Trends in International Mathematics and Science Study (TIMSS)

Website: [http://nces.ed.gov/timss/](http://nces.ed.gov/timss/)

1. OVERVIEW

The Trends in International Mathematics and Science Study (TIMSS) is a study of classrooms across the country and around the world. The National Center for Education Statistics (NCES), in the Institute of Education Sciences at the U.S. Department of Education, is responsible for the implementation of TIMSS in the United States. Beginning in 1995 and every 4 years thereafter, TIMSS has provided participating countries with an opportunity to measure students’ progress in mathematics and science achievement. Studies of students, teachers, schools, curriculum, instruction, and policy issues are 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 science are related in many respects and since there is broad interest among countries in students’ abilities in both subjects, TIMSS began to be conducted as an integrated assessment of both mathematics and science.

In 1995, TIMSS collected data on grades 3 and 4 as well as grades 7 and 8, and the final grade of secondary school (grade 12 in the United States), with 42 countries participating. In 1999, data were collected only for 8th-grade students, with 38 countries participating. For TIMSS 2003 and 2007, data were collected on grades 4 and 8, with 46 countries participating in 2003 and 58 countries participating in 2007.

In addition to the math and science assessments given to students, supplementary information is obtained through the use of student, teacher, and school questionnaires. Also, in 1995 and 1999, further component studies were implemented, including benchmark and video studies.

The TIMSS 1999 Benchmarking Study included states and districts or consortia of districts from across the United States that chose to participate. These states and districts completed the assessments and questionnaires following the same procedures developed for the participating countries. They then used the findings to assess their comparative international standing and to evaluate their mathematics and science programs in an international context.

For the TIMSS Videotape Study, designed as the first study to collect videotaped records of classroom instruction, representative samples of 8th-grade mathematics classes in 1995 and 1999 and science classes in 1999 were drawn and one lesson in each of the participating classrooms was videotaped. The analysis provides a more detailed context for understanding mathematics and science teaching and learning in the classroom.
**Purpose**

TIMSS is designed to measure student performance in mathematics and science against what is expected to be taught in school. This focus on school curriculum allows for two broad questions to be addressed through TIMSS: (1) How do mathematics and science education environments differ across countries, how do student outcomes differ, and how are differences in these 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 uses several types of instruments to collect data about students, teachers, schools, and national policies and practices that may contribute to student performance.

**Written assessment.** Assessments are developed to test students in various content areas within mathematics and science. For grade 4, the mathematics content areas are numbers; geometric shapes and measures; and data display. The grade 4 science content areas are Earth science; life science; and physical science. The grade 8 mathematics content areas are numbers; algebra; geometry; and data and chance. The grade 8 science content areas are biology; physics; chemistry; and Earth science.

In addition to being familiar with the mathematics and science content areas encountered in TIMSS, students are required to draw on a range of cognitive skills to successfully complete the assessment. TIMSS focuses on three cognitive domains in each subject: knowing, which covers the facts, procedures, and concepts students need to know; applying, which focuses on the ability of students to apply their knowledge and conceptual understanding to solve problems; and reasoning, which goes beyond solving routine problems to include unfamiliar situations and context that may require multi-step problem-solving.

After each TIMSS assessment cycle, approximately half of the items are publicly released, and replacement items that closely match the content of the original items are developed by international assessment and content experts. These new items are field tested and refined to the point where a variety of multiple choice and extended constructed-response items (i.e., items requiring written explanations from students) are chosen to be included in the TIMSS item pool.

Each student is asked to complete one booklet, made up of a subset of items taken from this item pool. No student answers all of the items in the item pool. The scoring of these booklets is accomplished through the use of a sophisticated and strict set of criteria that are implemented equally across all nations to ensure accuracy and comparability.

**Student background questionnaire.** Each student who takes the TIMSS assessment is asked to complete a questionnaire on issues including daily activities, family attributes, educational resources in the home, engagement in and beliefs about learning, instructional processes in the classroom, study habits, and homework.

