Trends in International Mathematics and Science Study (TIMSS)

Website: https://nces.ed.gov/timss/
Updated: January 2019

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 and other education systems 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 International Association for the Evaluation of Educational Achievement (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 and other education systems 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 (called TIMSS Advanced, which assessed only the most advanced grade 12 students in the United States), with 45 countries and other education systems participating. In 1999, data were collected only for eighth-grade students, with 38 countries and other education systems participating. Also in 1999, TIMSS was renamed from "*Third* International Mathematics and Science Study" to "*Trends* in International Mathematics and Science Study." In 2003, 2007, 2011, and 2015, TIMSS collected data at grades 4 and 8 from between 50 and 70 education systems in each administration, with differing numbers participating at each grade. In addition, TIMSS Advanced was conducted again internationally in 2008 and 2015, and measured advanced mathematics and physics achievement in the final year of secondary school across countries. The United States participated in the 1995 and 2015 TIMSS Advanced assessments. For a complete list of all countries and other education systems that participated in TIMSS in each administration by grade, see https://nces.ed.gov/timss/countries.asp.

In addition to the math and science assessments given to students, supplementary information is obtained through the use of student, teacher, and school questionnaires. 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. Findings were used to assess comparative international standings and to evaluate mathematics and science programs in an international context.

WORLDWIDE STUDY OF CLASSROOMS WITH MORE THAN 60 COUNTRIES AND OTHER EDUCATION SYSTEMS PARTICIPATING:

TIMSS tests a variety of subject and content areas:

- Grade 4 math: Number, geometric shapes and measures, data display
- Grade 8 math: Number, algebra, geometry, data and chance
- Grade 4 science: Life science, physical science, and Earth science
- Grade 8 science: Earth science, biology, chemistry, and physics

The TIMSS Videotape Study was the first study to collect videotaped records of classroom instruction; representative samples of eighth-grade mathematics classes in 1995 and 1999, and science classes in 1999, were drawn and one lesson was videotaped in each of the participating classrooms. These additional component studies provided a more detailed context for understanding mathematics and science teaching and learning in the classroom.

The designers of TIMSS 1995 chose to focus on curriculum as a broad explanatory factor underlying student achievement, and curriculum was considered to have three strands—which are also three influencers of student achievement: what society would like to see taught (the intended curriculum), what is actually taught (the implemented curriculum), and what the students learn (the attained curriculum). This view was first conceptualized for SIMS.

The first strand (the intended curriculum), reflects society's goals for teaching and learning, which in turn reflects the ideals and traditions of the greater society, although somewhat constrained by the resources of the education system. The second strand (the implemented curriculum), or what is being taught in the classroom is inspired by the intended curriculum but directly influenced by: 1) teacher education, training and experience; 2) by the nature and organizational structure of the school; 3) teacher interaction with colleagues; and 4) the composition of the student body. The last strand (the attained curriculum), or what the students actually learn, is influenced by both the intended and implemented curriculum as well as by the characteristics of individual students--such as abilities, attitudes, interests, and effort.

Because TIMSS 1999 essentially replicated the eighth-grade part of the 1995 study, much of the conceptual underpinnings of the 1999 and subsequent-year TIMSS studies are derived from this three-strand model of curriculum. The organization and coverage of the intended curriculum (first strand) was investigated through curriculum questionnaires that were completed by national research coordinators (NRCs) and by curriculum advisors. Although more modest in scope than the extensive curriculum analysis component of 1995, the 1999 and subsequent-year TIMSS questionnaires yield valuable information on the curricular intentions of participating countries and other education systems.

Data on the implemented curriculum (second strand) were collected as part of the 1999 and subsequent-year TIMSS surveys of student achievement. Questionnaires completed by the mathematics and science teachers of the students in the survey, and by the principals of their schools, provide information about the topics in mathematics and science that were taught, the instructional methods used in the classroom, the organizational structures that supported teaching, and the factors that were observed that may have facilitated or inhibited teaching and learning.

The student achievement survey provides information for the third strand, the attained curriculum. The wide-ranging mathematics and science tests that were administered to nationally representative samples of students provided not only a sound basis for international comparisons of student achievement, but a rich resource for the study of the attained curriculum in each country or other education system. Information about the student's characteristics, attitudes, beliefs, and experiences was collected from each participating student and helped to identify characteristics associated with learning and to provide a context for the attained curriculum.

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 educational environments and student outcomes differ across countries and other education systems, and how are the differences in student outcomes related to the differences in educational environments? (2) Are there patterns of relationships among contexts, inputs, and outcomes within countries and other education systems 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 NRC, or representative, of each participating country or other education system 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 and other education systems' curriculum, the NRCs are asked to indicate whether each topic covered in TIMSS is included in their countries and other education systems' intended curriculum through the fourth or eighth grade.

Encyclopedia. Beginning with TIMSS 2007, each participating country or other education system 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 eighth-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 eighth-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.

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 re-search plan called for each of the four topics to be studied in the fourth, eighth, and twelfth 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 studies. 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 and other education systems. 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.

In 2003, 2007, 2011, and 2015, several states again participated in TIMSS to benchmark their student performance internationally. In 2003, Indiana participated in TIMSS independently as well as part of the U.S. national sample; in 2007, Massachusetts and Minnesota did; and in 2011, Alabama, California, Colorado, Connecticut, Florida, Indiana, Massachusetts, Minnesota, and North Carolina did. In 2015, Florida participated as an international benchmark.

