1

# Third International Mathematics and Science Study: An Overview

Michael O. Martin
Dana L. Kelly
*Boston College*

## 1.1 INTRODUCTION

The Third International Mathematics and Science Study (TIMSS) is the largest and most ambitious international comparative study of student achievement to date. Under the auspices of the International Association for the Evaluation of Educational Achievement (IEA), TIMSS brought together educational researchers from more than 50 countries to design and implement a study of the teaching and learning of mathematics and science in each country.

TIMSS is a cross-national survey of student achievement in mathematics and science that was conducted at three levels of the educational system:

- The two adjacent grades with the largest proportion of 9-year-olds at the time of testing (third and fourth grades in many countries)
- The two adjacent grades with the largest proportion of 13-year-olds at the time of testing (seventh and eighth grades in many countries)
- The final year of secondary education

Forty-five countries took part in the survey (see Figure 1.1). The students, their teachers, and the principals of their schools were asked to respond to questionnaires about their backgrounds and their attitudes, experiences, and practices in the teaching and learning of mathematics and science.

A project of the magnitude of TIMSS necessarily has a long life cycle. Planning for TIMSS began in 1989; the first meeting of National Research Coordinators was held in 1990; data collection took place from the latter part of 1994 through 1995; the first international reports were released in November 1996 and June 1997, and further international reports will be issued through 1998. A large number of people contributed to the many strands that made up TIMSS. They came from all areas of educational assessment and included specialists in policy analysis, mathematics education, science education, curriculum design, survey research, test construction, psychometrics, survey sampling, and data analysis.

In addition to disseminating its findings as widely as possible, TIMSS aims to document fully the procedures and practices used to achieve the study goals. The *TIMSS Technical Report* series is an important part of this effort. Because of the long life cycle of TIMSS, and the involvement of so many individuals at its various stages, the *TIMSS*

**Figure 1.1 Countries Participating in TIMSS***

| | |
|---|---|
| • Argentina | • Korea, Republic of |
| • Australia | • Kuwait |
| • Austria | • Latvia |
| • Belgium† | • Lithuania |
| • Bulgaria | • Mexico |
| • Canada | • Netherlands |
| • Colombia | • New Zealand |
| • Cyprus | • Norway |
| • Czech Republic | • Philippines |
| • Denmark | • Portugal |
| • England | • Romania |
| • France | • Russian Federation |
| • Germany | • Scotland |
| • Greece | • Singapore |
| • Hong Kong | • Slovak Republic |
| • Hungary | • Slovenia |
| • Iceland | • South Africa |
| • Indonesia | • Spain |
| • Iran, Islamic Republic | • Sweden |
| • Ireland | • Switzerland |
| • Israel | • Thailand |
| • Italy | • United States |
| • Japan | |

* Argentina, Italy, and Indonesia were unable to complete the steps necessary for their data to appear in the TIMSS international reports or the TIMSS International Database. Mexico participated in the testing portion of TIMSS, but chose not to release its results.

† The Flemish and French educational systems in Belgium participated separately.

*Technical Report* is presented in several volumes, each documenting a major stage of the project and produced soon after the completion of that stage. Accordingly, *TIMSS Technical Report, Volume I: Design and Development* (Martin and Kelly, 1996) documents the study design and the development of TIMSS up to, but not including, the operational stage of main data collection.

This volume, *TIMSS Technical Report, Volume II: Implementation and Analysis*, describes the implementation of the design and the procedures underlying the analysis and reporting of data for two of the three TIMSS student populations (two adjacent grades with the most 9-year-olds and two adjacent grades with the most 13-year-olds). The results for these populations have been published in five volumes:

- *Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study*
- *Science Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study*
- *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*

- *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*
- *Performance Assessment in IEA's Third International Mathematics and Science Study*

These reports have been widely disseminated and are available on the internet (http://wwwcsteep.bc.edu/timss). The entire TIMSS international database containing the achievement and background data underlying these reports has been released and is available at the TIMSS website and through IEA Headquarters. The database is accompanied by a User's Guide and full documentation.

A third volume in the technical report series, to be published in 1998, will document the implementation and analysis for the assessment of students in their final year of secondary school.

This chapter provides an overview of the development and design of TIMSS, including the conceptual framework, student populations, instrument design, and management and organization of the study. This information is presented in detail in *TIMSS Technical Report, Volume I: Design and Development* (Martin and Kelly, 1996).[1] This chapter also describes the contents of the remaining chapters in this volume.

