Chapter 20: National Assessment of Educational Progress (NAEP)

1. OVERVIEW

The National Assessment of Educational Progress (NAEP) is mandated by Congress to assess the educational accomplishments of U.S. students and monitor changes in those accomplishments. As the only nationally representative and continuing assessment of what America's students know and can do in selected subject areas, NAEP serves as the "Nation's Report Card." The *main* NAEP regularly assesses the achievements of students in grades 4, 8, and 12 at the national level. The *state* NAEP assessed at both grades 4 and 8 in at least one subject in 1992, 1996, 1998, 2000, 2002, and 2003. In 2003 and beyond, State NAEP is planning to assess in at least two subjects, reading and mathematics, every 2 years at grades 4 and 8. The *trend* NAEP tracks national long-term trends in science, mathematics, and reading at ages 9, 13, and 17. It tracked writing proficiency trends at grades 4, 8, and 11 through 1999, when critical issues were identified with having so few writing prompts. The national assessments were first implemented in 1969 and were conducted on an annual or biennial basis through 1995, and annually since 1996. The state assessments have been administered biennially since 1990.

In 1988, Congress established the National Assessment Governing Board (NAGB) to provide policy guidance for the execution of NAEP. NAGB is composed of national and local elected officials, chief state school officers, classroom teachers, local school board members, leaders of the business community, and others. Specifically, it is charged by Congress to select subject areas to be assessed; identify appropriate achievement goals for each age group; develop assessment objectives; design the methodology of the assessment; and produce guidelines and standards for national, regional, and state comparisons.

Purpose

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

Components

NAEP comprises three separate assessments: *main national, main state,* and *trend*. Each of these assessments consists of four components: Elementary and Secondary School Students Survey; School Characteristics and Policies Survey; Teacher Survey; and Students with Disabilities or Limited English Proficiency (SD/LEP) Survey (for the *main* NAEP) or Excluded Student Survey (for the *trend* NAEP). In 1985, the Young Adult Literacy Study was also conducted nationally as part of NAEP, under a grant to the Educational Testing Service and Response Analysis Corporation; this study assessed the literacy skills of 21- to 25-year-olds. In addition, a High School Transcript Study is

BIENNIAL SURVEY OF A SAMPLE OF ELEMENTARY/ SECONDARY STUDENTS

Three assessments:

- Main National NAEP
- Main State NAEP
- Trend NAEP

Four component surveys:

- Elementary and Secondary School Students Survey
- School Characteristics and Policies Survey
- Teacher Survey
- SD/LEP Survey / Excluded Student Survey

periodically conducted as a component of NAEP. (See chapter 28.)

In 1996, 1998, and 2000, the national main and state assessments included a special study of the effects of accommodations on the performance of students with special needs. A subsample of students with disabilities or limited English proficiency was given special accommodations (e.g., extended testing time) during the assessment. A comparison subsample took the assessment under standard conditions. Both subsamples met the 1996 criteria for inclusion of special needs students in NAEP.

National-level Assessments. The *main national* NAEP and *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 (4th, 8th, 12th grades), and age/grade samples for the *trend* NAEP (age 9/grade 4; age 13/grade 8; age 17/grade 11). The test instruments for the two assessments are based on different frameworks, 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* NAEP follow the curriculum 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. Recent assessment instruments for the main NAEP have been kept stable for short periods of time, allowing short-term trends to be reported from 1990 through 2003.

To reliably measure change over longer periods of time, the *trend* NAEP must be used. For long-term trends, past procedures must be precisely replicated with each new assessment, and the survey instruments do not evolve with changes in curricula or educational practices. The instruments used today for the trend NAEP are identical to those developed in the mid-1980s. The trend NAEP allows measurement of trends from 1969 to the present.

State-level Assessments. The *main state* NAEP was implemented in 1990 on a trial basis and has been conducted biennially since that time. (The assessments from 1990 to 1994 are referred to as trial state assessments, or TSAs.) Participation of the states was completely voluntary until 2001. The reauthorization of the Elementary and Secondary Education Act, also referred to as the "No Child Left Behind" legislation, requires states that receive Title I funding to participate in state NAEP assessments in reading and mathematics at grades 4 and 8 every 2 years. State participation in other state NAEP subjects (i.e., science and writing) remains voluntary. 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 only in 1994, 1996, and 1998. This practice ended because of low participation rates. See below for the subject areas assessed.

Elementary and Secondary School Students Survey. The primary data collected by NAEP relate to student performance and educational experience as reported by students. Major assessment areas include: reading, writing, mathematics, science, civics, U.S. history, geography, social studies, and the arts.

In 1988, the *main national* NAEP assessed student performance in reading, writing, civics, and U.S. history, and conducted small special-interest assessments in geography and document literacy. In 1990, it assessed mathematics, writing, and science; in 1992, reading, mathematics, and writing; in 1994, reading, U.S. history, and world geography; and in 1996, science and mathematics. A probe of student performance in the arts at grade 8 was conducted in 1997. Reading, writing, and civics were assessed in 1998. (*Trend* NAEP was assessed in 1999.) In 2000, the main national NAEP assessed mathematics and science and, for 4th graders only, reading. In 2001, history and geography were assessed, and in 2002, reading and writing. In 2003, the assessments are in reading and mathematics for 4th and 8th graders.

The subjects assessed in *trend* NAEP are mathematics, science, reading, and until 1999, writing. The biennial assessments from 1988 through 1996 covered all subjects. The next trend assessment will be conducted in 2004 and then trend assessments are scheduled to be administered every 4 years.

Representative *main state-level* data were collected for the first time in the 1990 trial state assessment, when 8th-grade students were assessed in mathematics. In 1992, state-level data were collected in 4th-grade reading and mathematics, and in 8th-grade mathematics. In 1994, 4th-grade reading was assessed. In 1996, 4th-grade mathematics and 8th-grade mathematics and science were assessed. The 1998 NAEP collected state-level data in reading at grades 4 and 8, and writing at grade 8. The 2000 NAEP assessments covered mathematics and science, the 2002 assessments covered reading and writing, and the 2003 assessments cover reading and mathematics.

The student survey also asks questions about the student's background, as well as questions related to the subject area and the student's motivation in completing the assessment. *Student background questions* gather information about race/ethnicity, school attendance, academic expectations, and factors believed to influence academic performance, such as homework habits, the language spoken in the home, and the quantity of reading materials in the home. Some of these questions document changes that occur over time, and remain unchanged over assessment years.

Student subject-area questions gather three categories of information: time spent studying the subject, instructional experiences in the subject, and perceptions about the subject. Because these questions are specific to each subject area, they can probe in some detail the use of specialized resources such as calculators in mathematics classes.

*S*tudents are also asked how often they have been asked to write long answers to questions on tests or assignments that involved (this subject). In earlier assessments, students were also asked how many questions they thought they answered correctly, how difficult they found the assessment, how hard they tried on this test compared to how hard they had tried on most other tests or assignments they had taken that year in school, and how important it was to them to do well on this test. (In 2003, NAEP dropped the motivation questions.)

School Characteristics and Policies Survey. This survey collects supplemental data about school characteristics and school policies that can be used analytically to provide context for student performance issues. School data include: enrollment, absenteeism, dropout rates, curricula, testing practices, length of school day and year, school administrative practices, school conditions and facilities, size and composition of teaching staff, tracking policies, schoolwide programs and problems, availability of resources, policies for parental involvement, special services, and community services.

Teacher Survey. This survey collects supplemental data from teachers whose students are respondents to the assessment surveys. Part I of the Teacher Questionnaire covers background and general training, requesting information on the teacher's race/ethnicity, sex, age, years of teaching experience, certification, degrees, major and minor fields of study, coursework in education, coursework in specific subject areas, amount of in-service training, extent of control over instructional issues, and availability of resources for the classroom. Part II of the Teacher Questionnaire covers training in the subject area and classroom instructional practices, specifically the teacher's exposure to issues related to the subject and the teaching of the subject, pre- and in-service training, ability level of the students in the class, length of homework assignments, use of particular resources, and how students are assigned to particular classes.

SD/LEP Survey. This survey is completed in the *main* NAEP assessments by teachers of students selected to participate in NAEP but classified as having disabilities (SD) or classified as limited English proficient (LEP). Information is collected on the background and characteristics of each SD/LEP student and the reason for the SD/LEP classification, as well as whether these students receive accommodations in district or statewide tests. For SD students, questions ask about the student's functional grade levels and special education programs. For LEP students, questions ask about the student's native language, time spent in special language programs, and the 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, NAEP has allowed accommodations for both SD and LEP students.

Excluded Student Survey. This survey is completed in the *trend* NAEP for students who are sampled for the assessment but excluded by the school. Following exclusion criteria used in previous trend assessments, a school can exclude students with limited English-speaking ability, students who are educable mentally retarded, and students who are functionally disabled—if the school judges that these students are unable to "participate meaningfully" in the assessment. This survey is only completed for those students who are actually excluded from the assessment (whereas the SD/LEP Survey in the main assessment is also completed for participating students who are SD or LEP students—see above).

High School Transcript Study. Transcript studies have been conducted in 1987, 1990, 1994, 1998, and 2000. The studies collect information on current course offerings and course-taking patterns in the nation's schools. 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 from grade 12 students in selected schools from the NAEP sample. (For more information, see chapter 28, Other NCES Surveys and Studies.)

Special Studies. The 1998 assessment included three subsamples that used special procedures to study specific aspects of writing and civics. The special studies samples were drawn from the grade-only population. The three special studies consisted of: (1) Writing – 50: a sample of students in grades 8 and 12 who received 50-minute writing blocks in assessments sessions where no other writing format was administered; (2) Writing – Classroom: a sample of students in grades 4 and 8 who were assessed based on written assignments the students had completed as part of their regular school curriculum; and (3) Civics – Special Trend: a sample of students in grades 4, 8, and 12 who were assessed using the booklets and testing conditions used in the 1988 civics assessment.

Oral Reading Study Assessment. In 2002, NAEP conducted a special study on oral reading. The NAEP 2002 Oral Reading Study looked at how well the nation's 4th graders can read aloud a grade-appropriate story. NAEP assessed a random sample of 4th-grade students selected for the NAEP 2002 reading and writing assessments. 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 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, which is already operational. The primary goals of Mathematics Online (MOL) are to understand how computer delivery affects the measurement of NAEP math skills, to gain insights into the operational and logistical mechanics of computer-delivered assessments, and to evaluate the ability of 4th and 8th graders to deal with mathematics assessments delivered on computer. At grade 8, an additional goal is to investigate the technical feasibility of generating alternate versions of multiple-choice and constructed-response items using an "on-the-fly" (OTF) technology. MOL was field tested in 2002. The Writing Online (WOL) study is intended to help NAEP learn how computer delivery affects the measurement of NAEP performance-based writing skills, to gain insights into the operational and logistical mechanics of computer-delivered writing assessments, and to evaluate the ability of 8th graders to deal with writing assessments delivered on computer. WOL was field tested in 2002. The Problem Solving in Technology-Rich Environment (TRE) study was designed to develop an example set of modules to assess problem solving using technology. These example modules will use the computer to present multimedia tasks that cannot be delivered through conventional paper-and-pencil assessments, but which tap important emerging skills. TRE is being field tested in 2003.

Periodicity

Annual from 1969 to 1979 and, again, beginning in 1996; biennial in even-numbered years from 1980 to 1998. A probe of 8th graders in the arts area was conducted in 1997. State-level assessments, first initiated in 1990, follow the same schedule as the national assessments. Prior to 1990, NAEP was required to assess reading, mathematics, and writing at least once every 5 years. The previous legislation required assessments in reading and mathematics at least every 2 years, in science and writing at least every 4 years, and in history or geography and other subjects selected by the National Assessment Governing Board at least every 6 years. The No Child Left Behind Act requires NAEP to conduct national and state assessments at least once every 2 years in reading and mathematics in grades 4 and 8. In addition, in the future, NAEP will conduct a national assessment and may conduct a state assessment in reading and mathematics in grade 12 every 4 years starting in 2005. Finally, to the extent that time and money allow, NAEP will be conducted in grades 4, 8, and 12 at regularly scheduled intervals in additional subjects including writing, science, history, geography, civics, economics, foreign languages, and arts.

2. USES OF DATA

NAEP serves as the Nation's Report Card. It is the only ongoing, comparable, and representative assessment of what American students know and can do in several 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 data, available for many states since 1990, allow both state-to-state comparisons and comparisons of individual states with the nation as a whole.

During NAEP's history, more than 200 reports across 12 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?
- Education reform: What policy changes are being made by our nation's schools?

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 educational 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 releases nearly one-third of the questions to the interested public. 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 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 three levels are:

Basic. Partial mastery of 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. Superior performance. This level is only attained by a very small percentage of students (3–6 percent) at any of the three grade levels 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 *trend* assessments in science, mathematics, and reading. Grades 4 and/or 8 are usually assessed in the *state* NAEP; the number of grades has varied in the past, depending on 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 prior to 1994 and after 1998.

Sample Design

The sample for each NAEP assessment is selected using a complex multistage clustered design involving the sampling of students from selected schools within selected geographic areas, called primary sampling units (PSUs), across the United States. The sample designs for NAEP assessments have been similar since the mid-1980s. In 1983, student samples were expanded to include both age- and grade-representative populations. Since 1988, the samples have been drawn from the universe of 4th, 8th, and 12th graders for the Elementary and Secondary School Students Survey; from the teachers of those students for the Teacher Survey; and from the school administrators at those elementary and secondary schools for the School Characteristics and Policies Survey. In 1996, SD/LEP students were oversampled for a special study of SD/LEP inclusion; hence, exclusion rules and availability of accommodations were different than in previous studies. The national-level sample for each NAEP assessment contains approximately 7,000 to 10,000 students for each grade assessed—or 0.42 percent of the national student population for each grade.

NAEP's multistage sampling process involves the following steps:

(1) Selection of PSUs

(2) Selection of schools (public and nonpublic) within the selected PSUs

(3) Assignment of session types to schools

(4) Selection of students for session types within the selected schools

In 1996, the special study of SD/LEP inclusion required an additional step for the main assessments: the assignment of "sample types" to schools based on specific criteria for excluding students with limited English proficiency or severe disability, and the provision or nonprovision of accommodations. Results from this study indicated that revising the criteria for including students had little impact on the numbers of students included. Because of the lack of impact, the revised criteria for including students will be used in future assessments. Provision of accommodations was found to have a limited impact on performance results. NAEP made a full transition to providing allowable accommodations to all students who need them in 2002.

Selection of PSUs. In the first stage of sampling, the United States (the 50 states and the District of Columbia) is divided into geographic PSUs. The PSUs are classified into four regions (Northeast, Southeast, Central, and West), each containing about one-fourth of the U.S. population. In each region, PSUs are additionally classified as metropolitan or nonmetropolitan, resulting in eight subuniverses of PSUs.

For the 1998 main assessment, 94 PSUs were selected; 22 of these PSUs were designated as certainty units because of their size. Within each major stratum (subuniverse), further stratification was achieved by ordering the noncertainty PSUs according to several additional socioeconomic characteristics (e.g., median household income, educational level of residents over 25 years of age, demographic characteristics). One PSU was selected from each of the 72 noncertainty strata, with probability proportional to size (total population from the 1990 census). To enlarge the samples of Black and Hispanic students, thereby enhancing the reliability of estimates for these groups, PSUs from the high-minority strata were sampled at twice the rate of PSUs from the other strata. This was achieved by creating smaller strata with high-minority subuniverses.

There were no long-term trend NAEP samples in 1998; however, in 1996, when 94 PSUs were selected for the main assessment, 52 PSUs were selected for the long-term trend samples. Of these 52 trend PSUs, 10 selected with certainty because of their size, 6 were selected from the 12 remaining main sample certainty PSUs, and 36 were selected from the 72 noncertainty strata independently of the main sample selection.

Selection of schools. In the second stage of sampling, public schools (including Bureau of Indian Affairs—BIA—schools and Department of Defense Education Activity—DODEA—schools) and nonpublic schools (including Catholic schools) within each of the selected PSUs are listed according to the grades associated with the three age classes: age class 9 refers to age 9 or grade 4 in the trend NAEP or grade 4 in the main NAEP; age class 13 refers to age 13 or grade 8 in the trend NAEP or grade 8 in the main NAEP; age class 17 refers to age 17 or grade 11 in the trend NAEP or grade 12 in the main NAEP.

The school lists are obtained from two sources. Regular public, BIA, and DODEA schools are obtained from the school list maintained by Quality Education Data, Inc. (QED). Catholic and other nonpublic schools are obtained from the NCES Private School Survey. (See chapter 3.) To ensure that the state samples provide an accurate representation, public schools are stratified by urbanization, minority enrollment, and median household income. Nonpublic schools are stratified by type of control (e.g., parochial, private), urban status, and enrollment per grade. Once the stratification is completed, the schools within each PSU are assigned a probability of selection that is proportional to the number of students per grade in each school.

An independent sample of schools is selected separately for each age/grade so that some schools are selected for assessment of two age/grades and a few are selected for all three. Schools within each PSU are selected (without replacement) with probabilities proportional to assigned measures of size. Nonpublic schools and schools with high minority enrollment are oversampled.

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 nonpublic schools and schools with high minority enrollment are not oversampled. Schools are not selected for both main and long-term trend assessments at the same age/grade.

Assigning sample type to schools. As noted earlier, schools in the 1996 *main* assessments were assigned a "sample type" based on specific criteria for excluding students, with the goal of determining the effect of different exclusion criteria in NAEP assessments. Historically, a small proportion (less than 10 percent) of the sampled students have been excluded from NAEP assessments because they are SD/LEP students whom their local schools determined could not take the assessments. In recent years, increased attention has been given to including as many of these students as possible in NAEP assessments.

Three different sample types were assigned to the schools selected for the 1996 main assessment. For sample type 1 schools, the exclusion criteria for the main samples were identical to those used in 1990 and 1992. Sample type 2 schools used new inclusion criteria for SD and LEP students. In sample type 3 schools, the new inclusion criteria were used and, in addition, accommodations were offered to SD and LEP students. The specific criteria and availability of accommodations varied among the schools. The most frequently provided accommodations were small group administration, extended time (untimed testing), and, in mathematics, bilingual assessment booklets. Sample type was assigned separately for each grade.

In the 1998 national main and state reading assessments, sample types 2 and 3 were assigned to schools. The writing and civics assessments were administered to sample type 3 schools only.

Assignment of session types to schools. In the third stage of sampling, assessment sessions are assigned to the selected schools found to be in-scope, with three aims in mind. The first is to distribute students to the different session types (e.g., assessment in a particular academic subject or pilot test of new items) across the whole sample for each age class so that the target numbers of assessed students will be achieved. The second is to maximize the number of different session types that are administered within a given selected school without violating minimum session sizes. The third is to give each student an equal chance of being selected for a given session type regardless of the number of sessions conducted in the school. Beginning in 2002, for the main assessment, session types were no longer assigned to schools; rather, sessions all had a common session design so that multiple subjects can be spiraled across students.

Selection of students. The fourth 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 (grades 4, 8, and 11 for the writing assessment). The selection process differs slightly based on whether the sample of students is needed for the main national assessment, the long-term trend assessment, or the main state assessment. 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 grade-eligible and 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 example, to oversample Black and Hispanic students from public schools with low minority enrollment, as was done in 1998, after the initial sample was selected, the nonselected Black and Hispanic students were identified and listed. If the number of nonselected students was less than the number of selected students, then all nonselected Black and Hispanic students were assessed. Otherwise, Black and Hispanic students were sampled so that their overall within-school probability of selection was twice the rate of other students. Likewise in 1998, in each school where oversampling of SD/LEP students was to occur, the initial desired sample of students was drawn for each session assigned from the full list of eligible students. Among those students not selected for either of the two prior sampling operations for that school, the SD/LEP students were identified. A sample from among these was drawn, using a sampling rate that would achieve the double sampling rate required overall.

For schools assigned more than a single session type, which is the vast majority of schools, students are assigned to one of the various session types using specified procedures.

For each age class (separately for long-term trend and main samples), maxima are established as to the number of students who are to be selected for a given school. In those schools that, according to information on the sampling frame, have fewer eligible students than the established maxima, each eligible student enrolled at the school is selected in the sample for one of the sessions assigned to the school. In other schools, a sample of students is drawn and students are assigned to sessions as appropriate. No student is assigned to more than one session. The maximum sample sizes are established in terms of the number of grade-eligible students (by sample type in 1996) for the main samples, and in terms of the number of students in each age class for the trend samples.

The classroom-based writing study involved the random selection of one English/language arts classroom from each 4th- and 8th-grade school in which a writing assessment was to be conducted. At the same time, the students in that classroom were listed on a writing study linkage form so that the classroom students who also took the national writing assessment could be identified. The classroom's English/language arts teacher was asked to work with the students and have them select two examples of their best classroom writing. The students were asked to answer a few questions about each selection. The teachers completed an interview with the supervisor who collected the writing materials after the assessment.

Excluded students. Some students are excluded from the student sample because they are deemed unassessable 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 that should be excluded from the *main* assessments, school staff members are asked to identify those SD or LEP students who do not meet the NAEP inclusion criteria. School personnel are asked to complete an SD/LEP questionnaire for all SD and LEP students selected into the NAEP sample, whether they participate in the assessment or not. For the long-term *trend* assessments, excluded students are identified for each age class, and an Excluded Student Survey is completed for each excluded student.

For the special study of SD/LEP inclusion in the 1996 main assessment, oversampling procedures were applied to SD/LEP students at all three grades in sample types 2 and 3 for mathematics and in sample type 3 for science.

Main national and state NAEP sample sizes. Not all subject areas are assessed in every assessment year. In 1998, the main national NAEP assessed students in reading, writing, and civics at all three grades. The main state NAEP in 1998 assessed students in writing at grade 8 and in reading at grades 4 and 8. The total target sample size for the 1998 state assessments was 396,000 (132,000 for each grade and subject). The sample included students from an average of 225 schools per state. For the main national NAEP, the total target sample size was 132,000 students from 2,000 schools nationwide. Sample sizes by grade ranged from 8,000 to 13,000 in reading; from 20,000 to 26,000 in writing; and from 6,000 to 8,000 in civics. A separate civics trend sample included 2,000 students from each grade.

In comparison, the 1996 main national assessment, which tested mathematics and science at all three grade levels, required fewer than 100,000 students from about 1,800 schools. The state-level assessment, which tested only two grade levels, required a total sample of about 350,000 students from nearly 10,000 schools because of the number of states that participated.

Long-term trend NAEP sample sizes. The long-term trend assessment tested the same four subjects 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. The next long-term trend assessment will be administered in 2004, and then every 4 years thereafter (but not in the same years as the main assessments) in reading and mathematics.

Assessment Design

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

Development of framework and questions. NAGB 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 through a consensus process involving teachers, curriculum specialists, subject-matter specialists, school administrators, parents, and members of the general public.

Unlike earlier multiple-choice instruments, current instruments dedicate a majority of testing time to constructed-response questions that require students to compose written answers. Constructed-response questions provide a separate means of assessing ability, tapping recall not recognition.

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. For each state-level assessment, state curriculum and testing directors review the questions that will be included in the NAEP state component.

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 (e.g., 50 minutes to an hour and a half). By this method, different portions from the entire pool of cognitive questions are printed in separate booklets and administered to different but equivalent samples of students. About 2,600 students respond to each block of items.

The type of matrix sampling used by NAEP is called focused, balanced incomplete block (BIB) spiraling. The NAEP BIB design varies according to subject area.

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 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.

Reference dates. Data for the main national NAEP and main state NAEP are collected at overlapping times during winter. Data for the long-term trend NAEP are collected during fall for age 13/grade 8; during winter of the same school year for age 9/grade 4; and during spring for age 17/grade 11.

Data collection. Until 2002, NCES relied heavily on school administrators for the conduct of main state NAEP assessments. Beginning with the 2002 assessments, NAEP contract staff conduct all 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 in-person meetings where procedures are explained.

The questionnaires for the School Characteristics and Policies Survey, the Teacher Survey, and the SD/LEP Survey are sent to the schools ahead of the assessment date so that they can be collected when the assessment is administered. Questionnaires not ready at this time are retrieved later, either through a return visit by NAEP personnel or through the mail.