NAEP/TIMSS linking studies. NCES conducted a study in 2012 to link results from the 2011 National Assessment of Educational Progress (NAEP) to results from the 2011 TIMSS assessment. The goal of the NAEP-TIMSS Linking Study was to predict 2011 TIMSS mathematics and science scores at grade 8 for all U.S. states based on their NAEP performance for states, without incurring the costs associated with every state participating in TIMSS. The 2011 NAEP-TIMSS Linking Study represents the fourth study to link NAEP and TIMSS. The first study used results from the 1995 administration of TIMSS and the results from the 1996 NAEP. The second study used results from TIMSS 1999 and NAEP 2000. The third study examined linking 2007 NAEP and TIMSS data through the method of statistical moderation, by applying the correspondence between the national score distributions of NAEP and TIMSS to the state and district NAEP score distributions. The 2011 NAEP-TIMSS Linking Study was undertaken to improve on the previous attempts to link these two assessments. The 2011 NAEP-TIMSS Linking Study differed from the previous linking studies in several ways. First, unlike previous efforts, the 2011 NAEP-TIMSS Linking Study administered NAEP and TIMSS booklets at the same time under the same testing conditions. This provided examinee-level correlations between NAEP and TIMSS scores to improve the accuracy of the predicted TIMSS state scores. The second way that it differed is that a greater number of validation states, nine, participated in TIMSS, which meant a greater number of actual state TIMSS results were available to validate the predicted TIMSS average scores.

To validate the 2011 NAEP-TIMSS Linking Study, nine independent state samples were used from Alabama, California, Colorado, Connecticut, Florida, Indiana, Massachusetts, Minnesota, and North Carolina.

TIMSS Advanced studies. In 1995, TIMSS Advanced was first conducted to determine what students, who were finishing secondary school with special preparation in mathematics and science, know and can do. The achievement of these students was and continues to be of great interest to policy makers, as it may help determine a country's potential in global competitiveness. TIMSS Advanced has been repeated twice since 1995, in 2008 and 2015. The goal of TIMSS Advanced is to compare the performance of the most advanced final-year secondary students across participating countries in advanced mathematics and physics. The advanced mathematics content includes calculus; numbers, equations, and functions; validation and structure; probability and statistics; and geometry. The physics assessment includes questions about mechanics; electricity and magnetism; particle, quantum and other types of modern physics; heat; and wave phenomena. Sixteen countries participated in the 1995 TIMSS Advanced assessment; In the United States, the advanced mathematics assessment in 1995 was administered to a sample of 12th-grade students who had taken or were taking pre-calculus, calculus, AP calculus, or calculus and analytic geometry. The physics assessment in 1995 was administered to a sample of 12th-grade U.S. students who had taken or were taking at least one year-long course in physics, including physics I, physics II, advanced physics, and Advanced Placement physics. Nine education systems—all IEA member countries—participated in TIMSS Advanced 2015. In the United States, the advanced mathematics sample in 2015 was restricted to 12th-grade students who had taken or were taking a class teaching calculus to be more internationally comparable. The physics sample in 2015 was restricted to 12th-grade U.S. students who had taken or were taking a second-year physics class, an AP Physics course, or the equivalent. For additional information on TIMSS Advanced 1995 and 2015, please see *U.S. TIMSS 2015 and TIMSS Advanced 1995 & 2015 Technical Report and User's Guide* (Averett et al. 2017).

Periodicity
First conducted in the spring of 1995, TIMSS has been conducted every 4 years since then. TIMSS assessments are administered near the end of the school year in each country or other education system. In countries and other education systems in the Southern Hemisphere (where the school year typically ends in November or December) the assessment is conducted in October or November. In the Northern Hemisphere, the school year typically ends in June; so in these countries and other education systems the assessment is conducted in April, May, or June.

Data Availability
The U.S. National and International TIMSS data files are available at https://nces.ed.gov/timss/datafiles.asp. The TIMSS International Database, its user's guide and supplements, and supporting materials are available for download for each round of TIMSS from the TIMSS and PIRLS International Study Center website at https://timssandpirls.bc.edu/?_sm_au_=iMVrFHZZ1jJvfVTq.

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:

- How much mathematics and science have students learned?
- How well are students able to apply mathematics and science knowledge to problem solving?
- 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 or other education system, 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.

International Desired Population. The International Desired Population (also referred to as the International Target Population) is the target population which, in general, is defined for TIMSS at the lower grade as all students in their fourth year of formal schooling and, at the upper grade, all students in their eighth year of formal schooling. See the Target Population section below for more specifics.

National Desired Population. The stated objective in TIMSS is that the National Desired Population (also referred to as the National Target Population) within each country or education system be as close as possible to the International Desired Population. Using the International Desired Population as a basis, participating countries and other education systems must operationally define their populations for sampling purposes. Some countries and other education systems want to restrict coverage at the country or other education system level, by excluding remote regions or a segment of their country's education system, for example. Thus, the National Desired Population sometimes differs from the International Desired Population.

National Research Coordinators (NRCs). This is an official from each participating country or other education system appointed to implement national data collection and processing in accordance with international standards. In addition to selecting the sample of students, NRCs 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.