**Teacher questionnaire.** The teacher questionnaire is given to the mathematics and science teachers of the students assessed in the study. These questionnaires ask about topics such as attitudes and beliefs about teaching and learning, teaching assignments, class size and organization, topics covered in class, the use of various teaching tools, instructional practices, professional preparation, and continuing development.

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

**School questionnaire.** The principal or head administrator is also asked to complete a questionnaire for the school focused on community attributes, personnel, teaching assignments, policy and budget responsibilities, curriculum, enrollment, student behavior issues, instructional organization, and mathematics and science courses offered.

Information collected from students, their teachers and schools is summarized in composite indices focused, in particular, on the relationship between mathematics and science achievement and the home, classroom, and school environment.

**Curriculum questionnaire.** The national research coordinator, or representative, of each participating country is asked to complete a questionnaire focused on the policies and practices supported at the national level that may contribute to student performance. In addition, because the mathematics and science topics covered in the assessment may not be included in all countries’ curriculum, the national research coordinators are asked to indicate whether each topic covered in TIMSS is included in their countries’ intended curriculum through the fourth or eighth grade.
Encyclopedia. Beginning with TIMSS 2007, each participating country is asked to provide a written overview of the context in which mathematics and science instruction takes place, summarizing the structure of the education system, the mathematics and science curricula and instruction in primary and secondary grades, teacher education requirements, and the types of examinations and assessments employed to monitor success. The resulting chapters are compiled in a publication entitled the TIMSS Encyclopedia.

Videotape study. The 1995 TIMSS Videotape 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 in order to gather more in-depth information about the context in which learning takes place as well as 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.

The 1999 TIMSS Videotape 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 in 1995. The intended curriculum was composed of the mathematics and science instructional and learning goals as defined at the system level. The implemented curriculum was the mathematics and science curriculum as interpreted by teachers and made available to teachers. The attained curriculum was the mathematics and science content that students had 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 variations among the education 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 specifications derived from TIMSS mathematics and science curriculum frameworks.

To collect data about how the curriculum was implemented in classrooms, TIMSS administered a broad array of questionnaires. These questionnaires were administered at the country level on 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.

Ethnographic case studies. The case studies approach to understanding cultural differences in behavior has a long history in selected social science fields. Conducted only in 1995, the case studies were designed to focus on four key topics that challenge U.S. policymakers and to investigate how these topics were 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 validated and integrated the information gained from official sources with that obtained from teachers, students, and parents in order to ascertain the degree to which official policy reflected actual practice. The objective was to describe policies and practices in the nations under study that were 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 separated 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 benchmarking study.** In 1999, 13 states and 14 districts or 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 were representative of the student population in each of these states and districts. The findings from this project allowed these jurisdictions to assess their comparative international standing and judge their mathematics and science programs in an international context.

**NAEP/TIMSS linking study.** A subsample of students who took the 2000 state National Assessment of Educational Progress (NAEP) mathematics and science assessment also took the 1999 TIMSS assessment. (See *NAEP chapter* for more information.) This provided an opportunity to compare students’ performance on NAEP to their performance on TIMSS, and allowed for estimates of how states participating in the 2000 NAEP would have performed had they participated in TIMSS 1999. Results from the TIMSS 1999 Benchmarking Study were used to check the results of the linking study.

**Periodicity**
First conducted in 1995, TIMSS has been conducted every 4 years since then. Previous international math studies were conducted in 1964 and 1980–82; previous international science studies were conducted in 1970–71 and 1983–84.

### 2. USES OF DATA

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

- What are students’ attitudes toward mathematics and science?
- 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?
- How strongly are students motivated to learn, in general, and to the learning of mathematics and science, in particular?
- 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 about teaching them), and how are they related to the comparable opinions and attitudes of their students?
- What methods do teachers use to evaluate their students?
- If there are national curricula in a country, how specific are they, and what efforts are made to see that they are followed?