NCS Pearson produces the materials needed for NAEP assessments. NCS Pearson prints identifying bar codes and numbers for the booklets and questionnaires, preassigns 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 spiraled distribution process.

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 science and mathematics questions were administered using a paced audiotape. Beginning in 2004, the long-term trend assessments will be 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. Starting with the 2002 main national and state assessments, Westat has employed and trained about 3,000 field staff to carry out the assessments.

Westat ensures quality control across states by monitoring 25 percent of the sessions. Security of assessment materials and uniformity of administration are high priorities. (To date, there have been no reports from quality control monitors of serious breaches in procedures or major problems that could jeopardize the validity of the assessment.) 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.

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 scan 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. 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 and NCS Pearson develop 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 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 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 main state NAEP). In addition, the image 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 reliability of scoring is monitored during the coding process through (1) backreading, where table leaders review about 10 percent 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 25 percent of the items appearing only in the main national assessment and 6 percent of the items appearing in both the main national and state assessments, and a comparison of the two scores to give a measure of interscorer reliability. To monitor agreement across years, a random sample of 20–25 percent 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 reading, writing, and civics items from 1994 were rescored in 1998. See the table below.

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

Table 9. Sample score ranges and percent agreements for constructed-response reading items that were scored twice

	Dichotomously scored items		Polytomously scored items	
	Cohen Kappa	Percent agreement	Intraclass correlation	Percent agreement
1998 national main assessment reading items				
4th grade	0.705–0.970	87–98	0.821–0.957	78–91
8th grade	0.665–0.996	84–100	0.761–0.977	64–98
12th grade	0.596–0.967	83–100	0.668–0.992	66–97
1994 reading items rescored in 1998				
4th grade	0.722 to 0.944	86–96	0.855 to 0.968	78–92
8th grade	0.678 to 0.983	83–99	0.798 to 0.978	64–96
12th grade	0.535 to 0.952	76–98	0.698 to 0.974	62–95

SOURCE: Derived from tables in appendix C, Allen, Donoghue, and Schoeps, *The NAEP 1998 Technical Report* (NCES 2001–509).

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 data errors detected in the first phase of editing, compare the processed data with the original source document, and indicate whether the error is correctable or noncorrectable per the editing specifications. Suspect errors 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 errors are edited online. The edit criteria for each item in question appear on the screen along with the suspect item, and corrections are made immediately. Two different people view the same suspect data and operate on it separately, and 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 on 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 one and 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, the responses are weighted according to the sample design and population structure and then adjusted for nonresponse. This ensures that the students' representation in NAEP matches their actual proportion of the school population in the grades assessed. 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 student achievement (proficiency) in the various subject areas. The marginal maximum likelihood methodology is then used to estimate characteristics of the proficiency distributions. Estimates of cognitive ability are included in the NAEP database. Estimates of other variables are not included in the database.

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 student's responses is the inverse of the probability that the student would be selected for the sample. Through poststratification, the weighting ensures that the representation of certain subpopulations correspond 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 was selected for a particular assessment. This probability is the product of the following four factors:

- the probability that the PSU was selected;
- the conditional probability that the school was selected, given the PSU;
- the conditional probability, given the selected samples of schools in the PSU, that the school was allocated the specified assessment; and
- the conditional probability, given the school, that the student was selected for the assessment.

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 PSU strata. 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 to improve the stability of the adjustment factors.

Scaling. For purposes of summarizing item responses, ETS developed a scaling technique that has its roots in Item Response Theory (IRT) and the theories of imputation of missing data.

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 subsequent to the last question the student answered) and missing responses prior to the last observed response. Missing responses before the last observed response are considered intentional omissions. Missing responses at the end of the 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. Separate subscales are derived for each subject area. For the main assessments, the frameworks for the different subject areas dictate the number of subscales required. In the 1996 NAEP, five subscales were created for the main assessment in mathematics (one for each mathematics content strand), and three subscales were created for science (one for each field of science: earth, physical, and life). 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 separate scale is used for summarizing proficiencies at each age/grade level in each of the subject areas—science, mathematics, reading, and writing.

Within-grade vs. cross-grade scaling. 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-to-500 scale was established for all three grades in each assessment. In 1993, NAGB determined that future NAEP assessments should be developed using within-grade frameworks and be scaled accordingly. This both removes the constraint that the trait being measured is cumulative and eliminates 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 1996 long-term trend assessments converted to within-grade, using a 0 to 500 scale. The 1996 main science assessment was also developed within-grade, but adopted new scales ranging from 0 to 300. The 1996 main assessment in mathematics continued to use a cross-grade framework with a 0 to 500 scale. In 1998, reading assessments were scaled across grades, and writing and civics were scaled within-grade.

Linking of scales. Until 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 depends on those parts of the main national and state samples that represent 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. Beginning in 2002, the national sample is a subset of the state samples (except in those states that do not participate). Thus no equating is necessary.

Imputation. Up until NAEP's 2002 assessment, no statistical imputations have been generated for missing values in the teacher, school, or SD/LEP questionnaires, not 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 out of 110 questions in the assessment. Whether any given student got 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.

Because no student takes even a quarter of an assessment, NAEP does not—and cannot—calculate individual scores. 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 groups scores due to an inherently large component of measurement error. NAEP developed its 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 marginal maximum likelihood (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 not of scores for individuals.

NAEP conducts a special form of imputation during the third stage of its analysis procedures. The first stage requires estimating item response theory 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/LEP questionnaires, as well as geographical, sample frame, and school record information. The third stage involves calculating imputations 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. The five plausible values are calculated using the regression coefficients estimated in the second stage. 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 were implemented since 1990. For more detail, refer to earlier sections of this chapter.

- 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 (GPC) model was introduced to develop scales for the more complex constructed-response questions. The GPC 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, the 1996 trend assessments, 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).
- 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.

- In 2001, NAEP fixed the history and geography scales to have within grade scales, with mean of 150, like civics, science, and writing.
- With the expansion 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 is administering 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 is conducting pilot tests of candidate items for the next assessment 2 years in advance and field tests of items for precalibration 1 year in advance of data collection, thereby speeding up the scaling process. Third, NAEP is conducting bridge studies, administering tests both under the 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 with the 2002 assessments, a combined sample of public schools was selected for both state and national NAEP. Therefore, the national sample is a subset of the combined sample of students assessed in each participating state, plus an additional sample from the states that did not participate in the state assessment. This additional sample ensures that the national sample is representative of the total national student population.
- Beginning with the 2003 NAEP, each state must have participation from at least 85 percent—instead of from 70 percent—of the schools in the original sample in order to have results reported.

Future Plans

The next trend assessment will be administered in 2004, and then every 4 years thereafter. For the 21st century, NAEP is undergoing a full-scale redesign, and its assessment schedule is being placed on a more regular, predictable timetable. Main assessments are planned for annual administration (instead of every 2 years). Reading and mathematics will be assessed every 2 years in odd-numbered years; science and writing are planned to be assessed every 4 years (in the same years as reading and mathematics, but alternating with each other); and other subjects will be assessed at the national level in even-numbered years.

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., minority students or students attending nonpublic schools). Although only a very small percentage of the student population in each grade is assessed, NAEP estimates are accurate because they depend on the absolute number of students participating, not on the relative proportion of students.

Every activity in NAEP assessments is conducted with rigorous quality control, contributing to both 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 the 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 could result from either 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 the 1998 NAEP, the 1997 school list maintained by QED supplied the names of the regular public schools, Bureau of Indian Affairs schools, and DODEA schools. This list, however, did not include schools that opened between 1997 and the time of the 1998 NAEP. To be sure 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 were obtained from the NCES Private School Survey (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 PSS is very high—estimated at 96.5 percent for the 1995–96 PSS, which was used for the 1998 NAEP. (See chapter 3, section 5.) Prior to the 1996 NAEP, nonpublic schools were also obtained from telephone directories. This process was not repeated in 1996 because the PSS frame adequately supported the QED list.

Nonresponse error.

Unit nonresponse. For both the main NAEP and the trend NAEP, school response rates have generally declined over the years while student response rates have risen. The level of student participation has been consistently lower with each increment in student age and grade. At every age/grade level, the participation of students from nonpublic schools has exceeded that of students from public schools.

For the *main national assessments* in 1998, the unweighted school response rate across grades and subjects was 86 percent (after substitution). This reversed the small declines in national assessment school response rates that occurred between 1990 and 1996. The gains were most likely due to persistent efforts to convert refusals. Between 1990 and 1996, there was a small but steady decline in school response rates despite persistent efforts to convert uninterested schools and districts: from 88.3 to 85.8 percent at grade 4; from 86.7 to 81.9 percent at grade 8; and from 81.3 to 78.7 percent at grade 12. The reason most often given for school nonparticipation is the increase in required testing throughout the jurisdictions and the resulting difficulty in finding time to also conduct NAEP assessments.

Table 10, on the next page, provides weighted response rates for selected NAEP surveys.

Item nonresponse. Specific information about nonresponse for a particular item is available on NAEP summary data tables on the web.

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 the students. For example, the instruments may contain ambiguous definitions and/or questions that lead to different interpretations by the 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 the 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 2000 NAEP database was found to be more than accurate enough to support analyses. The observed error rates for the 2000 NAEP were comparable to those of past assessments. Error rates ranged from 8 errors per 10,000 responses for the Teacher Questionnaire to 44 errors per 10,000 responses for the School Characteristics and Policies Questionnaire.

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

Table 10. Weighted response rates for selected NAEP national (main sample) surveys

		School participation*	Student participation	Overall participation
1994 Reading	– age class 9	86.1	93.5	80.5
	– age class 13	82.9	91.1	75.5
	– age class 17	76.3	81.9	62.5
1996 Mathematics	– grade 4	82.3	95.3	78.4
	– grade 8	81.5	92.9	75.7
	– grade 12	76.2	82.3	62.7
1998 Reading	– grade 4	81.0	96.0	77.8
	– grade 8	76.7	92.7	71.1
	– grade 12	69.7	80.1	55.8

*Rates do not include substitutions.
SOURCE: Allen, Carlson, and Zelenak, *The NAEP 1996 Technical Report* (NCES 1999–452). Allen, Donoghue, and Schoeps, *The NAEP 1998 Technical Report* (NCES 2001–509). Allen, Kline, and Zelenak, *The NAEP 1994 Technical Report* (NCES 97–897).

mathematics assessments. These errors affected the mathematics scale scores reported in 1992 and the achievement level results reported in 1990 and 1992. NCES and NAGB evaluated the impact of these errors and subsequently reanalyzed 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 *NAEP 1996 Technical Report* (NCES 1999–452) and *NAEP 1996 Technical Report of the State Assessment Program in Mathematics* (NCES 97–951).

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* (NCES 97–897) and *Technical Report of the NAEP 1994 Trial State Assessment in Reading* (NCES 96–116).

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, beginning in 2002, they are 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.

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 have remained unchanged from those used in previous assessments. The 1996 trend instruments were identical to those used in the mid-1980s. Through implementation of additional procedures, the current year's data can be linked to even earlier years. The trend NAEP allows the measurement of trends back to 1969, the year of inception. For more detail on the linking of scales in the trend NAEP, refer to section 4, Scaling. The 2004 long-term trend NAEP is undergoing redesign. Bridge studies are planned to make the 2004 assessment comparable to earlier assessments.

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 IAEP. In 1992, results from the 1992 NAEP assessments in mathematics were successfully linked to those from the International Assessment of Educational

Progress (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 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 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 link, the results should be used cautiously.*

Comparisons with National Adult Literacy Survey (NALS). NAEP data can also be compared with results of NALS. The term “succeed consistently,” as it relates to literacy, means that a person at or above a given level of literacy has a certain percentage of a chance of correctly responding to a particular task. The criterion for the NAEP standard (65 percent) is less stringent than the NALS criterion (80 percent). Thus, if the NALS criterion were used for NAEP assessments, the proportions in the lower literacy levels would increase and the proportions in the higher levels would decrease. (See chapter 23 for a description of the NALS.)

Comparisons with IEA Reading Literacy Study. The picture of American students’ reading proficiency provided by NAEP assessments is less optimistic than that indicated by the International Association for the Evaluation of Educational Achievement’s (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, 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 they 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. Further, 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. Also, NAEP requires students to generate answers in their own words much more frequently than does the IEA. 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 abilities of students in the upper range. In contrast, the range of item difficulty on NAEP reading assessment 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:

Peggy Carr
Phone: (202) 502–7321
E-mail: peggy.carr@ed.gov

Steven Gorman
Phone: (202) 502–7347
E-mail: steven.gorman@ed.gov

Mailing Address:

National Center for Education Statistics
1990 K Street NW
Washington, DC 20006–5651

7. METHODOLOGY AND EVALUATION REPORTS

General

The NAEP Guide: A Description of the Content and Methods of the 1997 and 1998 Assessments, NCES 97–990, by J. Calderone, L.M. King, and N. Horkay, eds. Washington, DC: 1997.

The NAEP 1998 Technical Report, NCES 2001–509, by N.L. Allen, J.R. Donoghue, and T.L. Schoeps. Washington, DC: 2001.

The NAEP 1994 Technical Report, NCES 97–897, by N.L. Allen, D.L. Kline, and C.A. Zelenak. Washington, DC: 1996.

The NAEP 1996 Technical Report, NCES 1999–452, by N.L. Allen, J.E. Carlson, and C.A. Zelenak. Washington, DC: 1999.

NAEP 1992 Technical Report, NCES 94–490, by E.G. Johnson and J.E. Carlson. Washington, DC: 1994.

The 1998 High School Transcript Study User's Guide and Technical Report, NCES 2001–477, by S. Roey, N. Caldwell, K. Rust, E. Blumstein, T. Krenzke, S. Legum, J. Kuhn, M. Waksberg. Washington, DC: 2001.

Overview of NAEP Assessment Frameworks, NCES 94–412, by S. White. Washington, DC: 1994.

Procedures Guide for Transcript Studies, NCES Working Paper 1999–05, by M.N. Alt and D. Bradby. Washington, DC: 1999.

Technical Issues in Large-Scale Performance Assessment, NCES 96–802, by G. Phillips and A. Goldstein. Washington, DC: 1996.

Technical Report of the NAEP 1996 State Assessment Program in Mathematics, NCES 97–951, by N.L. Allen, F. Jenkins, E. Kulick, and C.A. Zelenak. Washington, DC: 1997.

Technical Report of the NAEP 1992 Trial State Assessment in Reading, NCES 94–472, by E.G. Johnson, J. Mazzeo, and D.L. Kline. Washington, DC: 1994.

Technical Report of the NAEP 1994 Trial State Assessment in Reading, NCES 96–116, by J. Mazzeo, N.L. Allen, and D.L. Kline. Washington, DC: 1995.

Technical Report: NAEP 1996 State Assessment Program in Science, NCES 98–480, by N.L. Allen, S.S. Swinton, S.P. Isham, and C.A. Zelenak. Washington, DC: 1998.

Uses of Data

Focus on NAEP: New Software Makes NAEP Data User-Friendly, NCES 97–045, by A. Vanneman. Washington, DC: 1997.

Interpreting NAEP Scales, NCES 93–421, by G. Phillips. Washington, DC: 1993.

Survey Design

ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results, NCES Working Paper 97–39, by L. Bay, L. Chen, B.A. Hanson, J. Happel, M.J. Kolen, T. Miller, M. Pommerich, J. Sconing, T. Wang, and C. Welch. Washington, DC: 1997.

Can State Assessment Data Be Used to Reduce State NAEP Sample Sizes?, NCES Working Paper 97–29, by D. McLaughlin. Washington, DC: 1997.

Exploring New Methods for Collecting Students' School-Based Writing, NCES 92–065, by E.H. Owen. Washington, DC: 1992.

The Inclusion of Students with Disabilities and Limited English Proficient Students in Large-Scale Assessments: A Summary of Recent Progress, NCES 97–482, by J.F. Olson and A.A. Goldstein. Washington, DC: 1997.

Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questionnaires), NCES Working Paper 97–32, by R.G. Niemi. Washington, DC: 1997.

Multiple Imputation for Nonresponse in Surveys, by D.B. Rubin. New York: John Wiley & Sons, 1987.

NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress, NCES Working Paper 97–31, by E.G. Johnson, S. Lazer, and C.Y. O'Sullivan. Washington, DC: 1997.

Optimal Rating Procedures and Methodology for NAEP Open-ended Items, NCES Working Paper 97–37, by R.J. Patz, M. Wilson, and M. Hoskens. Washington, DC: 1997.

Data Quality and Comparability

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

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

Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): A Technical Report, NCES 98–499, by E.G. Johnson. Washington, DC: 1998.

Model-Based Methods for Analysis of Data from 1990 NAEP Trial State Assessment, NCES 95–696, by D.A. Sedlacek. Washington, DC: 1995.

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

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, by S. Ingels and J. Taylor. Washington, DC: 1995.

Use of Person-Fit Statistics in Reporting and Analyzing National Assessment of Educational Progress Results, NCES 95–713, by D.A. Sedlacek. Washington, DC: 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, by D.A. Sedlacek. Washington, DC: 1995.

Chapter 21: Third International Mathematics and Science Study (TIMSS)

1. OVERVIEW

The Third International Mathematics and Science Study (TIMSS), sponsored by the International Association for the Evaluation of Educational Achievement (IEA), is a study of classrooms across the country and around the world. A half million students from 41 countries were tested in 30 different languages at five different grade levels to compare their mathematics and science achievement. Intensive studies of students, teachers, schools, curriculum, instruction, and policy issues were also carried out to understand the educational context in which learning takes place.

TIMSS represents the continuation of a long series of studies conducted by the IEA. The IEA conducted its First International Mathematics Study (FIMS) in 1964 and the Second International Mathematics Study (SIMS) in 1980–82. The First and Second International Science Studies (FISS and SISS) were carried out in 1970–71 and 1983–84, respectively. Since the subjects of mathematics and sciences are related in many respects and since there is broad interest among countries in students' abilities in both mathematics and science, the third studies (TIMSS) were conducted as an integrated effort.

TIMSS collected data from students in three separate populations. *Population 1*, in which 26 countries participated, consisted of students enrolled in the two adjacent grades that contained the largest proportion of 9-year-old students at the time of testing; in most countries, these were the 3rd and 4th grades. *Population 2*, in which 41 countries participated, consisted of students enrolled in the two adjacent grades that contained the highest proportion of 13-year-old students at the time of testing; in most countries, these were the 7th and 8th grades. *Population 3*, in which 23 countries participated, consisted of students in their final year of secondary education. As an additional option, countries could test special subgroups of these students: students having taken advanced courses in mathematics and students having taken courses in physics.

In 1999, a follow-up study called the Third International Mathematics and Science Study-Repeat (TIMSS-R) was conducted. The design of TIMSS-R makes it possible to track changes in achievement and certain background factors from the first TIMSS study. It incorporated an expanded videotape classroom study as well as a National Assessment of Educational Progress (NAEP)/TIMSS linking study to allow researchers to compare TIMSS results with those from NAEP. In addition, the TIMSS-R included a national Benchmarking Project, through which districts and states in the United States could compare their progress internationally as individual "nations." Unlike the first TIMSS, the 1999 TIMSS-R study focused only on 8th-grade students.

1995 WORLDWIDE STUDY OF CLASSROOMS, WITH FOLLOW-UP STUDY OF 8th GRADERS IN 1999; 41 COUNTRIES PARTICIPATED

TIMSS tested three populations:

- 9-Year-Olds/3rd and 4th Graders
- 13-Year-Olds/7th and 8th Graders
- Students in Final Year of Secondary Education

Purpose

The two broad questions that TIMSS addresses are: (1) How do mathematics and science educational environments differ across countries, how do student outcomes differ, and how are differences in those outcomes related to differences in mathematics and science education environments? (2) Are there patterns of relationships among contexts, inputs, and outcomes within countries that can lead to improvements in the theories and practices of mathematics and science education?

Components

TIMSS used several types of instruments to collect data about students, teachers, and schools. In addition, 8th graders in the United States, Japan, and Germany participated in a *videotape study*, in which actual classroom sessions were recorded, coded, and analyzed; this study was expanded to include seven nations in TIMSS-R. Various populations also participated in *curriculum studies* and *ethnographic case studies*. The United States sponsored two additional components of TIMSS-R: a Benchmarking Project and the NAEP/TIMSS-R Linking Study. The TIMSS-R *did not* include the performance assessment.

Written Assessment. Questionnaires were developed to test Population 1, 2, and 3 students in various content areas within mathematics and science. For Population 1, the mathematics content areas included: whole numbers; fractions and proportionality; measurement, estimation, and number sense; data representation, analysis, and probability; geometry; and patterns, relations, and functions. The Population 1 science content areas were earth science; life science; physical science; and environmental issues and the nature of science. The Population 2 mathematics content areas were fractions and number sense; geometry; algebra; data representation, analysis, and probability; measurement; and proportionality. The Population 2 science content areas were earth science; life science; physics; chemistry; and environmental issues and the nature of science. The Population 3 mathematics contents areas were numbers; measurement; geometry; proportionality; functions, relations, and equations; data, probability, and statistics; elementary analysis; and validation and structure. The Population 3 science contents were earth sciences; life sciences; physical sciences; science, technology, and mathematics; history of science; environmental issues; nature of science; and science and other disciplines. In addition, Population 3 students who had taken advanced mathematics were eligible for the advanced mathematics test, which included numbers and equations, calculus, geometry, probability and statistics, and validation and structure. Population 3 students who had taken physics were eligible for a physics test. Its contents were mechanics, electricity and magnetism, heat, wave phenomena, and modern physics—particle, quantum and astrophysics, and relativity.

TIMSS-R written assessment tests repeat the Population 2 content areas.

Student Background Questionnaire. The student questionnaire for Populations 1 and 2 asked about students' demographics and home environment, including academic activities outside of school, people living in the home, parental education (only at Population 2), books in the home, possessions in the home, and the importance of students' mothers, peers, and friends placed on different aspects of education. Students were also queried about their attitudes toward mathematics and science. The final sections of the questionnaires asked about classroom experiences in mathematics and science. Similar items were asked of students in TIMSS-R.

The student questionnaire administered to Population 3 students was similar in most respects to the Population 2 student questionnaires. The only differences were that Population 3 students were also queried as to their future plans, their programs of study, and the most advanced mathematics and science courses they had taken.

Teacher Questionnaire. The teacher questionnaires for Population 2 addressed four major areas: teachers' background, instructional practices, students' opportunity to learn, and teachers' pedagogic beliefs. There are separate questionnaires for teachers of mathematics and of science. Since most Population 1 teachers teach all subjects, a single teacher questionnaire was developed to address both mathematics and science. So as not to overburden the teachers, the classroom practice questions in the Population 1 teacher questionnaire pertain mostly to mathematics. However, teachers also were asked about how they spend their time in school and the atmosphere in their schools (e.g., teaching loads, collaboration policies, responsibilities for decision-making, and the availability of resources).

The teacher questionnaires were designed to provide information about the teachers of the student samples in TIMSS. The teachers who completed TIMSS questionnaires do not constitute a sample from any definable population of teachers. Rather, they represent the teachers of a national sample of students.

There was no teacher questionnaire administered to the teachers of students in Population 3.

The teacher questionnaire for TIMSS-R gathered data about topics such as attitudes and beliefs about teaching and learning, teaching assignments, class size and organization, topics covered, the use of various teaching tools, instructional practices, and participation in professional development.

School Questionnaire. The school questionnaires for each population sought information about the school's community, staff, students, curriculum and programs of study, and instructional resources and time. At Populations 1 and 2, the school questionnaires also ask about the number of years students are taught by the same teacher. A school questionnaire was to be completed by the principal, headmaster, or other administrator of each school that participated in TIMSS. Similar items were asked of principals in TIMSS-R.

