National Assessment of Educational Progress (NAEP)

Website: http://nces.ed.gov/nationsreportcard/
Updated: May 2017

1. OVERVIEW

SURVEY OF A SAMPLE OF ELEMENTARY/ SECONDARY STUDENTS

Two assessments:

- Main NAEP
- Long-Term Trend NAEP

Three foci:

- Main National NAEP
- Main State NAEP
- Trial Urban District NAEP

The National Assessment of Educational Progress (NAEP) is mandated by Congress to assess the educational achievement of U.S. students and monitor changes in those achievements. As the oldest nationally representative and continuing assessment of what America's students know and can do in subject areas, NAEP serves as the "Nation's Report Card." *Main* NAEP includes state and national assessments of students in grades 4, 8, and 12 in various subject areas and with varying periodicity. The *Trial Urban District Assessment*, or TUDA, assesses students in large urban districts at grades 4 and 8 in reading, mathematics, science, and writing. The *Long-Term Trend* Assessment is administered to students at ages 9, 13, and 17 in reading and mathematics and only at the national level. Results go back as far as 1971.

In 1988, Congress established the National Assessment Governing Board (referred to as the Governing Board or NAGB) to provide policy guidance for the execution of NAEP. The 26-member Governing Board 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. Its responsibilities include: select subject areas to be assessed; set appropriate student achievement levels; develop assessment objectives and test specifications; design the assessment methodology; and produce standards and procedures for interstate, regional, and national comparisons. NAEP is administered by the National Center for Education Statistics (NCES).

Purpose

To (1) monitor continuously the knowledge, skills, and performance of the nation's children; and (2) provide objective data about student performance at the national and regional levels, the state level (since 1990), and for large urban school districts (since 2002).

Components

NAEP comprises two separate assessments: *main* and *long-term trend.* Results for three geographic levels are reported in the *main* assessment: *national, state* and *urban school district*. The long-term trend assessment is conducted at the national level only. Two additional surveys, the High School Transcript Study (HSTS) and the National Indian Education Study (NIES), are conducted in conjunction with NAEP.

Since 1996, the main national and state assessments have provided accommodations for students with special needs, while the TUDA assessments have offered them since their inception. Long-term trend NAEP began offering accommodations in 2004.

National-level assessment. The *main national* NAEP and the *long-term trend* NAEP are both designed to report information for the nation and specific geographic regions of the country (Northeast, Southeast, Central, and West). However, these two assessments use separate samples of students from public and nonpublic schools: grade samples for the main national NAEP (grades 4, 8, and 12), and age samples for the long-term trend NAEP (ages 9, 13, and 17).

The test instruments for the main and long-term trend assessments are based on different frameworks; the student and teacher background questionnaires vary; and the results for the two assessments are reported separately. (See "Elementary and Secondary School Students Survey" below for the subject areas assessed.)

The assessments in the main national NAEP follow the frameworks developed by NAGB and use the latest advances in assessment methodology. The test instruments are flexible so they can be adapted to changes in curricular and educational approaches. Assessment instruments for the main NAEP reading and mathematics assessments have produced valid trend results from 1990 through 2015, except for changes introduced in 2009 in the mathematics assessment for grade 12 and the science assessment across grades. In 2005 and 2009 the Governing Board introduced changes in the NAEP mathematics framework for grade 12 in both the assessment content and administration procedures resulting in a break in trend. In 2009, a new framework was developed for science, also resulting in a break in trend. Since 2009, the mathematics, reading, science, and writing assessments were developed using the same framework, allowing the results to be compared from 2009 forward.

To reliably measure change over longer periods of time, the long-term trend NAEP must be used. Long-term trend NAEP allows measurement of trends since 1971 in reading and 1973 in mathematics.

State-level assessments. The *main state* NAEP was implemented in 1990 on a trial basis. Participation of the states was voluntary until 2003. The reauthorization of the Elementary and Secondary Education Act requires states that receive Title I funding to participate in state NAEP assessments in reading and mathematics at grades 4 and 8 every two years. State participation in other state NAEP subjects (i.e., science and writing) remains voluntary. Participation is voluntary for students for all NAEP assessments. Separate representative samples of students are selected for each jurisdiction to provide that jurisdiction with reliable state-level data concerning the achievement of its students. The state assessment included nonpublic schools in 1994, 1996, and 1998. This practice ended because of low participation rates.

The Trial Urban District Assessment. The *Trial Urban District Assessment (TUDA)* began assessing performance in selected large urban districts in 2002 in reading and writing; it continued in 2003 with reading and mathematics; in 2005 with reading, mathematics, and science; in 2007 with reading, mathematics and writing; in 2009 with reading, mathematics and science; in 2011 and 2013 with reading and mathematics; and in 2015 with reading, mathematics, and science. The program retains its trial status. The first TUDA occurred in 2002 for five urban districts. Nine urban districts participated in 2003, 10 districts participated in 2005 and 2007, 18 participated in 2009 and 21 participated in 2011, 2013, and 2015. The results for these districts are for public school students only. Beginning in 2009, the TUDA results include only those charter schools that the district is accountable for. Results for these districts are also compared with results for public school students in large central cities and the nation.

Results for District of Columbia public school students, normally included with NAEP's state assessment results, are also included in TUDA reports on mathematics for 2003, 2005, 2007, 2009, 2011, 2013, and 2015, as well as for reading for those same years plus 2002. Due to an insufficient sample size, the District of Columbia did not participate in the science assessments in 2005 and 2009, nor did it participate in the writing assessment in 2007 for the same reason.

Subjects assessed in NAEP. The primary data collected by NAEP relate to student performance and educational experience. Major assessment areas for main NAEP include reading, writing, mathematics, and science, which are assessed at the national, state, and district levels. At grades 4 and 8, mathematics and reading are assessed every two years; at grade 12, they are assessed every four years. Science and writing are assessed every four years in grades 4 and 8. Other subject areas (e.g., civics, U.S. history, geography, economics, technology and engineering literacy, and the arts) are assessed only at the national level, occur less frequently and, usually do not include all three grades.

The subjects assessed in long-term trend NAEP are mathematics and reading. Since 2004, the long-term trend assessments have been scheduled to be administered in mathematics and reading every 4 years. The most recent long-term trend assessment was conducted during the 2011-12 school year (fall for age 13; winter for age 9; spring for age 17).

Student survey questions. NAEP also asks questions about students' characteristics, as well as questions related to the subject area and students' motivation in completing the assessment. Information is gathered about race/ethnicity, school attendance, academic

expectations, and factors related to academic performance.

School Characteristics Survey. This survey collects supplemental data about school characteristics and policies that can be used analytically to provide context for student performance issues and are completed by the principal or assistant principal. The student remains the unit of analysis. Data are collected on such topics as demographic characteristics, classroom experiences, educational support, instructional practices, school policies, students with disabilities, and English-language learners.

Teacher Questionnaire. This survey collects supplemental data from teachers whose students are respondents to the assessment. The first part of the teacher questionnaire tends to cover background and general training, and includes items concerning years of teaching experience, certifications, degrees, major and minor fields of study, course work in education, course work in specific subject areas, the amount of in-service training, the extent of control over instructional issues, and the availability of resources for the classroom. Subsequent parts of the teacher questionnaire tend to cover training in the subject area, classroom instructional information, and teacher exposure to issues related to the subject and the teaching of the subject. Teacher questionnaires are completed by teacher at grade 4 and 8. NAEP typically does not collect teacher information for grade 12.

SD/ELL Survey. This survey is completed in the *main* NAEP assessments (and the *long-term trend* NAEP since 2004) by teachers of students who are selected to participate in NAEP and who are classified as either having disabilities (SD) or as English language learners (ELL). Information is collected on the characteristics of each SD/ELL student and the reason for the SD/ELL classification, as well as whether these students receive accommodations in district or statewide tests. For SD students, questions are asked about the student's functional grade level and special education programs. For ELL students, questions are asked about the student's native language, time spent in special language programs, and level of English language proficiency. This survey is used to determine whether the student should take the NAEP assessment. If any doubt exists about a student's ability to participate in the assessment, the student is included. Beginning with the 1996 assessments (2004 for *long-term* trend), NAEP has allowed accommodations for both SD and ELL students.

High School Transcript Study. Transcript studies have been conducted in 1987, 1990, 1994, 1998, 2000, 2005, and 2009. The studies collect information on current course offerings and course-taking patterns in the nation's schools for high-school graduates. Transcript data can be used to show course-taking patterns across years that may be associated with proficiency in subjects assessed by NAEP. Transcripts are collected for grade 12 students in schools in the NAEP sample who graduate from the school. (For more information, see the chapter on the High School Transcript Studies.)

National Indian Education Study. The National Indian Education Study (NIES) describes the condition of education for American Indian and Alaska Native (AI/AN) students in the United States. The study was conducted in 2005, 2007, 2009, 2011, and 2015 by NCES on behalf of the U.S. Department of Education, Office of Indian Education. After the 2011 administration of the study, NIES shifted its cycle from once every two years to once every four years. NIES is currently authorized under Executive Order 13592, "Improving American Indian and Alaska Native Educational Opportunities and Strengthening Tribal Colleges and Universities", which was signed in 2011 to improve education efforts for AI/AN students nationwide. Prior to 2011, NIES was authorized under Executive Order 13336.

NIES has two components: an assessment component and a survey questionnaire component. The assessment component of NIES is conducted through NAEP and provides in-depth information on the academic performance of 4th– and 8th–grade AI/AN students in reading and mathematics. The survey questionnaire component of NIES allows for the description of the educational experiences of the 4th– and 8th–grade AI/AN students who participate in the NAEP assessments. The survey focuses on the integration of native language and culture into school and classroom activities. NIES collects information through questionnaires for students, teachers, and schools.

Other NAEP Special Studies

In addition to the assessments, NAEP coordinates a number of related special studies. Such studies often involve special data collection procedures in the field, secondary analyses of NAEP results, and evaluations of various technical procedures. A full list and detailed description of each study can be found at http://nces.ed.gov/nationsreportcard/studies/**.** Examples of such studies are described below.

- Two special studies currently underway aim to understand what is feasible by way of

technology-based improvements for NAEP. The 2011 **Mathematics Computer Based Study** (MCBS) at grade 8 was conducted in part in anticipation of the Governing Board's 2017 date for online assessment, to inform the relative merits of static versus adaptive models for online assessment. An adaptive model has the potential to reduce measurement error, especially for higher- and lower-performing students, and to improve student engagement by administering items better tailored to students' individual ability levels. Another, the **Knowledge and Skills Appropriate** (KaSA) mathematics special study, considers blocks of new items specifically developed at each grade to better measure the knowledge and skills of lower ability students while still conforming to the content distribution specifications of the NAEP mathematics framework.

- ***Oral Reading Study.*** The NAEP 2002 Oral Reading Study looked at how well the nation's 4th- graders can read aloud a grade-appropriate story. The assessment provided information about a student's fluency in reading aloud and examined the relationship between oral reading accuracy, rate (or speed), fluency, and reading comprehension.

- ***Technology-Based Assessment (TBA) Project.*** TBA was a NAEP project in 2000 to 2003 and designed to explore the use of the computer as a tool to enhance the quality and efficiency of educational assessments. TBA was designed with five components—three empirical studies (Mathematics Online, Writing Online, and Problem Solving in Technology-Rich Environment); a conceptual paper (Computerized Adaptive Testing); and an online school and teacher questionnaire segment. The key questions being considered in the three studies were: How does the method of test delivery affect the inferences that can be drawn about students' mathematics or writing skills and do students perform differently between computer versus paper? How do different populations perform and does it vary because of the method of delivery? How are students with different levels of computer experience affected by a computer based assessment? Is the technology-based assessment more cost-effective and how might technological advances affect cost and timeliness? What are the logistical challenges of administering a NAEP assessment on computer?