### 3. KEY CONCEPTS

Key terms related to TIMSS are described below.

**National Desired Population.** The stated objective in TIMSS is that the National Desired Population within each country be as close as possible to the International Desired Population, which is the target population. (See “Target Population” under Section 4. Survey Design.) Using the International Desired Population as a basis, participating countries have to operationally define their populations for sampling purposes. Some national research coordinators have to restrict coverage at the country level, for example, by excluding remote regions or a segment of their country’s education system. Thus, the National Desired Population sometimes differs from the International Desired Population.

### 4. SURVEY DESIGN

**National Research Coordinators.** This is an 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, national research coordinators are 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 are also responsible for arranging the return of the testing materials from the school to the national center, preparing for and implementing the constructed-response item 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.

**Target Population**
The International Desired Population for all countries is defined as follows:

- **Grade 4:** All students enrolled in the grade that represents 4 years of schooling, counting from the 1st year of the International Standard Classification of Education (ISCED) Level 1, providing that the mean age at the time of testing is at least 9.5 years. For most countries, the target grade should be the fourth grade or its national equivalent. All students enrolled in the target grade, regardless of their age, belong to the international desired target population.

- **Grade 8:** All students enrolled in the grade that represents 8 years of schooling, counting from the 1st year of ISCED Level 1, providing that the mean age at the time of testing is at least 13.5 years. For most countries, the target grade should be the eighth grade or its national equivalent. All students enrolled in the target grade, regardless of their age, belong to the international desired target population.

Thus, TIMSS uses a grade-based definition of the target population.

**Sample Design**
Each country participating in TIMSS, like the United States, is required to draw random samples of schools. In the United States, a national probability sample is drawn for each study that has resulted in over 500 schools and approximately 33,000 students participating in 1995, approximately 220 schools and 9,000 students participating in 1999, approximately 480 schools and almost 19,000 students in 2003, and approximately 500 schools and over 20,000 students in 2007. This sample design ensures the appropriate number of schools and students are participating to provide a representative sample of the students in a specific grade in the United States as a whole.

The TIMSS sample design for each country and population is 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 has a chance of being selected, with a fixed probability of selection. These probabilities of selection are designed to be equal across eligible students as much as possible, but for a variety of reasons the probabilities of selection differ between students in most of the national samples.

**Written assessment.**
The TIMSS sample design is a two-stage stratified cluster sample, with schools as the first stage of selection and classrooms within schools as the second stage of selection. For the first time TIMSS 2007 included an optional third stage. The third-stage sampling units for TIMSS 2007 were students within sampled classrooms. Generally however, TIMSS chooses intact classrooms, so students are essentially chosen at the same stage as the classroom (i.e. the second stage).

Individual schools are selected with probability proportionate to size (PPS), size being the estimated number of students enrolled in the target grade. Prior to sampling, schools in the sampling frame can be assigned to a predetermined number of explicit or implicit strata. Substitution schools, selected to replace schools that refuse to participate, are identified simultaneously.

The classroom sampling design is intended to be an equal probability design with no subsampling in the classroom. However, a design based on a PPS sample of classrooms, with a fixed sample size of students selected within the sampled classroom, is permitted under the international guidelines. Exclusions can occur at the school level, the classroom level, or the student level. TIMSS participants are expected to keep such exclusions to no more than 10 percent of the National Desired Population.

The optional third-stage sampling unit for TIMSS 2007 was students within the sampled classrooms. While all students in a sampled classroom were to be selected for the assessment, it was possible for participating countries to sample a subgroup of students after consultation with Statistics Canada, the organization serving as the sampling referee.