Performance Assessment. The TIMSS performance assessment was administered at Populations 1 and 2 to a subsample of students in the upper grades that participated in the written assessment. The performance tasks permitted students to demonstrate their ability to make, record, and communicate observations; to take measurements or collect experimental data and present them systematically; to design and conduct a scientific investigation; or to solve certain types of problems. A set of 13 such hands-on activities was developed; 11 of these tasks were either identical or similar across populations, and 2 were different. Of these two, one task was administered to Population 1 (4th graders) and one was administered to Population 2 (8th graders).

Videotape Study. The videotape classroom study was designed as the first study to collect videotaped records of classroom instruction from national probability samples in Japan, Germany, and the United States to gather more in-depth information about the context in which learning takes place and also to enhance understanding of the statistical indicators available from the main TIMSS study. An hour of regular classroom instruction was videotaped in a subsample of 8th-grade mathematics classrooms (except in Japan, where videotaping was usually done in a different class, selected by the principal) included in the assessment phase of TIMSS in each of the three countries.

National-level univariate statistics were constructed to generate descriptive statistics for each country and a comparison was made between the mathematics achievement scores of classrooms in the main TIMSS samples and the subsample of classrooms selected for the video study.

The TIMSS-R Videotape Classroom Study was expanded in scope to examine national samples of 8th-grade mathematics and science instructional practices in seven nations: Australia, the Czech Republic, Hong Kong, Japan, the Netherlands, Switzerland, and the United States. Four countries—Australia, the Czech Republic, the Netherlands, and the United States—participated in both the mathematics and science components of the study. Hong Kong and Switzerland participated in only the mathematics component, and Japan in only the science component.

Curriculum Studies. Continuing the approach of previous IEA studies, TIMSS addressed three conceptual levels of curriculum. The intended curriculum is composed of the mathematics and science instructional and learning goals as defined at the system level. The implemented curriculum is the mathematics and science curriculum as interpreted by teachers and made available to teachers. The attained curriculum is the mathematics and science content that students have learned and their attitudes toward these subjects. To aid in interpretation and comparison of results, TIMSS also collected extensive information about the social and cultural contexts for learning, many of which are related to variation among educational systems.

To gather information about the intended curriculum, mathematics and science specialists within each participating country worked section by section through curriculum guides, textbooks, and other curricular materials to categorize aspects of these materials in accordance with detailed specification derived from TIMSS mathematics and science curriculum frameworks.

To collect data about how the curriculum is implemented in classrooms, TIMSS administered a broad array of questionnaires, which also collected information about the social and cultural contexts for learning. Questionnaires were administered at the country level about decision-making and organizational features within the education systems. The students who were tested answered questions pertaining to their attitudes toward mathematics and science, classroom activities, home background, and out-of-school activities. The mathematics and sciences teachers of sampled students responded to questions about teaching emphasis on the topics in the curriculum frameworks, instructional practices, textbook use, professional training and education, and their views on mathematics and science. The heads of schools responded to questions about school staffing and resources, mathematics and science course offerings, and support for teachers. In addition a volume was complied that presents descriptions of the educational systems of the participating countries.

Ethnographic Case Studies. The case studies approach to understanding cultural differences in behavior has a long history in selected social science fields. Given the goals of TIMSS, it was designed to focus on four key topics that challenge U.S. policymakers and investigate how these topics are dealt with in the United States, Japan, and Germany: implementation of national standards; the working environment and training of teachers; methods for dealing with differences in ability; and the role of school in adolescents' lives. Each topic was studied through interviews with a broad spectrum of students, parents, teachers, and educational specialists. The ethnographic approach permitted researchers to explore the topics in a naturalistic manner and to pursue them in greater or lesser detail, depending on the course of the discussion. As such, these studies both validate and integrate the information gained from official sources with that obtained from teachers, students, and parents in order to ascertain the degree to which official policy reflects actual practice. The objective is to describe policies and practices in the nations under study that are similar to, different from, or nonexistent in the United States.

In three regions in each of the three countries, the research plan called for each of the four topics to be studied in the 4th, 8th, and 12th grades. The specific cities and schools were selected "purposively" to represent different geographical regions, policy environments, and ethnic and socioeconomic backgrounds. Schools in the case studies were separate from schools in the main TIMSS sample. Where possible, a shortened form of the TIMSS test was administered to the students in the selected schools. The ethnographic researchers in each of the countries conducted interviews and obtained information through observations in schools and homes. Both native-born and nonnative researchers participated in the study to ensure a range of perspectives.

TIMSS-R Benchmarking Project. Twenty-seven states, districts and consortia of districts throughout the United States participated as their own "nations" in this project, following the same guidelines as the participating countries. The samples drawn for each of these states and districts are representative of the student population in each of these states and districts. The findings from this project allow these jurisdictions to assess their comparative international standing and judge their mathematics and science programs in an international context.

NAEP/TIMSS-R Linking Study. A subsample of students taking the 2000 state NAEP mathematics and science assessment also took the TIMSS-R assessment. (See chapter 20 for more information on NAEP.) This provides an opportunity to compare students' performance on NAEP to their performance on TIMSS-R, and allows for estimates of how states participating in NAEP 2000 would have performed had they participated in TIMSS-R. Results from the TIMSS-R Benchmarking Study are used to check the results of the linking study.

Periodicity

The Third International Mathematics and Science Study was conducted only once. Previous international math studies were conducted in 1964 and 1980–82; previous international science studies were conducted in 1970–71 and 1983–84. A follow-up study of 8th graders, using a similar design (but different students) was conducted in 1999. This follow-up study is called the Third International Mathematics and Science Study-Repeat (TIMSS-R).

2. USES OF DATA

The possibilities for specific research questions to be dealt with by TIMSS are numerous; however, the main research questions, focused at the student, the school or classroom, and the national or international levels, are illustrated below:

- How much mathematics and science have students learned?
- How well are students able to apply mathematics and science in problem-solving abilities?
- What are students' attitudes toward mathematics and science?
- How do gender differences in participation rates, course selection, and student outcomes differ across countries?
- What do teachers teach in their classrooms?
- What methods and materials do teachers use in teaching mathematics and science, and how are they related to student outcomes?
- What kinds of grouping practices, either within or between classrooms, are used, and how are those practices reflected in student outcomes and participation in subsequent mathematics and science courses?

- How strongly are students motivated to learning in general and to the learning of mathematics and science in particular? What are the sources of their motivation?
- What factors characterize the academic and professional preparation of teachers of mathematics and science?
- What are teachers' beliefs and opinions about the nature of mathematics and science and their teaching, and how are these related to comparable opinions and attitudes of their students?
- How do teachers evaluate their students?
- If there are national curricula in a country, how specific are they, and what efforts are made to see that the national curricula are followed?
- What proportions of students plan to study mathematics or science at the postsecondary level or to pursue mathematics or science-based careers?

Country-level outcomes are necessarily related to student- and classroom-level outcomes, and an important aspect of TIMSS is to identify the prime determinants of student outcomes, including the amount and quality of opportunity to learn and the intensity and perseverance of the students' motivation.

3. KEY CONCEPTS

Key terms related to TIMSS are described below.

Nationally Desired Population. The effective target population within each participating country. The stated objective in TIMSS was that the Nationally Desired Population within each country be as close as possible to the International Desired Population, which is the target population. (See below.) Using the International Desired Population as a basis, participating countries had to operationally define their populations for sampling purposes. Some National Research Coordinators had to restrict coverage at the county level, for example, by excluding remote regions or a segment of the educational system. Thus, the Nationally Desired Population sometimes differed from the International Desired Population.

National Research Coordinators (NRCs). The official from each participating country appointed to implement national data collection and processing in accordance with international standards. In addition to selecting the sample of students to be taken, NRCs were responsible for working with school coordinators, translating the test instruments, assembling and printing the test booklets, and packing and shipping the necessary materials to the sampled schools. They were also responsible for arranging the return of the testing materials from the school to the national center, preparing for and implementing the free-response scoring, entering the results into data files, conducting on-site quality assurance observations for a 10 percent sample of schools, and preparing a report on survey activities.

4. SURVEY DESIGN

Target Population

For TIMSS Populations 1 and 2, the International Desired Populations for all countries were defined as follows:

- Population 1: All students enrolled in the two adjacent grades that contain the largest proportion of 9-year-olds at the time of testing
- Population 2: All students enrolled in the two adjacent grades that contain the largest proportion of 13-year-olds at the time of testing

TIMSS used a grade-based definition of the target population at Populations 1 and 2. In a few cases, TIMSS components were administered only to the upper grade of these populations (i.e., the performance assessment was conducted at the upper grade and some background questions were asked of the upper grade students only). However, two adjacent grades were chosen to ensure extensive coverage of the same age cohort for most countries, thereby increasing the likelihood of producing useful age-based comparisons in addition to the grade-based analyses.

The intention of the assessment of final-year students (Population 3) was to measure what might be considered the "yield" of the elementary and secondary education systems of a country with regard to mathematics and science. This was accomplished by assessing the mathematics and science literacy of all students in the final year of secondary school, the advanced mathematics knowledge of students having taken advanced mathematics courses, and the physics knowledge of students having taken physics. The International Desired Population, then, was all students in the final year of secondary school, with students having taken advanced mathematics courses and students having taken physics courses as two overlapping subpopulations. Students repeating the final year were not part of the desired population. For each secondary education track in a country, the final grade of the track was identified as being part of the target population, allowing substantial coverage of students in their final year of schooling. For example, grade 10 could be the final year of a vocational program, and grade 12 the final year of an academic program. Both of these grade/track combinations are considered part of the target population, but grade 10 in the academic track is not.

For TIMSS-R, the international desired population consisted of all students in each participating nations who were enrolled in the upper of the two adjacent grades that contained the greatest proportion of 13-year-olds at the time of testing.

Sample Design

The TIMSS sample design for each country and population was intended to give a probability sample of all students within the target grades in the national school system (except for a small number of students allowed to be excluded as ineligible according to national criteria). Every eligible student in the country's school system had a chance of being selected, with a fixed probability of selection. These probabilities of selection were designed to be equal across eligible students as much as was possible, but for a variety of reasons the eligible students' probabilities of selection differ between students in most of the national samples.

Written Assessment. The TIMSS sample design was a two-stage cluster sample, with schools as the first stage of selection and classrooms within schools as the second stage of selection. The classroom sampling design was intended to be an equal probability design with no subsampling in the classroom. However, a design based on a probability proportionate to size (PPS) sample of classrooms with a fixed sample size of students selected within the sampled classroom was permitted under the international guidelines. Exclusions could occur at the school level, the student level, or both. TIMSS participants were expected to keep such exclusions to no more than 10 percent of the national desired population. Twenty of 23 participants in the Population 3 study achieved 100 percent coverage. The school sampling process was generally a stratified probability PPS sample, with the measure of size for a school equal to the number of students in the school in the two target grades for each population.

In the first stage of sampling, representative samples of schools were selected from sampling frames (comprehensive lists of all eligible students). TIMSS standards for sampling precision required that all population samples have an effective sample size of at least 400 students for the main criterion variables. To meet the standard, at least 150 schools were to be selected per target population. However, the clustering effect of sampling classrooms rather than students was also considered in determining the overall sample size for TIMSS. Because the magnitude of the clustering effect is determined by the size of the cluster and the intraclass correlation, TIMSS produced sample-design tables showing the number of schools to sample for a range of intraclass correlations and minimum-cluster-size values. Some countries needed to sample more than 150 schools. Countries, however, were asked to sample 150 schools even if the estimated number of schools to sample was less than 150.

The schools in each explicit stratum (e.g., geographical region, public/private, etc.) were listed in order of the implicit stratification variables, and then further sorted according to their measure of size. Of course, the stratification variables differed from country to country. Small schools were handled either through explicit stratification or through the use of pseudo-schools. In some very large countries, there was a preliminary sampling stage before schools were sampled in which the country was divided into primary sampling units.

In cases where a sampled school was unable to participate in the assessment, it was replaced by a replacement school. The mechanism for selecting replacement schools, established a priori, identified the next school on the ordered school-sampling list as the replacement for each particular sampled school. The school after that was a second replacement, should it be necessary. Using either explicit or implicit stratification variables and ordering of the school sampling frame by size ensured that any original sampled school's replacement would have similar characteristics.

In the second sampling stage, classrooms of students were sampled. Generally, in each school, one classroom was sampled from each target grade, although some countries opted to sample two classrooms at the upper grade in order to be able to conduct special analyses. Most participants tested all students in selected classrooms, and in these instances the classrooms were selected with equal probabilities. A few participants used a design based on a PPS sample of classrooms, with a fixed sample size of students selected within the sampled classrooms.

In an optional third sampling stage, participants with particularly large classrooms in their schools could decide to subsample a fixed number of students from each selected classroom. This was done using a simple random sampling method whereby all students in a sampled classroom were assigned equal selection probabilities.

For Population 3, in order to implement the TIMSS goal of assessing the mathematics and science literacy of all students while also assessing the advanced mathematics and physics knowledge of students with preparation in these subjects, it was necessary to develop a sampling design that ensured that students were stratified according to their level of preparation in mathematics and physics, so that appropriate test booklets could be assigned to them. Within each sampled school, students were classified according to a four-group classification scheme (i.e., students having studied neither advanced mathematics nor physics, students having studied physics but not advanced mathematics, students having studied advanced mathematics but not physics, and students having studied both advanced mathematics and physics), and 40 students were sampled at random, 10 from each of the four categories. If just three student types were present three samples of 13 students were drawn. In some tracked systems, schools frequently consisted of a single group. In these situations all 40 students were sampled from whichever group was appropriate.

The United States' national TIMSS design followed the international specifications described above for the three populations. Primary sampling units (PSUs) were sampled as the first stage of sampling with the PSUs defined as metropolitan statistical areas (MSAs), single counties, or groups of counties. There were 1,027 PSUs on the sampling frame with 11 of the PSUs taken as certainty selections (representing the 11 largest metropolitan areas) and 48 PSUs drawn from the remaining 1,016 PSUs, with probability proportionate to the 1990 population within the PSU. These PSUs were placed in eight primary strata. The 48 noncertainty PSUs were substratified by socioeconomic status and demographic characteristics that were found to be most highly related to educational achievement within the primary strata, as measured by aggregated assessment data from previous NAEP surveys. (For more information on NAEP, see chapter 20.)

For both the 11 certainty PSUs and the 48 sampled noncertainty PSUs, the measures of the size of the school were proportional to the target grade size in the school divided by the PSU probability of selection. In addition, schools in both types of PSUs with high percentages of Blacks, and Hispanics (greater than 15 percent of the population) were given doubled probabilities of selection. The school sample sizes for both Populations 1 and 2 were 220 schools.

Public and private schools were sampled from separate frames. The public school sample was drawn from the most recent Quality Education Data (QED) sampling frame. The private schools sample was drawn from the 1991–1992 Private School Universe Survey (PSS) file. (For more information on PSS, see chapter 3.)

The U.S. sample design within schools for Populations 1 and 2 consisted of an equal probability sample of two upper grade (4th- or 8th-grade) classrooms and one lower grade (3rd- or 7th-grade) classroom within the school. All eligible students in the classroom were designated to be in the sample (i.e., there was no subsampling of students in the U.S. sample). The extra sampled classroom in the upper grade beyond the international minimum was drawn for the purpose of permitting analyses that did not confound school effects and classroom effects for grades 4 and 8. Classrooms were sampled with equal probability for each target grade in each sampled school in the U.S. sample, in accord with international specifications. All students in the sampled classroom were taken in the TIMSS sample. The sample design was approximately self-weighting at the student level within particular subgroups of the schools.

Performance Assessment. For the performance assessment, TIMSS participants were to sample at least 50 schools from those already selected for the written assessment, and from each school a sample of either 9 or 18 upper-grade students already selected for the written assessment. This yielded a sample of about 450 students in the upper grade of Populations 1 and 2 (4th and 8th grades in most countries) in each country. For the performance assessment, in the interest of ensuring the quality of administration, countries could exclude additional schools if the schools had fewer than nine students in the upper grade or if the schools were in a remote region. The exclusion rate for the performance assessment was not to exceed 25 percent of the national desired population.

Teacher Questionnaire. The TIMSS database for each country includes questionnaire data from the teachers of the sampled classrooms, which can be linked to student assessment data in the classrooms. Any teacher linked as a mathematics or science teacher to any assessed student is eligible to receive a questionnaire. The classroom sample is drawn from a listing of mathematics classrooms, so that in most situations only one mathematics teacher is linked to each sampled classroom. If this single teacher is also only linked to single sampled classroom, then the teacher received a questionnaire for that single classroom.

This straightforward one-to-one linking does not always hold, however. In some cases, teachers may teach both mathematics and science to students in a sampled classroom, making them eligible to receive questionnaires for both subjects. For a single subject, a teacher may also teach multiple classrooms (e.g., the sampled classrooms for the school from both target grades).

For the U.S. TIMSS sample, a teacher was never asked to complete more than one questionnaire. In cases when a teacher taught both subject areas, the teacher was randomly assigned to receive a mathematics or science teachers' questionnaire. In cases when a teacher taught assessed students in one subject area in more than one classroom, the teacher was purposively assigned one classroom.

Each country was allowed to develop its own methodology for this process of assigning subjects and classrooms to teachers when the links were not straightforward due to the presence of one to many (or many to one) mappings.

Videotape Study. The sample for the TIMSS videotape study was assembled as a subsample of Population 2 students in Germany, Japan, and the United States. In the United States, schools were selected for the video study as follows: First, Population 2 TIMSS schools were listed in the order in which they were originally sampled. Using this ordering, pairs of schools were generated. Within each pair one of the two schools was randomly sampled (with each school having an equal probability of being sampled). The unsampled school in the pair was reserved as a potential replacement for the sampled school. A total of 109 pairs were assigned, with one school unpaired because one school of the original Population 2 sample of 220 schools had no 8th grade. The final videotape study sample size was 109. The unpaired school was not sampled. Within each sampled school, one 8th-grade classroom was selected with equal probability from the two TIMSS 8th-grade classrooms in the school. There was no sorting or stratification of classrooms by level of mathematics taught. In the event that the sampled teacher refused to be videotaped, the classroom was never replaced by the other 8th-grade classroom in the same school. Instead the entire school was replaced by its paired school.

The final TIMSS video sample in the United States consisted of 81 schools, of which 73 were public schools and 8 were private schools. The final video sample in Germany consisted of 100 schools, 15 of which were replacement schools. In Japan, 50 schools participated in the videotape study, 2 of which were replacement schools.

Sampling for the TIMSS-R videotape study was performed in two steps. The first step was to sample 100 schools in each country. The second step was to sample one mathematics classroom and one science classroom from each school. Sampling of schools in each country was performed using the same procedures being used in the TIMSS-R achievement study; most countries, however, did not videotape in the same schools in which the TIMSS-R assessment was conducted. Thus, linkage of the video study to the achievement study is only at the national level. A replacement school will be chosen for each of the 100 schools for each country. If the primary school refused to participate, its replacement school was invited to replace it. Within each school, one mathematics class and/or one science class was randomly selected for videotaping.

Assessment Design

The task of putting together the achievement item pools for the different TIMSS tests took more than 3 years to complete. The process necessitated building international consensus among NRCs, their national committees, mathematics and science experts, and measurement specialists. The NRCs from all participating countries worked to ensure that the items used in the tests were appropriate for their students and reflected their country's curricula. Because students in Population 3 were less likely to have been taught a comparable curriculum (due to some students' having taken advanced mathematics and physics classes), the design of written assessments for this population differs somewhat from that of Populations 1 and 2. As a result, Population 3 will be discussed separately.

The international versions of the test instruments and the student and school background questionnaires were developed in English and then translated into other languages by TIMSS countries. While the intent of TIMSS was to provide internationally comparable data for all variables, there were many contextual differences among countries so that the international version of the questions was not always appropriate in all countries. Therefore, the international versions of the questionnaires were designed to provide an opportunity for individual countries to modify some questions or response options in order to include the appropriate wording or options most consistent with their own national systems. Each item deviation or national adaptation was reviewed to determine whether the national data should be: deleted as not being internationally comparable, recoded to match the international version, or retained with some documentation describing modifications. Whenever possible, national data were retained to match as closely as possible the international version of the items and/or by documenting minor deviations.

For Populations 1 and 2, the test items were allocated to 26 different clusters. Also, at each population, the 26 clusters were assembled into eight booklets. Each student completed one booklet. At Population 1, the clusters were either 9 or 10 minutes in length. The core cluster, which was composed of five mathematics and five science multiple-choice items, was included in all booklets. Focus clusters appeared in at least three booklets, so that the items were answered by a relatively large fraction (three-eighths) of the student sample in each country. The breadth clusters, largely containing multiple-choice items, appeared in only one booklet each. The free-response clusters were each assigned to two booklets, so that items statistics of reasonable accuracy would be available. The booklet design for Population 2 is very similar to that for Population 1, differing only in the length and item content of the clusters.

Students in Population 3 were classified into four groups based on their preparation in mathematics and physics. Each student was characterized as having taken advanced mathematics or not, and as having taken physics or not. The assessment of these students was accomplished through a complex design that included four types of test booklets (nine booklets in total) that were distributed to students based on their academic preparation. The four types of test booklets were intended to yield proficiency estimates in mathematics and science literacy, advanced mathematics, and physics.

The TIMSS test design for Population 3 included 12 mutually exclusive clusters of items distributed among the four types of test booklets in a systematic fashion. The test booklets were rotated among students based on the student classification scheme so that each student completed one 90-minute test booklet.

TIMSS-R utilized the same assessment framework designed for TIMSS. Approximately one-third of the original 1995 TIMSS assessment items were kept secure so that they could be included in the 1999 TIMSS-R assessment to provide trend data. For the approximately two-thirds of items that were released to the public, a panel of international assessment and content experts and the NRCs of each participating country developed and reviewed replacement items that closely matched the content of the original items. Through this process, over 300 science and mathematics items were developed as potential replacement items, of which 277 items were carefully chosen to be field tested. Approximately 1,000 students per country participated in this field test. Of the 277 potential replacement items, 202 were selected based on the results of the field test.

Data Collection and Processing

Each country participating in TIMSS was responsible for collecting its national data and processing the materials in accordance with the international standards. In each country, a national research center and NRC were appointed to implement these activities. One of the main ways in which TIMSS sought to achieve uniform project implementation was by providing clear and explicit instructions on all operational procedures. Such instructions were provided primarily in the form of operations manuals, supported where possible by computer software systems that assisted NRCs in carrying out the specified filed operations procedures. Forms accompanying some of the manuals served to document the implementation of the procedures in each country. Many of these forms were used to track schools, students, and teachers, and to ensure proper linkage of schools, students, and teachers in the database.

Reference dates. All TIMSS testing was conducted at "the end of the school year." Because academic schedules differ across countries, this was not a set date for all countries, but was relative to each country's particular educational system. Most countries tested the mathematics and science achievement of their students at the end of the 1994–95 school year, most often in May and June of 1995. The three countries on a Southern Hemisphere school schedule (Australia, New Zealand, and South Africa) tested between August and December 1995, which was late in the school year in the Southern Hemisphere. Three countries (Iceland, Germany, and Lithuania) tested their final-year students (or a subset of them) at the end of the 1995–96 school year.

Likewise, TIMSS-R was conducted on two schedules. The Southern Hemisphere countries administered the survey from September to November, 1998, while the Northern Hemisphere countries did so from February to May, 1999.

Data collection. Each participating country was responsible for carrying out all aspects of the data collection, using standardized procedures developed for the study. Training manuals were created for school coordinators and test administrators that explained procedures for receipt and distribution of materials as well as for the activities related to the training sessions. The manuals covered procedures for test security, standardized scripts to regulate directions and timing, rules for answering students' questions, and steps to ensure that identification on the test booklets and questionnaires corresponded to the information on the forms used to track students.

Specific discussions of collection methods for the performance assessment and videotape study are provided below.

Performance Assessment. Specific procedures were established to ensure that the performance assessment was administered in as standardized a manner as possible across countries and schools. The NRC in each participating country was responsible for collecting the equipment and materials required for each of the performance assessment tasks, and for assembling a set of materials for each school. The tasks were designed to require only materials that were easy to obtain and inexpensive. Many of the pieces of "equipment" could be homemade; for example, one take required a balance that could be made from a coat hanger, plastic cups, and string. The Performance Assessment Administration Manual provided explicit instructions for setting up the equipment, described which tasks required servicing during administration, and contained instructions for recording information about the materials used that coders could refer to when scoring.