## 1.2 THE CONCEPTUAL FRAMEWORK FOR TIMSS

IEA studies have as a central aim the measurement of student achievement in school subjects, with a view to learning more about the nature and extent of student achievement and the context in which it occurs. The ultimate goal is to isolate the factors directly relating to student learning that can be manipulated through policy changes in, for example, curricular emphasis, allocation of resources, or instructional practices. Clearly, an adequate understanding of the influences on student learning can come only from careful study of the nature of student achievement and from the characteristics of the learners themselves, the curriculum they follow, the teaching methods of their teachers, and the resources in their classrooms and their schools. Such school and classroom features are of course embedded in the community and the educational system, which in turn are aspects of society in general.

The designers of TIMSS chose to focus on curriculum as a broad explanatory factor underlying student achievement (Robitaille and Garden, 1996). From that perspective, curriculum was considered to have three manifestations: what society would like to see taught (the intended curriculum), what is actually taught in the classroom (the implemented curriculum), and what the students learn (the attained curriculum). This conceptualization was first developed for the IEA's Second International Mathematics Study (Travers and Westbury, 1989).

The three aspects of the curriculum bring together three major influences on student achievement. The intended curriculum states society's goals for teaching and learning. These expectations reflect the ideals and traditions of the greater society, and are con-

[1] Appendix A contains the table of contents for *TIMSS Technical Report, Volume I: Design and Development.*

strained by the resources of the educational system. The implemented curriculum is what is taught in the classroom. Although presumably inspired by the intended curriculum, the actual classroom events are usually determined in large part by the classroom teacher, whose behavior may be greatly influenced by his or her own education, training, and experience, by the nature and organizational structure of the school, by interaction with teaching colleagues, and by the composition of the student body. The attained curriculum is what the students actually learn. Student achievement depends partly on the implemented curriculum and its social and educational context, and to a large extent on the characteristics of individual students, including ability, attitude, interests, and effort.

While the three-strand model of curriculum draws attention to three different aspects of the teaching and learning enterprise, it does have a unifying theme: the provision of educational opportunities to students. The curriculum, both as intended and as implemented, provides and delimits learning opportunities for students.

Considering the curriculum as a channel through which learning opportunities are offered to students leads to a number of general questions that can be used to organize inquiry about that process. In TIMSS, four general research questions helped to guide the development of the study:

- What are students expected to learn?
- Who provides the instruction?
- How is instruction organized?
- What have students learned?

The first of these questions concerns the intended curriculum, and is addressed in TIMSS by an extensive comparative analysis of curricular documents and textbooks from each participating country. The second and third questions address major aspects of the implemented curriculum: what are the characteristics of the teaching force in each country (education, experience, attitudes, and opinions), and how do teachers go about instructing their students (what teaching approaches do they use, and what curricular areas do they emphasize)? The final question deals with the attained curriculum: what have students learned, how does student achievement vary from country to country, and what factors are associated with student learning?

The study of the intended curriculum was a major part of the initial phase of the project. The TIMSS curriculum analysis consisted of an ambitious content analysis of curriculum guides, textbooks, and questionnaires completed by curriculum experts and education specialists. Its aim was a detailed rendering of the curricular intentions of the participating countries.

Data for the study of the implemented curriculum were collected as part of a large-scale international survey of student achievement. Questionnaires completed by the mathematics and science teachers of the students in the survey, and by the principals of their schools, provided information about the topics in mathematics and science that were taught, the instructional methods adopted in the classroom, the organizational structures that supported teaching, and the factors that were seen to facilitate or inhibit teaching and learning.

The student achievement survey provides data for the study of the attained curriculum. The wide-ranging mathematics and science tests that were administered to nationally representative samples of students at three levels of the educational system provide 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. Information about students' characteristics, and about their attitudes, beliefs, and experiences, comes from a questionnaire completed by each participating student. This information will help to identify the student characteristics associated with learning and provide a context for the study of the attained curriculum.

### 1.3 THE TIMSS CURRICULUM FRAMEWORKS

The TIMSS curriculum frameworks (Robitaille et al., 1993) were conceived early in the study as an organizing structure within which the elements of school mathematics and science could be described, categorized, and discussed. In the TIMSS curriculum analysis, the frameworks provided the system of categories by which the contents of textbooks and curriculum guides were coded and analyzed. The same system of categories was used to collect information from teachers about what mathematics and science they have taught. Finally, the system formed a basis for constructing the TIMSS achievement tests.