- ***Charter School Pilot Study***. NAEP conducted a pilot study of America's charter schools and their students as part of the 2003 NAEP assessments in reading and mathematics at the 4th-grade level. Charter schools are public schools of choice. They serve as alternatives to the regular public schools to which students are assigned. While there are many similarities between charter schools and other public schools, they do differ in some important ways, including the makeup of the student population and their location.

- ***School Composition and the Black-White Achievement Gap Study***. The School Composition and the Black-White Achievement Gap study was undertaken by the National Center for Education Statistics to present both descriptive and associative information on the relationships among the percentage of students in a school who were Black, the Black-White student academic performance gap, and student achievement. Data was primarily used from the NAEP 2011 mathematics grade 8 assessment.

Periodicity

Main NAEP assesses students at grades 4, 8, and 12 in various subject areas. Students are assessed at grades 4 and 8 in reading and mathematics every two years at both the national and state level. Similar assessments are given in science and writing approximately every four years. Grade 12 assessments are administered less frequently. Subjects other than these four are administered less frequently and at the national level only. The *Trial Urban District Assessment*, or TUDA, assesses students in large urban districts at grades 4 and 8 in connection with the assessments in reading, mathematics, science, and writing. The *NAEP Long-Term Trend* Assessment assesses students at ages 9, 13, and 17 in reading and mathematics only and only at the national level. Results go back as far as 1971.

2. USES OF DATA

NAEP is the largest ongoing, comparable, and representative assessment of what American students know and can do in various subject areas. Policymakers are keenly interested in NAEP results because they address national outcomes of education, specifically, the level of educational achievement. In addition, state-level and urban district-level data, available for many states since 1990 and for selected

large urban districts since 2002, allow both state-to-state and district-to-district comparisons, and comparisons of individual states with the nation as a whole (as well as comparisons of urban districts with large central cities and the nation).

During NAEP's history, a number of reports across various subject areas have provided a wealth of information on students' academic performance, learning strategies, and classroom experiences. Together with the performance results, the basic descriptive information collected about students, teachers, administrators, and communities can be used to address the following educational policy issues:

- *Instructional practices*. What instructional methods are being used?
- *Students-at-risk*. How many students appear to be at-risk in terms of achievement, and what are their characteristics? What gaps exist between at-risk categories of students and others?
- *Teacher workforce*. What are the characteristics of teachers of various subjects?

The NAEP Data Explorer, available at https://nces.ed.gov/nationsreportcard/tdw/database/data_tool.asp, allows extensive analysis of NAEP data. However, users should be cautious in their interpretation of NAEP results. While NAEP scales make it possible to examine relationships between students' performance and various background factors, the relationship that exists between achievement and another variable does not reveal its underlying cause, which may be influenced by a number of other variables. NAEP results are most useful when they are considered in combination with other knowledge about the student population and the education system, such as trends in instruction, changes in the school-age population, and societal demands and expectations.

NAEP materials such as frameworks and released questions also have many uses in the educational community. Frameworks present and explain what experts in a particular subject area consider important. Several states have used NAEP frameworks to revise their curricula. After most assessments, NCES publicly releases nearly one-third of the questions. Released constructed-response questions and their corresponding scoring guides have served as models of innovative assessment practices in the classroom.

3. KEY CONCEPTS

The achievement levels for NAEP assessments are defined below. For subject-specific definitions of achievement levels and additional terms, refer to NAEP technical reports, "report card" reports, and other publications.

Achievement levels. Starting with the 1990 NAEP, the National Assessment Governing Board (also referred to as 'Governing Board' or NAGB) developed achievement levels for each subject at each grade level to measure how well students' actual achievement matches the achievement desired of them. The 2001 reauthorization law requires that the achievement levels be used on a trial basis until the Commissioner of Education Statistics determines the levels are "reasonable, valid, and informative to the public" and so should be interpreted and used with caution. The Governing Board defines the three levels as follows:

- *Basic*. Partial mastery of the prerequisite knowledge and skills that are fundamental for proficient work at each grade.
- *Proficient*. Solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
- *Advanced*. This level signifies superior performance at each grade assessed.

4. SURVEY DESIGN

Target Population

Students enrolled in public and nonpublic schools in the 50 states and the District of Columbia who are deemed assessable by their school and classified in defined grade/age groups—grades 4, 8, and 12 for the *main national* assessments and ages 9, 13, and 17 for the *long-term trend* assessments in mathematics and reading. Grades 4 and/or 8 are usually assessed in the *state* assessments and *TUDA*; the number of grades assessed has varied in the past, depending on the availability of funding (although testing for 4th- and 8th-graders in reading and mathematics every 2 years is now required for states that receive Title I funds). Only public schools were included in the *state* NAEP and in *TUDA*.

Sample Design

For the national assessments, probability samples of schools and students are selected to represent the diverse student population in the United States. The numbers of schools and students vary from cycle to cycle, depending on the number of subjects and items to be assessed. A national sample will have sufficient schools and students to yield data for public schools in each of the four Census regions of the country, as well as results by sex, race/ethnicity, degree of urbanization of school location, parent education (for grades 8 and 12), and participation in the National School Lunch Program. A national sample of private schools is also selected for grades 4, 8, and 12. This sample is designed to produce national and regional estimates of student performance.

In the state assessment, a sample of schools and students is selected to represent a participating state. In a state, on average 2,500 students in approximately 100 public schools are selected per grade, per subject assessed. The selection of schools is random within classes of schools with similar characteristics; however, some schools or groups of schools (districts) can be selected for each assessment cycle if they are unique in the state. For instance, a particular district may be selected more often if it is located in the state's only major metropolitan area or has the majority of the state's Black, Hispanic, or other race/ethnicity population. Additionally, even if a state decides not to participate at the state level, schools in that state identified for the national sample will be asked to participate.

Typically, within each school, approximately 30 students per subject are selected randomly. Some of the students who are randomly selected are classified as students with disabilities (SD) or English-language learners (ELL). NAEP's goal is to assess all students in the sample, and this is done if at all possible.

NAEP's multistage sampling process involves the following steps:

- selection of schools (public and nonpublic) within strata;
- selection of students within the selected schools; and
- allocation of selected students to assessment subjects.

Selection of schools. In this stage of sampling, public schools in each state (including Bureau of Indian Education [BIE] schools serving grade 4 or 8 students and Department of Defense Education Activity [DoDEA] schools) and private schools in each state (including Catholic schools) are listed according to the grades associated with the three age classes: age class 9 refers to age 9 or grade 4 in the long-term trend NAEP (or grade 4 in the main NAEP); age class 13 refers to age 13 or grade 8 in the long-term trend NAEP (or grade 8 in the main NAEP); age class 17 refers to age 17 or grade 11 in the long-term trend NAEP (or grade 12 in the main NAEP).

The school lists are obtained from two sources. Regular public, BIE, and DoDEA schools are obtained from the school list maintained by NCES' Common Core of Data. Catholic and other nonpublic schools are obtained from the NCES Private School Universe Survey (PSS). To ensure that the state samples provide an accurate representation, public schools are stratified by urbanization, enrollment of Black, Hispanic, or other race/ethnicity students, state-based achievement scores, and median household income. Private schools are stratified by type (e.g., parochial, nonreligious), urban status, and enrollment per grade. Once the stratification is completed, the schools are assigned a probability of selection that is proportional to the number of students per grade in each school.

Prior to 2005, DoDEA overseas and domestic schools were reported separately. Starting with the 2005 assessments, all DoDEA schools, both domestic and overseas, were combined into one jurisdiction. In addition, the definition of the national sample changed in 2005; it now includes all of the overseas DoDEA schools.

The manner of sampling schools for the long-term trend assessments is very similar to that used for the main assessments. The primary difference is that in long-term trend nonpublic schools and schools with high enrollment of Black, Hispanic, or other race/ethnicity students are not oversampled. Schools are not selected for both main and long-term trend assessments at the same age/grade. The long-term trend assessments use a nationally representative sample and do not report results by state.

Selection of students. This stage of sampling involves random selection of national samples representing the entire population of U.S. students in grades 4, 8, and 12 for the main assessment and the entire population of students at ages 9, 13, and 17 for the long-term trend assessment. Some of the students who are randomly selected are classified as SD or ELL. A small number of students selected for participation are excluded because of limited English proficiency or severe disability.

To facilitate the sampling of students, a consolidated list is prepared for each school of all age-eligible students (long-term trend assessments) or all grade-eligible students (main assessments) for the age class for which the school is selected. A systematic selection of eligible students is made from this list—unless all students are to be assessed—to provide the target sample size.

For each age class (separately for long-term trend and main samples), measures of size are established as to the number of students who are to be selected for a given school. In those schools that, according to information in the sampling frame, have fewer eligible students than the final measures of size, each eligible student enrolled at the school is selected in the sample. In other schools, a sample of students is drawn. The measures of size are established in terms of the number of grade-eligible students for the main samples, and in terms of the number of students in each age class for the trend samples.

Excluded students. Some students are excluded from the student sample because they are deemed unable to participate meaningfully by school authorities. The exclusion criteria for the main samples differ somewhat from those used for the long-term trend samples. In order to identify students who should be excluded from the main assessments, school staff members are asked to identify those SD or ELL students who do not meet the NAEP inclusion criteria. School personnel are asked to complete an SD/ELL questionnaire for all SD and ELL students selected into the NAEP sample, whether they participate in the assessment or not. Prior to 2004, for the long-term trend assessments, excluded students were identified for each age class, and an Excluded Student Survey was completed for each excluded student. Beginning in 2004, both long-term trend and main NAEP assessments use identical procedures. In 2010, the Governing Board revised its policy on inclusion. The current policy defines specific inclusion goals for NAEP samples. At the national, state, and district levels, the goal is to include 95 percent of all students selected for the NAEP samples, and 85 percent of those in the NAEP sample who are identified as SD or ELL.

Main national NAEP sample sizes. In 2011, the main national and state NAEP assessed students in reading and mathematics at grades 4 and 8 and in science at grade 8. In addition, the writing assessment was administered to a national sample at grades 8 and 12. The main national mathematics assessment sampled 214,200 grade 4 students and 180,400 grade 8 students; the reading assessment sampled 222,200 grade 4 students and 174,700 grade 8 students. The science assessment sampled 124,200 grade 8 students. The main national writing assessment sampled 24,600 grade 8 students. For 2013, the main national mathematics assessment sampled 186,500 grade 4 students and 170,100 grade 8 students, while the main national reading assessment sampled 190,400 grade 4 students and 171,800 grade 8 students. In 2015, the main national mathematics assessment sampled 139,900 grade 4 students and 136,900 grade 8 students, while the main national reading assessment sampled 139,100 grade 4 students and 136,500 grade 8 students.

TUDA sample sizes. In 2011, 2013, and 2015, twenty-one urban districts (including District of Columbia) participated in TUDA in mathematics and reading. The sample of students in the participating TUDA school districts is an extension of the sample of students who would usually be selected as part of the state and national samples. The sample design for TUDA districts provides for oversampling. These extended samples allow reliable reporting of student groups within these districts.

Results for students in the TUDAs are included with those for states and the nation with appropriate weighting. For example, the data for students tested in the Chicago sample are used to report results for Chicago, but also contribute to Illinois' estimates (and, with appropriate weights, to national estimates). Chicago has approximately 20 percent of the students in Illinois; therefore Chicago will contribute 20 percent, and the rest of the state will contribute 80 percent, to Illinois' results.