Students were required to move from station to station around a room to perform the tasks assigned to them. The administrator was responsible for overseeing the activities, keeping time, directing students to their stations, maintaining and replenishing equipment as necessary, and collecting the students' work. The administrator also provided advance instruction regarding certain materials and equipment, for tasks where the use of the equipment was not what was being measured. Administrators did not provide instruction on other procedures nor answer any other questions related to the activities required for the tasks.

To facilitate the students' movements around the room and keep track of where each should be, each student was given a routing card, prepared at the TIMSS national center. The routing cards stated the rotation scheme and sequence number of that student, his or her identifying information, and the stations to which the student was to go and in what order.

At each station, students performed the assigned task. This involved performing the designated activities, answering questions, and documenting their work in booklets (one booklet per task per student). Students had 30 minutes to work at each station. When students had finished their work at a station (or when time had expired), they handed their completed booklets to the administrator.

The performance assessment was not conducted in TIMSS-R.

Videotape Study. It was intended that TIMSS videotaping be spread out evenly over the school year. In Germany and the U.S. this goal was accomplished by employing a single videographer in each country to tape over an 8-month period, from October 1994 through May 1995. It was not possible to implement the same plan in Japan, due to the starting time of the school year in Japan and the necessity of coordinating the videotaping with the test administration. As a result, videotaping in Japan was compressed primarily into a 4-month period, from November 1994 though February 1995, with a few lessons taped in March.

Two kinds of data were collected in the TIMSS videotape study: videotapes and questionnaires. Supplementary materials deemed helpful for understanding the lesson (e.g., copies of textbook pages or worksheets) were also collected. Each classroom was videotaped once on a date convenient for the teacher. One complete lesson, as defined by the teacher, was videotaped in each classroom. Teachers were initially contacted by a project coordinator in each country who explained the goals of the study and scheduled the date and time for videotaping. Because teachers knew when the taping would take place, it was understood that they would attempt to prepare in some way for the event. In order to cut down somewhat on the variability in preparation methods across teachers, all participating teachers were given a common set of instructions, asking them not to make any special preparations for the taped class (e.g., by making special materials, planning special lessons, or practicing the lesson ahead of time). On the appointed day the videographer arrived at the school and videotaped the lesson. After the taping each teacher was given a questionnaire and an envelope in which to return it. The purpose of the questionnaire was to assess how typical the lesson was according to the teacher and to gather contextual information important for understanding the contents of the videotape.

All videotaping was done in real time, using a single camera. The camera was turned on at the beginning of the class, and not turned off until the lesson was over. In order to ensure comparability between videotapes, videographers were asked to adhere to two basic principles in choosing what to tape. The first principle required videographers to assume the perspective of an ideal student in the class and to aim the camera toward the object of focus of an ideal student at any given time.

An ideal student was defined as one who is always attentive to the lesson at hand and always occupied with the learning tasks assigned by the teacher, one who will attend to individual work when assigned to work alone, will attend to the teacher when she or he addresses the class, and will attend to peers when they ask questions or present their work or ideas to the whole class. In cases where different students in the same class are engaged in different activities, the ideal student is assumed to be doing whatever the majority of students are doing.

The second principle required videographers to capture everything the teacher did to instruct the class, regardless of the activities of the ideal student. Usually, this principle was in agreement with the first principle: whenever the ideal student is attending to the teacher, both principles would have the camera pointed at the teacher. However, there are times when the two principles are in conflict. In order to develop a set of standardized procedures for such instances, the three videographers were trained over the course of two intensive training seminars that lasted a total of 14 days. Tests conducted both during the training seminars and later during data collection revealed that videotaping methods were indeed comparable.

The TIMSS-R data collection methods differed in several respects from those used for TIMSS. Two cameras were used, instead of one, to videotape each lesson. One of the cameras focused primarily on the teacher, but was also used to capture close-ups of students' work during periods when students were working independently. The second camera was stationary. It was placed at the front of the room facing the students in order to capture students' interactions with the teacher and/or with each other during the lesson.

Editing. To maintain equality among countries, very little optical scanning and no image processing of item responses was permitted. All student test information was recorded in the student booklets or on separate coding sheets, and similar procedures were used for the questionnaires. Entry of the achievement and background data was facilitated by the *International Codebooks*, and the DataEntryManager software program.

The background questionnaires were stored with the various tracking forms so that the data entry staff could control the number of records to enter and transcribe the necessary information during data entry. NRCs were asked to arrange for double-entry of a random sample of at least 5 percent of the test instruments and questionnaires to gauge the error rate. An error rate of 1 percent was considered acceptable.

After entering data files in accordance with the international procedures, countries submitted their data files to the IEA Data Processing Center. There, TIMSS data underwent an exhaustive cleaning process designed to identify, document, and correct deviations for the international instruments, file structures, and coding schemes. The process also emphasized consistency of information with national data sets and appropriate linking among the many data files. The national centers were contacted regularly throughout the cleaning process and were given multiple opportunities to review the data for their countries. As a result of this review process, several items were identified as not being international comparable in certain countries and were deleted from the international data files and from the analyses for the international reports. In certain instances, recodes were performed on the cognitive items as a result of the item review.

Estimation Methods

Once TIMSS data are scored and compiled, the responses are weighted according to the sample design and population structure and then adjusted for nonresponse. This ensures that countries' representation in TIMSS is accurately assessed. The analyses of TIMSS 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 student achievement (proficiency) in the various subject areas. The methodology of multiple imputations (plausible values) is then used to estimate characteristics of the proficiency distributions. Although imputation is conducted for the purpose of determining plausible values, no imputations are included in the TIMSS database.

Weighting. Appropriate estimation of population characteristics based on TIMSS samples requires that the TIMSS sample design be taken into account in all analyses. This is accomplished in part by assigning a weight to each respondent, where the sampling weight properly accounts for the sample design, takes into account any stratification or disproportional sampling of subgroups, and includes adjustments for nonresponse.

There are four types of sampling weights available for use with TIMSS data: student weights, school weights, student-teacher weights, and teacher weights. In all of these cases, weighted totals, means, and percentages using these weights are unbiased estimates of “weighted” national population totals, with the number of target grade students as the weight.

Student weights. The student sampling weights in TIMSS have two primary components: a student base weight and a nonresponse adjustment. The student base weight is the reciprocal of the student’s probability of selection into the TIMSS sample, and is a product of up to three factors, reflecting the three stages of student sampling: the school selection probability, the classroom selection probability, and (if classroom subsampling has occurred) the student selection probability within selected classrooms. In most country samples, there is both school and student nonresponse. This nonresponse affects any estimators in that the effective sample size of both schools and students is reduced, increasing sampling variance. In addition, if there are systematic differences between the respondents and the nonrespondents, there will also be a bias of unknown size and direction in any estimators. This bias is partially adjusted for in TIMSS samples through the use of weighting adjustments multiplied to the student base weights.

Three versions of the students’ sampling weight are provided in the user database. All three give the same figures for statistics such as means and proportions, but vary for statistics such as totals and population sizes. In addition to the total weight, described above, there are *House weights* and *Senate weights* for each student (the names are derived from an analogy with the U.S. legislative system). *House weights* are a set of weights based on the total sample size of each country, to be used when estimates across countries are computed or significance tests performed. The transformation of the weights will be different within each country, but in the end, the sum of the house-weight variables within each country will total to the sample size for that country. The house-weight variable is proportional to the total weight for that variable by the ratio of the sample size divided by the size of the population. These sampling weights can be used when the user wants the actual sample size to be used in performing significance tests.

Senate weights are a set of weights based on a constant scalar, to be used when estimates across countries are computed or significance tests performed. The transformation of the weights will be different within each country, but in the end, the sum of the senate-weight variables within each country will total to a fixed value (1000 in Populations 1 and 2, where two grades were sampled, and 500 in Population 3). The senate-weight variable, within each country, is proportional to the total weight by the ratio of 1000 (or 500) divided by the size of the population estimate. These sampling weights can be used when cross-national comparisons are required and the user wants to have each country contribute the same amount to the comparison, regardless of the size of the population.

Teacher weights. The teacher weight is a teacher-classroom weight, and so is greater than 0 for a classroom only if the teacher filled out a questionnaire for that classroom. The teacher-classroom weight is equal to the summation of the student-teacher weights for students linked to that classroom (for that assessment).

Student-teacher weights. The U.S. TIMSS public use file includes student-teacher weights and student-teacher replicate weights. These are aggregated into the teacher weights described above. Two student-teacher weights are assigned to each assessed student in U.S. TIMSS: a mathematics assessment weight and a science assessment weight. A student-teacher weight for a particular student and assessment is set to 0 if a teacher’s questionnaire was not filled out for that student’s assessment classroom. This occurred in the following situations: the teacher taught both mathematics and science and was randomly assigned to the other assessment; the teacher was assigned no classroom because of all his/her classrooms had fewer than five TIMSS-assessed students; the teacher was assigned a questionnaire classroom but not the student” classroom; the teacher refused to answer the questionnaire.

Population 3 advanced mathematics/physics adjustment factors. Student weights for Population 3 are similar to the Population 1 and 2 weights; but an additional set of weights was created to reflect the fact that some respondents had taken advanced mathematics or physics courses, or both. Weights were developed as the inverse of the probabilities that a student received a mathematics/physics literacy booklet, an advanced mathematics booklet, or a physics booklet. If a student was not assessed on these items, the value of the weight was set to 0. As a result, the total, house, and senate weights in Population 3 for each math or science assessment are the product of the base weight (the inverse of the school selection probability multiplied by the inverse of the student selection probability), the nonresponse adjustment factor, the literacy adjustment factor, the advanced mathematics adjustment factor, and the physics adjustment factor.

The internationally-defined weighting specifications for TIMSS-R require that each assessed student’s sampling weight should be the product of (1) the inverse of the school's probability of selection, (2) an adjustment for school-level nonresponse, (3) the inverse of the classroom's probability of selection, and (4) an adjustment for student-level nonresponse.

Scaling. The principal method by which student achievement is reported in TIMSS is through scale scores derived using IRT scaling. IRT is used to estimate students' average proficiency for the nation, for various subgroups of interest within the nation (e.g., those defined by age, race/ethnicity, sex), and for the states and territories. TIMSS utilized a one parameter IRT model to produce score scales that summarized the achievement results.

In 1999, the TIMSS-R assessment had five scales describing mathematics content strands and six scales for describing fields of science. The 1995 TIMSS data were rescaled using a three-parameter IRT model, to match the procedures used to scale the 1999 TIMSS-R data. After careful study of the rescaling process, the International Study Center concluded that the fit between the original TIMSS data and the rescaled TIMSS data met acceptable standards. However, as a result of rescaling, the average achievement scores of some nations changed from those initially reported in 1996.

Imputation. No imputations are generated for missing values in teacher, school, or questionnaires for any TIMSS data file. However, multiple imputation techniques have been applied to create *plausible values* for students' proficiency scores. The data include a set of five plausible values for each student in each of the assessed areas. Plausible values improve the estimation of population parameters at the cost of additional computational requirements.

Plausible values were developed during the analysis of the 1983–84 NAEP data in order to improve estimates of population distributions. In the TIMSS survey design, students are presented with separate blocks of exercises, each block consisting of both mathematics and science problems. Since each student attempts only a small portion of the total TIMSS test in each subject, attempts to estimate proficiency distributions are affected by the imprecision of the measurement. During the estimation phase, plausible values for content-area scale scores are generated for each student participating in the assessment. The plausible values technology estimates five possible scores for each student, which ensures that the estimates of the average performance of subpopulations and the estimates of variability in those estimates are more accurate and appropriate than if only a single score were estimated for each student.

The process of drawing plausible values from the predictive distribution of proficiency values is called "conditioning." Plausible values are computed separately for each population. They are based on the student's responses to the items going into the scale and on the values of a set of background variables that are important for the reporting of proficiency scores. The variables used to calculate plausible values for a given assessment scale or group of scales include a broad spectrum of background, attitude, and experiential variables and composites of such variables.

Rubin (1987) proposes that this process be carried out several times—that is, multiple imputations—so that the uncertainty associated with imputation can be quantified.

Future Plans

Another international assessment—Trends in International Mathematics and Science Study—is currently planned for 2003, and will survey both 4th- and 8th-grade students. Subsequent follow ups are planned at 4-year intervals thereafter.

5. DATA QUALITY AND COMPARABILITY

In addition to setting high standards for data quality, the TIMSS International Study Center has tried to ensure the overall quality of the study through a dual strategy of support to the national centers and quality control checks.

Despite the efforts taken to minimize error, any sample survey as complex as TIMSS has the possibility of error. Below are discussed possible sources of error in TIMSS.

Sampling Error

With complex sampling designs that involve more than simple random sampling, as in the case of TIMSS where a multistage cluster design was used, there are several methods for estimating the sampling error of a statistic that avoid the assumption of simple random sampling. One such method is the jackknife repeated replication (JRR) technique. The particular application of the JRR technique used in TIMSS is termed a paired selection model because it assumes that the sampled population can be partitioned into strata, with the sampling in each stratum consisting of two primary sampling units (PSUs) selected independently. Following this first-stage sampling, there may be any number of subsequent stages of selection that may involve equal or unequal probability selection of the corresponding elements. The TIMSS design called for a total of 150 schools for the target population. These schools constituted the PSUs in most countries, and were paired sequentially after sorting by a set of implicit stratification variables. This resulted in the implicit creation of 75 strata, with two schools selected per stratum.

Imputation error. The variance introduced by imputation of missing data must be considered when using plausible values to estimate standard errors for proficiency estimates. The general procedure for estimating the imputation variance using plausible values is as follows: first estimate the statistic (t), each time using a different set of the plausible values (M). The statistics t_m can be anything estimable from the data, such as a mean, the difference between means, percentiles, etc. If all of the (M=5) plausible values in the TIMSS database are used, the parameter will be estimated five times, once using each set of plausible values. Each of these estimates will be called t_m, where m=1,2,…,5. Once the statistics are computed the imputation variance is then computed as:

$$Var_{imp} = \left(1 + {}^{1}\!/\!_{M}\right) \cdot Var\left(t_m\right)$$

where M is the number of plausible values used in the calculation and $Var(t_m)$ is the variance of the estimates computed using each plausible value.

Nonsampling Error

Due to the particular situations of individual TIMSS countries, sampling and coverage practices had to be adaptable, in order to ensure an internationally comparable population. As a result, nonsampling errors in TIMSS can be related to both coverage error and nonresponse. Measurement error was also a nontrivial issue in administering TIMSS, as different countries had different mathematics and science curricula. These potential sources of error are discussed in detail below.

Coverage error. The stated objective in TIMSS was that the effective population, the population actually sampled by TIMSS, be as close as possible to the International Desired Population. Yet, because a purpose of TIMSS was to study the effects of different international curricula and pedagogical methods on mathematics and science learning, participating countries had to operationally define their population for sampling purposes. Some NRCs had to restrict coverage at the country level, for example, by excluding remote regions or a segment of the educational system. In these few situations, countries were permitted to define a national desired population that did not include part of the International Desired Population. Exclusions could be based on geographic areas or language groups. Most countries participating in the Population 3 (20 out of 24) had 100 percent coverage, after sample exclusions. Among the four countries with incomplete coverage, the coverage rate ranged from 50 percent for Latvia to 84 percent for Lithuania.

To provide a better curricular match, several Population 2 countries elected to test students in the 7th and 8th grades (the two grades tested by most countries), even though that meant not testing the two grades with the most age-eligible students. This led to the students in these four countries being somewhat older than those in the other countries. The majority of countries in all sample populations satisfied the international guidelines for sample participation rates, grade selection, and sampling procedures.

Nonresponse error.

Unit nonresponse. Unit nonresponse error results from nonparticipation of schools and students. Weighted and unweighted response rates were computed for each participating country by grade, at the school level, and at the student level. Overall response rates (combined school and student response rates) were also computed.

The minimum acceptable school-level response rate, before the use of replacement schools, was set at 85 percent. This criterion was applied to the unweighted school-level response rate. Both weighted and unweighted school-level response rates were reported, with and without replacement schools. It was generally the case that weighted and unweighted response rates were similar.

Like the school-level response rate, the minimum acceptable student-level response rate was set at 85 percent. This criterion was applied to the unweighted student-level response rate. Both weighted and unweighted student level response rates were calculated. The weighted student-level response rate is the sum of the inverse of the selection probabilities for all participating students divided by the sum of the inverse of the selection probabilities for all eligible students.

Measurement error. Measurement error is introduced into a survey when its test instruments do not accurately measure the knowledge or aptitude they are intended to assess. The largest potential source of measurement error in TIMSS results from differences in the mathematics and science curricula across participating countries. In order to minimize the effects of measurement error, TIMSS carried out a special test called the Test-Curriculum Matching Analysis (TCMA). Each country was asked to identify, for each item, whether the topic of the item was intended in the curriculum for the majority of the students.

Data Comparability

The data collected for TIMSS in 1995 and the data collected for TIMSS-R in 1999 are comparable because comparability was built into the design and implementation. Through a careful process of review, analysis, and refinement, the assessment and questionnaire items were purposefully developed and field tested for similarity and for reliable comparisons between TIMSS and TIMSS-R. After careful review of all available data, including a test for item reliability between old and new items, the TIMSS and TIMSS-R assessments were found to be very similar in format, content, and difficulty level. Moreover, TIMSS and TIMSS-R data are on the same 8th-grade scale to allow for reliable comparisons between the two 8th-grade cohorts over time. Procedures for conducting the assessments were the same.

Findings from comparisons between the results of TIMSS and TIMSS-R, however, cannot be interpreted to indicate the success or failure of mathematics and science reform efforts within a particular country, such as the United States. TIMSS-R was designed to specifications detailed in the TIMSS curriculum frameworks. International experts developed the TIMSS curriculum frameworks to portray the structure of the intended school mathematics and science curricula from many nations, not specifically the United States Thus, when interpreting the findings, it is important to take into account the mathematics and science curricula likely encountered by U.S. students in school. TIMSS and TIMSS-R results are most useful when they are considered in light of other knowledge about education systems, including not only curricula, but also factors such as trends in education reform, changes in the school-age populations, and societal demands and expectations.

The ability to compare data across different countries constitutes a considerable part of the purpose behind TIMSS. As a result, it was crucial to ensure that items developed for use in one country were functionally identical to those used in other countries. Because questionnaires were originally developed in English and later translated into the language of each of the TIMSS countries, some differences do exist in the wording of questions. NRCs from each country reviewed the national adaptations of individual questionnaire items and submitted a report to the IEA Data Processing Center. In addition to the translation verification steps used for all TIMSS test items, a thorough item review process was used to further evaluate any items that were functioning differently in different countries according to the international item statistics. In certain cases, items had to be recoded or deleted entirely from the international database as a result of this review process.

6. CONTACT INFORMATION

For content information about TIMSS, contact:

Patrick Gonzales
Phone: (202) 502–7346
E-mail: patrick.gonzales@ed.gov

Mailing Address:

National Center for Education Statistics
1990 K Street NW
Washington, DC 20006–5651

7. METHODOLOGY AND EVALUATION REPORTS

Most of the technical documentation for TIMSS is published by Boston College. The U.S. Department of Education, National Center for Education Statistics, is the source of several additional references listed below; these publications are indicated with an NCES number.

General

Pursuing Excellence: Comparisons of International Eighth-Grade Mathematics and Science Achievement from a U.S. Perspective, 1995 and 1999, NCES 2001–028, by P. Gonzales, C. Calsyn, L. Jocelyn, K. Mak, D. Kastberg, S. Arafeh, T. Williams, and W. Tsen. Washington, DC: 2000.

Uses of Data

Linking the National Assessment of Educational Progress (NAEP) and The Third International Mathematics and Science Study (TIMSS): A Technical Report, NCES 98–499, by E.G. Johnson. Washington, DC: 1998.

Linking The National Assessment of Educational Progress (NAEP) and The Third International Mathematics and Science Study (TIMSS): Eighth-Grade Results, NCES 98–500, by E.G. Johnson and A. Siegendorf. Washington, DC: 1998.

User's Guide for the Third International Mathematics and Science Study (TIMSS) and U.S. Augmented Data Files, by B. Chaney, L. Jocelyn, D. Levine, T. Mule, L. Rizzo, K. Rust, S. Roey, T. Williams, and S. Warren. Rockville, MD: 1998.

Survey Design

Multiple Imputation for Nonresponse in Surveys, by D.B. Rubin. New York: John Wiley & Sons, 1987.

TIMSS International Study Center, Boston College, *TIMSS Technical Report: Volume I: Design and Development,* by M.O. Martin and D.L. Kelly (eds.). Chestnut Hill, MA: 1996.

TIMSS International Study Center, Boston College, *TIMSS Technical Report: Volume II: Implementation and Analysis Primary and Middle School Years,* by M.O. Martin and D.L. Kelly (eds.). Chestnut Hill, MA: 1998.

TIMSS International Study Center, Boston College, *TIMSS Technical Report: Volume III: Final Year of Secondary School,* by M.O. Martin and D.L. Kelly (eds.). Chestnut Hill, MA: 1998.

Data Quality and Comparability

TIMSS International Study Center, Boston College, *Quality Assurance in Data Collection*, by M.O. Martin and I.V.S. Mullis (eds.). Chestnut Hill, MA: 1996.

Chapter 22: IEA Reading Literacy Study

1. OVERVIEW

The International Association for the Evaluation of Educational Achievement (IEA) Reading Literacy Study was conducted during the 1990–91 school year in 32 countries around the world. The International Steering Committee (ISC), the International Coordinating Center (ICC), and the National Research Coordinators of each of the participating countries developed the assessment instruments, assessment procedures, and scaled scores used to report the results and oversaw the conduct of the study internationally. Nationally representative samples of the classes in the grades with the most 9-year-old and 14-year-old students were directed to read and respond to a broad range of materials over two testing periods. The U.S. component involved 7,200 4th-grade students and 3,800 9th-grade students at 332 public and private schools, distributed in 227 districts across 31 states and the District of Columbia.

STUDY OF 9-YEAR-OLD AND 14-YEAR-OLD STUDENTS IN 32 COUNTRIES

IEA Reading data collected through:

- Reading Assessment
- Student Questionnaire
- Teacher Questionnaire
- Principal Questionnaire
- National Questionnaire

Purpose

To (1) develop internationally valid instruments for measuring reading literacy suitable for establishing internationally comparable literacy levels in each of the participating countries; (2) describe on one international scale the literacy profiles of 9- and 14-year-olds in school in each of the participating countries; (3) describe the reading habits of the 9- and 14-year-olds in each participating country; and (4) identify the home, school, and societal factors associated with the literacy levels and reading habits of the 9-year-olds in school.

Components

The IEA Reading Literacy Study used a reading assessment instrument and four sets of questionnaires (for students, their teachers, their principals, and the nation) developed by committees working under the International Sampling Coordinator. The instruments were designed so that the same content would be used in all participating countries in the appropriate languages for those countries.

Reading Literacy Tests. Two reading assessments were developed to measure the reading proficiency of 9- and 14-year-olds. The assessments were designed to provide scaled scores that reflect students' understanding of three types of text: narrative prose (continuous text materials in which the writer's aim was to tell a story, whether fact or fiction), expository prose (continuous text materials designed to describe or explain things), and documents (structured tabular texts, such as forms, charts, labels, graphs, lists, and sets of instructions). The assessments include questions that tapped six types of reading processes: verbatim, paraphrase, inference, main theme, locating information, and following directions.

Questionnaires. The four sets of questionnaires—student, teacher, principal, and national—were designed to collect data about those factors that are known to influence reading achievement and that might vary across nations. These data could best be described in terms of two dimensions: to whom and to what they referred. In the case of the who dimension, the data describe students, their families, their teachers, and their schools. On the what dimension, the data describe their attributes, the kinds of environments provided, the forms of instruction used, and the reading behaviors they exhibited.

Student Questionnaires included items on student/parent background information such as parent's educational level, language spoken at home, student reading activities, etc. There were separate questionnaires for 4th and 9th graders.