TIMSS standards for sampling precision require a minimum of 4,000 students to be assessed per grade. To meet the standard, at least 150 schools are selected.
The schools in each explicit stratum (geographical region, public/private, etc.) are listed in order of the implicit stratification variables and then further sorted according to their measure of size. The stratification variables differ from country to country. Small schools are handled either through explicit stratification or through the use of pseudo-schools. In some very large countries, there is a preliminary sampling stage before schools are sampled in which the country is divided into primary sampling units.

In cases where a sampled school is unable to participate in the assessment, a replacement school is used. The replacement school is the next school on the ordered school-sampling list as the replacement for each particular sampled school. The school after that is a second replacement, should it be necessary. Using either explicit or implicit stratification variables and ordering of the school-sampling frame by size ensures that any original sampled school’s replacement has similar characteristics.

In the second stage of sampling, classrooms of students are sampled. Generally, in each school, one classroom is sampled from each target grade, although some countries opt to sample two classrooms at the upper grade in order to be able to conduct special analyses. Most countries test all students in selected classrooms, and in these instances the classrooms are selected with equal probabilities. A few participants use a design based on a PPS sample of classrooms, with a fixed sample size of students selected within the sampled classrooms. Participants with particularly large classrooms in their schools can decide to subsample a fixed number of students from each selected classroom. This is done using a simple random sampling method whereby all students in a sampled classroom are assigned equal selection probabilities.

In the United States, TIMSS 2007 used a two-stage stratified cluster sampling design based on the 2006 NAEP school sampling frame. The United States did not use the optional third stage of sampling (i.e. students within classrooms) for TIMSS 2007. (Time constraints related to recruitment activities required sample selection before the 2007 frame became available.) For this purpose the sampling frame, though not explicitly stratified, was implicitly stratified by four categorical variables: type of school (public or private); region of the country (Northeast, Central, West, Southeast); community type (eight levels); and percentage of Black, Hispanic, and other race/ethnicity students (above or below 15 percent of the student population).

The first stage of the design used a systematic PPS technique to select schools for the original sample. That is, schools were selected with a probability proportionate to the school’s estimated enrollment of fourth- or eighth-grade students. Enrollment data for public schools were taken from the 2003–04 Common Core of Data (CCD), and data for private schools were taken from the 2003–04 Private School Universe Survey (PSS). For each original school selected, the two adjacent schools in the sampling frame, and within the same implicit stratum, were designated as the first and second replacement schools. The first substitute followed the original sample school in the frame listing and the second substitute preceded it. Substitute schools were designed to be used only if an original school refused to participate. In this situation the first substitute was to be contacted first, with the second substitute contacted only if the first substitute also refused to participate. Additionally, one sampled school was not allowed to substitute for another, and a given school could not be assigned to substitute for more than one sampled school.

An initial sample of 300 schools was selected at each grade level. Ineligible schools among these reduced the grade 4 sample to 290 schools and the grade 8 sample to 290 schools.

At each grade level, the U.S. sample design within schools consisted of an equal probability sample of two classrooms. In schools with a single eligible classroom, that classroom was selected with certainty. All eligible students in the classroom were designated to be in the sample (although generally the option for sub sampling did exist, there was no subsampling of students in the TIMSS 2007 U.S. sample).

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 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 a single sampled classroom, then the teacher receives 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 the U.S. TIMSS 2007 sample, a teacher was not asked to complete more than one questionnaire. In cases where a teacher taught both subject areas, the teacher was provided a specially designed questionnaire that included questions for both mathematics and science teachers.

In general, each country is allowed to develop its own methodology for this process of assigning subjects and classrooms to teachers when the links are not straightforward due to the presence of one to many (or many to one) mappings.

**Assessment Design**

TIMSS is a cooperative effort involving representatives from every country participating in the study. For TIMSS 2007, the development effort began with a revision of the frameworks that were used to guide the construction of the assessment. The frameworks were updated to reflect changes in the curriculum and instruction of participating countries. Extensive input from experts in mathematics and science education, assessment, and curriculum, and representatives from national education centers around the world contributed to the final shape of the frameworks used in 2007. Maintaining the ability to measure change over time is an important factor in constantly revising the frameworks.