Long-term trend NAEP sample sizes. The long-term trend assessment tested mathematics, reading, science, and writing across years through 1999, using relatively small national samples. Samples of students were selected by age (9, 13, and 17) for mathematics, science, and reading, and by grade (4, 8, and 11) for writing. Students within schools were randomly assigned to either mathematics/science or reading/writing assessment sessions subsequent to their selection for participation in the assessments. In 2004, science and writing were removed from the trend assessments; the trend assessments are now scheduled to be administered in mathematics and reading every 4 years (but not in the same years as the main assessments). In 2004, approximately 24,100 students took the modified[1] reading assessment, while about 14,000 took the bridge[2] reading assessment. In 2004,

[1] The modified assessment included new items and features, representing the new design.
[2] The bridge assessment replicates the assessment given in the previous assessment year.

approximately 22,400 students took the modified mathematics assessment, while about 14,700 took the bridge mathematics assessment. The latest long-term trend assessment was conducted during the 2011–12 school year (fall for age 13; winter for age 9; spring for age 17), with over 17,000 students at each age being assessed in either reading or mathematics. For the 2007–08 assessment, approximately 26,600 students were assessed in reading and 26,700 students assessed in mathematics.

NIES sample sizes. The NIES survey questionnaire sample is designed to produce information representative of the target population of all fourth- and eighth-grade AI/AN students in the United States. In 2005, the survey questionnaire sample included about 5,600 eligible students at approximately 550 schools located throughout the United States. The sample consisted of approximately 84 percent public, 4 percent private, and 12 percent BIE schools (unweighted). In 2007, the NIES survey questionnaire sample included about 12,900 AI/AN students at approximately 1,900 schools at grade 4 and 14,600 AI/AN students at 2,000 schools at grade 8 located throughout the United States. The sample consisted of approximately 94 percent public, 1 percent private, and 5 to 6 percent BIE schools at grades 4 and 8 (as well as a small number of DoDEA schools). All BIE schools were part of the sample. In 2009, the NIES survey questionnaire sample consisted of about 12,300 grade 4 students in approximately 2,300 schools and approximately 10,400 students in grade 8 at about 1,900 schools. In 2011, the NIES survey questionnaire sample consisted of about 10,200 grade 4 students in approximately 1,900 schools and approximately 10,300 students in grade 8 at about 2,000 schools.

The samples of AI/AN students participating in the 2011 NAEP reading and mathematics assessments, upon which the student performance results are based (and which also comprises the assessment component of NIES), represent augmentations of the sample of AI/AN students who would usually be selected to participate in NAEP. This allows more detailed reporting of performance for this group.

In 2005, seven states had sufficient samples of AI/AN students to report state-level data: Alaska, Arizona, Montana, New Mexico, North Dakota, Oklahoma, and South Dakota. In 2007, a total of 11 states had sufficiently large samples, with Minnesota, North Carolina, Oregon, and Washington being added to the original seven selected states for 2005. In 2009, results were also reported for Utah, resulting in state-level reporting for a total of 12 states. In 2011, results are reported for the same 12 states. While 6 of the 12 states had sufficient AI/AN students without oversampling, schools in 6 states were oversampled in 2011: Arizona, Minnesota, North Carolina, Oregon, Utah, and Washington.

Assessment Design

Since 1988, the Governing Board has selected the subjects for the main NAEP assessments. NAGB also oversees the creation of the frameworks that underlie the assessments and the specifications that guide the development of the assessment instruments.

Development of framework and questions. The Governing Board uses an organizing framework for each subject to specify the content that will be assessed. This framework is the blueprint that guides the development of the assessment instrument. The framework for each subject area is determined with input from teachers, curriculum specialists, subject-matter specialists, assessment experts, policy makers, and members of the general public.

Unlike earlier multiple-choice instruments, current instruments dedicate a significant amount of testing time to constructed-response questions that require students to compose written answers.

The questions and tasks in an assessment are based on the subject-specific frameworks. They are developed by teachers, subject-matter specialists, and testing experts under the direction of NCES and its contractors. For each subject-area assessment, a national committee of experts provides guidance and reviews the questions to ensure that they meet the framework specifications. Items are also reviewed by NAGB. For each state-level assessment, teachers, state curriculum and assessment specialists review the NAEP questions.

Matrix sampling. Several hundred questions are typically needed to reliably test the many specifications of the complex frameworks that guide NAEP assessments. However, administering the entire collection of cognitive questions to each student would be far too time-consuming to be practical. Matrix sampling allows the assessment of an entire subject area within a reasonable amount of testing time, in most cases 50 minutes for paper-pencil administered assessments and 60 minutes for computer administered assessments. By this method, different portions from the entire pool of cognitive questions are printed in separate booklets or assembled separately on the computer and administered to different samples of students.

In matrix sampling, NAEP uses a focused balanced incomplete block or partial balanced incomplete

block (BIB or pBIB) design. The NAEP BIB design varies according to subject area. A BIB spiraling design ensures that students receive different interlocking sections of the assessment, enabling NAEP to check for any unusual interactions that may occur between different samples of students and different sets of assessment questions. This procedure assigns blocks of questions in a manner that "balanced" the positioning of blocks across booklets and "balanced" the pairing of blocks within booklets according to content. The balancing is "incomplete" because not all blocks are matched to all other blocks. The "spiraling" aspect of this procedure cycles the booklets for administration so that, typically, any group of students will receive approximately the target proportions of different types of booklets.

Data Collection and Processing

Since 1983, NCES has conducted NAEP through a series of contracts, grants, and cooperative agreements with the Educational Testing Service (ETS) and other contractors. ETS is directly responsible for developing the assessment instruments, analyzing the data, and reporting the results. Westat, Inc. selects the school and student samples, trains assessment administrators, and manages field operations (including assessment administration and data collection activities). NCS Pearson is responsible for printing and distributing the assessment materials and for scanning and scoring students' responses. Contractors are subject to change in future contracts.

Reference dates/testing window. Data for the main national NAEP and main state NAEP are collected from the last week in January through the first week in March. Data for the long-term trend NAEP are collected during the fall for age 13; during the winter of the same school year for age 9; and during the spring for age 17.

Data collection. Before 2002, NCES had relied heavily on school personnel to administer NAEP assessments. Beginning with the 2002 assessments, however, NAEP field staff has administered NAEP assessment sessions. Obtaining the cooperation of the selected schools requires substantial time and energy, involving a series of mailings that includes letters to the chief state school officers and district superintendents to notify the sampled schools of their selection; additional mailings of informational materials; and introductory online or in-person meetings where procedures are explained.

The corresponding teacher and school questionnaires are available online ahead of the NAEP assessment (typically more than six weeks before the assessment window begins).

NCS Pearson produces the materials needed for NAEP assessments. NCS Pearson prints identifying barcodes and numbers for the booklets and questionnaires, pre-assigns the booklets to testing sessions, and prints the booklet numbers on the administration schedule. These activities improve the accuracy of data collection and assist with the BIB spiraled distribution process. With the introduction of technology-based assessments (TBA), all responses will be collected electronically.

Assessment exercises are administered either to individuals or to small groups of students by specially trained field personnel. For all three ages in the long-term trend NAEP, the mathematics questions were administered using a paced audiotape before 2004. Since 2004, the long-term trend assessments have been administered through test booklets read by the students.

For the long-term trend assessments, Westat hires and trains approximately 85 field staff to collect the data. For the 2009 main national and state assessments, Westat hired and trained about 7,000 field staff to conduct the assessments.

After each session, Westat staff interview the assessment administrators to receive their comments and recommendations. As a final quality control step, a debriefing meeting is held with the state supervisors to receive feedback that will help improve procedures, documentation, and training for future assessments.

For the NIES survey questionnaire, NCES data collection contractor staff visit the schools to administer survey questionnaires. Students complete the questionnaires in group settings proctored by study representatives. In order to decrease the possibility that survey responses might be adversely affected by students' reading levels, the questions are read aloud to all grade 4 students and to grade 8 students whom school staff think might need assistance. In addition, the study representatives are available to answer any questions that students have as they work on the questionnaires.

For both NIES and NAEP, teachers and school administrators were asked to complete the questionnaires on their own. While the vast majority of teachers and schools complete these questionnaires online, there is a paper questionnaire option for those that need it.

Data processing. NCS Pearson handles all receipt control, data preparation and processing, scanning, and

scoring activities for NAEP. Using an optical scanning machine, NCS Pearson staff scans the multiple-choice selections, the handwritten student responses, and other data provided by students, teachers, and administrators. An intelligent data entry system is used for resolution of the scanned data, the entry of documents rejected by the scanning machine, and the entry of information from the questionnaires (for those that complete a paper questionnaire). An image-based scoring system introduced in 1994 virtually eliminates paper handling during the scoring process. This system also permits online monitoring of scoring reliability and creation of recalibration sets.

ETS develops focused, explicit scoring guides with defined criteria that match the criteria emphasized in the assessment frameworks. The scoring guides are reviewed by subject-area and measurement specialists, the instrument development committees, NCES, and NAGB to ensure consistency with both question wording and assessment framework criteria. Training materials for scorers include examples of student responses from the actual assessment for each performance level specified in the guides. These exemplars help scorers interpret the scoring guides consistently, thereby ensuring the accurate and reliable scoring of diverse responses.

The image-based scoring system allows scorers to assess and score student responses online. This is accomplished by first scanning the student response booklets, digitizing the constructed responses, and storing the images for presentation on a large computer monitor. The range of possible scores for an item also appears on the display; scorers click on the appropriate button for quick and accurate scoring. The image-based scoring system facilitates the training and scoring process by electronically distributing responses to the appropriate scorers and by allowing ETS and NCS Pearson staff to monitor scorer activities consistently, identify problems as they occur, and implement solutions expeditiously. The system also allows the creation of calibration sets that can be used to prevent drift in the scores assigned to questions. This is especially useful when scoring large numbers of responses to a question (e.g., more than 30,000 responses per question in the state NAEP). In addition, the image-based scoring system allows all responses to a particular exercise to be scored continuously until the item is finished, thereby improving the validity and reliability of scorer judgments. The newer computer-based assessments do not require scanning.

The reliability of scoring is monitored during the coding process through (1) backreading, where scoring supervisors review a portion of each scorer's work to confirm a consistent application of scoring criteria across a large number of responses and across time; (2) daily calibration exercises to reinforce the scoring criteria after breaks of more than 15 minutes; and (3) a second scoring of some of the items appearing only in the main national assessment, as well as some of the items appearing in both the main national and state assessments (and a comparison of the two scores to give a measure of inter-rater reliability). To monitor agreement across years, a random sample of responses from previous assessments (for identical items) is systematically interspersed among current responses for rescoring. If necessary, current assessment results are adjusted to account for any differences.

To test scoring reliability, constructed-response item score statistics are calculated for the portion of responses that are scored twice. Cohen's Kappa is the reliability estimate used for dichotomized items and the intraclass correlation coefficient is used as the index of reliability for nondichotomized items. Scores are also constructed for items that are rescored in a later assessment. For example, some 2007 reading and mathematics items were rescored in 2009.

Editing. The first phase of data editing takes place during the keying or scanning of the survey instruments. Machine edits verify that each sheet of each document is present and that each field has an appropriate value. The edit program checks each booklet number against the session code for appropriate session type, the school code against the control system record, and other data fields on the booklet cover for valid ranges of values. It then checks each block of the document for validity, proceeding through the items within the block. Each piece of input data is checked to verify that it is of an acceptable type, that the value falls within a specified range of values, and that it is consistent with other data values. At the end of this process, a paper edit listing of data errors is generated for nonimage and key-entered documents. Image-scanned items requiring correction are displayed at an online editing terminal.