Teacher Questionnaires were used to collect information on school and classroom policy, instructional approaches used by the teacher, and the teacher's educational background and experience.

School Questionnaires were completed by the school principal or person designated by the school principal on school demographics, school policies and resources, and evaluation of instruction. One questionnaire was to be obtained from each participating school.

The *National Questionnaire*, completed by the national research team, was used to collect data about the national system, and requested data on standard demographic characteristics, available resources, and practices related to reading achievement.

Periodicity

The IEA Reading Literacy Study was conducted in 1991. The Progress in Reading Literacy Study (PIRLS) was administered in 2001 and tested just 4th-grade students.

2. USES OF DATA

Beyond the usual reporting of reading literacy in NCES compendia (e.g., *Digest of Education Statistics*, *Youth Indicators)*, NCES released four volumes concerning the IEA Reading Literacy Study. These include a technical report, a methodological report, a summary of findings, and a set of collected papers. Among the issues discussed in these reports are sampling for international comparative studies in education, the development and interpretation of reading literacy scales, the study of various effects (e.g., classroom, school, community, family) on reading literacy, and instructional practice in teaching reading.

3. KEY CONCEPTS

Some of the key concepts related to the IEA Reading Literacy Study are described below.

Types of text. Scaled scores were developed to reflect students' understanding of three types of text:

Narrative prose. Continuous text materials in which the writer's aim was to tell a story, whether fact or fiction. They are normally designed to entertain or involve the reader emotionally; they are written in the past tense, and usually have people or animals as their main theme;

Expository prose. Continuous text materials designed to describe or explain something. The subjects of such text are usually things, but they may be written in the present or the past; the style is typically impersonal, highlighting such features as definitions, causes, classifications, functions, contrasts, and examples, rather than a moving plot with climax; and

Documents. Structured tabular texts, such as forms, charts, labels, graphs, lists, and sets of instructions where the reading requirements typically involve locating information or following directions, rather than continuous reading of connected text.

4. SURVEY DESIGN

Target Population

Within each of the participating countries, nationally representative samples were to be drawn based on two internationally defined target populations: (1) *Population A*: All students attending school on a full-time basis at the grade level in which most students 9 years old (during the 1st week of the 8th month of the school year) are enrolled; and (2) *Population B*: All students attending school on a full-time basis at the grade level in which most students 14 years old (during the 1st week of the 8th month of the school year) are enrolled.

Within the United States, these definitions were implemented and modified in the following ways: (1) *Population A*: All students attending school on a full-time basis at the grade 4 level in the 50 states and the District of Columbia, during the 1990–91 school year, who, in the opinion of school personnel, are capable of taking the test; and (2) *Population B*: All students attending school on a full-time basis at the grade 9 level in the 50 states and the District of Columbia, during the 1990–91 school year, who, in the opinion of school personnel, are capable of taking the test.

A number of practical sampling issues in the United States necessitated some additional departures from the procedures proposed in the IEA sampling manual (Ross 1991). First, because the geographic dispersion of schools made it fiscally impossible to consider collecting data from a stratified random sample of schools, the sample size was increased to offset the additional clustering effects introduced by the three-stage sampling frame designed to facilitate data collection. Second, because the United States lacks a single set of national policies that would control such factors as entrance age, retention in grades, and placement in mainstream classes, study designers in the United States could not identify a single grade with a clean majority of the target population. Hence, the national target population was defined so that the modal grade for each desired age group was chosen. These modal grades contained more than 50 percent (i.e., a majority) of students of the relevant age in each case.

Sample Design

The sample for the IEA Reading Literacy Study was selected using a complex multistage clustered design involving the sampling of intact classes from selected schools within selected geographic areas, called primary sampling units (PSUs), across the United States.

The structure of the sampling design differed somewhat from the models suggested by the international referee (Ross 1991). The United States adopted the approach, approved by the referee, of arranging for personnel from outside the school system to administer the assessments. This approach was taken to maximize school participation by minimizing the burden on schools and to assist in maintaining uniformly high standards of assessment administration throughout the sample by using field workers who were trained as a group by study staff. In most other countries, school personnel administered the assessments in the interest of minimizing costs.

The basic U.S. sample plan called for sampling intact classrooms and/or classes. For grade 4, if a sample school had fewer than an estimated 50 4th-grade students, all were included. In schools with 50 or more 4th graders, two classrooms were taken at random. For grade 9, in schools with fewer than an estimated 25 9th-grade students, all were included. Otherwise, the plan called for taking one classroom (typically, the language arts class). The number of students in the grade was estimated by dividing the total enrollment, as reported on the 1989 Quality Education Data (QED) file, by the grade span of the school.

The multistage sampling process for the IEA Reading Literacy Study involved the following steps:

(1) Selection of PSUs

(2) Selection of schools (public and nonpublic) within the selected PSUs

(3) Selection of intact classrooms and/or classes within the selected schools

Selection of PSUs. In the first stage of sampling, the United States (the 50 states and the District of Columbia) were divided into the geographic PSUs used by the National Assessment of Educational Progress (NAEP), which are counties (or independent cities) and groups of counties with a minimum population of 60,000 as of the 1980 Census. The counties composing metropolitan areas are kept together; other aggregations avoid mixing urban and rural counties. Since IEA specifications did not require certain estimates by subgroups (such as minorities) that were mandated by NAEP, the NAEP PSUs were restratified for use in the IEA study. The first level stratification was by NAEP region (four geographic strata) and two degrees of urbanization strata (Metropolitan Statistical Area—MSA—and non-MSA). In addition, the Southeast and West regions were stratified by percent minority, those with less than 20 percent minorities in one class and those with 20 percent or more in another.

Fourteen PSUs were of sufficiently large size that it was appropriate to include them in the sample with certainty. Minorities (outside of the large cities, included with certainty) are relatively less prevalent in the Northeast and the Central regions, so the minority stratification was not used in those regions. The high minority, non-MSA stratum in the West contained so few PSUs that it was combined with the low minority, non-MSA stratum. It was possible to subdivide them by percent minority in the second stage of stratification.

A sample of 50 PSUs in total was drawn according to the above allocation. Sampling weights equal to the inverse of the probabilities of selection were attached to them.

Selection of schools. The schools in the sampled PSUs were extracted from the QED file and were substratified by stage II strata. The two stage II stratifying variables were type of control (public schools in one class; private schools in the other class) and enrollment in the 4th grade for Population A or the 9th grade for Population B.

The schools were put into three classes at Population A and two classes at Population B on the basis of their estimated grade enrollment. A relatively thin sample of small schools was drawn to increase the efficiency of the design, since the per-student assessment costs for such schools were high. This had the effect of increasing the weights of these schools so that their effect on national projections was proportionate to the total enrollment of the stratum.

The sample of 200 schools from each population was allocated to the deeply stratified universe in proportion to the number of students in the given grade projected from the sampled PSUs, since, at the time the sample was drawn, total counts for the universe were not available in time to meet the deadline for the design work. This required a later adjustment in the sampling weights, as is discussed later in this section.

As required by the sampling referee, checks were made on the selected sample of schools and their base weights to ensure that the samples had been drawn without error. By stratum, the weighted measures of size of the selected schools were summed and then compared with the total of the measures of size for the stratum. They agreed exactly in each case, as was appropriate.

Selection of intact classroom and/or classes. As schools agreed to participate in the IEA study, they were sent a Fourth/Ninth Grade Class List Form asking for names and identifying information for all eligible classes within that school. This Class List Form was used to select the sample of the class(es) participating in the study.

Data Collection and Processing

The National Center for Education Statistics (NCES) began its efforts to gain support for the IEA Reading Literacy Study through presentations to the Council of Chief State School Officers' (CCSSO) Education Information Advisory Council (EIAC). EIAC endorsed the study and encouraged its members to participate fully in all activities.

According to the specifications of IEA, those who would conduct the Reading Literacy Study should first obtain permission to test in the schools. In the United States, because the school system is decentralized and locally autonomous, this requirement necessitated adherence to a protocol of contacting several levels of government officials: chief state school officers, local district superintendents, building principals, and classroom teachers.

The IEA Reading Literacy Study was administered by Westat, under a contract administered by NCES. Westat selected the schools in the sample and made the necessary contacts with state, district, and school administrators to obtain permissions to test in these schools. It also recruited, trained, and supervised the field assessment staff, and received the completed materials.

Reference dates. Data for the IEA Reading Literacy Study were collected in February and March, 1991.

Data collection. The ICC specifications permitted participating countries to choose field administrators from a range of categories, including classroom teachers, school administrators, and nonschool personnel. The U.S. study team felt that the study would be better served by creating a field staff that was in no way associated with the schools themselves. The primary benefit would be that the assessment administrators could be trained together and would subsequently administer the test to all students in a standardized manner. In addition, using study staff rather than school personnel would reduce the burden of response and might thereby increase the rate of participation.

Subsequently, Westat hired and trained a field staff of 45 assessment administrators and two supervisors to administer and collect the data. Each assessment administrator met with a coordinator at each school to schedule the assessments and make appropriate arrangements. At this time, it was also determined which students appearing on the class roster should be identified as "excluded" on the Administration Schedule. For this study, a student was excluded from the assessment only for the following two reasons: (1) a student was enrolled in a special education program and had an Individual Educational Plan (IEP) that specifically prohibited pencil-and-paper assessment; or (2) a student was non-English speaking and had been enrolled in a mainstream English class for less than 2 years. In total, 183 students were excluded from the grade 4 sample and 18 students from the grade 9 sample.

Each set of classroom sessions involved approximately 25 students, each of whom completed the Reading Literacy Test and the Student Questionnaire.

Data was collected on approximately 7,200 students in the 4th grade and 3,800 students in the 9th grade, with 167 schools participating at grade 4 and 165 at grade 9. Both public and private schools were included, distributed in 227 districts across 31 states and the District of Columbia. Three hundred 4th-grade and 160 9th-grade teachers also provided data for the study, as did 332 school administrators.

Data processing. Those materials returned directly to Westat included the School Questionnaire, the Teacher

Questionnaire, and the Student Questionnaires. The assessment administrators sent the Reading Literacy Tests to Data Recognition Corporation (DRC) for coding, keying, verifying, and basic editing.

The data keying at Westat used a 100 percent verification system. All data were entered twice by different operators and then compared. Any differences were resolved, with the supervisor adjudicating difficult cases. After keying, additional machine editing was used to detect and resolve range and logic errors.

DRC had two major tasks: key entry of the responses to the Reading Literacy Test items (DRC also used a 100 percent verification system), and scoring the open-ended writing responses included in the Reading Literacy Tests. Each essay was read by two readers independently and scored; if the scores differed, a third resolving reading was done by a task leader. Scoring was monitored closely, with daily reports produced for each reader indicating the number of papers read, the percentage of exact, adjacent, and nonadjacent agreement with the other readers of the same papers, the tendency of the disagreement, and the score point distribution. The area of scrutiny was inconsistency, or drift from an established standard. Throughout the project, readers scored sample papers at rangefinding meetings in order to validate and recalibrate the criteria. Retraining was ongoing to secure continued familiarity with and adherence to the scoring criteria and to prevent roomwide drift as the project progressed. Legibility issues were addressed implicitly in the open-ended question scoring process.

The scorers of the open-ended items were experienced in scoring similar questions for other large-scale assessments. They were generally high school teachers who were provided training for scoring open-ended questions for this study.

Editing. The first phase of data editing took place during the keying of the questionnaires and literacy assessments. The 100 percent verification process required all data to be entered twice by different operators and then compared. Discrepancies were corrected, and in the case of difficult cases, were adjudicated by the supervisor.

In the second phase of data editing, a machine-edit program was used to detect and resolve as many errors as possible prior to delivering the data for more complex interfile editing and statistical data quality analyses. The errors detected by machine editing were of two general types: (1) range errors, in which response values fell outside a predetermined acceptable range; and (2) logic errors, in which there were some inconsistencies between response values. These included improperly followed skip patterns, data inconsistencies among two or more variables, and addition checks where values of a group of variables were to sum to a known value.

Creating the files. The study produced eight U.S. files in all. Two were reading test data for each population. In addition, a file was created for each population for the Student, Teacher, and School Questionnaires.

These eight U.S. files were combined and reformatted in accordance with the specifications provided by ICC to produce six ICC international format files. The U.S. Teacher and School Questionnaire files were mapped onto ICC versions; the U.S. Student Questionnaire and Reading Literacy Test files were mapped onto a single ICC student file for each population. While only a few of the questions in the U.S. questionnaires were asked with the same wording and response alternatives as their analogues in the ICC version, the data, nonetheless, were to go to the ICC in the format of its questionnaires.

The ICC supported its questionnaires with software for data entry, record editing, range checks, ID checks across files, and logic and consistency checks, including skip patterns and intra- and interfile checks. When the data were converted to ICC format and these checking programs were run, almost all of the errors occurred in cases where a prescribed range was violated by a legitimate, if unusual, value, or a consistency check was violated by a combination of such values. Essentially the data did not require further editing in order to conform to ICC standards.

As part of the agreement to participate in IEA Reading Literacy Study, each participating country, including the United States, had granted IEA permission to release its data to individuals or organizations desiring to perform secondary analyses. To avoid disclosure problems, the U.S. files submitted to IEA were considered public use data files, and extensive analyses were performed to ensure that individual respondents could not be identified.

Estimation Methods

Once IEA data were scored and compiled, the responses were weighted according to the sample design and population structure and then adjusted for nonresponse. This ensured that the students' representation in the IEA Reading Literacy Study matched their actual proportion in the school population for the grades assessed.

Weighting. The weighting of the national IEA sample reflected the probability of selection for each student in the sample, adjusted for nonresponse. The weight assigned to a student's responses was the inverse of the probability that the student would be selected for the sample. Through poststratification, weighting ensured that the representation of certain subpopulations corresponded to figures from the Current Population Survey (CPS) and also accounted for the low sampling rates that occurred for very small schools. Thus, properly weighted IEA data provided results that reflect the representative performances of the entire nation and of the subpopulations of interest. The following provides an overview of the steps involved in deriving the sampling weights.

Applying the secondary stratification only to the schools in the initial sample of NAEP PSUs, after weighting the characteristics of the schools in the sampled PSUs by the inverse of the probabilities of selection of those PSUs, introduced sampling error in the estimates of the substratum totals. Since the time that the design was set, it has been possible to tabulate the entire QED file by the characteristics that define the substrata. This made it possible to adjust the sample weights so that the number of schools in the selected sample would weight up to the number of schools in the QED tape within each substratum—a straightforward poststratification procedure.

The enrollments in the sampled schools were multiplied by the school weights and compared with estimated enrollments for the 4th and 9th grades produced by the CPS. The differences were judged to be large enough that a second adjustment to the sampling weights was made so that the estimated enrollments in the two grades would equal the CPS estimates within each NAEP region.

The two weight adjustments automatically corrected for school nonresponse to the survey. In making the first adjustment, the weighted number of sampled schools was adjusted to equal the number of schools listed in the QED file, with no account taken of the number of schools that had closed.

The student weights within each school reflected both the subsampling of classrooms in the school and the individual student nonresponse within the school. That is, the school weight was multiplied by the number of classrooms in the school and divided by the number of classrooms sampled. This weight was multiplied by the number of students in the selected classrooms and divided by the number of responding students to produce the student weights.

Scaling. For purposes of summarizing item responses, the ISC developed procedures for creating international scaled scores based on the Rasch model, the one-parameter item response theory (IRT) model. The underlying principle of IRT is that, when a number of items require similar skills, the regularities observed across patterns of response can often be used to characterize both respondents and tasks in terms of a relatively small number of variables.

The ICC performed all tasks related to scaling of the Reading Literacy Tests (i.e., calibrated items and estimated student abilities). Calibration of items and estimation of abilities were performed separately for each of the three reading literacy domains (narrative, expository, and document). Item difficulties were estimated on the basis of responses of a random sample of students selected from all participating countries. This international calibration sample consisted of 10,790 students for grade 4 and 10,772 for grade 9.

The ICC deleted a total of six items for grade 4 and seven items for grade 9 that did not fit the international calibration sample. Rasch analysis was performed within each participating country, setting the item difficulties derived on the international calibration sample as known parameters. Item fit was also examined within each participating country. If an item was found not to fit the Rasch model in a given country, that item was not included in estimating student abilities within the country under consideration. Based on the invariance properties of the Rasch model (i.e., examinee ability estimation is independent of the particular set of items administered from a calibrated pool), the ICC derived reading literacy ability estimates for students within each participating country and placed them on a common scale. For ease of use, the logit scale was transformed such that the international mean and standard deviation were 500 and 100, respectively, for each reading literacy domain.

Since the international mean and standard deviation were arbitrarily set, the scale scores across the domains are not equated. Similarly, the scale scores across the two populations are not equated either.

Imputation. The IEA study employed a combination of a hot-deck imputation procedures and deterministic imputations to assign values for missing responses for the data items. Hot-deck (using Wesdeck) imputation procedures were used to handle missing responses for most items. For some of the remaining items, the missing responses were completed from information available in other data sources; for some items, it was possible to deduce the missing response from the responses to other items on the questionnaire; and for other items, the overall modal response for respondents was assigned for all missing responses. The latter technique, which was employed for operational expediency, was used only when the item nonresponse rate was very small.

Future Plans

The IEA plans to continue its study of reading literacy through PIRLS, an assessment of 4th graders on a recurring basis.

5. DATA QUALITY AND COMPARABILITY

The U.S. component of the IEA Reading Literacy Study had to report accurate results for populations of students and subgroups of these populations (e.g., minority students or students attending nonpublic schools). Although only a very small percentage of the student population in each grade were assessed, IEA Reading Literacy Study estimates are accurate because they depend on the absolute number of students participating, not on the relative proportion of students.

Every activity in IEA Reading Literacy Study assessments was conducted with rigorous quality control. All questions underwent 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 IEA Reading Literacy Study data were collected and processed was monitored closely. Westat ensured uniformity of procedures through training, supervision, and quality control monitoring. (See section 4 for more detail on quality control procedures.)

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

Sampling Error

The primary component of uncertainty in the IEA Reading Literacy Study is due to sampling only a small number of students relative to the whole population. This accounts for the variability of estimates of percentages of students having certain background characteristics or answering a certain cognitive question correctly.

Because the IEA Reading Literacy Study used complex sampling procedures, a jackknife replication procedure was used to estimate standard errors. A set of jackknife replicate weights was developed for each assessed student.

Because of the effects of clustering and unequal probabilities of selection in the IEA Reading Literacy Study, in most cases the design effect is greater than 1. This means that the sample design is generally less efficient than simple random sampling, although it is more cost-effective.

Nonsampling Error

While there is the possibility of some coverage error in the IEA Reading Literacy Study, 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 the IEA Reading Literacy Study, coverage error could result from either the sampling frame of schools being incomplete or from the schools' failure to include all the students on the lists from which grade samples were drawn. The IEA Reading Literacy Study, while conducted in 1991, used the 1989 QED school list for the names of the regular public and private schools. This list, however, did not include schools that opened between 1989 and the time of the 1991 IEA Reading Literacy Study. The weighting adjustment for school nonresponse to the survey considered schools closed between 1989 and 1991 as nonresponding schools. Apparently there was no check by the assessment administrators to verify the inclusion of all students on the lists provided them.

Nonresponse error. Unit nonresponse error results from nonparticipation of schools and students. Item nonresponse error results from students who participate but do not answer every question.

Unit nonresponse. The unweighted school response rate across public and private sectors was 87 percent for the grade 4 schools and 86 percent for the grade 9 schools. These rates exceeded the international requirement of at least 85 percent for each grade. At the student level, about 7 percent of the grade 4 students and 14 percent of the grade 9 students were unit nonrespondents. Weighting class adjustments were used to compensate for unit nonresponse at both the school and student levels. There were responses from all teachers and administrators (100 percent response rate) on the teacher and administrator questionnaires, so no adjustments were necessary to compensate for unit nonresponse on these two sets of data.

Item nonresponse. Item nonresponse to the questionnaire items occurred when a student who completed the reading performance test failed to complete an item on the student background questionnaire, or when a teacher or principal failed to complete an item on the questionnaires that they completed. The level of item nonresponse was generally low, but some items were not answered by 10 percent or more of the respondents.

Data Comparability

Since the IEA Reading Literacy Study was by definition an international study involving 32 countries, it allows comparisons between participating countries. Additionally, the results of the IEA Reading Literacy Study should be comparable with those of the NAEP Reading assessments. Trend comparisons are available through PIRLS.

Comparisons with other countries. In contrast to the poor showing of American students in other international comparisons, in reading, at least, American students were among the best of the 32 nations involved in the study. With the exception of Finland, no country consistently outperformed the United States. It should be noted that these 32 nations are a self-selected group that are neither a representative sample of all nations nor of our principal trading partners (e.g., Japan, the United Kingdom, and Mexico were not included). However, among these are 18 members of the Organization for Economic Cooperation and Development (OECD), and the average of the OECD countries is a benchmark against which measurements of the overall American performance, as well as particular American subpopulations, can be compared. This has been done in the NCES report *Reading Literacy in the United States: Findings from the IEA Reading Literacy Study* (NCES 96–258). The NCES report *Reading Literacy in an International Perspective: Collected Papers from the IEA Reading Literacy Study* (NCES 97–875) contains nine papers addressing issues regarding reading literacy, focusing on outcomes in literacy achievement, instructional practices in reading, and school climate. Several of these papers limit their analysis to a nine-country focus of eight European nations and the United States.

Comparisons with NAEP Reading assessments. The finding that the results of the IEA study were more optimistic in their portrayal of the reading proficiency of American students than the results of the NAEP assessments has generated additional study comparing the two assessments in an effort to determine the reason for these differences. (See chapter 20.)

Comparisons with PIRLS. The PIRLS data collection was scheduled for 2001 to coincide with the 10th anniversary of the IEA Reading Literacy Study to provide an opportunity for countries that participated in the earlier study to obtain a measure of change from 1991. The United States was among the countries that participated in the PIRLS trend study, in which the 1991 test and student questionnaire were administered to a sample of PIRLS students.

Content changes. For PIRLS in 2001, the general thrust of the assessment was the same, although the frameworks were modified and new test items were developed.

Design changes. Given that a large number of countries which are participating in PIRLS are also participating in the OECD Program for International Assessment (PISA), the older cohort has been eliminated. Only one age/grade level is being tested.

6. CONTACT INFORMATION

For content information on the IEA Reading Literacy Study, contact:

Eugene Owen
Phone: (202) 502–7422
E-mail: eugene.owen@ed.gov

Mailing Address:

National Center for Education Statistics
1990 K Street NW
Washington, DC 20006–5651

7. METHODOLOGY AND EVALUATION REPORTS

General

Reading Literacy in the United States: Technical Report, NCES 94–259, by M. Binkley and K. Rust (eds). Washington, DC: 1994.

Survey Design

Sampling Manual for the IEA International Study of Reading Literacy, by K.N. Ross. University of Hamburg, Hamburg, Germany: International Coordinating Center, IEA International Study of Reading Literacy, 1991.

Data Quality and Comparability

Methodological Issues in Comparative Educational Studies: The Case of the IEA Reading Literacy Study, NCES 94–469, by M. Binkley, K. Rust, and M. Winglee (eds). Washington, DC: 1995.

Chapter 23: National Adult Literacy Survey (NALS)

1. OVERVIEW

The National Adult Literacy Survey (NALS) was initiated to fill the need for accurate and detailed information on the English literacy skills of America's adults. In accordance with a congressional mandate, it provides the most detailed portrait that has ever been available on the condition of literacy in this nation—and on the unrealized potential of its citizens.

The 1992 National Adult Literacy Survey is the third and largest assessment of adult literacy funded by the federal government and conducted by the Educational Testing Service (ETS). The two previous efforts were: (1) the 1985 Young Adult Literacy Assessment (funded as an adjunct to the National Assessment of Educational Progress—see chapter 20); and (2) the Department of Labor's 1990 Workplace Literacy Survey. Building on these two earlier surveys, literacy for the NALS is defined along three dimensions—prose, document, and quantitative—designed to capture an ordered set of information-processing skills and strategies that adults use to accomplish a diverse range of literacy tasks encountered in everyday life. The background data collected in NALS provide a context for understanding the ways in which various characteristics are associated with demonstrated literacy skills.