**Test development.** As part of the TIMSS dissemination strategy, approximately one-half of the items at each grade are released for public use. To replace assessment items that have been released, countries submit items for review by subject-matter specialists, and additional items are written to ensure that the content, as explicated in the frameworks, is covered adequately. Items are reviewed by an international Science and Mathematics Item Review Committee and field tested in most of the participating countries. Results from the field tests are used to evaluate item difficulty, how well items discriminate between high- and low-performing students, the effectiveness of distracters in multiple-choice items, scoring suitability and reliability for constructed-response items, and evidence of bias toward or against individual countries or in favor of boys or girls.

**Instrument design.** TIMSS 2007 included booklets containing assessment items as well as questionnaires submitted to principals, teachers, and students. The assessment booklets were constructed such that not all of the students responded to all of the items, which is consistent with the design of other large-scale assessments, such as NAEP. To keep the testing burden to a minimum, and to ensure broad subject-matter coverage, TIMSS 2007 used a rotated block design that included both mathematics and science items. That is, students encountered both mathematics and science items during the assessment.

The U.S. 2007 fourth-grade assessment consisted of 14 booklets, each requiring approximately 72 minutes of response time. The 14 booklets were rotated among students, with each participating student completing only 1 booklet. The mathematics and science items were assembled into 14 blocks, or clusters, of items, with each block containing either mathematics or science items. The secure, or trend, items were included in 3 blocks, with the other 11 blocks containing replacement items. Each of the 14 booklets contained a total of 6 blocks.

The U.S. 2007 eighth-grade assessment consisted of 18 booklets, each requiring approximately 90 minutes of response time. The 18 booklets were rotated among students, with each participating student completing only 1 booklet. The mathematics and science items were assembled into 14 blocks, or clusters, of items, with each block containing either mathematics or science items. The secure, or trend, items were included in 3 blocks, with the other 11 blocks containing replacement items. Each of the 18 booklets contained a total of 4 blocks. As part of the design process, it was necessary to ensure that the booklets showed a distribution across the mathematics and science content domains as specified in the frameworks.

**Data Collection and Processing**

**Data collection.** TIMSS 2007 emphasized the use of standardized procedures in all countries. Each country collected its own data, based on comprehensive manuals and trainings provided by the international project team to explain the survey’s implementation, including precise instructions for the work of school coordinators and scripts for test administrators to use in testing sessions. Test administration in the United States was carried out by professional staff trained according to the international guidelines. School staff was asked only to assist with listings of students,
identifying space for testing in the school, and specifying any parental consent procedures needed for sampled students.

Each country was responsible for conducting quality control procedures and describing this effort in the national research coordinator’s report documenting procedures used in the study. In addition, the TIMSS International Study Center considered it essential to monitor compliance with the standardized procedures. National research coordinators were asked to nominate one or more persons unconnected with their national center, such as retired school teachers, to serve as quality control monitors for their countries. The International Study Center developed manuals for the monitors and briefed them in 2-day training sessions about TIMSS 2007, the responsibilities of the national centers in conducting the study, and their own roles and responsibilities.

Data entry and cleaning. Responsibility for data entry is taken by the national research coordinator from each participating country. The data collected for TIMSS 2007 were entered into data files with a common international format, as specified in the Manual for Entering the TIMSS 2007 Data. Data entry was facilitated by the use of common software available to all participating countries (WinDEM). The software facilitated the checking and correction of data by providing various data consistency checks. After data entry, the data were sent to the IEA Data Processing Center (DPC) in Hamburg, Germany, for cleaning. The DPC checked that the international data structure was followed; checked the identification system within and between files; corrected single-case problems manually; and applied standard cleaning procedures to questionnaire files. Results of the data cleaning process were documented by the DPC. This documentation was then shared with the national research coordinator with specific questions to be addressed. The national research coordinator then provided the DPC with revisions to coding or solutions for anomalies. The DPC then compiled background univariate statistics and preliminary classical and Rasch Item Analysis.