In the second phase of data editing, experienced editing staff review the errors detected in the first phase, compare the processed data with the original source document, and indicate whether the error is correctable or noncorrectable per the editing specifications. Suspect items found to be correct as stated, but outside the edit specifications, are passed through modified edit programs. For nonimage and key-entered documents, corrections are made later via key-entry. For image-processed documents, suspect items are edited online. The edit criteria for each item in

question appear on the screen along with the item, and corrections are made immediately. Two different people view the same suspect item and operate on it separately; a "verifier" ensures that the two responses are the same before the system accepts that item as correct.

For assessment items that must be paper-scored rather than scored using the image system (as was the case for some mathematics items in the 1996 NAEP), the score sheets are scanned on a paper-based scanning system and then edited against tables to ensure that all responses were scored with only one valid score and that only raters qualified to score an item were allowed to score it. Any discrepancies are flagged and resolved before the data from that scoring sheet are accepted into the scoring system.

In addition, a count-verification phase systematically compares booklet IDs with those listed in the NAEP administration schedule to ensure that all booklets expected to be processed were actually processed. Once all corrections are entered and verified, the corrected records are pulled into a mainframe data set and then re-edited with all other records. The editing process is repeated until all data are correct.

Estimation Methods

Once NAEP data are scored and compiled, data from schools and students are weighted according to the sample design and population structure and then adjusted for nonresponse. This ensures that results of the assessments are fully representative of the target populations. The analyses of NAEP data for most subjects are conducted in two phases: scaling and estimation. During the scaling phase, item response theory (IRT) procedures are used to estimate the measurement characteristics of each assessment question. During the estimation phase, the results of the scaling are used to produce estimates of score scale score distributions for groups of students in the various subject areas applying Marginal maximum likelihood (MML) methodology.

Weighting. The weighting for the national and state samples reflects the probability of selection for each student in the sample, adjusted for school and student nonresponse. The weight assigned to a school's or student's response is the inverse of the probability that the student would be selected for the sample. Prior to 2002, poststratification was used to ensure that the results were representative of certain subpopulations corresponding to figures from the U.S. Census and the Current Population Survey (CPS).

Student base weights. The base weight assigned to a student is the reciprocal of the probability that the student would be selected for a particular assessment. This probability is the product of the following two factors:

- the conditional probability that the school would be selected, given the strata; and
- the conditional probability, given the school, that the student would be selected within the school.

Nonresponse adjustments of base weights. The base weight for a selected student is adjusted by two nonresponse factors. The first factor adjusts for sessions that were not conducted. This factor is computed separately within classes formed by the first three digits of strata (formed by crossing the major stratum and the first socioeconomic characteristic used to define the final stratum). Occasionally, additional collapsing of classes is necessary to improve the stability of the adjustment factors, especially for the smaller assessment components. The second factor adjusts for students who failed to appear in the scheduled session or makeup session. This nonresponse adjustment is completed separately for each assessment. For assessed students in the trend samples, the adjustment is made separately for classes of students based on subuniverse and modal grade status. For assessed students in the main samples, the adjustment classes are based on subuniverse, modal grade status, and race class. In some cases, nonresponse classes are collapsed into one class to improve the stability of the adjustment factors.

NIES survey questionnaire weighting. For the survey questionnaire component of NIES, the school probability of selection is a function of three factors: NAEP selection, the probability of being retained for the survey questionnaire component of NIES, and the number of AI/AN students in the NAEP sample per school. Nonresponse adjustments at the school level attempt to mitigate the impact of differential response by school type (public, private, and BIE), region, and estimated percentage enrollment of AI/AN students. For student weights, nonresponse adjustments take into account differential response rates based on student age (above age for grade level or not) and English language learner status. In order to partially counteract the negative impact of low private school participation, a poststratification adjustment is applied to the NIES survey questionnaire weights. The relative weighted proportions of students from public, private, and BIE schools, respectively, are adjusted to match those from the data of the assessment component of NIES. This not only ensured greater consistency between the findings of the two NIES components, but since the

proportions of students are more reliably estimated from the NIES assessment data (which involved a far larger school sample than the survey questionnaire), this weight adjustment increases the accuracy and reliability of the NIES survey questionnaire results.

Scaling. For purposes of summarizing item responses, a scaling technique that has its roots in IRT procedures and the theories of imputation of missing data are used.

The first step in scaling is to determine the percentage of students who give various responses to each cognitive, or subject-matter, question and each background question. For cognitive questions, a distinction is 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 are considered intentional omissions. Missing responses at the end of a block are generally considered "not reached" and treated as if the questions had not been presented to the student. In calculating response percentages for each question, only students classified as having been presented that question are used in the analysis. Each cognitive question is also examined for differential item functioning (DIF). DIF analyses identify questions on which the scores of different subgroups of students at the same ability level differ significantly.

Development of scales. For the main assessments, the frameworks for the different subject areas dictate the number of subscales required. In the 2009 NAEP, five subscales were created for the main assessment in mathematics in grades 4 and 8 (one for each mathematics content strand), and three subscales were created for science (one for each field of science: Earth, physical, and life). Generally, a composite scale is also created as an overall measure of students' performance in the subject area being assessed (e.g., mathematics). The composite scale is a weighted average of the separate subscales for the defined subfields or content strands. For the long-term trend assessments, a single scale is used for summarizing proficiencies at each age in mathematics and reading.

Within-grade vs. cross-grade scaling. The reading and mathematics main NAEP assessments were developed with a cross-grade framework, where the trait being measured was conceptualized as cumulative across the grades of the assessment. Accordingly, a single 0−500 scale was established for all three grades in each assessment. In 1993, however, the Governing Board determined that future NAEP assessments should be developed using within-grade frameworks and be scaled accordingly. This both removed the constraint that the trait being measured is cumulative and eliminated the need for overlap of questions across grades. Any questions that happen to be the same across grades are scaled separately for each grade, thus making it possible for common questions to function differently in the separate grades.

The 1994 history and geography assessments were developed and scaled within grade, according to NAGB's new policy. The scales were aligned so that grade 8 had a higher mean than grade 4 and grade 12 had a higher mean than grade 8. The 1994 reading assessment, however, retained a cross-grade framework and scaling. All three main assessments in 1994 used scales ranging from 0 to 500.

The 2008 long-term trend assessments remained cross-age, using a 0−500 scale. The 2009 main science assessment was developed within-grade, but adopted new scales ranging from 0 to 300. The 2005 main assessment in mathematics continued to use a cross-grade framework with a 0−500 scale in grades 4 and 8, but used a 0–300 within-grade scale for 12th grade. In 1998, reading, writing and civics assessments were scaled within-grade.

Linking of scales. Before 2002, results for the main state assessments were linked to the scales for the main national assessments, enabling state and national trends to be studied. Equating the results of the state and national assessments depended on those parts of the main national and state samples that represented a common population: (1) the state comparison sample—students tested in the national assessment who come from the jurisdictions participating in the state NAEP; and (2) the state aggregate sample—the aggregate of all students tested in the state NAEP. Since 2002, the national sample has been a superset of the state samples (except in those states that do not participate).

Imputation. Until the 2002 NAEP assessment, no statistical imputations were generated for missing values in the teacher, school, or SD/ELL questionnaires, or for missing answers to cognitive questions. Most answers to cognitive questions are missing by design. For example, 8th-grade students being assessed in reading are presented with, on average, 21 of the 110 assessment items. Whether any given student gets any of the remaining 89 individual questions right or wrong is not something that NAEP imputes. However, since 1984, multiple imputation techniques have been used to create plausible values. Once created, subsequent users can analyze these plausible values with common software packages to

obtain NAEP results that properly account for NAEP's complex item sampling designs.

Trying to use partial scores based on the small proportion of the assessment to which any given student is exposed would lead to biased results for group scores due to an inherently large component of measurement error. NAEP developed a process of group score calculation in order to get around the unreliability and noncomparability of NAEP's partial test forms for individuals. NAEP estimates group score distributions using MML estimation, a method that calculates group score distributions based directly on each student's responses to cognitive questions, not on summary scores for each student. As a result, the unreliability of individual-level scores does not decrease NAEP's accuracy in reporting group scores. The MML method does not employ imputations of answers to any questions or of scores for individuals.

Imputation is performed in three stages. The first stage requires estimating IRT parameters for each cognitive question. The second stage results in MML estimation of a set of regression coefficients that capture the relationship between group score distributions and nearly all the information from the variables in the teacher, school, or SD/ELL questionnaires, as well as geographical, sample frame, and school record information. The third stage involves the imputation that is designed to reproduce the group-level results that could be obtained during the second stage.

NAEP's imputations follow Rubin's (1987) proposal that the imputation process be carried out several times, so that the variability associated with group score distributions can be accurately represented. NAEP estimates five plausible values for each student. Each plausible value is a random selection from the joint distribution of potential scale scores that fit the observed set of response for each student and the scores for each of the groups to which each student belongs. Estimates based on plausible values are more accurate than if a single (necessarily partial) score were to be estimated for each student and averaged to obtain estimates of subgroup performances. Using the plausible values eliminates the need for secondary analysts to have access to specialized MML software and ensures that the estimates of average performance of groups and estimates of variability in those averages are accurate.

Recent Changes

Several important changes have been implemented since 1990.

- Beginning with the 1990 mathematics assessment, NAGB established three reporting levels for reporting NAEP results: basic, proficient, and advanced.
- In 1990, state assessments were added to NAEP. The 1990 to 1994 assessments are referred to as trial state assessments.
- In 1992, a generalized partial-credit model (GPCM) was introduced to develop scales for the more complex constructed-response questions. The GPCM model permits the scaling of questions scored according to multipoint rating schemes.
- In 1993, NAGB determined that future NAEP assessments should have within-grade frameworks and scales. The 1994 main history and geography assessments followed this new policy, as did the 1996 main science assessment, and the 1998 writing assessment. Mathematics and reading in the main NAEP will continue to have cross-grade scales until further action by NAGB (and a parallel change in the trend assessment), except for mathematics at grade 12, which was removed from cross-grade scales and reported in a within-grade scale in 2005.
- In 1994, the new image-based scoring system virtually eliminated paper handling during the scoring process. This system also permits scoring reliability to be monitored online and recalibration methods to be introduced.
- The 1996 main NAEP included new samples for the purpose of studying greater inclusion of SD/LEP students and obtaining data on students eligible for advanced mathematics or science sessions.
- In 1997, there was a probe of student performance in the arts.
- New assessment techniques included: open-ended items in the 1990 mathematics assessment; primary trait, holistic, and writing mechanics scoring procedures in the 1992 writing assessment; the use of calculators in the 1990, 1992, 1996, and 2000 mathematics assessments; a special study on group problem solving in the 1994 history assessment; and a special study in theme blocks in the 1996 mathematics and science assessments.
- Beginning in 1998, testing accommodations were provided in the NAEP reading assessments; in this transition to a more inclusive NAEP, administration procedures were introduced that

allowed the use of accommodations (e.g., extra time, individual rather than group administration) for students who required them to participate. During this transition period, reading results in 1998 were reported for two separate samples: one in which accommodations were not permitted and one in which accommodations were permitted. Beginning in 2002, accommodations were permitted for all reading administrations.

- In 1999, NAGB discontinued the long-term trend assessment in writing for technical reasons. More recently, NAGB decided that changes were needed to the design of the science assessment and, given recent advances in the field of science, to its content. As a result, the science long-term trend assessment was not administered in 2003-04 or in subsequent administrations.