NALS is the first national study of literacy for *all* adults since the Adult Performance Level Surveys conducted in the early 1970s. It is also the first in-person literacy assessment involving the prison population. A second adult literacy survey, the National Assessment of Adult Literacy (NAAL), is planned for 2003.

Purpose

To (1) evaluate the English language literacy skills of adults (16 years and older) living in households or prisons in the United States; (2) relate the literacy skills of the nation's adults to a variety of demographic characteristics and explanatory variables; and (3) compare the results with those from the 1985 Young Adult Literacy Assessment and the 1990 Workplace Literacy Survey.

Components

The 1992 survey consisted of one component that was administered to three different representative samples: a national household sample; supplemental state household samples for 12 states (California, Florida, Illinois, Indiana, Iowa, Louisiana, New Jersey, New York, Ohio, Pennsylvania, Texas, Washington); and a national sample of federal and state prison inmates. Responses from the national, state, and prison samples were combined to yield the best possible performance estimates.

PERIODIC SURVEY OF A SAMPLE OF ADULTS LIVING IN HOUSEHOLDS OR PRISONS:

Assesses literacy skills:

- Prose
- Document
- Quantitative

Collects background data on:

- Demographics
- Education
- Labor Market Experiences
- Income
- Activities

National Adult Literacy Survey. The 1992 survey assessed the literacy skills of a representative sample of the U.S. adult population using simulations of three kinds of literacy tasks that adults would ordinarily encounter in daily life (prose, document, and quantitative literacy). The data were collected through in-person interviews with adults who were living in households, or federal or state prisons. Adults were defined as individuals 16 years or older for the national and prison samples, and 16 to 64 years of age for the state samples. In addition to the cognitive tasks, the personal interview gathered information on demographic characteristics, language background, educational background, reading practices, and labor market experiences. To ensure comparability across all samples, the literacy tasks assessed were the same for all three samples. Background data varied somewhat between the household and prison samples—labor force questions were irrelevant to prisoners, and questions about criminal behavior and sentences were relevant only to prisoners.

Literacy assessment. The pool of literacy tasks used to measure adult proficiencies consisted of 165 literacy questions—41 prose, 81 document, and 43 quantitative. To ensure that valid comparisons could be made by linking the scales to those of the 1985 Young Adult Literacy Assessment, 85 tasks from that survey were included in the 1992 survey. An additional 80 new tasks were developed specifically to complement and enhance the original 85 tasks. The literacy tasks administered in NALS varied widely in terms of materials and content. The six major context/content areas were: home and family; health and safety; community and citizenship; consumer electronics; work; and leisure and recreation. Each adult was given a subset (about 45) of the total pool of assessment tasks to complete. Each of the tasks extended over a range of difficulty on the three literacy scales. The new tasks were designed to simulate the way in which people use various types of materials and to require different strategies for successful performance.

The responses to the literacy assessment were pooled and reported by proficiency scores, ranging from 0 to 500, on three separate scales, one each for prose, document, and quantitative literacy. By examining the overall characteristics of individuals who performed at each literacy level on each scale, it is possible to identify factors associated with higher or lower proficiency in reading and using prose, documents, and quantitative materials.

Background information. Background information collected for the state and household samples included data on *background and demographics*—country of birth, languages spoken or read, access to reading materials, size of household, educational attainment of parents, age, race/ethnicity, and marital status; *education*—highest grade completed in school, current aspirations, participation in adult education classes, and education received outside the country; *labor market experiences*—employment status, recent labor market experiences, and occupation; *income*—personal and household; and *activities*—voting behavior, hours spent watching television, frequency and content of newspaper reading, and use of literacy skills for work and leisure. Respondents from each of the 12 participating states were also asked 5 state-specific questions.

To address issues of particular relevance to the prison population, a separate background questionnaire was developed for the prison sample. This instrument drew questions from the 1991 Survey of Inmates of State Correctional Facilities, sponsored by the Department of Justice's Bureau of Justice Statistics. The background questionnaire for the prison population addressed the following major topics: general and language background; educational background and experience; current offenses and criminal history; prison work assignments and labor force participation prior to incarceration; literacy activities and collaboration; and demographic information.

Periodicity

NALS was conducted in 1992. A second adult literacy study is scheduled for 2003.

2. USES OF DATA

Results from NALS provide the most detailed portrait that has ever been available on the condition of literacy in this nation and on the unrealized potential of its citizens. NALS data provide vital information to policymakers, business and labor leaders, researchers, and citizens. The survey results can be used to:

- describe the levels of literacy demonstrated by the adult population as a whole and by adults in various subgroups (e.g., those targeted at risk, prison inmates, and older adults);
- characterize adults' literacy skills in terms of demographic and background information (e.g., reading characteristics, education, and employment experiences);
- profile the literacy skills of the nation's workforce;
- compare assessment results from the current study with those from the 1985 Young Adult Literacy Assessment;

- interpret the findings in light of information-processing skills and strategies, so as to inform curriculum decisions concerning adult education and training; and
- increase understanding of the skills and knowledge associated with living in a technological society.

3. KEY CONCEPTS

Some of the key concepts related to the literacy assessment are described below. See the NALS Electronic Codebook or appendices of NALS reports for lists and descriptions of variables.

Literacy. The ability to use printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential. This definition goes beyond simply decoding and comprehending text to include a broad range of information-processing skills that adults use in accomplishing the range of tasks associated with work, home, and community contexts.

Prose Literacy. The ability to locate information contained in expository or narrative prose in the presence of related but unnecessary information, find all the information, integrate information from various parts of a passage of text, and write new information related to the text. Expository prose consists of printed information in the form of connected sentences and longer passages that define, describe, or inform, such as newspaper stories or written instructions. Narrative prose tells a story, but is less frequently used by adults in everyday life than by school children, and did not occur as often in the text presented in NALS as prose literacy tasks. Prose varies in its length, density, and structure.

Document Literacy. The ability to locate information in documents, repeat the search as many times as needed to find all the information, integrate information from various parts of a document, and write new information as requested in appropriate places in a document, while screening out related but inappropriate information. Documents differ from prose text in that they are more highly structured. Documents consist of structured prose and quantitative information in complex arrays arranged in rows and columns, such as tables, data forms, and lists (simple, nested, intersected, or combined); in hierarchical structures, such as tables of contents or indexes; or in two-dimensional visual displays of quantitative information, such as graphs, charts, and maps.

Quantitative Literacy. The ability to use quantitative information contained in prose or documents (specifically the ability to locate quantities while screening out related but unneeded information), repeat the search as many times as needed to find all the numbers, integrate information from various parts of a text or document, infer the necessary arithmetic operation(s), and perform arithmetic operation(s). Quantities can be located in either prose texts or in documents. Quantitative information may be displayed visually in graphs, maps, or charts, or it may be displayed numerically using whole numbers, fractions, decimals, percentages, or time units (hours and minutes).

Literacy Scales. Three scales used to report the results for prose, document, and quantitative literacy. These scales, each ranging from 0 to 500, are based on those established for the 1985 Young Adult Literacy Assessment. The scores on each scale represent degrees of proficiency along that particular dimension of literacy. The literacy tasks administered in the 1992 survey varied widely in terms of materials, content, and task requirements, and thus in difficulty. A careful analysis of the range of tasks along each scale provides clear evidence of an ordered set of information-processing skills and strategies along each scale. To capture this ordering, each scale was divided into five levels that reflect this progression of information-processing skills and strategies: Level 1 (0 to 225), Level 2 (226 to 275), Level 3 (276 to 325), Level 4 (326 to 375), and Level 5 (376 to 500). Level 1 comprised those adults who could consistently succeed with Level 1 literacy tasks but not with Level 2 tasks, as well as those who could not consistently succeed with Level 1 tasks and those who were not literate enough in English to take the test at all. Adults in Levels 2 through 4 were consistently able to succeed with tasks at their level but not with the next more difficult level of tasks. Adults in Level 5 were consistently able to succeed with Level 5 tasks.

Succeed Consistently. Indicates that a person at or above a given level of literacy has at least an 80 percent chance of correctly responding to a particular task. This 80 percent criterion is more stringent than the 65 percent standard used in the National Assessment of Educational Progress (NAEP—see chapter 20) for measuring what school children know and can do.

4. SURVEY DESIGN

The 1992 NALS was designed and administered by the Educational Testing Service (ETS). A subcontract was awarded to Westat, Inc. for sampling and field data collection. A committee of experts from business and industry, labor, government, research, and adult education worked with the ETS staff to develop the definition of literacy that underlies NALS, as well as to prepare the assessment objectives that guided the selection and construction of assessment tasks. In addition to this Literacy Definition Committee, a Technical Review Committee was formed to help ensure the soundness of the assessment design, the quality of the data collected, the integrity of the analyses conducted, and the appropriateness of the interpretations of the final results. The prison survey was developed in consultation with the Bureau of Justice Statistics and the Federal Bureau of Prisons. The survey design for the 1992 survey is described below.

Target Population

The target population for the national household sample consisted of adults 16 years and older in the 50 states and the District of Columbia who, at the time of the survey, resided in private households or college dormitories. The target population for the supplemental state household sample consisted of individuals 16 to 64 years of age who, at the time of the survey, resided in private households or college dormitories in the participating state (California, Florida, Illinois, Indiana, Iowa, Louisiana, New Jersey, New York, Ohio, Pennsylvania, Texas, or Washington). Individuals residing in other institutions—nursing homes, group homes, or psychiatric facilities—were not included in the household samples. The target population for the prison sample consisted of adults 16 years or older who were in state or federal prisons at the time of the survey; those held in local jails, community-based facilities, or other types of institutions were not included.

Sample Design

Because this 1992 survey was designed to provide data representative at the national level (including prison inmates) and at the state level for participating states, it included three different samples: a national household sample, supplemental state household samples for 12 states, and a supplemental national sample of state and federal prison inmates.

Household samples. The sample design for the national and state household samples involved a four-stage stratified area sample: (1) the selection of primary sampling units (PSUs) consisting of counties or contiguous groups of counties; (2) the selection of segments (within the selected PSUs) consisting of census blocks or groups of contiguous census blocks; (3) the selection of households within the segmented samples; and (4) the selection of age-eligible individuals within each selected household. The sample design requirements called for an average cluster size of seven interviews (i.e., seven completed background questionnaires per segment). In addition, a reserve sample at the household level of approximately 5 percent of the size of the main sample was selected and set aside in case of shortfalls due to unexpectedly high vacancy and nonresponse rates.

One national area sample was drawn for the national household sample, and 12 independent state-specific area samples were drawn from the 12 states participating in the supplemental state samples. The sample designs used for all 13 samples were similar, with one major difference. In the national sample, Black and Hispanic respondents were sampled at about double the rate of the remainder of the population to assure reliable estimates of their literacy proficiencies, whereas the state samples used no oversampling.

The first stage of sampling involved the selection of PSUs. A national sampling frame of 1,404 PSUs was constructed primarily from 1990 census data, stratified on the basis of region, metropolitan status, percent Black, percent Hispanic, and whenever possible, per capita income. Using this frame, 101 PSUs were selected for the national sample. The national frame of PSUs, subdivided at state boundaries if needed, was used to construct individual state frames for the supplemental state sample; a sample of 8 to 12 PSUs was selected within each of the given states. All PSUs were selected with probability proportional to the PSU's 1990 population.

The second stage of sampling involved the selection of segments within the selected PSUs. The Bureau of Census' Topologically Integrated Geographical Encoding and Referencing (TIGER) System File was used for the production of segment maps. The segments were selected with probability proportional to size where the measure of size for a segment was a function of the number of year-round housing units within the segment. The oversampling of Blacks and Hispanic respondents for the national sample was carried out at the segment level, where segments were classified as high minority (segments more than 25 percent Black or Hispanic population) or not high minority.

The third stage of sampling involved the selection of households within the segmented samples. Westat field staff visited all selected segments in the fall of 1991 and prepared lists of all housing units within the boundaries of each segment as determined by the 1990 census block maps. The lists were used to construct the sampling frame for households. Households were selected with equal probability within each segment, except for White, non-Hispanic households in high minority segments in the national sample, which were subsampled so that the sampling rates for White, non-Hispanic respondents would be about the same overall.

The fourth stage of sampling involved the selection of one or two adults within each selected household during the data collection phase of the survey. One person was selected at random from households with fewer than four eligible members; two persons were selected from households with four or more eligible members. Using a screener, the interviewer constructed a list of age-eligible household members (16 and older for the national sample; 16 to 64 for the state sample) for each selected household. The interviewers, who were instructed to list the eligible household members in descending order by age, then identified one or two household members to interview, based on computer-generated sampling messages that were attached to each questionnaire in advance.

Prison sample. There were two stages of selection for the prison sample. The first stage involved the selection of state or federal correctional facilities. The sampling frame for the correctional facilities was based on the 1990 census of federal and state prisons, updated in mid-1991. The facility frame was stratified prior to sample selection on the basis of type of facility (federal or state prison), region of country, inmate gender composition, and type of security. A sample of 88 facilities and a reserve sample of 8 facilities was then drawn from the frame based on probability proportional to size, where the measure of size for a given facility was equal to the inmate population. The second stage of sampling involved the selection of inmates within each selected facility, using a list of names obtained from the facility administrators. An average of 12 inmates were selected from each facility based on a probability inversely proportional to their facility's inmate population (up to a maximum of 22 interviews in a facility), so that the product of the first and second stage probabilities would be constant.

Assessment Design

Building on the 1985 Young Adult Literacy Assessment and the 1991 Workplace Literacy Survey, the NALS Technical Committee adopted the definition of literacy and the literacy scales—prose, document, and quantitative—used in the previous surveys. The materials were selected to represent a variety of contexts and contents: home and family; health and safety; community and citizenship; consumer electronics; work; and leisure and recreation.

BIB spiraling. The survey design gave each respondent a subset of the total pool of literacy tasks, while at the same time ensuring that each of the 165 tasks was administered to a nationally representative sample of the adult population. The design most suitable for this purpose is a variant of standard matrix sampling called balanced incomplete block (BIB) design.

Literacy tasks were assigned to blocks or sections that could be completed in about 15 minutes, and these blocks were then compiled into booklets so that each block appeared in each position (first, middle, and last) and each block was paired with every other block. Thirteen blocks of simulation tasks were assembled into 26 unique booklets, each of which contained four blocks of tasks: the core (same for all exercise booklets), and three cognitive blocks. Each booklet could be completed in about 45 minutes.

Pretests. A field test of the national household sample was conducted in the spring of 1991 using a sample of 2,000 adults drawn from 16 PSUs. The purposes of the field test were to evaluate the impact of incentives on response rates, performance, and survey costs; to evaluate newly developed literacy exercises for item bias and testing time; and to evaluate the administration and appropriateness of the background questions. As a result of the field test, some of the literacy tasks and their scoring guides were revised or dropped from the final assessment.

For the prison sample, a small pretest was conducted at the Roxbury Correctional Institution in Hagerstown, Maryland. This pretest was designed to evaluate the ease of administration of the survey instruments, survey administration time, within-facility procedures, and inmate reaction to the survey. The pretest demonstrated that several changes to the background questionnaire would facilitate administration. Administrative procedures were also refined to reflect lessons learned during the pretest.

Data Collection and Processing

The survey data were collected through in-person household or prison interviews during the first eight months of 1992. As field operations were completed, the data were shipped to ETS for processing. Further description follows.

Reference dates. Respondents answered the employment status and weekly wages questions for the week before the survey was administered.

Data collection. During January and February of 1992, field interviewers, supervisors, and editors received extensive training both in general and survey-specific interview techniques. The NALS field period began in February 1992, immediately following the completion of the first interviewer training sessions, and lasted 28 weeks, until the end of August. All three survey sample groups were worked simultaneously (except for the state of Florida where data were not collected until 1993). Except for a small, experimental "no incentive" group, all household participants who completed as much of the assessment as their skills allowed received $20 for their time. More than 400 trained interviewers visited about 44,000 households to select and interview almost 31,000 adults. In addition, over 1,147 prison inmates at 87 facilities were interviewed.

Each survey participant was asked to spend approximately one hour responding to survey questions and tasks. Data collection instruments included the screener (designed to enumerate household members and select survey respondents), the background questionnaire, and the literacy exercise booklets. Answering the screener and background questionnaire required no reading or writing skills; to ensure standardized administration, the questions on each were read to respondents in English or Spanish and the answers recorded by the assessment interviewer. Each of the exercise booklets had a corresponding interview guide, with specific instructions to the interviewer for directing the exercise booklet. Reading and writing skills in the English language were required to complete the exercise booklet. When a sampled respondent did not complete any or all of the survey instruments, the interviewer was required to complete a noninterview report form. Field supervisors reviewed the noninterview forms to determine the case's potential for conversion, and the data collected on the form were processed for nonresponse analysis.

Following the completion of an interview, interviewers edited all materials for legibility and completeness. The interviewers sent their completed work to their regional supervisors for a complete edit of the instruments, quality control procedures, and any required data retrieval. As these tasks were completed, the cases were shipped to ETS for processing.

During the data collection process, two special quality control procedures were implemented to identify any households or dwellings missed during the listing phase: the missing structure procedure and the missed dwelling unit procedure. These procedures were used to give these missed structures and dwelling units a chance of selection at time of data collection.

The field effort occurred in three overlapping stages:

(1) *Initial phase*. Each area segment was assigned by the regional supervisor to an interviewer, who followed certain rules in making a prescribed number of calls (a maximum of four was used) to every sampled dwelling in the segment.

(2) *Reassignment phase*. Cases that did not result in completed interviews during the initial phase were reviewed by the regional supervisor, and a subset was selected for reassignment to another interviewer in the same PSU or an interviewer from a nearby PSU.

(3) *Special nonresponse conversion phase*. The home office assembled a special traveling team of the most experienced or productive interviewers to perform a nonresponse conversion effort, under the supervision of a subset of the field supervisors.

Data processing. Coding and scoring staff underwent intensive training prior to the actual coding/scoring. A scoring supervisor monitored both the coding of the questionnaires and the scoring of the exercise booklets. The background questionnaire was designed to be read by a computerized scanning device. Nearly all the simulation tasks contained in the exercise booklet were open-ended; with scoring guides as examples, responses to these items were classified as correct, incorrect, or omitted by trained readers. Responses from the screener and scores from the exercise booklets were transferred to scannable answer sheets. Each survey instrument's scannable forms were batched and sent to the scanning department at regular intervals. As the different instruments were processed, the data were transferred to a database on the main ETS computer for editing.

Editing. Several quality control procedures related to data collection were used during the field operation: an interviewer field edit, a complete edit of all documents by a trained field editor, validation of 10 percent of each interviewer's closeout work, and field observation of both supervisors and interviewers. Additional edits were done during data processing. These included an assessment of the internal logic and consistency of the data received. Discrepancies were corrected whenever possible. The background questionnaires were also checked to make sure that the skip patterns had been followed and all data errors were resolved. In addition, a random set of exercise booklets was selected to provide an additional check on the accuracy of transferring information from booklets and answer sheets to the database.

Estimation Methods

Weighting was used in the 1992 NALS, prior to the calculation of base weights. Responses to the literacy tasks were scored using item response theory (IRT) scaling. A multiple imputation procedure based on plausible values methodology was used to estimate the literary proficiencies of individuals who completed literacy tasks. An innovative approach was implemented to impute missing cognitive data in order to minimize distortions in the population proficiency estimates due to nonresponse to the literacy booklet.

Weighting. Full sample and replicate weights were calculated for survey respondents who completed the exercise booklet; those who could not start the exercises because of a language barrier, a physical or mental barrier, or a reading or writing barrier; and those who refused to complete the exercises but had completed background questionnaires. Demographic variables critical to the weighting were recoded and imputed, if necessary, prior to the calculation of base weights. (See Imputation below.) Separate sets of weights were computed for the incentive and "no incentive" samples.

Household samples. A base weight was computed for each eligible record. The base weight initially was computed as the reciprocal of the product of probabilities of selection for a respondent at the PSU, segment, dwelling unit, and person levels. The final base weight included adjustments to reflect the selection of the reserve sample, the selection of missed dwelling units, and the chunking process conducted during the listing of the segments, and to account for the subsample of segments assigned to the "no incentive" experiment and the subsampling of respondents within households. The base weights for each sample were then poststratified to known 1990 census population totals, adjusted for undercount. This first-level stratification provided sampling weights with lower variation and adjusted for nonresponse. State records were poststratified separately from national records to provide a common base for applying composite weighting factors; population totals were calculated separately for each distinct group.

Composite weights were developed so that NALS data could be used to produce both state and national statistics. For the household samples, a composite weight was computed as the product of the poststratified base weight and a compositing factor which combined the national and state sample data in an optimal manner, considering the differences in sample design, sample size, and sampling error between the two sampled groups. Up to four different compositing factors were used in each of the 11 participating states, and a pseudo factor (equal to one) was used for all persons 65 and older and for all national sample records from outside the 11 participating states.

To compute the final sample weights, the composite weights were adjusted to known 1990 census counts (adjusted for undercount), using a poststratification raking ratio adjustment. The cells used for raking were defined to the finest combination of age, race/ethnicity, sex, education, and geographic indicators (e.g., MSA vs. non-MSA) that the data would allow. Raking adjustment factors were calculated separately for each of the state samples and then for the remainder of the United States.

The above steps used to create the final sample weights were repeated for 60 strategically constructed subsets of the household sample to create a set of replicate weights to be used for variance estimation using the jackknife method.

Prison sample. Base weights for the prison respondents were constructed to be equal to the reciprocal of the product of the selection probabilities for the facility and the inmate within the facility. These weights were then nonresponse-adjusted to reflect both facility and inmate nonresponse. To compute the final sample weights, the resulting nonresponse-adjusted weights were then raked to agree with independent estimates for certain subgroups of the prison population. The above procedures were repeated for 45 strategically constructed subsets of the prison sample to create a set of replicate weights to be used for variance estimation using the jackknife method.

Scaling. Since NALS used a variant of matrix sampling and since different respondents received different sets of tasks, it would be inappropriate to report its results using conventional scoring methods based on the number of correct responses. The literacy assessment results are reported using IRT scaling, which assumes some uniformity in response patterns when items require similar skills. Such uniformity can be used to characterize both examinees and items in terms of a common scale attached to the skills, even when all examinees do not take identical sets of items. Comparisons of items and examinees can then be made in reference to a scale, rather than to the percent correct. IRT scaling also allows the distributions of examinee groups to be compared.

The results of the 1992 literacy assessment are reported on three scales (prose, document, and quantitative) that were established for the 1985 Young Adult Literacy Assessment. Separate IRT linking and scaling were carried out for each of the three domains, using the three-parameter logistic (3PL) scaling model from item response theory. This is a mathematical model for estimating the probability that a particular person will respond correctly to a particular item from a single domain of items. The probability is given as a function of a parameter characterizing the proficiency of that person, and three parameters characterizing the properties of that item. Item parameters needed for the 3PL scaling model were estimated by linking each of the literacy scales used in the 1992 survey to the 1985 Young Adult Literacy Assessment scales.

Imputation. Imputation was performed prior to weighting on missing demographic items considered critical to weighting. Literacy proficiencies of respondents were estimated using a multiple imputation procedure based on plausible values methodology. Missing cognitive data were also imputed.

Demographic data. Demographic variables critical to the weighting (race/ethnicity of the head of household; sex, age, race/ethnicity, and education of the respondent) were recoded and collapsed to required levels, and imputed, if necessary, prior to the calculation of base weights. Data from the background questionnaire were preferred for all items except race/ethnicity of the head of household, which was collected on the screener. For the few cases in which the background questionnaire measure was missing, the screener measure was generally available and was used as a direct substitute. The amount of missing data remaining after substitution was small, making the imputation task fairly straightforward. A standard (random within class) hot-deck imputation procedure was performed for particular combinations of fields that were missing. Imputation flags were created for each of the five critical fields to indicate whether data were originally reported or were based on substitution or imputation. The imputed values were used only for the sample weighting process.