4. SURVEY DESIGN

Target Population

Student population. The International Desired Population for all countries and other education systems 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 and other education systems, the target grade is 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 and other education systems, the target grade is 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.

Teacher population. The target population is all mathematics and science teachers linked to the selected students. Note that these teachers are not a representative sample of teachers within the country or other education system. Rather, they are the mathematics and science teachers who teach a representative sample of students in two grades within the country or other education system.

School population. The target population is all eligible schools containing either one or more fourth-grade classrooms or either one or more eighth-grade classrooms. Although participating education systems were expected to include all students in the International Target Population, sometimes it was not feasible to include all of these students because of geographic or linguistic constraints specific to the country or other education system. Thus, each participating education system had its own National desired Target Population, which was the International Target Population reduced by the exclusions of those sections of the population that were not possible to assess. Working from the National Target Population, each participating education system had to operationalize the definition of its population for sampling purposes: i.e., identify their National Defined target Population. While each education system's National Defined Population ideally coincides with its National Target Population, in reality, there may be additional exclusions (e.g., of regions or school types) due to constraints of operationalizing the assessment.

Student – Advanced population. The students who participated in TIMSS Advanced were students in their final year of high school who had taken or were taking advanced mathematics or physics courses. For the list of courses considered advanced mathematics and physics courses for the purposes of TIMSS Advanced, please see *Trends in International Mathematics and Science Study (TIMSS), U.S. TIMSS 2015 and TIMSS Advanced 1995 & 2015 Technical Report and User's Guide* (Averett et al., 2017).

Sample Design

Each country or other education system participating in TIMSS, like the United States, is required to draw random samples of schools. Table TIMSS-1 displays the number of students who participated in TIMSS in the United States by year. It is not feasible to assess every fourth- and eighth-grade student in all participating countries and other education systems; therefore, a representative sample of fourth- and eighth-grade students is selected using a two-stage stratified cluster design.

First-stage sampling units. In the first stage of sampling, sampling statisticians selected individual schools with a probability-proportionate-to-size (PPS) approach, which means that each school's probability of selection is proportional to the estimated number of students enrolled in the target grade. Prior to sampling, statisticians assigned schools in the sampling frame to a predetermined number of explicit or implicit strata. Then, sampling staff identified sample schools using a PPS systematic sampling method. Statisticians also identified substitute schools (schools to replace original sampled school that refused to participate). The original and substitute schools were identified simultaneously.

Second-stage sampling units. In the second stage of sampling, statisticians selected classrooms within sampled schools using sampling software. The software uses a sampling algorithm for selecting classes that standardized the class sampling across schools and assures uniformity in the class selection procedures across participants. The software identified at least one classroom from a list of eligible classrooms that sampling staff prepared for each target grade. In various countries and other education systems, including the United States, more than one eligible classroom per target grade per school was selected when possible. All students in sampled classrooms were selected for assessment.

TIMSS guidelines call for a minimum of 150 schools to be sampled per grade, with a minimum of 4,000 students assessed per grade. The basic sample design of one classroom per target grade per school was designed to yield a total sample of approximately 4,500 students per population. Countries and other education systems with small class sizes or less than 30 students per school were directed to consider sampling more schools, more classrooms per school, or both, to meet the minimum target of 4,000 tested students.

TIMSS Advanced guidelines call for a minimum of 120 schools to be sampled, with a minimum of 3,600 students assessed per subject. Some countries selected complete classrooms of advanced students, while other countries, like the United States, randomly selected eligible advanced students for each subject.

In the United States and most other participating countries and education systems, the target populations of students corresponded to the fourth and eighth grades. In sampling these populations, TIMSS used a two-stage stratified cluster sampling design. The U.S. sampling frame was explicitly stratified by three categorical stratification variables: percentage of students eligible for free or reduced-price lunch, school control (public or private), and region of the country (Northeast, Central, West, Southeast). The U.S. sampling frame was implicitly stratified (that is, sorted for sampling) by two categorical stratification variables: community type or "locale" (four categories) and minority status (i.e., above or below 15 percent of the student population).

The first stage made use of a systematic PPS technique to select schools for the original sample from a sampling frame based on the 2015 NAEP school sampling frame. Data for public schools were taken from the Common Core of Data (CCD), and data for private schools were taken from the Private School Universe Survey (PSS). In addition, for each original school selected, the two neighboring schools in the sampling frame were designated as substitute schools. The first school following the original sample school was the first substitute and the first school preceding it was the second substitute. If an original school refused to participate, the first substitute was contacted. If that school also refused to participate, the second substitute was contacted. There were several constraints on the assignment of substitutes. 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. Furthermore, substitutes were required to be in the same implicit stratum as the sampled school.

The second stage consisted of selecting intact mathematics classes within each participating school. Schools provided lists of fourth- or eighth-grade classrooms. Within schools, classrooms with fewer than 15 students were collapsed into pseudo-classrooms, so that each classroom in the school's classroom sampling frame had at least 20 students. An equal probability sample of two classrooms was identified from the classroom frame for the school. In schools where there was only one classroom, this classroom was selected with certainty. All students in sampled classrooms and pseudo-classrooms were selected for assessment.

