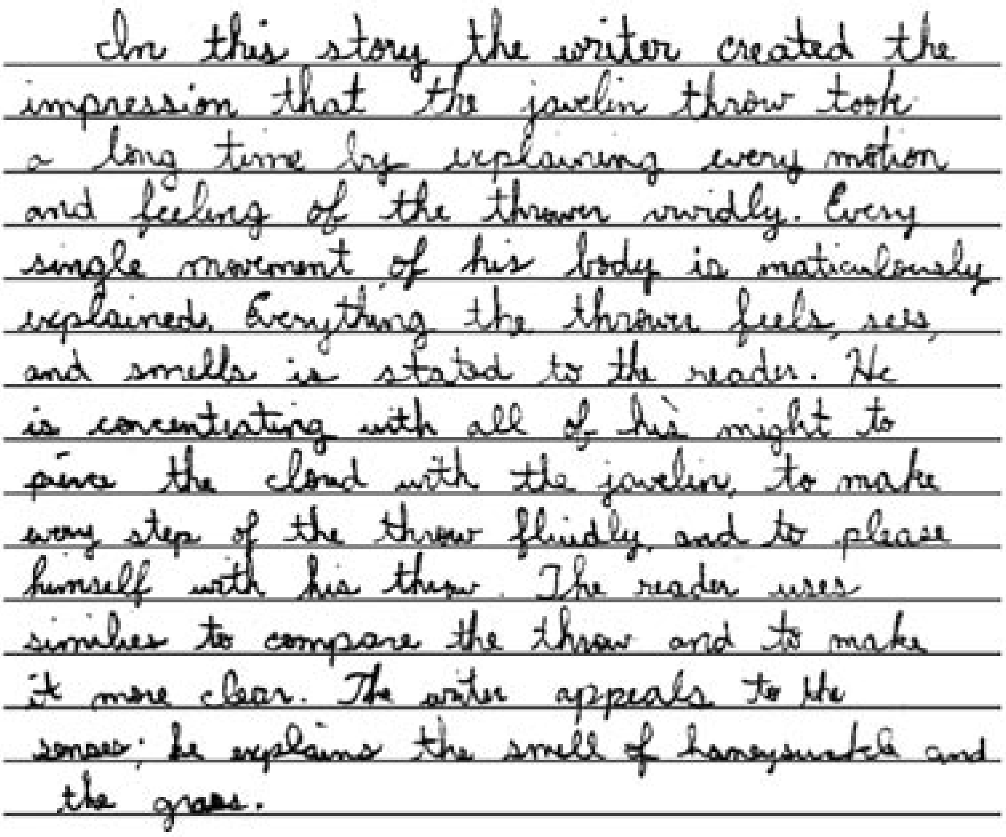
Sample question 8 is a constructed-response question, which asked students to explain how the author of the passage created an effect. Responses to this task were rated according to a four-level scoring guide in one of the following score categories: “Elaborated Interpretation,” “Satisfactory Interpretation,”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

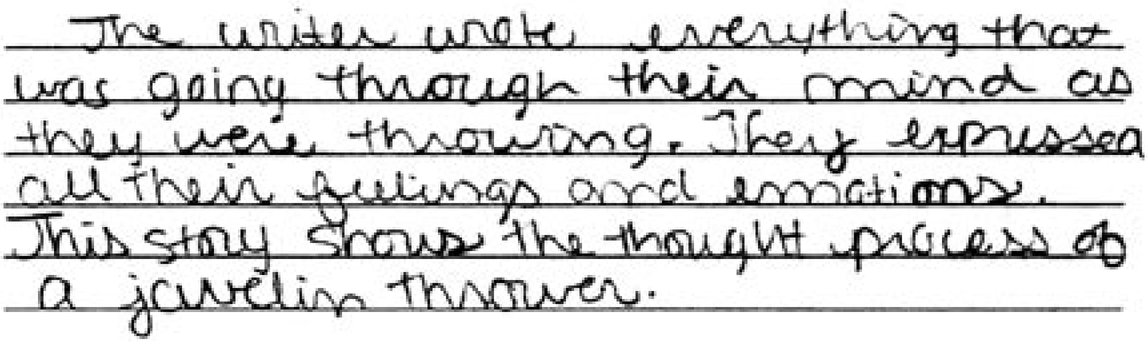


In this story the writer created the impression that the javelin throw took a long time by explaining every motion and feeling of the thrower vividly. Every single movement of his body is maticulously explained. Everything the thrower feels, sees, and smells is stated to the reader. He is concentrating 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