The TIMSS curriculum frameworks have their antecedents in the content-by-cognitive-behavior grids used in earlier studies (e.g., Travers and Westbury, 1989) to categorize curriculum units or achievement test items. A content-by-cognitive-behavior grid is usually represented as a matrix, or two-dimensional array, where the horizontal dimension represents a hierarchy of behavior levels at which students may perform, while the vertical dimension specifies subject-matter topics or areas. Individual items or curriculum units are assigned to a particular cell of the matrix. These grids facilitate comparisons of curricula and the development of achievement tests by summarizing curriculum composition and test scope.

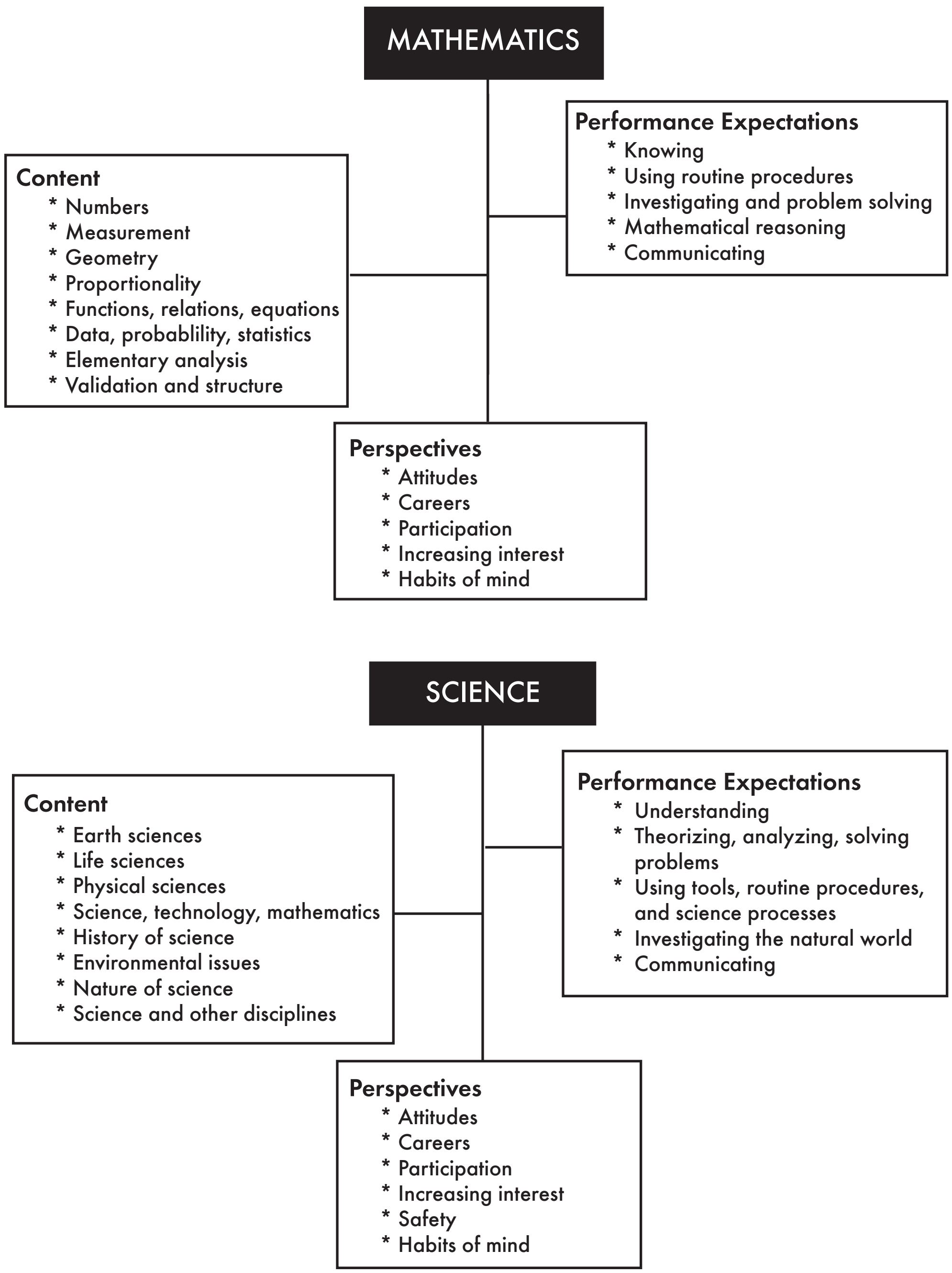
The TIMSS curriculum frameworks are an ambitious attempt to expand the concept of the content-by-cognitive-behavior grids.