In this way, the overall sample design for the United States results in an approximately self-weighting sample of students, with each fourth- or eighth-grade student having a roughly equal probability of selection. Note that in small schools, a higher proportion of the classes (and therefore of the students) is selected, but this higher rate of selecting students in small schools is offset by a lower selection rate of small schools, as schools are selected with probability proportional to size.

U.S. TIMSS Advanced National Sample. In TIMSS Advanced 2015, the U.S. sample followed a two-stage sampling process with the first stage a sample of schools, and the second stage a sample of advanced students within schools rather than classrooms. The overlap with the 2015 NAEP grade 12 school sample was minimized for the 2015 TIMSS Advanced. The NAEP grade 12 sample was selected before the TIMSS Advanced sample. Thus the overlap between the samples was minimized when the TIMSS sample was selected.

A single TIMSS Advanced school sample consisting of 348 schools containing at least one twelfth-grade class was selected for both subjects (advanced mathematics and physics). The schools were selected with probability proportionate to the school's estimated grade enrollment of twelfth-graders from the 2015 NAEP school frame. Within each sampled school, separate student samples were selected for advanced mathematics and physics in systematic equal probability samples. The student sampling algorithm was designed to meet international guidelines with a target of 4,000 each in advanced mathematics and physics for a total of 8,000 sampled advanced students.

Assessment Design

With the large number of mathematics and science items, it was not possible for every student to respond to all items. To ensure broad subject-matter coverage without overburdening individual students, TIMSS assessments use a matrix-sampling technique that assigns each assessment item to one of a set of item blocks, and then assembles student test booklets by combining the item blocks according to a balanced design. Each student takes one booklet containing both mathematics and science items. Thus, the same students participated in both the mathematics and science testing.

In 1995 and 1999, the assessment consisted of eight booklets, each requiring 90 minutes of response time. Each participating student was assigned one booklet only. In accordance with the design, the mathematics and science items were assembled into 26 clusters. The secure trend items were in clusters A through H, and items replacing the released 1995 items in clusters I through Z. Eight of the clusters were designed to take 12 minutes to complete; 10 clusters, 22 minutes; and 8 clusters, 10 minutes.

In the TIMSS 2003 assessment design, the 313 fourth-grade mathematics and science items and the 383 eighth-grade items were divided among 28 item blocks at each grade, 14 mathematics blocks and 14 science blocks, with each block containing either mathematics items only or science items only. This general block design was the same for both grades, although the planned assessment time per block was 12 minutes for fourth grade and 15 minutes for eighth grade. There were 12 student booklets at each grade level, with six blocks of items in each booklet. To enable linking between booklets, each block appears in two, three, or four different booklets. The assessment time for each student was 72 minutes at fourth grade (six 12-minute blocks) and 90 minutes at eighth grade (six 15-minute blocks). The booklets were organized into two three-block sessions with a break between the parts.

The 2003 assessment was the first TIMSS assessment in which calculators were permitted, and so it was important that the design allow students to use calculators when working on the new 2003 items. However, because calculators were not permitted in TIMSS 1995 or 1999, the 2003 design also had to ensure that students did not use calculators when working on trend items from these assessments. The solution was to place the blocks containing trend items in Part I of the test booklets, to be completed without calculators before the break. After the break, calculators were allowed for the new items. To provide a more balanced design, however, and have information about differences with calculator access, two mathematics trend blocks and two science trend blocks also were placed in Part II of one booklet each. Note that calculators were allowed only at the eighth grade, and not at the fourth grade.

The TIMSS 2007 assessment contained 353 items at the fourth grade, including 179 in mathematics and 174 in science. At the eighth grade there were 429 items, 215 in mathematics and 214 items in science. At both grades, the TIMSS 2007 assessment involved assembling the items into 14 blocks of mathematics items and 14 blocks of science items, and then assembling the blocks into 14 booklets, each one including 2 blocks of mathematics items and 2 blocks of science items assembled according to a very careful rotated design. Each student was administered a single booklet.

In 2011, the fourth-grade assessment consisted of 14 booklets, each requiring approximately 72 minutes. The assessment was given in two 36-minute parts, with a 5- to 10-minute break in between. The student questionnaire was given after the second part of the assessment. Although it was untimed, it was allotted approximately 30 minutes for response time. The 14 booklets were rotated among students, with each participating student completing 1 booklet only. The mathematics and science items were each assembled separately into 14 blocks, or clusters, of items. Each of the 14 TIMSS 2011 booklets contained 4 blocks in total. Each block contained either mathematics items or science items only and each block occurred twice across the 14 books. For each subject, the secure, or trend, items used in prior assessments were included in 8 blocks, with the other 6 blocks containing new item.

The 2011 eighth-grade assessment followed the same pattern and consisted of 14 booklets, each requiring approximately 90 minutes of response time. The assessment was given in two 45-minute parts, with a 5- to 10-minute break in between. As in fourth grade, the student questionnaire was given after the second part of the assessment, and was allotted approximately 30 minutes of response time. The 14 booklets were rotated among students, with each participating student completing 1booklet only. The mathematics and science items were assembled into 14 blocks, or clusters, of items. Each block contained either mathematics items or science items only. For each subject, the secure, or trend, items used in prior assessments were included in 8 blocks, with the other 6 blocks containing new items. Each of the 14 TIMSS 2011 booklets contained 4 blocks in total. The TIMSS booklets administered in the state samples were exactly the same as those administered in the national sample.