- With the expansion and redesign of NAEP under the No Child Left Behind Act, NAEP's biennial state-level assessments are being administered by contractor staff (not local teachers). The newly redesigned NAEP has four important features. First, NAEP administers tests for different subjects (such as mathematics, science, and reading) in the same classroom, thereby simplifying and speeding up sampling, administration, and weighting. Second, NAEP conducts pilot tests of candidate items for the next assessment and field tests of items for precalibration in advance of data collection, thereby speeding up the scaling process. Third, NAEP conducts bridge studies, administering tests both under new and the old conditions, thereby providing the possibility of linking old and new findings. Finally, NAEP is adding additional test questions at the upper and lower ends of the difficulty spectrum, thereby increasing NAEP's power to measure performance gaps.

- Beginning in 2002, the NAEP national sample for main national assessment was obtained by aggregating the samples from each state, rather than by obtaining an independently selected national sample. Prior to 2002, separate samples were drawn for the NAEP main national and state assessments.

- In 2002, TUDA began assessing performance in five large urban districts with reading and writing assessments. TUDA continued in 2003 in nine large urban districts with reading and mathematics and in 2005 in 10 large urban districts with reading, mathematics, and science. As of 2013, 21 urban school districts were included in the TUDA program.

- Beginning with the 2003 NAEP, each state must have participation from at least 85 percent—instead of 70 percent—of the schools in the original sample in order to have its results published.

- In 2003 and 2005, Puerto Rico participated in the NAEP assessment of mathematics. However, Puerto Rico was excused from the NAEP assessment of reading in English because Spanish is the language of instruction in Puerto Rico. NCES also administered the 2007 mathematics assessment in Puerto Rico. In 2007, a representative sample of approximately 2,800 students in 100 schools was assessed at both grade 4 and at grade 8. In 2011, public school students in Puerto Rico at grades 4 and 8 participated in a research study using a Spanish-language version of the National Assessment of Educational Progress (NAEP) in mathematics. This was not a full assessment, so results were not reported until they could be verified with the 2013 assessment.

- In 2004, several changes were implemented to the NAEP long-term trend assessments to reflect changes in NAEP policy, maintain the integrity of the assessments, and increase the validity of the results obtained. The changes to the assessment instruments include: removal of science items; inclusion of students with disabilities and English language learners; replacement of items that used outdated contexts; creation of a separate background questionnaire; elimination of "I don't know" as a response option for multiple-choice items; and use of assessment booklets that pertain to a single subject area (whereas in the past, a single assessment booklet may have contained both reading and mathematics items).

- In 2005, NAGB introduced changes in the NAEP mathematics framework for grade 12 in both the assessment content and administration procedures. One of the major differences between the 2005 assessment and previous assessments at grade 12 is the five content areas were collapsed into four areas, with geometry and measurement being combined. In addition, the assessment included more questions on algebra, data analysis, and probability to reflect changes in high school mathematics standards

and coursework. The overall average mathematics score in 2005 was set at 150 on a 0–300 scale.

- In 2006, economics was assessed at grade 12 for the first time. The NAEP economics assessment results present a broad view of how well our nation's students at grade 12 understand economics and have knowledge of the workings of domestic and international economics. More than 11,000 grade 12 students in approximately 600 public and private schools across the nation were assessed. A within-grade scale was developed, with the overall average economics score in 2006 set at 150 on a 0–300 scale.

- In 2009, the reading framework changed to include more emphasis on literary and informational texts, a redefinition of reading cognitive processes, a systemic assessment of vocabulary knowledge, and the addition of poetry to grade 4. Results from special analyses conducted in 2009 determined that, even with these changes to the assessment, results could continue to be compared to those from earlier assessments.

- In 2009, TUDA was expanded to 18 large urban districts, assessing reading, mathematics and science. In addition, 11 states were assessed in reading and mathematics at grade 12 on a trial basis. In 2011, TUDA expanded to 21 large urban districts, assessing reading and mathematics.

- In 2009, interactive computer tasks in science were administered online at grades 4, 8, and 12. These tasks consisted of simulations for the students to draw inferences and conclusions about a problem.

- In 2011, NAEP administered its first computer-based assessment in writing at grades 8 and 12. A pilot test of students at grade 4 was also conducted in 2012, and the empirical correlations observed between performance and the contextual and demographic factors largely supported the predictions as specified in the conceptual model, including the key prediction that the differential effects of the computer on the writing performance of high- and non-high-performing fourth-graders would be related to their prior exposure to writing on the computer.

- In 2015, NAEP began a phased approach to transition its paper-and-pencil assessments to digital-based assessments and delivery, starting with a pilot test for mathematics, reading, and science assessments using the latest technology tools. Results are not available at this time.

Future Plans

Main NAEP assessments are scheduled for annual administration. Reading and mathematics are assessed every 2 years in odd-numbered years; science and writing are scheduled to be assessed every 4 years (in the same years as reading and mathematics, but alternating with each other); and other subjects are assessed at the national level in even-numbered years. A new, computer-based assessment, Technology and Engineering Literacy, was piloted in 2013, and a full assessment was conducted at grade 8 in 2014. NAEP broadly defines technological and engineering literacy as the capacity to use, understand, and evaluate technology as well as to understand technological principles and strategies needed to develop solutions and achieve goals. For the full NAEP Assessment schedule, see https://nces.ed.gov/ationsreportcard/about/calendar.aspx.

The NAEP program is in the midst of transitioning all of its assessments to digitally based content and delivery. Beginning in 2017, the NAEP mathematics, reading, and writing assessments will be administered to students throughout the nation in NAEP-provided tablets. Some questions may include multimedia such as audio and video, other questions may allow the use of embedded technological features (such as an onscreen calculator) to form a response. Additional subjects will be administered on tablets in 2018 and 2019. NCES will also pilot science interactive computer tasks (ICTs) and hands-on-tasks (HOTs).

To continue moving the NAEP program forward, a summit of diverse experts in assessment, measurement, cognition, and technology was convened in August 2011 and January 2012. These experts discussed and debated ideas for the future of NAEP. NCES convened its most recent workshop in January 2013. State and district assessment staff met to develop and prioritize recommendations for NAEP.

NIES is shifting from a two-year administration cycle to a four-year administration cycle. The most recent NIES administration was conducted in 2015.

5. DATA QUALITY AND COMPARABILITY

As the Nation's Report Card, NAEP must report accurate results for populations of students and subgroups of these populations (e.g., Black, Hispanic,

or other race/ethnicity, or students attending private schools).

Every activity in NAEP assessments is conducted with rigorous quality control, contributing both to the quality and comparability of the assessments and their results. All questions undergo extensive reviews by subject-area and measurement specialists, as well as careful scrutiny to eliminate any potential bias or lack of sensitivity to particular groups. The complex process by which NAEP data are collected and processed is monitored closely. Although each participating state is responsible for its own data collection for the main state NAEP, Westat ensures uniformity of procedures across states through training, supervision, and quality control monitoring.

With any survey, however, there is the possibility of error. The most likely sources of error in NAEP are described below.

Sampling Error

Two components of uncertainty in NAEP assessments are accounted for in the variability of statistics based on scale scores: (1) the uncertainty due to sampling only a small number of students relative to the whole population; and (2) the uncertainty due to sampling only a relatively small number of questions. 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, a jackknife replication procedure is used to estimate standard errors. While the jackknife standard error provides a reasonable measure of uncertainty about student data that can be observed without error, each student in NAEP assessments typically responds to so few questions within any content area that the scale score for the student would be imprecise. It is possible to describe the performance of groups and subgroups of students because, as a group, all students are administered a wide range of items.

NAEP uses MML procedures to estimate group distributions of scores. However, the underlying imprecision that makes this step necessary adds an additional component of variability to statistics based on NAEP scale scores. This imprecision is measured by the imputed variance, which is estimated by the variance among the plausible values drawn from each student's posterior distribution of possible scores. The final estimate of the variance is the sum of the sampling variance and the measurement variance.

Nonsampling Error

While there is the possibility of some coverage error in NAEP, the two most likely types of nonsampling error are nonresponse error due to nonparticipation and measurement error due to instrumentation defects (described below). The overall extent of nonsampling error is largely unknown.

Coverage error. In NAEP, coverage error can result either from the sampling frame of schools being incomplete or from the schools' failure to include all the students on the lists from which grade or age samples are drawn. For NAEP, the most recent and available school list maintained by CCD supplies the names of the regular public schools, BIE schools, and DoDEA schools. This list, however, does not include schools that opened between the time of the creation of the list and the time of the NAEP assessment. To ensure that students in new public schools were represented, each sample district in NAEP was asked to update lists of schools with newly eligible schools.

Catholic and other nonpublic schools in NAEP were obtained from the most recent and available PSS. PSS uses a dual-frame approach. The list frame (containing most private schools in the country) is supplemented by an area frame (containing additional schools identified during a search of randomly selected geographic areas around the country). Coverage of private schools in the PSS is very high. (See the PSS chapter)

Nonresponse error. *Unit nonresponse.* Standards established by the National Assessment Governing Board require that school participation rates for the original state/jurisdiction samples need to be at least 85 percent for results to be reported. In 2015, all 52 states and jurisdictions met this participation rate requirement with participation rates of 93 to 100 percent. Before substituting new schools for originally sampled schools that declined to participate, the weighted national school participation rates for the 2015 reading and mathematics assessments were 97 percent for grade 4 (100 percent for public schools, 61 percent for private schools, and 83 percent for Catholic schools), and 96 percent for grade 8 (99 percent for public schools, 56 percent for private schools, and 80 percent for Catholic schools). In 2015, the school participation rates for private schools at both grades 4 and 8 did not meet the criteria so their results are not reportable. Nonresponse bias analyses were conducted for the private school samples at both grades 4 and 8 in 2015. The results of the nonresponse bias analyses showed that including substitute schools and adjusting the sampling weights to account for school nonresponse were partially

effective in reducing the potential for nonresponse bias. For example, the potential bias for Catholic schools and non-Catholic private schools was reduced after the substitution and nonresponse adjustments. However, some variables examined in the analyses still indicated potential bias after nonresponse adjustments. For instance, larger schools were somewhat less represented in the final private school samples at both grades.

For the 2014 U.S. History and Geography assessments, the weighted national school participation rates for public and private schools was 90 percent and the weighted student participation rates was 94 percent at grade 8. For the 2014 Civics assessment, the weighted national school participation rate for public and private schools was 89 percent and the weighted student participation rate was 94 percent at grade 8.

For the 2012 Economics assessment, the weighted participation rate for public schools exceeded the 85 percent threshold; however, a nonresponse bias analysis was conducted for the public school sample because there were no participating public schools in Texas, which makes up approximately 9 percent of the public schools nationally. The original responding public school sample differed from the entire sample of eligible schools with respect to several variables; for instance, public schools in the south Census region were underrepresented in the responding sample. Including substitute schools in the assessment sample was not effective in reducing potential bias, as no substitute schools in Texas participated. However, adjusting the sampling weights to account for school nonresponse resulted in the reduction of potential nonresponse bias.

In the 2009 reading and mathematics assessments, all 52 states and jurisdictions met participation rate standards at both grade 4 and grade 8. The national school participation rates for public and private schools combined were 97 percent at grades 4 and grade 8. Student participation rates were 95 percent at grade 4 and 93 percent at grade 8. Participation rates needed to be 70 percent or higher to report results separately for private schools. While the participation rate for private schools did meet the standard in 2009, it did not always meet the standard in previous assessment years. See table NAEP-1 for more details.