Literacy proficiency estimation (plausible values). A multiple imputation procedure based on plausible values methodology was used to estimate respondents' literacy proficiency in the 1992 NALS. When analyzing the distribution of proficiencies in a group of persons, more efficient estimates can be obtained from a sample design similar to that used in this 1992 survey. Such designs solicit relatively few cognitive responses from each sampled respondent but maintain a wide range of content representation when responses are summed for all respondents.

In the 1992 survey, all proficiency data were based on two types of information: responses to the background questions and responses to the cognitive items. As an intermediate step, a functional relationship between the two sets of information was calculated for the total sample, and this function was used to obtain unbiased proficiency estimates for population groups with reduced error variance. Possible values for a respondent's proficiency were sampled from a posterior distribution that is the product of two functions: the conditional distribution of proficiency given the pattern of background variables, and the likelihood function of proficiency given the pattern of responses to the cognitive items. Since exact matches of background responses are quite rare, NALS used more than 200 principal components to summarize the background information, capturing more than 99 percent of the variance. More detailed information on the plausible values methodology used in the 1992 survey is available in the *Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey* (NCES 2001–467).

Cognitive data. New procedures were implemented in the 1992 NALS to minimize distortions in the population proficiency estimates due to nonresponse to the literacy booklets. When a sampled individual decided to stop the assessment (answered less than five literacy items per scale), the interviewer used a standardized nonresponse coding procedure to record the reason why the person was stopping. This information was used to classify nonrespondents into two groups: (1) those who stopped the assessment for literacy-related reasons (e.g., language difficulty, mental disability, or reading difficulty not related to a physical disability), and (2) those who stopped for reasons unrelated to literacy (e.g., physical disability or refusal). About half of the individuals did not complete the assessment for reasons related to their literacy skills; the other respondents gave no reason for stopping, or gave reasons unrelated to their literacy.

To represent the range of implied causes of missing literacy responses, the imputation procedure selected relied on background variables and self-reported reasons for nonresponse, in addition to the functional relationship between background variables and proficiency scores for the total population. It treated "consecutively missing" data from the literacy booklet instrument differently depending on whether the nonrespondents' reasons were related or unrelated to their literacy skills: (1) those who gave literacy-related reasons were treated as wrong answers, based on the assumption that they could not have correctly completed the literacy tasks, whereas (2) those who gave no reason or cited reasons unrelated to literacy skills for not completing the assessment were essentially ignored (considered not reached), since it could not be assumed that their answers would have been either correct or incorrect. The proficiencies of such respondents were inferred from the proficiencies of other adults with similar characteristics using the plausible values methodology described above.

Future Plans

A second survey, the National Assessment of Adult Literacy (NAAL) is planned for 2003.

5. DATA QUALITY AND COMPARABILITY

The NALS sampling design and weighting procedures assured that participants' responses could be generalized to the population of interest. In addition, NCES conducted special evaluation studies to examine issues related to the quality of NALS. These studies included: (1) a study of the role of incentives in literacy survey research; (2) an evaluation of its sample design and composite estimation; and (3) an evaluation of the construct validity of the adult literacy scales.

Sampling Error

In the 1992 survey, the use of a complex sample design, adjustments for nonresponse, and poststratification procedures resulted in dependence among the observations. Therefore, a jackknife replication method was used to estimate the sampling variance. The mean square error of replicate estimates around their corresponding full sample estimate provides an estimate of the sampling variance of the statistic of interest. The replication scheme was designed to produce stable estimates of standard errors for national and prison estimates as well as for the 12 individual states.

The advantage of compositing the national and state samples during sample weighting was the increased sample size, which improved the precision of both the state and national estimates. However, biases could be present because the national PSU sample strata were not designed to maximize the efficiency of state-level estimates.

Nonsampling Error

The major source of nonsampling error in the 1992 NALS was nonresponse error; special procedures were developed to minimize potential nonresponse bias based on how much of the survey the respondent completed. Other possible sources of nonsampling error were random measurement error and systematic error due to interviewers, coders, or scorers.

Coverage error. Coverage error could result from either the sampling frame of households or prisons being incomplete or from a household's or prison's failure to include all adults 16 years and older on the lists from which the sampled respondents were drawn. Special procedures and edits were built into NALS to review both listers' and interviewers' ongoing work and to give any missed structures and/or dwelling units a chance of selection at data collection. However, just as all other household personal interview surveys have persistent undercoverage problems, the 1992 survey had problems in population coverage due to interviewers not gaining access to households in dangerous neighborhoods, locked residential apartment buildings, and gated communities.

Nonresponse error.

Unit nonresponse. Since three survey instruments—screener, background questionnaire, and exercise booklet—were required for the administration of the survey, it was possible for a household or respondent to refuse to participate at the time of the administration of any one of these instruments. Because the screener and background questionnaire were read to the survey participants in English or Spanish, but the exercise booklet required reading and writing in the English language, it was possible to complete the screener or background questionnaire but not the exercise booklet, and vice versa. Thus, response rates were calculated for each of the three instruments for the household samples. For the prison sample, there were only two points at which a respondent could not respond—at the administration of the background questionnaire or exercise booklet.

The response rate to the background questionnaire was 80.5 percent. For the household samples, the response rates exclude individuals who were not paid incentives. Also excluded are the respondents to the Florida state survey, which had a delayed administration.

The combined national and state household target sample in the 1992 NALS included 43,783 representative housing units, of which 5,405 were vacant. Approximately 89 percent of the occupied households completed a screener.

The household sample screening effort identified a total of 30,806 eligible respondents, of which 24,939 (81.0 percent unweighted) completed the background questionnaire. For the prison sample, 87 of the 88 sampled facilities participated in the survey. Of the 1,340 inmates selected, 1,147 (85.6 percent unweighted) completed the background questionnaire.

For the occupied households, "refusal or breakoff" was the most common explanation for nonresponse to the screener and background questionnaire. The second most common explanation was "not at home after maximum number of calls." Nonresponse also resulted from language, physical, and mental problems. Housing units or individuals who refused to participate before any information was collected about them, or who did not answer a sufficient number of background questions, were never incorporated into the database. Because these individuals were unlikely to know that the survey intended to assess their literacy, it was assumed that their reason for not completing the survey was not related to their level of literacy.

Literacy assessment booklets were considered complete if at least five items were answered on each scale. A total of 24,944 household sample members were classified as eligible for the exercise booklet. Of these, 88.6 percent completed the booklet and another 6.1 percent partially completed the exercise. Of the 1,147 eligibles in the prison sample, 86.8 percent completed the booklet and another 9.3 percent partially completed it.

There were reasons to believe that the literacy performance data were missing more often for adults with lower levels of literacy than for adults with higher levels. Field test evidence and experience with surveys indicated that adults with lower levels of literacy were more likely than adults with higher proficiencies either to decline to respond to the survey at all or to begin the assessment but not complete it. Ignoring this pattern of missing data would have resulted in overestimating the literacy skills of adults in the United States. Therefore, to minimize bias in the proficiency estimates due to nonresponse to the literacy assessment, special procedures were developed to impute the literacy proficiencies of nonrespondents who completed fewer than five literacy tasks.

Item nonresponse. For each background questionnaire, staff verified that certain questions providing critical information for weighting and data analyses had been answered, namely education level, employment status, parents' level of education, race, and sex. If a response was missing, the case was returned to the field for data retrieval. Therefore, item response rates for completed background questionnaires were quite high, although they varied by type of question. Questions asking country of origin (first question in the booklet) and sex (last question in the booklet) had nearly 100 percent response rates, indicating that most respondents attempted to complete the entire questionnaire. Response rates were lower, however, for questions about income and educational background.

The electronic codebook provides counts of item nonresponse. These, however, have to be considered in terms of the number of adults that were offered each task, because a great deal of the missing data is missing by design.

Measurement error. All background questions and literacy tasks underwent extensive review by subject area and measurement specialists, as well as scrutiny to eliminate any bias or lack of sensitivity to particular groups. Special care was taken to include materials and tasks that were relevant to adults of widely varying ages. During the test development stage, the tasks were submitted to test specialists for review, part of that involved checking the accuracy and completeness of the scoring guide. After preliminary versions of the assessment instruments were developed and after the field test was conducted, the literacy tasks were closely analyzed for bias or "differential item functioning." The goal was to identify any assessment tasks that were likely to underestimate the proficiencies of a particular subpopulation, whether it be older adults, females, or Black or Hispanic adults. Any assessment item that appeared to be biased against a subgroup was excluded from the final survey. The coding and scoring guides also underwent further revisions after the first responses were received from the main data collection.

Interviewer error checks. Several quality control procedures related to data collection were used during the field operation: an interviewer field edit, a complete edit of all documents by a trained field editor, validation of 10 percent of each interviewer's closeout work, and field observation of both supervisors and interviewers.

Coding/scoring error checks. In order to monitor the accuracy of coding, the questions dealing with country of birth, language, wages, and date of birth were checked in 10 percent of the questionnaires by a second coder. For the industry and occupation questions, 100 percent of the questionnaires were recoded by a second coder. Twenty percent of all the exercise booklets were subjected to a reader reliability check, which entailed a scoring by a second reader. There was a high degree of reader reliability across tasks—ranging from 88.1 to 99.9 percent—with an average agreement of 97 percent. For 133 out of 165 open-ended tasks, the agreement between the two readers was above 95 percent.

Data Comparability

One of the major goals of this survey was to compare its results to the 1985 Young Adult Literacy Assessment and other large assessment studies.

Comparisons with the 1985 Young Adult Literacy Assessment. Comparisons are possible because the sample design, item pool, and methodology used in the 1985 Young Adult Literacy Assessment and the 1992 survey were very similar. Literacy tasks for each survey were developed using the same definition of literacy, and a subset of identical tasks was administered in both assessments. Scoring guides were the same for both surveys. Both gave nearly identical incentive payments to participants ($15 in 1985 and $20 in 1992). The literacy scales used in the two surveys were linked so that the scores could be reported on a common scale.

Nevertheless, there were some differences in procedures for the two surveys. For example, missing responses to the literacy tasks were handled differently. In the 1985 Young Adult Literacy Assessment, individuals who could not answer six core literacy tasks and those who spoke only Spanish were excluded from the analyses. In the 1992 survey, however, a special procedure was used to impute literacy proficiencies for literacy-related nonrespondents.

Due to such procedural differences, direct comparisons of the results of the two surveys are not simple and straightforward. However, because the 1992 sample is more inclusive than the 1985 sample, subsamples that have more exact counterparts in the 1985 survey can be selected. For instance, the initial report from the 1992 NALS presented data, using no subsample matching, indicated that young adults in 1992 were somewhat less literate than their predecessors in 1985. However, when a comparison was made between matched subsamples of the 1985 and 1992 survey respondents based on reasons for nonresponse, the proficiency differences decreased significantly. Furthermore, results from partition analysis of the two surveys' matched subsamples—based on change due to variations in demographic characteristics versus change not related to demography—suggest that most of the observed declines in the average literacy skills of young adults over time can be accounted for by shifts in the composition of the population and by changes across the assessments in the rules used to include or exclude nonrespondents.

Comparisons with the 1993 GED. Comparisons between NALS and GED examinees are explored in *The Literacy Proficiencies of GED Examinees: Results from the GED-NALS Comparison Study* (by Janet Baldwin, Irwin S. Kirsch, Don Rock, and Kentaro Yamamoto; American Council on Education and Educational Testing Service: 1993). The GED Tests and NALS instruments have a considerable degree of overlap in what they measure. Both assess skills that appear to represent verbal comprehension and reasoning, or the ability to understand, analyze, interpret, and evaluate written information and apply fundamental principles and concepts. Despite the considerable degree of overlap, the two instruments also measure somewhat different skills. For example, the GED Tests seem to tap unique dimensions of writing mechanics and mathematics, while the adult literacy scales appear to tap unique dimensions of document literacy. In addition, the evidence shows that there are no differences in the average prose, document, or quantitative literacy skills of those adults who terminated their schooling at the high school or GED level.

6. CONTACT INFORMATION

For content information on the National Adult Assessments of Literacy, contact:

Andrew J. Kolstad
Phone: (202) 502–7374
E-mail: andrew.kolstad@ed.gov

Mailing Address:
National Center for Education Statistics
1990 K Street NW
Washington, DC 20006–5651

7. METHODOLOGY AND EVALUATION REPORTS

General

Adult Literacy in America: *A First Look at the Findings of the National Adult Literacy Survey*, NCES 93–275, by I.S. Kirsch, A. Jungeblut, and L. Jenkins. Washington, DC: 1993.

Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey, NCES 2001–457, by I. Kirsch, K. Yamamoto, N. Norris, D. Rock, A. Jungeblut, P. O'Reilly, A. Campbell, L. Jenkins, A. Kolstad, M. Berlin, L. Mohadjer, J. Waksberg, H. Goksel, J. Burke, S. Rieger, J. Green, M. Klein, P. Mosenthal, and S. Baldi. Washington, DC: 2000.

Survey Design

Assessing Literacy: The Framework for the National Adult Literacy Survey, NCES 92–113, by A. Campbell and I.S. Kirsch. Washington, DC: 1992.

Chapter 24: International Adult Literacy Survey (IALS)

1. OVERVIEW

1994 INTERNATIONAL STUDY OF ADULT LITERACY

IALS collected:

- Background Assessments
- Literacy Assessments

The 1994 International Adult Literacy Survey (IALS) represented a first attempt to assess the literacy skills of entire adult populations in a framework that provided data comparable across cultures and languages. This collaborative project was designed to inform both education and labor market policy and program development activities in participating countries. The international portion of the study was carried out under the auspices of an International Steering Committee chaired by Canada, with each participating country holding a seat on the committee along with representatives from the Organization for Economic Cooperation and Development (OECD), European communities, and the United Nations Educational, Scientific and Cultural Organization.

In the United States, IALS is the fourth assessment of adult literacy funded by the federal government and conducted by the Educational Testing Service (ETS). The three previous efforts were: (1) the 1992 National Adult Literacy Survey (see chapter 23); (2) the Department of Labor's (DOL) 1990 Workplace Literacy Survey; and (3) the 1985 Young Adult Literacy Survey (funded as an adjunct to the National Assessment of Educational Progress—see chapter 20). In order to maximize the comparability of estimates across countries, the IALS study chose to adopt the National Adult Literacy Survey methodology and scales. Literacy was defined along three dimensions—prose, document, and quantitative. These were designed to capture an ordered set of information-processing skills and strategies that adults use to accomplish a diverse range of literacy tasks encountered in everyday life. The background data collected in IALS provide a context for understanding the ways in which various characteristics are associated with demonstrated literacy skills.

IALS was originally conducted in seven countries (Canada, Germany, the Netherlands, Poland, Sweden, French- and German-speaking Switzerland, and the United States). A second phase was subsequently conducted in five additional countries (Australia, Flemish-speaking Belgium, Great Britain, New Zealand, and Ireland), and in a final phase included an additional 10 countries. This chapter will focus on the first phase, in which the United States participated.

Purpose

To (1) develop scales that would permit comparisons of the literacy performance of adults (16 and older) with a wide range of abilities; (2) if such an assessment could be created, describe and compare the demonstrated literacy skills of adults in different countries.

Components

Each IALS country was given a set of model administration manuals and survey instruments as well as guidelines for adapting and translating the survey instruments. IALS instruments consisted of three parts: (1) a background questionnaire, which collected demographic information about respondents; (2) a set of core literacy tasks, which screened out respondents with very limited literacy skills; and (3) a main booklet of literacy tasks, used to calibrate literacy levels.

Background Questionnaire. The background questionnaire collected information on languages spoken or read; parents' educational attainment and employment; labor force experiences—employment status, recent labor force experiences, and occupation; reading and writing at work and looking for work; participation in adult education classes—courses taken, financial support, purpose; reading and writing in daily life (excluding work or school); family literacy—children's reading habits, the household's access to reading materials, hours spent watching television; and household information—total income and sources of income. The background questionnaire was to be administered in about 20 minutes.

Literacy Assessment—Core Literacy Tasks and Main Literacy Tasks. One hundred and fourteen tasks were grouped into three scales and divided into seven blocks (labeled A through G), which in turn were compiled into seven test booklets (numbered 1 through 7). Each booklet contained three blocks of tasks and was designed to take about 45 minutes to complete. Respondents began the cognitive part of the assessment by performing a set of six "core" tasks. Only those who were able to perform at least two of the six core tasks correctly (93 percent of respondents) were given the full assessment.

Periodicity

The first phase of data collection for the IALS was conducted during the autumn of 1994 in Canada, Germany, the Netherlands, Poland, Sweden, Switzerland (French and German-speaking cantons), and the United States. Data were collected from a second group of countries—Australia, Flemish-speaking Belgium, Great Britain, New Zealand, and Ireland—in 1995–96. Data were collected from a third group of countries in 1997–98. No second administration is planned.

2. USES OF DATA

IALS is designed to inform both educational and labor market policy and program development activities in participating countries. The primary objectives of the study are:

- To shed light on the relationship between microeconomic variables—such as individual literacy, educational attainment, labor market participation and employment, and macroeconomic issues—such as competitiveness, growth, and restructuring;
- To identify subpopulations that are economically and socially disadvantaged by their literacy skill profiles; and
- To establish the comparability of assessments of adult literacy.

IALS data provide comparable information about the activities and outcomes of educational systems and institutions in participating countries. Such data can lead to improvements in accountability and policymaking. These data are increasingly relevant to policy formation due to the growing political, economic, and cultural ties between countries.

3. KEY CONCEPTS

Some of the key concepts related to the IALS literacy assessment are described below.

Literacy. The ability to use printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential.

Prose Literacy. The ability to read and use texts of varying levels of difficulty which are presented in sentence and paragraph form, including editorials, news stories, poems, and fiction.

Document Literacy. The knowledge and skills required to locate and use information contained in formats such as job applications, payroll forms, transportation schedules, maps, tables, and graphics.

Quantitative Literacy. The knowledge and skills required to apply arithmetic operations, either alone or sequentially, to numbers embedded in printed materials, such as balancing a checkbook, calculating a tip, completing an order form, or determining the amount of interest on a loan from an advertisement.

Literacy Scales. The three scales used to report the results for prose, document, and quantitative literacy. These scales, each ranging from 0 to 500, are based on those established for the Young Adult Literacy Survey, the DOL's Workplace Literacy Survey, and the National Adult Literacy Survey. The scores on each scale represent degrees of proficiency along that particular dimension of literacy. The scales make it possible not only to summarize the literacy proficiencies of the total population and of various subpopulations, but also to determine the relative difficulty of the literacy tasks administered in IALS.

The literacy tasks administered in IALS varied widely in terms of materials, content, and task requirements, and thus in difficulty. A careful analysis of the range of tasks along each scale provides clear evidence of an ordered set of information-processing skills and strategies along each scale. To capture this ordering, each scale was divided into five levels that reflect this progression of information-processing skills and strategies: Level 1 (0 to 225), Level 2 (226 to 275), Level 3 (276 to 325), Level 4 (326 to 375), and Level 5 (376 to 500). Level 1 comprised those adults who could consistently succeed with Level 1 literacy tasks but not with Level 2 tasks, as well as those who could not consistently succeed with Level 1 tasks and those who were not literate enough to take the test at all. Adults in Levels 2 through 4 were consistently able to succeed with tasks at their level but not with the next more difficult level of tasks. Adults in Level 5 were consistently able to succeed with Level 5 tasks. The use of three parallel literacy scales makes it possible to profile and compare the various types and levels of literacy demonstrated by adults in different countries and by subgroups within those countries.

4. SURVEY DESIGN

Statistics Canada and ETS, a private testing organization in the United States, coordinated the development and management of IALS. These organizations were assisted by national research teams from the participating countries in developing the survey design. The survey design for the 1994 IALS is described below.

Target Population

The IALS target population was the civilian, noninstitutionalized population aged 16 to 65 in each country; however, countries were also permitted to sample older adults, and several did so. All IALS samples excluded full-time members of the military and people residing in institutions such as prisons, hospitals, and psychiatric facilities.

For the United States, the target population consisted specifically of civilian noninstitutionalized residents aged 16 to 65 years in the 50 states and the District of Columbia, excluding members of the armed forces on active duty, those residing outside the United States, and those with no fixed household address (i.e., the homeless or residents of institutional group quarters such as prisons and hospitals).

Sample Design

IALS was designed to provide data representative at the national level. Each country that participated in IALS agreed to draw a probability sample that would accurately represent its civilian, noninstitutionalized population aged 16 to 65. The final IALS sample design criteria specified that each country's sample should result in at least 1,000 respondents, the minimum sample size needed to produce reliable literacy proficiency estimates. Given the different sizes of the population of persons aged 16 to 65 in the countries involved, sample sizes varied considerably from country to country (ranging from 1,500 to 8,000 per country), but sample sizes were sufficiently large in all cases to support the estimation of reliable IRT item parameters.

IALS countries were strongly encouraged to select high-quality probability samples because the use of probability designs would make it possible to produce unbiased estimates for individual countries and to compare these estimates across the countries. Because the available data sources and resources were different in each of the participating countries, however, no single sampling methodology was imposed. Each IALS country created its own sample design. All countries used probability sampling for at least some stages of their sample designs, and some used probability sampling for all stages of sampling. Sampling designs were approved by expert review.

The sample for the United States was selected from a sample of individuals in housing units who were completing their final round of interviews for the Current Population Survey (CPS) in March, April, May, and June 1994. These housing units were included in the CPS for their initial interviews in December 1992 and January, February, and March 1993. The CPS is a large-scale continuous household survey of the civilian noninstitutionalized population aged 15 and over.

The sample was selected from housing units undergoing their final CPS interviews in March–June, 1994. The frame for the CPS consisted of 1990 Decennial Census files, which are continually updated for new residential construction and are adjusted for undercount, births, deaths, immigration, emigration, and changes in the armed forces.

The CPS sample is selected using a stratified multistage design. Housing units that existed at the time of the 1990 Population Census were sampled from the Census list of addresses. Housing units that did not exist at that time were sampled from lists of new construction when available and otherwise by area sampling methods. Occupants of housing units that came into existence between the time of the CPS sample selection and the time of the IALS fieldwork had no chance of being selected for IALS.

The IALS sample was confined to 60 of the 729 CPS primary sampling units (PSUs). Within these 60 PSUs, all persons aged 16 to 65 years of age in the sampled housing units were classified into 20 cells defined by race/ethnicity and education. Within each cell, persons were selected for IALS with probability proportional to their CPS weights, with the aim of producing an equal probability sample of persons within cells. A total of 4,901 persons was selected for IALS. IALS interviews were conducted in October and November 1994.

Assessment Design

The success of IALS depended on the development and standardized application of a common set of survey instruments. The test framework explicitly followed the precedent set by the National Adult Literacy Survey, basing the test on United States definitions of literacy along three dimensions—prose literacy, document literacy, and quantitative literacy—but extending the instruments into an international context. Study managers from each participating country were encouraged to submit materials such as news articles and documents that could be used to create tasks with the goal of building a new pool of literacy tasks that could be linked to established scales. IALS team field tested 175 tasks and identified 114 that were valid across cultures. Approximately half of these tasks were based on materials from outside North America. (However, each respondent was administered only a fraction of the pool of tasks, using a variant of matrix sampling.)

Each IALS country was given a set of model administration manuals and survey instruments as well as graphic files containing the pool of IALS literacy items with instructions to modify each item by translating the English text to its own language without altering the graphic representation. Certain rules governed the item modification process. For instance, some items required respondents to perform a task that was facilitated by the use of keywords. The keyword in the question might be identical, similar but not exactly the same, or a synonym of the word used in the body of the item, or respondents might be asked to choose among multiple keywords in the body of the item, only one of which was correct. Countries were required to preserve these conceptual associations during the translation process. Particular conventions used in the items—for example, currency units, date formats, and decimal delimiters—were adapted as appropriate for each country.