During the TIMSS data collection period in spring 2011, TIMSS added state-level assessments in the nine states noted above (under Benchmarking studies) as well as a separate national sample of students who were also administered a set of braided booklets that included both NAEP and TIMSS blocks of items. During the TIMSS assessment window, the braided booklets were designed to appear as similar as possible to a regular TIMSS assessment booklet and were administered under the same conditions as TIMSS. The inclusion of two additional samples of students who were administered the braided booklets during the NAEP and TIMSS assessment windows was similar to the braided booklet design used for NAEP studies to maintain NAEP trends in 2009 reading and twelfth-grade mathematics.

The 2015 fourth and eighth-grade assessments replicated the 2011 assessment design, with all students completing one of 14 booklets, lasting either 72 or 90 minutes. Math and sciences items were separated in to 28 blocks of items and booklets consisted of 4 blocks.

The 2015 Advanced assessments consisted of 6 booklets for advanced mathematics and 6 booklets for physics, each requiring approximately 90 minutes. Booklets consisted of 3 blocks and students completed only one booklet. After the cognitive assessment, students completed a 30-minute questionnaire designed to provide information about their backgrounds, attitudes, and experiences in school.

Data Collection and Processing

Data collection. Across administrations, TIMSS emphasized the use of standardized procedures for countries and schools. Each country or other education system 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 or other education system was responsible for conducting quality control procedures and describing this effort in the NRC'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. NRCs 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 and other education systems. For each administration of TIMSS, the International Study Center developed manuals for the monitors and briefed them in training sessions about TIMSS, about the responsibilities of the national centers in conducting the study, and about their own roles and responsibilities.

Data entry and cleaning. Responsibility for data entry is taken by the NRC from each participating country or other education system. The data collected for TIMSS were entered into data files with a common international format. Data entry was facilitated by the use of common software available to all participating countries and other education systems. The software also facilitates 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 NRC with specific questions to be addressed. The NRC then provided the DPC with revisions to coding or solutions for anomalies. The DPC then compiled background univariate statistics and preliminary test scores using classical item analysis and item response theory (IRT).

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 and other education systems' 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, 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. Responses from the groups of students were assigned sampling weights to adjust for over- or under-representation during the sampling of a particular group. The use of sampling weights is necessary for the computation of sound, nationally representative estimates. The weight assigned to a student's responses is the inverse of the probability that the student is selected for the sample. When responses are weighted, none are discarded, and each contributes to the results for the total number of students represented by the individual student assessed. Weighting also adjusts for various situations (such as school and student nonresponse) because data cannot be assumed to be randomly missing. The internationally defined weighting specifications for TIMSS 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. All TIMSS 1995, 1999, 2003, 2007, 2011, and 2015 analyses are conducted using sampling weights. For 2011, though the national and state samples share schools, the samples are not identical school samples and, thus, weights are estimated separately for the national and state samples. A detailed description of this process is provided in the *Methods and Procedures in TIMSS and PIRLS 2011* (Martin & Mullis, 2011).

Scaling. TIMSS assessments use IRT procedures to produce scale scores that summarize 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. For TIMSS assessments after 1995, the propensity of students to answer questions correctly was estimated with a two-parameter IRT model for dichotomous constructed response items, a three-parameter IRT model for multiple choice response items, and a generalized partial credit IRT model for polytomous constructed response items. The two- and three-parameter models have been used in scaling the TIMSS assessments since 1999, because they more accurately account for the differences among items due to the ability to discriminate between students of high and low ability. A one-parameter model was used for scaling in the TIMSS 1995 assessment, and continues to be used to provide initial estimates of item difficulty.

Using IRT, the difficulty of each item, or item category, is deduced using information about how likely it is for students to get some items correct (or to get a higher rating on a constructed response item) versus other items. Once the parameters of each item are determined, the ability of each student can be estimated even when different students have been administered different items. The scale scores assigned to each student are estimated using *multiple imputation and plausible values*, with input from the IRT results. In the estimation process, achievement scores are expressed on a standardized logit scale ranging from -4 to +4; In order to make the scores more meaningful and to facilitate their interpretation, the scores for the first year (1995) are transformed to a scale of 0 to 1000, with a mean of 500 and a standard deviation of 100. Once this is done, subsequent waves of assessment can be linked to this metric, as next described.

To make scores from the second (1999) wave of data comparable to the first (1995) wave, two steps had to be taken. First, the 1995 and 1999 data for countries and education systems that participated in both years were scaled together to estimate item parameters. Ability estimates for all students (those assessed in 1995 and those assessed in 1999) based on the new item parameters were then estimated. To put these jointly calibrated 1995 and 1999 scores on the 1995 metric, a linear transformation was applied such that the jointly calibrated 1995 scores have the same mean and standard deviation as the original 1995 scores. Such a transformation also preserves any differences in average scores between the 1995 and 1999 waves of assessment. In order for scores resulting from subsequent waves of assessment (2003, 2007, 2011, and 2015) to be made comparable to 1995 scores (and to each other), the two steps above are applied sequentially for each pair of adjacent waves of data: two adjacent years of data are jointly scaled, then resulting ability estimates are linearly transformed so that the mean and standard deviation of the prior year is preserved. As a result, the transformed scores are comparable to all previous waves of the assessment and longitudinal comparisons between all waves of data are meaningful.