In the 2007 reading and mathematics assessments, all 52 states and jurisdictions met participation rate standards at both grades 4 and 8. The national school participation rates for public and private schools combined were 98 percent at grade 4 and 97 percent at grade 8. Student participation rates were 95 percent at grade 4 and 92 percent at grade 8. The participation rate for private schools met the reporting standard of 70 percent in 2007.

In the 2005 reading and mathematics assessments at grade 12, participation standards were met for public schools but not for private schools. At the student level, response rates at grade 12 fell below 85 percent for students in both public and private schools. A nonresponse bias analysis showed significant differences between responding and nonresponding public school students in terms of gender, race/ethnicity, age, and English language learner identification. Although the differences are quite small, it is unlikely that nonresponse weighting adjustments completely accounted for these differences.

In the 2011 survey questionnaire component of NIES, questionnaires were completed by about 10,200 AI/AN grade 4 students from 1,900 schools and 10,300 AI/AN grade 8 students from 2,000 schools. Also responding to the survey were about 3,000 grade 4 teachers, 4,600 grade 8 teachers, 1,900 grade 4 school administrators and 2,000 grade 8 school administrators associated with these students.

Some school administrators responded for both grades 4 and 8. The weighted school response rates were 97 percent at grade 4 and 98 percent at grade 8. The weighted student response rates were 86 percent at grade 4 and 84 percent at grade 8.

Because the weighted student response rate for grade 8 was below 85 percent, a student nonresponse bias analysis was conducted. The analysis showed that the responding grade 8 sample was different from the original sample with respect to geographical distribution across regions, states, and types of location; gender; relative age; school density; and proportions of students with disabilities (SD) and English language learners (ELL). Both SD and ELL students were underrepresented in the responding sample. After weighting adjustments were made to account for differences in the response rates by student groups, the only evidence of remaining bias was the slight underrepresentation of AI/AN students with disabilities and students from low density schools (population less than 25 percent AI/AN). The effect on survey estimates seems likely to be very slight, since the distribution of the final student sample matches closely with that of the original sample.

Table NAEP-1. Weighted school and student response rates for selected NAEP national assessments, by assessment and grade: 2006–2015

	School participation[1]		Student participation
Assessment and grade	Student weighted	School weighted	Student weighted
2015 Mathematics			
Grade 4	97	89	94
Grade 8	96	83	92
2015 Reading			
Grade 4	97	89	94
Grade 8	96	83	92
2013 Mathematics			
Grade 4	97	90	95
Grade 8	97	85	93
2013 Reading			
Grade 4	97	90	95
Grade 8	97	85	93
2012 Trend			
Age 9	87	82	95
Age 13	88	81	93
Age 17	84	80	88
2012 Economics			
Grade 12	87	84	85
2011 Mathematics			
Grade 4	97	92	95
Grade 8	98	88	93
2011 Reading			
Grade 4	97	92	95
Grade 8	98	88	93
2011 Science			
Grade 8	97	88	93
2011 Writing			
Grade 8	97	87	94
Grade 12	94	89	87
2008 Trend			
Age 9	96	91	95
Age 13	95	89	94
Age 17	90	85	88
2007 Writing			
Grade 8	97	87	92
Grade 12	89	83	80
2007 Reading			
Grade 4	98	92	95
Grade 8	97	87	92
2007 Mathematics			
Grade 4	98	92	95
Grade 8	97	87	92
2006 Economics			
Grade 12	79	78	73

[1] Participation rates do not include substitutions.

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2015 Mathematics and Reading Assessments, 2013 Mathematics and Reading Assessments, 2012 Trend Assessment, 2011 and 2009 Mathematics, Reading and Science Assessments, 2011 Writing Assessment, 2008 Trend Assessment, 2007 Writing, Reading and Mathematics Assessments, 2012 and 2006 Economics Assessments.

In the 2007 survey questionnaire component of NIES, questionnaires were completed by about 10,400 grade 4 students from 1,700 schools and 11,300 grade 8 students from 1,800 schools. Also responding to the survey were about 3,000 grade 4 teachers, 4,600 grade 8 teachers, 1,700 grade 4 school administrators and 1,800 grade 8 school administrators associated with these students. Some school administrators responded for both grades 4 and 8. The weighted student response rates were 85 percent at grade 4 and 82 percent at grade 8. The weighted school response rates were 88 percent at grade 4 and 90 percent at grade 8.

In the 2005 survey questionnaire component of NIES, questionnaires were completed by about 2,600 grade 4 students and 2,500 grade 8 students at approximately 480 schools. Also responding to the survey were about 480 grade 4 teachers, 820 grade 8 teachers, 240 grade 4 principals, and 230 grade 8 principals associated with these students. Some principals responded for both grades 4 and 8. The weighted student response rates were 95 percent at grade 4 and 91 percent at grade 8. The weighted school response rates were 87 percent at grade 4 and 93 percent at grade 8.

In the 2008 long-term trend assessments, private school participation rate at age 17 was 61 percent, below the standard for reporting. However, Catholic school participation rates at all three ages (88, 94, and 76 percent at ages 9, 13, and 17, respectively) met the reporting standards.

In the 2004 long-term trend reading and mathematics assessments, the overall response rate (the product of the weighted school participation rate before substitution and the weighted student participation rate) 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 than from Catholic schools. In addition, nonrespondent schools in the long-term trend assessment had a lower percentage of Black students than participating schools. 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 than participating schools at age 17. At the student level at age 17, some bias was shown for race/ethnicity, free lunch eligibility, and disability status.

Item nonresponse. In almost all NAEP Item Response Theory (IRT) analyses, missing responses at the end of a block of items are considered *not reached items* and are treated as if they had not been presented to the respondent. Occasionally, extended constructed-response items are the last item in a block. Because considerably more effort is required of the student to answer these items, nonresponse to an extended constructed-response item at the end of a block is considered an intentional omission (and scored as the lowest category) unless the student also did not respond to the item immediately preceding that item. In that case, the extended constructed-response item considered not reached is treated as if it had not been presented to the student. In the case of the national main and state writing assessment, there is a single constructed-response item in each separately-timed block. In the writing assessment when a student does not respond to the item or when the student provides an off-task response, the response also is treated as if the item had not been administered.

Missing responses to items before the last observed response in a block are considered intentional omissions. If the omitted item is a multiple-choice item, the missing response is treated as fractionally correct at the value of the reciprocal of the number of response alternatives. If the omitted item is not a multiple-choice item, the missing response is scored so that the response is in the lowest category.

Measurement error. Nonsampling error can result from the failure of the test instruments to measure what is being taught and, in turn, what is being learned by students. For example, the instruments may contain ambiguous definitions and/or questions that lead to different interpretations by students. Additional sources of measurement error are the inability or unwillingness of students to give correct information and errors in the recording, coding, or scoring of data.

To assess the quality of the data in the final NAEP database, survey instruments are selected at random and compared, character by character, with their records in the final database. As in past years, the 2008 NAEP data-base was found to be more than accurate enough to support analyses.

The observed error rates for the 2008 NAEP were comparable to those of past assessments. Error rates ranged from 0 errors per 10,000 responses for some of the student booklets to 78 errors per 10,000 responses for the students with disabilities questionnaire in Arts.

Revised results. Following the 1994 NAEP assessment, two technical problems were discovered in the procedures used to develop the scale and achievement levels for the 1990 and 1992 mathematics assessments. These errors affected the mathematics scale scores

reported for 1992 and the achievement-level results reported for 1990 and 1992.

NCES evaluated the impact of these errors and subsequently reanalyzed data and reported the revised results from both mathematics assessments. The revised results for 1990 and 1992 are presented in the 1996 mathematics reports. For more detail on these problems, see *The NAEP 1996 Technical Report* (Allen, Carlson, and Zelenak 1999) and the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics* (Allen et al. 1997).

There were also problems related to reading scale scores and achievement levels. These errors affected the 1992 and 1994 NAEP reading assessment results. The 1992 and 1994 reading data have been reanalyzed and reissued in revised reports. For more information, refer to *The NAEP 1994 Technical Report* (Allen, Kline, and Zelenak 1996) and the *Technical Report of the NAEP 1994 Trial State Assessment in Reading* (Mazzeo, Allen, and Kline 1995).

Data Comparability

NAEP allows reliable comparisons between state and national data for any given assessment year. By linking scales across assessments, it is possible to examine short-term trends for data from the main national and state NAEP and long-term trends for data from the long-term trend NAEP.

Main national vs. main state comparisons. NAEP data are collected using a closely monitored and standardized process, which helps ensure the comparability of the results generated from the main national and state assessments. The main national NAEP and main state NAEP use the same assessment booklets, and, since 2002, they have been administered in the same sessions using identical procedures.

Short-term trends. Although the test instruments for the main national assessments are designed to be flexible and thus adaptable to changes in curricular and educational approaches, they are kept stable for shorter periods (up to 12 years or more) to allow analysis of short-term trends. For example, through common questions, the 1996 main national assessment in mathematics was linked to both the 1992 and 1994 assessments.

For 2005 and 2009, the Governing Board adopted a new mathematics framework for grade 12 to reflect changes in high school standards and coursework. In addition, changes were made in booklet design and calculator-use policy for the one-third of the assessment in which calculators were allowed. As a result of these changes, the 2005 and 2009 results could not be placed on the previous NAEP scale and are not compared to results from previous years.

Long-term trends. In order to make long-term comparisons, the long-term trend NAEP uses different samples than the main national NAEP. Unlike the test instruments for the main NAEP, the long-term instruments in mathematics and reading have remained relatively unchanged from those used in previous assessments. The 2004 long-term trend instruments were almost identical to those used in the 1970s. The long-term trend NAEP allows the measurement of educational progress since 1971 in reading and 1973 in mathematics. For more detail on the linking of scales in the long-term trend NAEP, see "Scaling" in section 4 above.

The long-term trend assessment was updated in several ways in 2004 (e.g., inclusion of SD/ELL students). To ensure the comparability of the new assessment and the previous assessments, a bridge study was performed.

Linking to non-NAEP assessments. Linking results from the main state assessments to those from the main national assessments has encouraged efforts to link NAEP assessments with non-NAEP assessments.

Linking to state assessment. NAEP data can be used to map state proficiency standards in reading and mathematics onto the appropriate NAEP scale. The mapping exercise was carried out for data from the 2004–05 and 2006–07 academic years at both grades 4 and 8. Although there is an essential ambiguity in any attempt to place state standards on a common scale, the ranking of the NAEP score equivalents to the states' proficiency standards offers an indicator of the relative stringency of those standards. The current report highlights the findings of the study using state data from the 2008–09 academic year and the 2009 NAEP grades 4 and 8 reading and mathematics assessments. It also examines changes in mapping results over time by comparing the 2009 results with those from the 2005 and 2007 NAEP administrations.

Most states' proficiency standards are at or below NAEP's definition of *Basic* performance. For example, in grade 4 reading, 35 of the 50 states included in the analysis set standards for proficiency (as measured on the NAEP scale) that were lower than the scale score for *Basic* performance on NAEP and another 15 were in the NAEP *Basic* range. In grade 8 reading, 16 of 50 states set standards that were lower than the cut-point for *Basic* performance on NAEP and another 34 were in the NAEP *Basic* range.

These results should be employed cautiously, as differences among states in apparent stringency can be due, in part, to reasonable differences in the assessment frameworks, the types of item formats employed, and the psychometric characteristics of the tests. Moreover, there is some variation among states in the proportion of NAEP sample schools that could be employed in the analysis.