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. The TIMSS international design provides for two categories of sampling weights. The first category is designed to be used when schools, classrooms, or students are the unit of analysis. The second category is designed to be used in analyses where teachers, or both teachers and students, are the units of analysis.

First category. Sampling weights in the first category consist of school, classroom, and student weights, along with a combined student weight that is the product of these weights. The school weight is, essentially, the inverse of the probability of a school being sampled in the first stage of the sampling design. A school-level nonresponse adjustment is applied to compensate for any sampled schools that did not participate and were not replaced. This adjustment is calculated independently for each explicit stratum.

Classroom weights reflect the probability of the sampled classroom(s) being selected from among all the classrooms in the school at the target grade level. This classroom weight is calculated independently for each participating school. If a sampled classroom in a school does not participate, or if the participation rate among students in a classroom falls below 50 percent, a classroom-level participation adjustment is made to the classroom weight. If one (or more) selected classrooms in a school do not participate, the classroom participation adjustment is computed at the explicit stratum level rather than at the school level to reduce the risk of bias.

In the first category, student sampling weights are set at 1.0 since intact classrooms are sampled and each student in the sampled classrooms is certain of selection. A nonresponse adjustment is applied to adjust for sampled students who do not take part in the testing. This adjustment is calculated independently for each sampled classroom. An overall student sampling weight is provided as well and is calculated as the product of the school, class, and student weights described above.

In addition, TIMSS provides “house” and “senate” weights, which are scaled versions of the overall student weight just described. 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 data 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. The senate-weight variable, within each country, is proportional to the total weight for that variable by the ratio of the fixed value divided by the size of the population estimate. These sampling weights can be used when cross-national comparisons are required and the data user wants to have each country contribute the same amount to the comparison, regardless of the size of the population.

**Second category.** 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 sum of the student-teacher weights (see discussion below) for students linked to a classroom for a particular assessment.

Sampling weights in this second category are provided to facilitate analyses in which student and teacher data are analyzed together. TIMSS does not provide for a sample of teachers. Rather, the teachers in question are those who teach the sample of TIMSS students. As a consequence, analyses involving teachers have to be viewed as student-level analyses. Accordingly, teacher weights and student-teacher weights are derived from the overall student weight and are designed to accommodate the fact that students may have more than one teacher. Teacher weights are calculated by dividing the sampling weight for a student by the number of teachers that the student has. Separate mathematics and science student-teacher weights are developed by dividing the student sampling weight by, respectively, the number of mathematics teachers and the number of science teachers that the student has.

**Scaling.** TIMSS 1995, 1999, 2003, and 2007 used IRT procedures to produce scale scores that summarized the achievement results. With this method, the performance of a sample of students in a subject area or subarea can be summarized on a single scale or a series of scales, even when different students are administered different items. Because of the reporting requirements for TIMSS and because of the large number of background variables associated with the assessment, a large number of analyses have to be conducted. The procedures TIMSS uses for the analyses are developed to produce accurate results for groups of students while limiting the testing burden on individual students. Furthermore, these procedures provide data that can be readily used in secondary analyses. IRT scaling provides estimates of item parameters (e.g., difficulty, discrimination) that define the relationship between the item and the underlying variable measured by the test. IRT model parameters are estimated for each test question, with an overall scale being established as well as scales for each predefined content area specified in the assessment framework. For example, the TIMSS 2007 8th-grade mathematics assessment had four scales describing mathematics content strands, and the science assessment had scales for four fields of science.