To facilitate the joint calibration of scores from adjacent years of assessment, common test items are included in successive administrations. This also enables the comparison of item parameters (difficulty and discrimination) across administrations. If item parameters change dramatically across administrations, they are dropped from the current assessment so that scales can be more accurately linked across years. In this way even if the average ability levels of students in countries and education systems participating in TIMSS changes over time, the scales still can be linked across administrations.

Scaling for TIMSS Advanced follows a similar process, using data from the 1995, 2008, and 2015 administrations. More detailed information can be found in the *Methods and Procedures in TIMSS 2015* (Martin, Mullis, and Hooper 2016a) and in *Methods and Procedures in TIMSS Advanced 2015* (Martin, Mullis, and Hooper 2016b).

Imputation and plausible values. 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 accommodate this situation, during the scaling process plausible values were estimated to characterize students participating in the assessment, given their background characteristics. Plausible values are imputed values and not test scores for individuals in the usual sense. In fact, they are biased estimates of the proficiencies of individual students. Plausible values do, however, provide unbiased estimates of population characteristics (e.g., means and variances of demographic subgroups).

Plausible values represent what the performance of an individual on the entire assessment might have been, had it been observed. The values are estimated as random draws (usually five) 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. Differences between plausible values drawn for a single individual quantify the degree of error (the width of the spread) in the underlying distribution of possible scale scores that could have caused the observed performances.

Scale anchoring and international benchmarks. International benchmarks for achievement were developed in an attempt to provide a concrete interpretation of what the scores on the TIMSS mathematics and science achievement scales mean (for example, what it means to have a scale score of 513 or 426). To describe student performance at various points along the TIMSS mathematics and science achievement scales, TIMSS uses scale anchoring to summarize and describe student achievement above four thresholds on the mathematics and science scales—*Advanced* (625), *High* (550), *Intermediate* (475), and *Low* (400) international benchmarks. The TIMSS Advanced uses *Advanced* (625), *High* (550), and *Intermediate* (475) international benchmarks.

Scale anchoring involves selecting benchmarks (scale points) on the TIMSS achievement scales to be described in terms of student performance. Once benchmark scores have been chosen, items are identified that students are likely to score highly on. The content of these items describes what students at each benchmark level of achievement know and can do. To interpret the content of anchored items, these items are grouped by content area within benchmarks and reviewed by mathematics and science experts. These experts focus on the content of each item and describe the kind of mathematics or science knowledge demonstrated by students answering the item correctly. The experts then provide a summary description of performance at each anchor point leading to a content-referenced interpretation of the achievement results. Prior to TIMSS 2003, benchmarks were defined primarily in terms of percentiles.

Recent Changes

TIMSS is in the process of transitioning to a digitally-based assessment called eTIMSS. Beginning in 2019 TIMSS items will be administered via computers or tablets. In addition to items typically found in a paper-and-pencil test, eTIMSS will include *new* innovative problem solving and inquiry tasks, known as PSIs. These PSIs are designed to simulate real world and laboratory situations where students can integrate and apply process skills and content knowledge to solve mathematics problems and conduct scientific experiments or investigations. The new PSI tasks involve visually attractive, interactive scenarios that present students with adaptive and responsive ways to follow a series of steps toward a solution. About half of the education systems, including the United States, are expected to participate in eTIMSS in the 2019 cycle, while the other half will administer TIMSS in a paper and pencil format as in previous assessments.

Future Plans

The next TIMSS data collection will take place in spring 2019.

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. However, just as most surveys with complexity similar to TIMSS would have the possibility of error, potential sources of error are next discussed.

Estimates produced using data from TIMSS are subject to two types of error—nonsampling and sampling errors. Nonsampling errors can be due to errors made in collecting and processing data. Sampling errors can occur because the data were collected from a sample rather than a complete census of the population.

Sampling Error

Sampling errors arise when a sample of the population, rather than the whole population, is used to estimate some statistic. Different samples from the same population would likely produce somewhat different estimates of the statistic in question. This fact means that there is a degree of uncertainty associated with statistics estimated from a sample. This uncertainty is referred to as sampling variance and is usually expressed as the standard error of a statistic estimated from sample data. The approach used for calculating standard errors in TIMSS is jackknife repeated replication (JRR). Standard errors can be used as a measure for the precision expected from a particular sample.

Confidence intervals provide a way to make inferences about population statistics in a manner that reflects the sampling error associated with the statistic, using the standard error. Assuming a normal distribution, the population value of this statistic can be inferred to lie within the confidence interval in 95 out of 100 replications of the measurement on different samples drawn from the same population. For example, the average mathematics score for the U.S. eighth-grade students was 518 in 2015, and this statistic had a standard error of 3.1. Therefore, it can be stated with 95 percent confidence that the actual average of U.S. eighth-grade students in 2015 was between 512 and 524 (1.96 x 3.1 = 6.1; confidence interval = 518 plus and minus 6.1).

Nonsampling Error

Nonsampling error is a term used to describe variations in the estimates that may be caused by population coverage limitations, nonresponse bias, and measurement error, as well as data collection, processing, and reporting procedures. The sources of nonsampling errors are typically problems like unit and item nonresponse, differences in respondents' interpretations of the meaning of the survey questions, response differences related to the particular time the survey was conducted, and mistakes in data preparation.