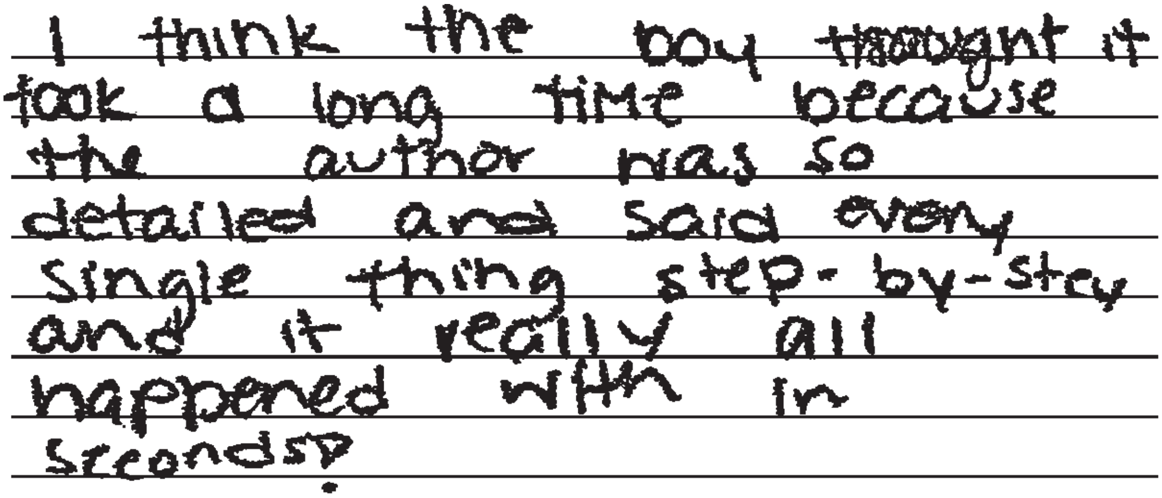
Sample “Satisfactory Interpretation” response. The following sample response is rated “Satisfactory Interpretation” because it explains how the javelin throw seems to take a long time with a specific example from the passage.



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

	CITY	
YEAR	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

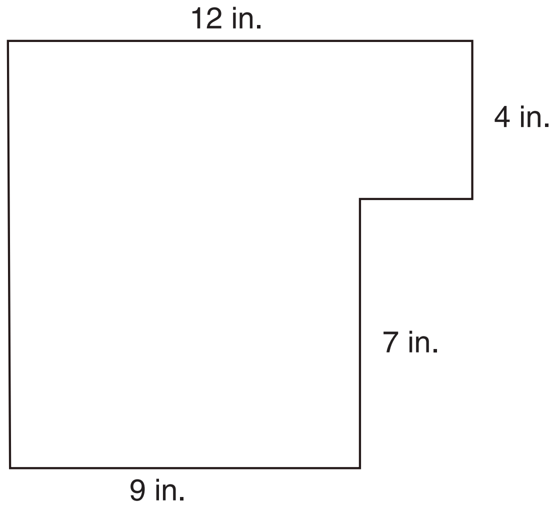
● 75

Ⓓ 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 × 11) + (3 × 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 × 11) – (3 × 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				2004 mathematics trend bridge assessment			
		Position				Position		
		1	2	3		1	2	3
Ages 9/13	Book 1	Reading	Mathematics block 1	Science	Book 1	Reading	Mathematics block 1	New mathematics block 1
	Book 2	Mathematics block 2	Science	Reading	Book 2	Mathematics block 2	New mathematics block 1	New mathematics block 2
	Book 3	Science	Reading	Mathematics (Calculator)				
Age 17	Book 1	Mathematics block 1	Mathematics block 2	Science	Book 1	Mathematics block 1	Mathematics block 2	New mathematics block 1
	Book 2	Science	Science	Mathematics (calculator)				

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

										2004	
Age	1971	1975	1980	1984	1988	1990	1992	1994	1996	1999	Bridge	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

								2004	
Age	1978	1982	1986	1990	1992	1994	1996	1999	Bridge	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

	Bridge assessment		Modified assessment	
Item response types	Multiple-choice items	Constructed-response items	Multiple-choice items	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

	Bridge assessment		Modified assessment	
Item response types	Multiple-choice items	Constructed-response items	Multiple-choice items	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.