**Imputation and plausible values.** Although multiple imputation techniques are applied to create plausible values for student proficiency scores, with one exception, imputations were not generated for missing values in the TIMSS 2007 teacher, school, or student questionnaire data files. The single exception refers to a U.S.-only variable in the school file, the principal’s report of the percentage of students eligible for free- or reduced-price lunch. For public schools, missing values for this variable were replaced by information obtained from the CCD. Analogous information was not available for private schools. Subsequently, analyses were undertaken to ensure that confidentiality was maintained.

During the scaling phase, plausible values are used to characterize scale scores for students participating in the assessment. To keep student burden to a minimum, TIMSS administers a limited number of assessment items to each student; too few to produce accurate content-related scale scores for each student. To account for this, for each student, TIMSS generates five possible content-related scale scores that represent selections from the distribution of content-related scale scores of students with similar backgrounds who answer the assessment items the same way. The plausible-values technology is one way to ensure that the estimates of the average performance of student populations and the estimates of variability in these estimates are more accurate than those determined.
through traditional procedures, which estimate a single score for each student.

While constructing plausible values, careful quality control steps ensure that the subpopulation estimates based on these plausible values are accurate. Plausible values are constructed separately for each national sample. TIMSS uses the plausible-values methodology to represent what the true performance of an individual might have been, had it been observed. This is done by using a small number of random draws from an empirically derived distribution of score values based on the student’s observed responses to assessment items and on background variables. Each random draw from the distribution is considered a representative value from the distribution of potential scale scores for all students in the sample who have similar characteristics and identical patterns of item responses. The draws from the distribution are different from one another to quantify the degree of precision (the width of the spread) in the underlying distribution of possible scale scores that could have caused the observed performance. The TIMSS plausible values function like point estimates of scale scores for many purposes, but they are unlike true point estimates in several respects. They differ from one another for any particular student, and the amount of difference quantifies the spread in the underlying distribution of possible scale scores for that student. Because of the plausible-values approach, secondary researchers can use the TIMSS data to carry out a wide range of analyses.

Scale anchoring. Beginning with TIMSS 2003, the percentage of students in each country performing at each of four international benchmarks of performance are reported. The benchmarks are selected to represent the range of performance of students internationally. The four benchmarks selected to represent points along the scale are advanced (set at 625), high (550), intermediate (475), and low (400). Using these points along the TIMSS scale, a scale anchoring analysis is conducted to describe student performance in terms of what they know and can do. The scale anchoring process involves a statistical component, which identifies assessment items that discriminate between points on the scale, and expert judgment, in which subject-matter specialists examine the items that anchor at different points along the scale and generalize about students’ knowledge and understanding.

Future Plans
The next TIMSS data collection will take place in spring 2011. In addition, a new effort to link national and international assessments will be initiated in 2011 so that states can compare their own students’ performance against international benchmarks. The linking study is intended to enable NCES to project state-level scores on the TIMSS using data from the National Assessment of Educational Progress (NAEP).

In the linking study, two representative national samples will be tested on their knowledge of mathematics and science by taking both the NAEP and TIMSS assessments. One sample of 10,000 eighth-graders will take combined test booklets in the winter of 2011 as part of NAEP. The other sample of 7,500 eighth-graders will take combined test booklets in the spring of 2011 as part of TIMSS. The relationships between the two assessments of mathematics and science that are found in these two samples will permit state-level projections of how the students in the 50 states and the District of Columbia that took NAEP would have performed in eighth-grade mathematics and science on TIMSS, with scores that can be compared to those of other countries. Data from a number of states that have agreed to administer TIMSS 2011 to state representative samples will be compared to the projected scores to ensure the accuracy of the linking projections.

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 providing support to the national centers and performing quality control checks.

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

Sampling Error
With complex sampling designs that involve more than the simple random sampling of students, as in the case of the stratified multistage design used in TIMSS 2007, where students were clustered within schools, 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 primary sampling units can be paired in a manner consistent with the sampling design, with each pair regarded as members of a pseudo-stratum for variance estimation purposes.
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.