Unit nonresponse. Unit nonresponse error results from nonparticipation of schools and students. Weighted and unweighted response rates are computed for each participating country or other education system 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 and other education systems, 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 and other education systems. 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 TIMSS-2 shows the weighted unit level response rates for the U.S. for the data collections of 1995, 1999, 2003, 2007, 2011, 2015 for grades 4 and 8, and 2015 for grade 12 advanced 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 and other education systems. In order to minimize the effects of measurement error, TIMSS carries out a special test called the Test-Curriculum Matching Analysis. Each country or other education system 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.

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 or other education system, 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. (See also NAEP/TIMSS linking study discussion above.)

The ability to compare data across different countries and other education systems 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 or other education system are functionally identical to those used in other countries and other education systems. Because questionnaires are originally developed in English and later translated into the language of each of the TIMSS countries and other education systems, some differences do exist in the wording of questions. NRCs from each country or other education system 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 and other education systems 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.

Table TIMSS-1. Number of U.S. students and schools participating in TIMSS, by year: 1995–2015

Year	Schools	Students
1995[1]	730	22,088
1999[2]	221	9,072
2003	480	18,741
2007	496	15,273
2011	870	23,046
2015[3]	902	26,136

[1] Includes grades 3, 4, 7, and 8.
[2] TIMSS did not collect data from grade 4 in 1999.
[3] Includes TIMSS Advanced schools.
NOTE: Includes grades 4 and 8 unless otherwise noted.
SOURCE: TIMSS publications NCES 2003-075, NCES 2006-058, NCES 2009-012, NCES 2013-046, and NCES 2018-020, available at https://nces.ed.gov/pubsearch/getpubcats.asp?sid=073; and *TIMSS Technical Report: Volume II,* available at https://timssandpirls.bc.edu/timss1995i/TIMSSPDF/TR2book.pdf.

Table TIMSS-2. TIMSS weighted U.S. unit-level response rates, by level, year, and grade: 1995–2015

Year and grade	School	Student	Overall
1995			
Fourth grade	85	94	80
Eighth grade	85	92	78
1999			
Fourth grade	†	†	†
Eighth grade	90	94	85
2003			
Fourth grade	82	95	78
Eighth grade	78	94	73
2007			
Fourth grade	89	95	84
Eighth grade	83	93	77
2011			
Fourth grade	84	95	80
Eighth grade	87	94	81
2015			
Fourth grade	85	96	81
Eighth grade	84	94	78
Twelfth grade Advanced Math	76	87	66
Twelfth grade Physics	68	85	58

† Not applicable.

NOTE: TIMSS did not collect data from grade 4 in 1999.
SOURCE: TIMSS publications NCES 2003-075, NCES 2006-058, NCES 2009-012, NCES 2013-046, and NCES 2018-020, available at https://nces.ed.gov/pubsearch/getpubcats.asp?sid=073; and *TIMSS Technical Report: Volume II,* available at https://timssandpirls.bc.edu/timss1995i/TIMSSPDF/TR2book.pdf.

6. CONTACT INFORMATION

For content information on TIMSS, contact:

Lydia Malley
Phone: (202) 245-7266
E-mail: lydia.malley@ed.gov

Mailing Address
National Center for Education Statistics
Institute of Education Sciences
Potomac Center Plaza
550 12th Street, SW
Washington, DC 20202

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

Gonzales, P., Calsyn, C., Jocelyn, L., Mak, K., Kastberg, D., Arafeh, S., Williams, T., and Tsen, W. (2000). *Pursuing Excellence: Comparisons of International Eighth-Grade Mathematics and Science Achievement from a U.S. Perspective, 1995 and 1999* (NCES 2001-028). National Center for Education Statistics, U.S. Department of Education. Washington, DC. https://nces.ed.gov/pubs2001/2001028.pdf

Gonzales, P., Guzmán, J. C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., and Williams, T. (2004). *Highlights From the Trends in International Mathematics and Science Study (TIMSS) 2003* (NCES 2005-005). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. https://nces.ed.gov/pubs2005/2005005.pdf

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., and Brenwald, S. (2008). *Highlights From TIMSS 2007: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context* (NCES 2009-001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. https://nces.ed.gov/pubs2009/2009001.pdf

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., and Jenkins, F. (2012). *Highlights from TIMSS 2011: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context* (NCES 2013-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. https://nces.ed.gov/pubs2013/2013009.pdf

Provasnik, S., Malley, L., Stephens, M., Landeros, K., Perkins, R., and Tang, J. (2016). *Highlights from TIMSS and TIMSS Advanced 2015: Mathematics and Science Achievement of U.S. Students in Grades 4 and 8 and in Advanced Courses at the End of High School in an International Context* (NCES 2017-002). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. https://nces.ed.gov/pubs2017/2017002.pdf

U.S. Department of Education, National Center for Education Statistics. (2000). *Mathematics and Science in the Eighth Grade: Findings from the Third International Mathematics and Science Study*, NCES 2000-014, Washington, DC. https://nces.ed.gov/pubs2000/2000014.pdf