Linking to the International Assessment of Educational Progress (IAEP). In 1992, results from the 1992 NAEP assessment in mathematics in grade 8 were successfully linked to those from IAEP of 1991. Sample data were collected from U.S. students who had been administered both instruments. The relation between mathematics proficiency in the two assessments was modeled using regression analysis. This model was then used as the basis for projecting IAEP scores from non-U.S. countries onto the NAEP scale. The relation between the IAEP and NAEP assessments was relatively strong and could be modeled well. The results, however, should be considered only in the context of the similar construction and scoring of the two assessments. Further studies should be initiated cautiously, even though the path to linking assessments is now better understood.

Linking to TIMSS. The success in linking NAEP to the IAEP sparked an interest in linking the results from the 1996 NAEP assessments in mathematics and science in grade 8 to those from the Third International Mathematics and Science Study (TIMSS) of 1995. The data from this study became available at approximately the same time as the 1996 NAEP data for mathematics and science. Because the two assessments were conducted in different years and no students responded to both assessments, the regression procedure that linked NAEP and IAEP assessments could not be used. The results from grade 8 NAEP and TIMSS assessments were instead linked by matching their score distributions. A comparison of the linked results with actual results from states that participated in both assessments suggested that the link was working acceptably. The results from U.S. students were linked to those of their academic peers in more than 40 other countries. As with the IAEP linked results, these results should be used cautiously.

A second study attempted to link the 2000 grade 8 NAEP assessments in mathematics and science to the 1999 grade 8 TIMSS (which also assessed mathematics and science). The primary linkage used a projection method, which drew data from a sample of students to whom both assessments were administered. The linkage found that the projections were substantially off the mark. A secondary linkage, based on nationally reported numbers using a statistical moderation approach, provided a fairly weak linkage; the moderation linkage did a decent job of projecting TIMSS scores from NAEP scores in the 12 states that participated in both studies, but failed to predict the TIMSS score in the linking sample.

The analyses showed that the TIMSS assessments functioned differently in the linking sample than they did in the national and state samples. A recent study (Phillips 2009) shows that it is possible to make comparisons between TIMSS 2007 and NAEP 2007. For more details, please refer to *The Second Derivative: International Benchmarks in Mathematics for U.S. States and School Districts* (Phillips 2009).

Comparisons with TIMSS. Studies were undertaken to compare the content of two fourth- and eighth-grade assessments in mathematics and science: the NAEP 2000 assessment and the TIMSS 2003 assessment. The comparison study drew upon information provided by the developers of the assessments, as well as data obtained from an expert panel convened to compare the frameworks and items from the two assessments on various dimensions.

For science, the content comparisons between NAEP and TIMSS reveal some key differences in the topics covered, grade-level correspondence, and the characteristics of the item pools on other dimensions. All of these factors together may result in differences in student performance, and it is important to consider these differences when interpreting the results from the different assessments.

Differences in the science content included in each assessment can be seen at both the framework level and in the pool of items developed based on these frameworks. Even in content areas where there is considerable overlap of the frameworks (such as life science and Earth science), a closer examination of the topics and specific objectives covered by the items in each assessment reveals some important differences. In comparison to NAEP, whose framework was developed in the context of the U.S. system, the TIMSS framework reflects a consensus across many countries. Some of the differences in curricula across these countries are reflected in the frameworks and in the differences in content of the two assessments. In particular, the inclusion in TIMSS of separate content areas in chemistry, physics, and environmental science results in broader topic coverage in some areas. While there is a considerable overlap in the topics included in some content areas, the items included in each assessment place different emphases at the topic level. In addition, the "hands-on" tasks in NAEP provide

complementary information to the pencil-and-paper portions of both assessments, enabling the measurement of student performance in this area of knowing and doing science.

With respect to mathematics, a comparison of the frameworks revealed considerable agreement on the general boundaries and basic organization of mathematics content, with both assessments including five main content areas corresponding to traditional mathematics curricular areas: number, measurement, geometry, data, and algebra. Both the NAEP and TIMSS frameworks also include dimensions that define a range of cognitive skills and processes that overlap the two assessments. Despite these apparent similarities at the broadest level, a closer examination of the items in each assessment reveals different emphases at the topic and subtopic levels, as well as some differences in grade-level expectations across mathematics topics.

In 2012, NCES conducted a study to link results from the 2011 NAEP to results from the 2011 TIMSS assessment. The goal of the 2011 NAEP-TIMSS Linking Study was to predict 2011 TIMSS mathematics and science scores at grade 8 for all U.S. states based on their NAEP performance for states, without incurring the cost associated with every state participating in TIMSS. (For results of the 2011 NAEP-TIMSS Linking Study see http://nces.ed.gov/TIMSS/naeplink.asp.)

The new NAEP and TIMSS components were created to complement each other and made it possible to avoid interference with the respective study's normal operations.

Comparisons with PIRLS. In 2003, NCES released results for both the 2001 Progress in International Reading Literacy Study (PIRLS) fourth-grade assessment and the 2002 NAEP fourth-grade reading assessment. In anticipation of questions about how these two assessments compare, NCES convened an expert panel to compare the content of the PIRLS and NAEP assessments and determine if they are measuring the same construct. This involved a close examination of how PIRLS and NAEP define reading, the texts used as the basis for the assessments, and the reading processes required of students in each. The comparison suggests that there is a great deal of overlap in what the two assessments are measuring. While they do seem to define and measure the same kind of reading, PIRLS is an easier assessment than NAEP, with more text-based tasks and shorter, less complex reading passages. The similarities and differences between the two are discussed below.

The comparison revealed that, overall, the NAEP and PIRLS reading assessments are quite similar. Both define reading similarly, as a constructive process. Both use high-quality reading passages and address similar purposes for which young children read (for literary experience and information). Both call for students to develop interpretations, make connections across text, and evaluate aspects of what they have read. Finally, both have a similar distribution of multiple-choice and constructed-response items: in each, about half of the items are constructed-response items.

While the two assessments have similar definitions of reading and assess many of the same aspects of it, a closer look at how the domain is operationalized by each revealed some important differences. NAEP places more emphasis than PIRLS on having students taking what they have read and connecting it to other readings or knowledge. PIRLS places a greater emphasis than NAEP on text-based reading skills and interactions, including items that ask students to locate information in the text, make text-based inferences and interpretations, and evaluate aspects of the text.

The PIRLS reading passages are, on average, about half the length of the NAEP reading passages. PIRLS readability formulas indicate that the passages used in PIRLS are less complex than those used in NAEP. The classification of items also revealed differences in how the two frameworks function. The panel had an easier time classifying PIRLS and NAEP items by the PIRLS framework categories than by the NAEP framework categories. For more information on the similarities and differences between PIRLS and NAEP, see *A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments* (Binkley and Kelly 2003).

Comparisons with the International Association for the Evaluation of Educational Achievement's (IEA) Reading Literacy Study. The picture of American students' reading proficiency provided by NAEP assessments is less optimistic than that indicated by the IEA Reading Literacy Study. This can be explained by the following:

1. *The basis for reporting differs considerably between the two assessments.* With the IEA study, students are compared against other students and not against a standard set of criteria on knowledge, as in NAEP. Much of NAEP reporting is based on comparisons between actual student performance and desired performance (what students are expected to do).

2. *NAEP and IEA assess different aspects of reading.* More than 90 percent of the IEA items

assess tasks covered in only 17 percent of NAEP items. Furthermore, virtually all of the IEA items are aimed solely at literal comprehension and interpretation, while such items make up only one-third of NAEP reading assessments.

3. *NAEP and IEA differ in what students must do to demonstrate their comprehension.* More interpretive and higher level thinking is required to reach the advanced level in NAEP than in the IEA study. Also, NAEP requires students to generate answers in their own words much more frequently than does the IEA study. Moreover, the IEA test items do not cover the entire expected ability range. Many American students answer every IEA item correctly, making it impossible to distinguish between the abilities of students in the upper range. In contrast, the range of item difficulty on NAEP reading assessments exceeds the ability of most American students, so differences in the abilities of students in the upper range can be distinguished easily.

Despite the differences between these two assessments, there is a high probability that, if students from other countries were to take NAEP, the rank ordering or relative performance of countries would be about the same as in the IEA findings. This assumption is based on the theoretic underpinnings of item response theory and its application to the test scaling used for both the IEA Reading Literacy Study and the NAEP reading assessment.

6. CONTACT INFORMATION

For content information on NAEP, contact:

Linda Hamilton
Phone: (202) 245-6360
E-mail: linda.hamilton@ed.gov

Mailing Address:

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
Potomac Center Plaza
550 12th Street SW
Room 4098
Washington, D.C. 20202

7. METHODOLOGY AND EVALUATION REPORTS

General

Allen, N.L., Carlson, J.E., and Zelenak, C.A. (1999). *The NAEP 1996 Technical Report* (NCES 1999-452). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=1999452

Allen, N.L., Donoghue, J.R., and Schoeps, T.L. (2001). *The NAEP 1998 Technical Report* (NCES 2001-509). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2001509

Allen, N.L., Jenkins, F., Kulick, E., and Zelenak, C. A. (1997). *Technical Report of the NAEP 1996 State Assessment Program in Mathematics* (NCES 97-951). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=97951

Allen, N.L., Kline, D.L., and Zelenak, C.A. (1996). *The NAEP 1994 Technical Report* (NCES 97-897). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=97897

Allen, N.L., Swinton, S.S., Isham, S.P., and Zelenak, C.A. (1998). *Technical Report: NAEP 1996 State Assessment Program in Science* (NCES 98-480). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=98480

Alt, M.N., and Bradby, D. (1999). *Procedures Guide for Transcript Studies* (NCES Working Paper 1999-05). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=199905

Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., and Chan, D. (2015). *School Composition and the Black–White Achievement Gap* (NCES 2015-018). National Center for Education Statistics, U.S. Department of Education, Washington, DC.

http://nces.ed.gov/nationsreportcard/subject/studies/pdf/school_composition_and_the_bw_achievement_gap_2015.pdf

Braun, H., Jenkins, F., and Grigg, W. (2006). *Comparing Private Schools and Public Schools Using Hierarchical Linear Modeling* (NCES 2006-461). U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office. http://nces.ed.gov/nationsreportcard/pubs/studies/2006461.aspx

Calderone, J., King, L.M., and Horkay, N. (1997). *The NAEP Guide: A Description of the Content and Methods of the 1997 and 1998 Assessments* (NCES 97-990). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=97990

Johnson, C. (2004). *Nation's Report Card: An Overview of NAEP* (NCES 2004-552). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2004552

Jonhson, E.G., and Carlson, J.E. (1994). *NAEP 1992 Technical Report* (NCES 94-490). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=94490

Jonhson, E.G., Mazzeo, J., and Kline, D.L. (1994). *Technical Report of the NAEP 1992 Trial State Assessment in Reading* (NCES 94-472). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=94472

Lutkus, A.D., Mazzeo, J., Zhang, J., Jerry, L., Barton, K.E., and Zenisky, A. (2003). *Including Special-Needs Students in the NAEP 1998 Reading Assessment, Part I, Comparison of Overall Results With and Without Accommodations* (NCES 2003-467). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003467

Mazzeo, J., Allen, N.L., and Kline, D.L. (1995). *Technical Report of the NAEP 1994 Trial State Assessment in Reading* (NCES 96-116). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=96116

Moran, R., and Rampey, B. (2008). *National Indian Education Study 2007 - Part II: The Educational Experiences of American Indian and Alaska Native Students in Grades 4 and 8* (NCES 2008-458). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/nationsreportcard/pdf/studies/2008458.pdf