To ensure that the adaptation process did not compromise the psychometric integrity of the items, each country's test booklets were carefully reviewed for errors of adaptation. Countries were required to correct all errors found. However, this review was imperfect in two important respects. First, it is clear that countries chose not to incorporate a number of changes that were identified during the course of the review, believing that they "knew better." Second, the availability of empirical data from the study has permitted the identification of several additional sources of task and item difficulty that were not included in the original framework, which was based on research by Irwin Kirsch of ETS and Peter Mosenthal of Syracuse University. (See "Exploring Document Literacy: Variables Underlying the Performance of Young Adults," by I.S. Kirsch and P.B. Mosenthal, in *Reading Research Quarterly 25*: 5–30.) Item adaptation guidelines and item review procedures associated with subsequent rounds of IALS data collection were adapted to reflect this additional information.

The model background questionnaires contained two sets of questions: mandatory questions, which all countries were required to include; and optional questions, which were recommended but not required. Countries were not required to field literal translations of the mandatory questions, but were asked to respect the conceptual intent of each question in adapting it for use. Countries were permitted to add questions to their background questionnaires if the additional burden on respondents would not reduce response rates. Statistics Canada reviewed all background questionnaires except Sweden's before the pilot survey and offered comments and suggestions to each country.

Data Collection and Processing

IALS data for the first round of countries were collected through in-person household interviews in the fall of 1994. Each country mapped its national dataset into a highly structured, standardized record layout which it sent to Statistics Canada. Further description follows.

Reference dates. Respondents answered questions about jobs they may have held in the 12 months before the survey was administered.

Data collection. Statistics Canada and ETS coordinated the development and management of IALS. Participating countries were given model administration manuals and survey instruments as well as guidelines for adapting and translating the survey instruments and for handling nonresponse codings.

Countries were permitted to adapt these models to their own national data collection systems, but they were required to retain a number of key features: (1) respondents were to complete the core and main test booklets alone, in their homes, without help from another person or from a calculator; (2) respondents were not to be given monetary incentives for participating; (3) despite the prohibition on monetary incentives, interviewers were provided with procedures to maximize the number of completed background questionnaires, and were to use a common set of coding specifications to deal with nonresponse. This last requirement was critical. Because noncompletion of the core and main task booklets was correlated with ability, background information about nonrespondents was needed in order to impute cognitive data for these persons.

IALS countries were instructed to obtain at least a background questionnaire from sampled individuals. All countries participating in IALS instructed interviewers to make callbacks at households that were difficult to contact.

In general, the survey was carried out in the national language. In Canada, respondents were given a choice of English or French, and in Switzerland, samples drawn from French-speaking and German-speaking cantons were required to respond in those respective languages. When respondents could not speak the designated language, attempts were made to complete the background questionnaire so that their literacy level could be estimated and the possibility of distorted results would be reduced. In the United States, the test was given in English, but a Spanish version of the background questionnaire and bilingual interviewers were available to assist individuals whose native language was not English.

Survey respondents spent approximately 20 minutes answering a common set of background questions concerning their demographic characteristics, educational experiences, labor market experiences, and literacy-related activities. Responses to these background questions made it possible to summarize the survey results using an array of descriptive variables, and also increased the accuracy of the proficiency estimates for various subpopulations. After answering the background questions, the remainder of respondents' time was spent completing a booklet of literacy tasks designed to measure their prose, document, and quantitative skills. Most of these tasks were open-ended, requiring respondents to provide a written answer.

In the United States, the IALS interview period was from October to November 1994. IALS was conducted by 149 Census Bureau interviewers. All of them had at least 5 days of interviewer training. They were given a 1-day training on IALS and were provided with substantial training and reference materials based on the Canadian training package. They also performed a day of field training under the supervision of a regional office supervisor. Each interviewer had an average workload of 33 interviews, and the average number of response interviews per interviewer was 21. They were supervised by six regional supervisors who reviewed and commented on their work.

Before data collection, a letter was sent to the selected addresses describing the upcoming survey. The survey was limited to 90 minutes. If a respondent took more than 20 minutes per block, the interviewer was instructed to move the respondent on to the next block.

Data processing. As a condition of their participation in IALS, countries were required to capture and process their files using procedures that ensured logical consistency and acceptable levels of data capture error. Specifically, countries were advised to conduct complete verification of the captured scores (i.e., enter each record twice) in order to minimize error rates. One hundred percent keystroke validation was needed. Specific details about scoring are provided in a separate section below.

To create a workable comparative analysis, each IALS country was required to map its national dataset into a highly structured, standardized record layout. In addition to specifying the position, format, and length of each field, this International Record Layout included a description of each variable and indicated the categories and codes to be provided for that variable. Upon receiving a country's file, Statistics Canada performed a series of range checks to ensure compliance to the prescribed format. When anomalies were detected, countries corrected the problems and submitted new files. Statistics Canada did not, however, perform any logic or flow edits, as it was assumed that participating countries performed this step themselves.

Editing. Most countries followed IALS guidelines, verifying 100 percent of their data capture operation. The two countries that did not comply with this recommendation conducted sample verifications, one country at 20 percent and the other at 10 percent. Each country coded and edited its own data, mapping its national dataset into the detailed International Record Layout, which included a description of each variable and indicated the categories and codes to be provided for that variable. Industry, occupation, and education were coded using the standard international coding schemes: the International Standard Industrial Classification (ISIC), the International Standard Occupational Classification (ISOC), and the International Standard Classification of Education (ISCED). Coding schemes were provided for open-ended items; the coding schemes came with specific instructions so that coding error could be contained to acceptable levels.

Scoring. Respondents' literacy proficiencies were estimated based on their performance on the cognitive tasks administered in the assessment. Because the open-ended items used in IALS elicited a large variety of responses, responses had to be grouped in order to summarize the performance results. As they were scored, responses to IALS open-ended items were classified as correct, incorrect, or omitted. The models employed to estimate ability and difficulty were predicated on the assumption that the scoring rubrics developed for the assessment were applied in a consistent fashion within and between countries. To reinforce the importance of consistent scoring, a meeting of national study managers and chief scorers was held prior to the commencement of scoring for the main study. The group spent 2 days reviewing the scoring rubrics for all the survey items. Where this review uncovered ambiguities and situations not covered by the guides, clarifications were agreed to collectively, and these clarifications were then incorporated into the final rubrics. To provide ongoing support during the scoring process, Statistics Canada and ETS maintained a joint scoring hotline. Any scoring problems encountered by chief scorers were resolved by this group, and decisions were forwarded to all national study managers. Study managers conducted intensive scoring training using the scoring manual and discussed unusual responses with scorers. They also offered additional training to some scorers, as needed, to raise their accuracy to the level achieved by other scorers.

To maintain coding quality within acceptable levels of error, each country undertook to rescore a minimum of 10 percent of all assessments. Where significant problems were encountered, larger samples of a particular scorer's work were to be reviewed and, where necessary, their entire assignments rescored. Countries were not required to resolve contradictory scores in the main survey (as they had been in the pilot), since outgoing agreement rates were far above minimum acceptable tolerances.

Since there could still be significant differences in the consistency of scoring between countries, countries agreed to exchange at least 300 randomly selected booklets with another country sharing the same test language. In all cases where serious discrepancies were identified, countries were required to rescore entire items or discrepant code pairs.

Intra-country rescoring. A variable sampling ratio procedure was set up to monitor scoring accuracy. At the beginning of scoring, almost all responses were rescored to identify inaccurate scorers and to detect unique or difficult responses that were not covered in the scoring manual. After a satisfactory level of accuracy was achieved, the rescoring ratio was dropped to a maintenance level to monitor the accuracy of all scorers. Average agreements were calculated across all items. Precautions were taken to ensure that the first and second scores were truly independent.

Intercountry rescoring. To determine intercountry scoring reliabilities for each item, the responses of a subset of examinees were scored by two separate groups. Usually, these scoring groups were from different countries. Intercountry score reliabilities were calculated by Statistics Canada, then evaluated by ETS. Based on the evaluation, every country was required to introduce a few minor changes in scoring procedures. In some cases, ambiguous instructions in the scoring manual were found to be causing erroneous interpretations and therefore lower reliabilities.

Using the intercountry score reliabilities, researchers could identify poorly constructed items, ambiguous scoring criteria, erroneous translations of items or scoring criteria, erroneous printing of items or scoring criteria, scorer inaccuracies, and, most important, situations in which one country consistently scored differently from another. In the latter circumstance, scorers in one country may consistently rate a certain response as being correct while those in another country score the same response as incorrect. ETS and Statistics Canada examined scoring carefully to identify situations in which scorers in one country were consistently rating a certain response as being correct while those in another country were scoring the same response as incorrect. Where a systematic error was identified in a particular country, the original scores for that item were corrected for the entire sample.

Estimation Methods

Weighting was used in the 1994 IALS to adjust for sampling and nonresponse. Responses to the literacy tasks were scored using IRT scaling. A multiple imputation procedure based on plausible values methodology was used to estimate the literacy proficiencies of individuals who completed literacy tasks.

Weighting. IALS countries used different methods for weighting their samples. Countries with known probabilities of selection could calculate a base weight using the probability of selection. To adjust for unit nonresponse, all countries poststratified their data to known population counts, and a comparison of the distribution of the age and sex characteristics of the actual and weighted samples indicates that the samples were comparable to the overall populations of IALS countries. Another commonly used approach was to weight survey data to adjust the rough estimates produced by the sample to match known population counts from sources external to IALS. This “benchmarking” procedure assumes that the characteristics of nonrespondents are similar to those of respondents. It is most effective when the variables used for benchmarking are strongly correlated with the characteristic of interest—in this case, literacy levels. For IALS, the key benchmarking variables were age, employment status, and education. All of IALS countries benchmarked to at least one of these variables. The United States used education.

Weights for the United States IALS included two components. The first assigned weights to CPS respondents, and the second assigned weights to IALS respondents.

The CPS weighting scheme was a complex one involving three components: basic weighting, noninterview adjustment, and ratio adjustment. The basic weighting compensated for unequal selection probabilities. The noninterview adjustment compensated for nonresponse within weighting cells created by clusters of PSUs of similar size; Metropolitan Statistical Area (MSA) clusters are subdivided into central city areas, and the balance of the MSA and non-MSA clusters are divided into urban and rural areas. The ratio adjustment made the weighted sample distributions conform to known distributions on such characteristics as age, race, Spanish origin, sex, and residence.

The weights of persons sampled for IALS were adjusted to compensate for the use of the four rotation groups, the sampling of the 60 PSUs, and the sampling of persons within the 60 PSUs. The IALS noninterview adjustment compensated for sampled persons for whom no information was obtained because they were absent, refused to participate, had a short-term illness, had moved or had experienced an unusual circumstance that prevented them from being interviewed. Finally, the IALS ratio adjustment ensured that the weighted sample distributions across a number of education groups conformed to March 1994 CPS estimates of these numbers.

Scaling (item response theory). The scaling model used in IALS was the two-parameter logistic model from item response theory.

Items developed for IALS were based on the framework used in three previous large-scale assessments: the Young Adult Literacy Survey (YALS), the DOL survey, and the National Adult Literacy Survey. As a result, IALS items shared the same characteristics as the items in these earlier surveys. The English version of IALS items were reviewed and tested to determine whether they fit into the literacy scales in accordance with the theory and whether they were consistent with the National Adult Literacy Survey data. Quality control procedures for item translation, scoring, and scaling followed the same procedures used in the National Adult Literacy Survey and extended the methods used in other international studies.

Identical item calibration procedures were carried out separately for each of the three literacy scales: prose, document, and quantitative literacy. Using a modified version of Mislevy and Bock’s 1982 BILOG computer program—see *BILOG: Item analysis and test scoring with binary logistic models*, Scientific Software—the two-parameter logistic IRT model was fit to each item using sample weights. BILOG procedures are based on an extension of the marginal-maximum-likelihood approach described by Bock and Aitkin in their 1981 *Psychometrika* article, “Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm.”

Most of the items administered in IALS were successful from a psychometric standpoint. However, despite stringent efforts at quality control, some of the assessment items did not meet the criteria for inclusion in the final tabulation of results. Specifically, in carrying out the IRT modeling used to create the three literacy scales, researchers found that a number of assessment items had significantly different item parameters across IALS countries.

Imputation. A respondent had to complete the background questionnaire, pass the core block of literacy tasks, and attempt at least five tasks per literacy scale in order for researchers to be able to estimate his or her literacy skills directly. Literacy proficiency data were imputed for individuals who failed or refused to perform the core literacy tasks and for those who passed the core block but did not attempt at least five tasks per literacy scale. Because the model used to impute literacy estimates for nonrespondents relied on a full set of responses to the background questions, IALS countries were instructed to obtain at least a background questionnaire from sampled individuals. IALS countries were also given a detailed nonresponse classification to use in the survey.

Literacy proficiencies of respondents were estimated using a multiple imputation procedure based on plausible values methodology. Special procedures were used to impute missing cognitive data.

Literary proficiency estimation (plausible values). A multiple imputation procedure based on plausible values methodology was used to estimate respondents' literacy proficiency in the 1994 IALS. When a sampled individual decided to stop the assessment, the interviewer used a standardized nonresponse coding procedure to record the reason why the person was stopping. This information was used to classify nonrespondents into two groups: (1) those who stopped the assessment for literacy-related reasons (e.g., language difficulty, mental disability, or reading difficulty not related to a physical disability); and (2) those who stopped for reasons unrelated to literacy (e.g., physical disability or refusal). About 45 percent of the individuals did not complete the assessment for reasons related to their literacy skills; the other respondents gave no reason for stopping, or gave reasons unrelated to their literacy.

When individuals cited a literacy-related reason for not completing the cognitive items, this implies that they were unable to respond to the items. On the other hand, citing reasons unrelated to literacy implies nothing about a person's literacy proficiency. Based on these interpretations, IALS adapted a procedure originally developed for the National Adult Literacy Survey to treat cases in which an individual responded to fewer than five items per literacy scale, as follows: (1) if the individual cited a literacy-related reason for not completing the assessment, then all consecutively missing responses at the end of the block of items were treated as wrong; and (2) if the individual cited reasons unrelated to literacy for not completing the assessment, then all consecutively missing responses at the end of a block were treated as "not reached."

Proficiency values were estimated based on respondents' answers to the background questions and the cognitive items. As an intermediate step, the functional relationship between these two sets of information was calculated, and this function was used to obtain unbiased proficiency estimates with reduced error variance. A respondent's proficiency was calculated from a posterior distribution that was the multiple of two functions: a conditional distribution of proficiency, given responses to the background questions; and a likelihood function of proficiency, given responses to the cognitive items.

Recent Changes

Since IALS was a onetime assessment, there are no changes to report.

Future Plans

There are no plans to conduct IALS again. However, a new survey called the International Study of Adults (ISA, also known as ALL) is being administered in 2003. The aspects of this survey that address literacy build on methodologies used in IALS.

5. DATA QUALITY AND COMPARABILITY

The literacy tasks contained in IALS and the adults asked to participate in the survey were samples drawn from their respective universes. As such, they were subject to some measurable degree of uncertainty. IALS implemented procedures to minimize both sampling and nonsampling errors. The IALS sampling design and weighting procedures assured that participants' responses could be generalized to the population of interest. Scientific procedures employed in the study design and the scaling of literacy tasks permitted a high degree of confidence in the resulting estimates of task difficulty. Quality control activities continued during interviewer training, data collection, and processing of the survey data.

In addition, special evaluation studies were conducted to examine issues related to the quality of IALS. These studies included: (1) an external evaluation of IALS methodology; (2) an examination of how similar or different the sampled persons were from the overall population; (3) an evaluation of the extent to which the literacy levels of the population in the database for each nation were predictable based on demographic characteristics; (4) an examination of the assumption of unidimensionality; and (5) an evaluation of the construct validity of the adult literacy scales.

Sampling Error

Because IALS employed probability sampling, the results were subject to sampling error. Although small, this error was rather higher in IALS than in most studies because the cost of surveying adults in their homes is so high. Most countries simply could not afford large sample sizes.

Each country provided a set of replicate weights for use in a Jackknife variance estimation procedure.

There were three situations in which nonprobability-based sampling methods were used: France and Germany used "random route" procedures for selecting households into their samples, and Switzerland used an alphabetic sort to select one member of each household. However, based on the available evidence, it is not believed that these practices introduced significant bias into the survey estimates.

In 1998, the UK Office of National Statistics coordinated the European Adult Literacy Review, a split-sample survey intended, in part, to measure the effects of sampling methods on the IALS results. This follow-up survey compared an IALS sample design with an alternative, standardized "best practice" design. Although certain differences were noted between the two samples, the IALS sample design was not confirmed to be inferior to the "best practice" design.

Nonsampling Error

The key sources of nonsampling error in the 1994 IALS were differential coverage across countries and nonresponse bias, which occurred when different groups of sampled individuals failed to participate in the survey. Other potential sources of nonsampling error included deviations from prescribed data collection procedures, and errors of logic which resulted from mapping idiosyncratic national data into a rigid international format. Scoring error, associated with scoring open-ended tasks reliably within and between countries, also occurred. Finally, because IALS data were collected and processed independently by the various countries, the study was subject to uneven levels of commonplace data capture, data processing, and coding errors.

Three studies were conducted to examine the possibility of nonresponse bias. Because the sampling frames for Canada and the United States contained information about the characteristics of sampled individuals, it was possible to compare the characteristics of respondents and nonrespondents, particularly with respect to literacy skill profiles. The Swedish National Study Team also commissioned a nonresponse follow-up study.

Coverage error. The design specifications for IALS stated that in each country the study should cover the civilian, noninstitutional population aged 16–65. It is the usual practice to exclude the institutional population from national surveys because of the difficulties in conducting interviews in institutional settings. Similarly, it is not uncommon to exclude certain other parts of a country's population that pose difficult survey problems (e.g., persons living in sparsely populated areas). The intended coverage of the surveys generally conformed well to the design specifications: each of IALS countries attained a high level of population coverage, ranging from a low of 89 percent in Switzerland to 99 percent in the Netherlands and Poland. However, it should be noted that actual coverage is generally lower than the intended coverage because of deficiencies in sampling frames and sampling frame construction (e.g., failures to list some households and some adults within listed households). In the United States, for example, comparing population sizes estimated from the survey with external benchmark figures suggests that the overall coverage rate for the CPS (the survey from which the IALS sample was selected) is about 93 percent, but that it is much lower for certain population subgroups (particularly young Black male adults).

Nonresponse error. For IALS, several procedures were developed to reduce biases due to nonresponse, based on how much of the survey the respondent completed.

Unit nonresponse. The definition of a respondent for IALS was a person who partially or fully completed the background questionnaire. Unweighted response rates varied considerably from country to country, ranging from a high of 69 percent (Canada, Germany) to 45 percent (the Netherlands), with four countries in the 55–60 percent range.

In the United States, which had a response rate of 60 percent, nonresponse to IALS occurred for two reasons: (1) some individuals did not respond to the CPS; and (2) some of the CPS respondents selected for IALS did not respond to IALS instruments. In any given month, nonresponse to the CPS is typically quite low, around 4 to 5 percent. Its magnitude in the expiring rotation groups employed for IALS selection is not known. About half of the CPS nonresponse is caused by refusals to participate, while the remainder is caused by temporary absences, other failures to contact, inability of persons contacted to respond, and unavailability for other reasons.

A sizeable proportion of the nonresponse to the IALS background questionnaire was attributable to persons who had moved. For budgetary reasons, it was decided that persons who were not living at the CPS addresses at the time of IALS interviews would not be contacted. This decision had a notable effect on the sample of students, who are sampled in dormitories and other housing units in the CPS only if they do not officially reside at their parents' homes. Those who reside at their parents' homes are included in the CPS at that address, but because most of these students were away at college during the IALS interview period (October to November 1994), they could not respond to IALS.

The high level of nonresponse for college students could cause a downward bias in the literacy skill-level estimates. This group represents only a small proportion of the United States population, however, so the potential bias is likely to be quite small. Further, comparison of IALS results to the U.S. National Adult Literacy Survey data discounts this as a major source of bias.

Item nonresponse. The weighted percentage of omitted responses for the United States IALS ranged from 0 to 18 percent.

Not-reached responses were classified into two groups: nonparticipation immediately or shortly after the background information was collected, and premature withdrawal from the assessment after a few cognitive items were attempted. The first type of not-reached response varied a great deal across countries according to the frames from which the samples were selected. The second type of not-reached response was due to quitting the assessment early, resulting in incomplete cognitive data. Not-reached items were treated as if they provided no information about the respondent's proficiency, so they were not included in the calculation of likelihood functions for individual respondents. Therefore, not-reached responses had no direct impact on the proficiency estimation for subpopulations. The impact of not-reached responses on the proficiency distributions was mediated through the subpopulation weights.

Measurement error. Assessment tasks were selected to ensure that, among population subgroups, each literacy domain (prose, document, and quantitative) was well covered in terms of difficulty, stimuli type, and content domain. The IALS item pool was developed collectively by participating countries. Items were subjected to a detailed expert analysis at ETS and vetted by participating countries to ensure that the items were culturally appropriate and broadly representative of the population being tested. For each country, experts who were fluent in both English and the language of the test reviewed the items and identified ones that had been improperly adapted. Countries were asked to correct problems detected during this review process. To ensure that all of the final survey items had a high probability of functioning well, and to familiarize participants with the unusual operational requirements involved in data collection, each country was required to conduct a pilot survey. Although the pilot surveys were small and typically were not based strictly on probability samples, the information they generated enabled ETS to reject items, to suggest modifications to a few items, and to choose good items for the final assessment. ETS's analysis of the pilot survey data and recommendations for final test design were presented to and approved by participating countries.

Data Comparability

While most countries closely followed the data collection guidelines provided, some did deviate from the instructions. First, two countries (Sweden and Germany) offered participation incentives to individuals sampled for their survey. The incentive paid was trivial, however, and it is unlikely that this practice distorted the data. Second, the doorstep introduction provided to respondents differed somewhat from country to country. Three countries (Germany, Switzerland, and Poland) presented the literacy test booklets as a review of the quality of published documents rather than as an assessment of the respondent's literacy skills. A review of these practices suggested that they were intended to reduce response bias and were warranted by cultural differences in respondents' attitudes toward being tested. Third, there were differences across the countries in the way in which interviewers were paid. No guidelines were provided on this subject, and the study teams therefore decided what would work best in their respective countries. Fourth, several countries adopted field procedures that undermined the objective of obtaining completed background questionnaires for an overwhelming majority of selected respondents.

This project was designed to produce data comparable across cultures and languages. After one of the countries in the first round raised concerns about the international comparability of the survey data, Statistics Canada decided that the IALS methodology should be subjected to an external evaluation. In the judgment of the expert reviewers, the considerable efforts that were made to develop standardized survey instruments for the different nations and languages were successful, and the data obtained from them should be broadly comparable.

However, the standardization of procedures with regard to other aspects of survey methodology was not achieved to the extent desired, resulting in several weaknesses. Nonresponse proved to be a particular weakness, with generally very high nonresponse rates and variation in nonresponse adjustment procedures across countries. For some countries the sample design was problematic, resulting in some unknown biases. The data collection and its supervision differed between participating countries, and some clear weaknesses were evident for some countries. The reviewers felt that the variation in survey execution across countries was so large that they recommended against publication of comparisons of overall national literacy levels. They did, however, despite the methodological weaknesses, recommend that the survey results be published. They felt that the instruments developed for measuring adult literacy constituted an important advance, and the results obtained for the instruments in the first round of IALS were a valuable contribution to the field. They recommended that the survey report focus on analyses of the correlates of literacy (e.g., education, occupation, and age) and the comparison of these correlates across countries. Although these analyses might also be distorted by methodological problems, they believed that the analyses were likely to be less affected by these problems than were the overall literacy levels.

6. CONTACT INFORMATION

For content information on IALS, contact:

Eugene Owen
Phone: (202) 502–7422
E-mail: eugene.owen@ed.gov

Mailing Address:

National Center for Education Statistics
1990 K Street NW
Washington, DC 20006–5651

7. METHODOLOGY AND EVALUATION REPORTS

Adult Literacy in OECD Countries: Technical Report on the First International Adult Literacy Survey, NCES 98–053, T.S. Murray, I.S. Kirsch, and L.B. Jenkins (eds.). Washington, DC: 1997.