Survey Design

Averett, C., Ferraro, D., Tang, J., Erberber, E., and Stearns P. (2017). *Trends in International Mathematics and Science Study (TIMSS), U.S. TIMSS 2015 and TIMSS Advanced 1995 & 2015 Technical Report and User's Guide* (NCES 2018-020). U.S. Department of Education. Washington, DC: National Center for Education Statistics, Institute of Education Sciences. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018020

Ferraro, D., and Van de Kerckhove, W. (2006). *Trends in International Mathematics and Science Study (TIMSS) 2003 Nonresponse Bias Analysis* (NCES 2007-044). U.S. Department of Education. Washington, DC: National Center for Education Statistics. https://nces.ed.gov/pubs2007/2007044.pdf

Garnier, H.E., Lemmens, M., Druker, S.L., and Roth, K.J. (2011). *Third International Mathematics and Science Study 1999 Video Study Technical Report Volume 2:*

Science (NCES 2011-049). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. https://nces.ed.gov/pubs2011/2011049.pdf

Kastberg, D., Roey, S., Ferraro, D., Lemanski, N., and Erberber, E. (2013). *U.S. TIMSS and PIRLS 2011 Technical Report and User's Guide* (NCES 2013-046). U.S. Department of Education. Washington, DC: National Center for Education Statistics. https://nces.ed.gov/pubs2013/2013046_1.pdf

Martin, M. O., and Kelly, D. L. (Eds.). (1996). *TIMSS Technical Report: Volume I: Design and Development.* Boston College, International Study Center. Chestnut Hill, MA. https://timss.bc.edu/timss1995i/TIMSSPDF/TRall.pdf

Martin, M. O., and Kelly, D. L. (Eds.). (1997). *TIMSS Technical Report, Volume II: Implementation and Analysis, Primary and Middle School Years*. Boston College, International Study Center. Chestnut Hill, MA. https://timssandpirls.bc.edu/timss1995i/TIMSSPDF/TR2book.pdf

Martin, M. O., and Kelly, D. L. (Eds.). (1998). *TIMSS Technical Report: Volume III: Implementation and Analysis, Final Year of Secondary School*. Boston College, International Study Center. Chestnut Hill, MA. https://timssandpirls.bc.edu/timss1995i/TIMSSPDF/TR2book.pdf

Martin, M. O., Gregory, G. D., and Stemler, S. E. (Eds.). (2000). *TIMSS 1999 Technical Report*. Boston College, International Study Center. Chestnut Hill, MA. https://timss.bc.edu/timss1999i/pdf/T99_TR.book.pdf

Martin, M. O., Mullis, I. V. S., and Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 Technical Report: Findings From IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades*. Boston College, International Study Center. Chestnut Hill, MA. https://timss.bc.edu/PDF/t03_download/T03TECHRPT.pdf

Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/methods/index.html

Martin, M.O., Mullis, I.V.S., and Hooper, M. (Eds.). (2016a). *Methods and Procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://timss.bc.edu/publications/timss/2015-methods.html

Martin, M.O., Mullis, I.V.S., and Hooper, M. (Eds.). (2016b). *Methods and Procedures in TIMSS Advanced 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://timss.bc.edu/publications/timss/2015-a-methods.html

Olson, J. F., Martin, M. O., and Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 Technical Report*. Boston College International Study Center, Chestnut Hill. https://timssandpirls.bc.edu/TIMSS2007/PDF/TIMSS2007_TechnicalReport.pdf

Rubin, D. B. (2008). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons. http://www.worldcat.org/title/multiple-imputation-for-nonresponse-in-surveys/oclc/800810137&referer=brief_results

Williams, T., Ferraro, D., Roey, S., Brenwald, S., Kastberg, D., Jocelyn, L., Smith, C., and Stearns, P. (2009). *TIMSS 2007 U.S. Technical Report and User Guide* (NCES 2009-012). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. https://nces.ed.gov/pubs2009/2009012.pdf

U.S. Department of Education, National Center for Education Statistics. *Third International Mathematics and Science Study 1999 Video Study Technical Report Volume I: Mathematics,* NCES (2003-012), by Jennifer Jacobs, Helen Garnier, Ronald Gallimore, Hilary Hollingsworth, Karen Bogard Givvin, Keith Rust, Takako Kawanaka, Margaret Smith, Diana Wearne, Alfred Manaster, Wallace Etterbeek, James Hiebert, and James Stigler. Washington, DC: 2003. https://nces.ed.gov/pubs2003/2003012.pdf

Uses of Data

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

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

National Center for Education Statistics, U.S. Department of Education. Washington, DC. https://nces.ed.gov/pubs98/98500.pdf

Kastberg, D., Roey, S., Jocelyn, L., and Williams, T. (2005). *Trends in International Mathematics and Science Study (TIMSS) 2003 Data Analysis User's Guide* (NCES 2006-058). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. https://nces.ed.gov/timss/pdf/CB2003_users_guide.pdf

Rutkowski, L., González, E., Joncas, M., and von Davier, M. (2010). *International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting.* Educational Researcher 39 (2), 142-151. https://journals.sagepub.com/doi/pdf/10.3102/0013189X10363170

Data Quality and Comparability

Martin, M. O., and Mullis, I. V. S. (Eds.). (1996). *Quality Assurance in Data Collection.* Boston College, International Study Center. Chestnut Hill, MA. https://timssandpirls.bc.edu/timss1995i/TIMSSPDF/QAall.pdf