**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 five 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 \( t \), where \( m=1, 2, ..., 5 \). Once the statistics are computed, the imputation variance is then computed as

\[
Var_{imp} = (1 + \frac{1}{M})Var(t_m)
\]

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

**Coverage error.** The stated objective in TIMSS is 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 is to study the effects of different international curricula and pedagogical methods on mathematics and science learning, participating countries have to operationally define their population for sampling purposes. Some national research coordinators have to restrict coverage at the country level, for example, by excluding remote regions or a segment of their country’s education system. In these few situations, countries are permitted to define a National Desired Population that does not include part of the International Desired Population. Exclusions can be based on geographic areas or language groups.

**Nonresponse error.** Unit nonresponse error results from nonparticipation of schools and students. Weighted and unweighted response rates are 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) are also computed.

The minimum acceptable school-level response rate for all countries, before the use of replacement schools, is set at 85 percent. This criterion is applied to the unweighted school-level response rate. However, both weighted and unweighted school-level response rates are calculated, with and without replacement schools. It is generally the case that weighted and unweighted response rates are similar.

Like the school-level response rate, the minimum acceptable student-level response rate is set at 85 percent for all countries. This criterion is applied to the unweighted student-level response rate. However, both weighted and unweighted student-level response rates are 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.

Table 15 shows the unweighted unit level response rates for the data collections of 1995, 1999, 2003, and 2007 for grades 4 and 8.

**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 carries out a special test called the Test-Curriculum Matching Analysis. Each country is asked to identify, for each item, whether the topic of the item is in the curriculum of the majority of the students.

**Data Comparability**

Through a careful process of review, analysis, and refinement, the assessment and questionnaire items are purposefully developed and field tested for similarity and for reliable comparisons between survey years. After careful review of all available data, including a test for reliability between old and new items, the TIMSS assessments are found to be very similar in format, content, and difficulty level across years.

<table>
<thead>
<tr>
<th>Year and grade</th>
<th>School</th>
<th>Student</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th grade</td>
<td>86</td>
<td>94</td>
<td>81</td>
</tr>
<tr>
<td>8th grade</td>
<td>84</td>
<td>92</td>
<td>77</td>
</tr>
<tr>
<td>1999</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th grade</td>
<td>†</td>
<td>†</td>
<td>†</td>
</tr>
<tr>
<td>8th grade</td>
<td>90</td>
<td>93</td>
<td>84</td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th grade</td>
<td>83</td>
<td>95</td>
<td>78</td>
</tr>
<tr>
<td>8th grade</td>
<td>78</td>
<td>94</td>
<td>73</td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th grade</td>
<td>89</td>
<td>95</td>
<td>84</td>
</tr>
<tr>
<td>8th grade</td>
<td>83</td>
<td>93</td>
<td>77</td>
</tr>
</tbody>
</table>

† Not available. TIMSS did not collect data from grade 4 in 1999.


Findings from comparisons between the results of TIMSS, 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. International experts develop 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 results are most useful, however, when they are considered in light of knowledge about education systems that include curricula, but also factors in trends in education reform, changes in 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 is crucial to ensure that items developed for use in one country are functionally identical to those used in other countries. Because questionnaires are 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. National research coordinators from each country review the national adaptations of individual questionnaire items and submit 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 is used to further evaluate any items that are functioning differently in different countries according to the international item statistics. In certain cases, items have to be recoded or deleted entirely from the international database as a result of this review process.

6. CONTACT INFORMATION

For content information about the TIMSS project, contact:

Patrick Gonzales
Phone: (415) 920-9229
E-mail: patrick.gonzales@ed.gov

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

7. METHODOLOGY AND EVALUATION REPORTS

Most of the technical documentation for TIMSS is published by the International Study Center at 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 by an NCES number.

General


Uses of Data


Survey Design


Data Quality and Comparability