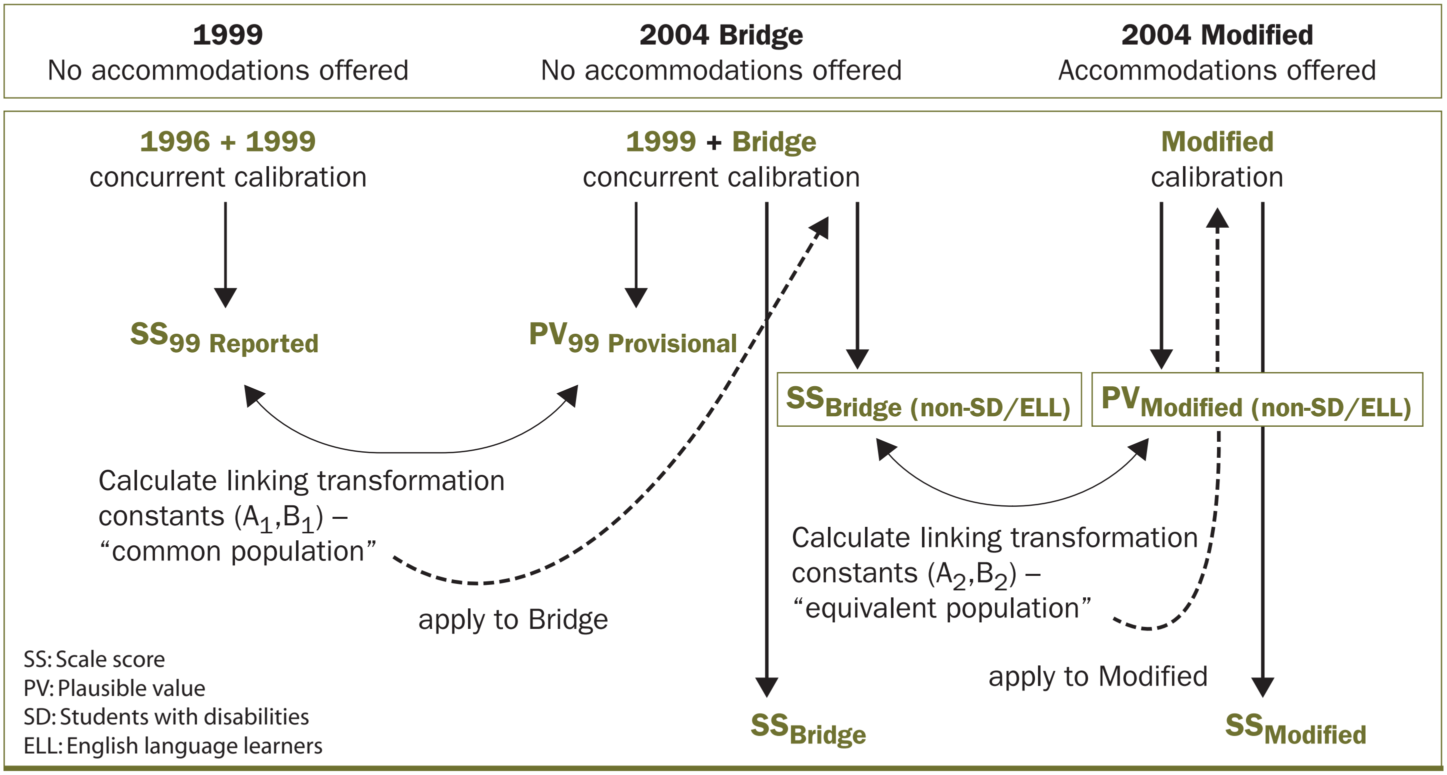
The reading and mathematics trend assessments are composed of three types of questions: multiple-choice, short constructed-response (scored either dichotomously or allowing for partial credit), and extended constructed-response (scored according to a partial-credit model). Prior to 2004, all constructed-response 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 $PV_{99\ \text{Provisional}}$) 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:

Average ± 2 standard errors =
256 ± 2 (1.2) = 256 ± 2.4 =
256 – 2.4 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:

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

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 $100 - P < P_{\text{lim}}$. 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 × (5 – 1)/5 = 0.04 = 4 percent, 20 percent for the group 1 comparison would be compared to 0.05 × (5 – 2)/5 = 0.03 = 3 percent, 4 percent for the group 3 comparison would be compared to 0.05 × (5 – 3)/5 = 0.02 = 2 percent, and 1 percent for the group 2 comparison (actually slightly smaller than 1 prior to rounding) would be compared to 0.05 × (5 – 4)/5 = 0.01 = 1 percent. The 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 counter-intuitive 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

	Age 9		Age 13		Age 17	
Characteristic	Bridge	Modified	Bridge	Modified	Bridge	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

	Age 9		Age 13		Age 17	
Characteristic	Bridge	Modified	Bridge	Modified	Bridge	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