National Center for Education Statistics (2015). *The Nation's Report Card: 2015 Mathematics and Reading Assessments* (NCES 2015-136). Institute of Education Sciences, U.S. Department of Education, Washington, D.C. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2015136

National Center for Education Statistics (2015). *The Nation's Report Card: 2014 U.S. History, Geography, and Civics at Grade 8* (NCES 2015-112). Institute of Education Sciences, U.S. Department of Education, Washington, D.C. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2015112

National Center for Education Statistics (2013a). *The Nation's Report Card: A First Look: 2013 Mathematics and Reading* (NCES 2014-451). Institute of Education Sciences, U.S. Department of Education, Washington, D.C. http://nces.ed.gov/nationsreportcard/subject/publications/main2013/pdf/2014451.pdf

National Center for Education Statistics (2013b). *The Nation's Report Card: A First Look: 2013 Mathematics and Reading Trial Urban District Assessment* (NCES 2014-466). Institute of Education Sciences, U.S. Department of Education, Washington, D.C. http://nces.ed.gov/nationsreportcard/subject/publications/main2013/pdf/2014466.pdf

National Center for Education Statistics (2012). *NAEP: Looking Ahead Leading Assessment into the Future*. http://nces.ed.gov/nationsreportcard/about/future_of_naep.asp

National Center for Education Statistics (2012). *National Indian Education Study 2011* (NCES 2012–466). Institute of Education Sciences, U.S. Department of Education, Washington, D.C. http://nces.ed.gov/nationsreportcard/pdf/studies/2012466.pdf

Patricia, D. (2005). *The Nation's Report Card: An Introduction to The National Assessment of Educational Progress (NAEP)* (NCES 2005-454). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005454

Perie, M., Vanneman, A., and Goldstein, A. (2005). *Student Achievement in Private Schools: Results From NAEP 2000–2005* (NCES 2006-459). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, D.C. http://nces.ed.gov/nationsreportcard/pdf/studies/2006459.pdf

Persky, H., Brent, S.A., and Janice, A. (2003). *Assessing the Arts: Selected NAEP Tasks and Scoring Guides for Grades 4 and 12 1997 Field Test. Dance, Music, Theatre, and Visual Art* (NCES 2003-452). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003452

Phillips, G., and Goldstein, A. (1996). *Technical Issues in Large-Scale Performance Assessment* (NCES 96-802). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs/96802.pdf

Roey, S., Caldwell, N., Rust, K., Blumstein, E., Krenzke, T., and Legum, S. (2001). *The 1998 High School Transcript Study User's Guide and Technical Report* (NCES 2001-477). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2001/2001477.pdf

Roey, S., Caldwell, N., Rust, K., Hicks, L., Lee, J., Perkins, R., Blumstein, E., and Brown, J. (2005). *The High School Transcript Study: The 2000 High School Transcript Study User's Guide and Technical Report* (NCES 2005-483). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2005/2005483.pdf

National Center for Education Statistics (2013c). *The Nation's Report Card: Economics 2012* (NCES 2013 –453). Institute of Education Sciences, U.S. Department of Education, Washington, D.C. http://nces.ed.gov/nationsreportcard/subject/publications/main2012/pdf/2013453.pdf

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2007). *Mapping 2005 State Proficiency Standards Onto the NAEP Scales* (NCES 2007-482). Washington, DC. http://nces.ed.gov/nationsreportcard/pubs/studies/2007482.asp

White, S. (1994). *Overview of NAEP Assessment Frameworks* (NCES 94-412). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=94412

White, S., Kim, Y., Chen, J., and Liu, F. (2015). *Performance of fourth-grade students in the 2012 NAEP computer-based writing pilot assessment: Scores, text length, and use of editing tools* (NCES 2015-119). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/nationsreportcard/subject/writing/pdf/2015119.pdf

Uses of Data

Phillips, G. (1993). *Interpreting NAEP Scales* (NCES 93-421). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=93421

Vanneman, A. (1997). *Focus on NAEP: New Software Makes NAEP Data User Friendly* (NCES 97-045). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=97045

Vanneman, A., Hamilton, L., Baldwin Anderson, J., and Rahman, T. (2009). *Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress*, (NCES 2009-

455). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/nationsreportcard/pdf/studies/2009455.pdf

Survey Design

Bay, L., Chen, L., Hanson, B.A., Happel, J., Kolen, M.J., Miller, T., Pommerich, M., Sconing, J., Wang, T., and Welch, C. (1997). *ACT's NAEP Redesign Project: Assessment Design Is the Key to Useful and Stable Assessment Results* (NCES Working Paper 97-39). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs97/9739.pdf

Bock, D.R., and Zimowski, M.F. (2003). *NAEP Validity Studies: Feasibility Studies of Two-Stage Testing in Large-Scale Educational Assessment: Implications for NAEP* (NCES Working Paper 2003-14). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200314.pdf

Chromy, J.R. (2003). *NAEP Validity Studies: The Effects of Finite Sampling on State Assessment Sample Requirements* (NCES Working Paper 2003-17). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200317.pdf

Durán, R.P. (2003). *NAEP Validity Studies: Implications of Electronic Technology for the NAEP Assessment* (NCES Working Paper 2003-16). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200316.pdf

Hedges, L.V., Konstantopoulos, S., and Thoreson, A. (2003). *NAEP Validity Studies: Computer Use and Its Relation to Academic Achievement in Mathematics, Reading, and Writing* (NCES Working Paper 2003-15). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200315.pdf

Hedges, L.V., and Vevea, J.L. (2003). *NAEP Validity Studies: A Study of Equating in NAEP* (NCES Working Paper 2003-13). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200313.pdf

Jaeger, R.M. (2003). *NAEP Validity Studies: Reporting the Results of the National Assessment of Educational Progress* (NCES Working Paper 2003-11). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200311.pdf

Jakwerth, P.R., Stancavage, R.B., and Reed, E.D. (2003). *NAEP Validity Studies: An Investigation of Why Students Do Not Respond to Questions* (NCES Working Paper 2003-12). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200312.pdf

Johnson, E.G., Lazer, S., and O'Sullivan, C.Y. (1997). *NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress* (NCES Working Paper 97-31). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs97/9731.pdf

Levine, R., Huberman, M., and Buckner, K. (2002). *The Measurement of Instructional Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth- and Eighth-Grade Students and Teachers to Questionnaire Items* (NCES Working Paper 2002-06). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2002/200206.pdf

McLaughlin, D. (1997). *Can State Assessment Data Be Used to Reduce State NAEP Sample Sizes?* (NCES Working Paper 97-29). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs97/9729.pdf

Mullis, V. (2003). *NAEP Validity Studies: Optimizing State NAEP: Issues and Possible Improvements* (NCES Working Paper 2003-09). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200309.pdf

Niemi, R.G. (1997). *Innovative Solutions to Intractable Large-Scale Assessment (Problem 2: Background Questionnaires)* (NCES Working Paper 97-32). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

http://nces.ed.gov/pubs97/9732.pdf

Olson, J.F., and Goldstein, A.A. (1997). *The Inclusion of Students With Disabilities and Limited English Proficient Students in Large-Scale Assessments: A Summary of Recent Progress* (NCES 97-482). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs97/97482.pdf

Owen, E.H. (1992). *Exploring New Methods for Collecting Students' School-Based Writing* (NCES 92-065). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=92065

Patz, R.J., Wilson, M., and Hoskens, M. (1997). *Optimal Rating Procedures and Methodology for NAEP Open-Ended Items* (NCES Working Paper 97-37). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs97/9737.pdf

NCS Pearson, D.P., and Garavaglia, D.R. (2003). *NAEP Validity Studies: Improving the Information Value of Performance Items in Large-Scale Assessments* (NCES Working Paper 2003-08). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200308.pdf

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Stancavage, F.B. (2003) *NAEP Validity Studies: An Agenda for NAEP Validity Research Validity Research* (NCES Working Paper 2003-07). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200307.pdf

Weston, T.J. (2003). *NAEP Validity Studies: The Validity of Oral Accommodation in Testing* (NCES Working Paper 2003-06). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200306.pdf

Data Quality and Comparability

Bandeira de Mello, V., Blankenship, C., and McLaughlin, D.H. (2009). *Mapping State Proficiency Standards Onto NAEP Scales: 2005-2007* (NCES 2010-456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/nationsreportcard/pdf/studies/2010456.pdf

Binkley, M., and Kelly, D. (2003). *A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments* (NCES Working Paper 2003-10). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200310.pdf

DeVito, P.J., and Koenig, J.A. (2001). *NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting.* Washington, DC: National Research Council.

Hoffman, G.R., Becker, D.E., and Wise, L. (2003). *NAEP Quality Assurance Checks of the 2002 Reading Assessment Results of Delaware* (NCES Working Paper 2003-19). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2003/200319.pdf

Ingels, S., and Taylor, J. (1995). *National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data* (NCES Working Paper 95-06). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs95/9506.pdf

Johnson, E.G. (1998). *Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): A Technical Report* (NCES 98-499). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs98/98499.pdf

Johnson, E., Cohen, J., Chen, W.H., Jiang, T., and Zhang, Y. (2005). *2000 NAEP–1999 TIMSS Linking Report* (NCES Working Paper 2005-01). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2005/200501.pdf

Kitmitto, S., and Bandeira de Mello, V., (2008). *Measuring the Status and Change of NAEP State Inclusion Rates for Students with Disabilities* (NCES 2009–453). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

http://nces.ed.gov/nationsreportcard/pubs/studies/2009453.asp

Neidorf, T.S., Binkley, M., Gattis, K., and Nohara, D. (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006-029). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2006/2006029.pdf

Neidorf, T.S., Binkley, M., and Stephens, M. (2006). *Comparing Science Content in the National Assessment of Educational Progress (NAEP) 2000 and Trends in International Mathematics and Science Study (TIMSS) 2003 Assessments* (NCES 2006-026). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubs2006/2006026.pdf

Pellegrino, J.W., Jones, L.R., and Mitchell, K.J. (1999). *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress*. Washington, DC: National Research Council.

Phillips, G. W (2009), *The Second Derivative: International Benchmarks in Mathematics for U.S. States and School Districts*. Washington, DC: American Institutes for Research.

Raju, N.S., Pellegrino, J.W., Bertenthal, M.W., Mitchell, K.J., and Jones, L.R. (2000). *Grading the Nation's Report Card: Research From the Evaluation of NAEP.* Washington, DC: National Research Council.

Sedlacek, D.A. (1995). *Model-Based Methods for Analysis of Data From 1990 NAEP Trial State Assessment* (NCES 95-696). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=95696

Sedlacek, D.A. (1995). *Use of Person-Fit Statistics in Reporting and Analyzing National Assessment of Educational Progress Results* (NCES 95-713). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=95713

Sedlacek, D.A. (1995). *Using HLM and NAEP Data to Explore School Correlates of 1990 Mathematics and Geometry Achievement in Grades 4, 8, 12—Methodology and Results* (NCES 95-697). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=95697

National Center for Education Statistics (2014). *The Nation's Report Card: Mathematics and Reading Grade 12 Assessments*. Institute of Education Sciences, U.S. Department of Education, Washington, D.C. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014087

National Center for Education Statistics (2013d). *The Nation's Report Card: Trends in Academic Progress 2012* (NCES 2013-456). Institute of Education Sciences, U.S. Department of Education, Washington, D.C. http://nces.ed.gov/nationsreportcard/subject/publications/main2012/pdf/2013456.pdf

U.S. Department of Education, National Center for Education Statistics (2008). *Treatment of Missing Responses in NAEP.* http://nces.ed.gov/nationsreportcard/tdw/analysis/2000_2001/scaling_missing.asp.