> For the purposes of TIMSS, curriculum consists of the concepts, processes, and attitudes of school mathematics and science that are intended for, implemented in, or attained during students' schooling experiences. Any piece of curriculum so conceived – whether intended, implemented, or attained, whether a test item, a paragraph in an "official" curriculum guide, or a block of material in a student textbook – may be characterized in terms of three parameters: subject-matter content, performance expectations, and perspectives or context (Robitaille et al., 1993, p.43).

Subject-matter content, performance expectations, and perspectives constitute the three dimensions, or aspects, of the TIMSS curriculum frameworks. *Subject-matter con-*

*tent* refers simply to the content of the mathematics or science curriculum unit or test item under consideration. *Performance expectations* are a reconceptualization of the earlier cognitive-behavior dimension. Their purpose is to describe, in a non-hierarchical way, the many kinds of performance or behavior that a given test item or curriculum unit might elicit from students. The *perspectives* aspect is relevant to analysis of documents such as textbooks, and is intended to permit the categorization of curricular components according to the nature of the discipline as reflected in the material, or in the context within which the material is presented.

**Figure 1.2 The Major Categories of the TIMSS Curriculum Frameworks**



Each of the three aspects is partitioned into a number of categories, which are partitioned into subcategories, which are further partitioned as necessary. The curriculum frameworks (the major categories are shown in Figure 1.2) were developed separately for mathematics and science. Each framework has the same general structure, and includes the same three aspects: subject-matter content, performance expectations, and perspectives.[2]

## 1.4 THE TIMSS CURRICULUM ANALYSIS

The TIMSS analysis of the intended curriculum focused on curriculum guides, textbooks, and experts as the sources of information about each country's curricular intentions. The investigation of variations in curricula across countries involved three major data collection efforts: (1) a detailed page-by-page document analysis of curriculum guides and selected textbooks; (2) mapping (or tracing) the coverage of topics in the TIMSS frameworks across textbook series and curriculum guides for all pre-university grades; and (3) collecting questionnaire data designed to characterize the organization of the educational system, the decision-making process regarding learning goals, and the general contexts for learning mathematics and science.

In the document analysis, the participating countries partitioned the curriculum guides and textbooks into homogeneous blocks and coded the substance of each block according to the TIMSS frameworks. The document analysis provided detailed information for the grades studied, but does not allow tracing the full continuum of topic coverage through all the grades in the pre-university system. Information on continuity of coverage was obtained by tracing topics through the curriculum from the beginning of schooling to the end of secondary school. The topic tracing for TIMSS included two procedures. In the first, curriculum experts within each country characterized the points at which instruction is begun, ended, and concentrated on for all topics in the frameworks. In this effort, each topic was treated discretely even though many of the topics are related in terms of their specification in the learning goals. Therefore, for six topics each within mathematics and the sciences, a second tracing procedure was used, based on the curriculum guides that specified how subtopics fit together in the coverage of a topic as a whole. The twelve topics were selected as being of special interest to the mathematics and science education communities. Taken together, the two tracing procedures offer both breadth, covering all topics across all grades, and depth in terms of covering a limited number of topics across all grades (Beaton, Martin, and Mullis, 1997).

The TIMSS curriculum analysis was conducted by the Survey of Mathematics and Science Opportunities (SMSO) project of Michigan State University, under the direction of William H. Schmidt. The initial results of this study are available in two volumes: *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Mathematics* (Schmidt et al., 1996) and *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Science* (Schmidt et al., 1997).

[2] The complete TIMSS curriculum frameworks can be found in Robitaille et al. (1993).

## 1.5 THE STUDENT POPULATIONS

TIMSS chose to study student achievement at three points in the educational process: at the earliest point at which most children are considered old enough to respond to written test questions (Population 1); at a point at which students in most countries have finished primary education and are beginning secondary education (Population 2); and at the end of secondary education (Population 3). The question whether student populations should be defined by chronological age or grade level in school is one that faces all comparative surveys of student achievement. TIMSS addressed this issue by defining (for Populations 1 and 2) the target population as the pair of adjacent grades that contains the largest proportion of a particular age group (9-year-olds for Population 1, and 13-year-olds for Population 2). Most cross-country comparisons in TIMSS are based on grade levels, since educational systems are organized around grade levels; but it is also possible to make cross-country comparisons on the basis of student age for countries where the pair of adjacent grades contains a high percentage of the age cohort.

The student populations in TIMSS are defined below.

- Population 1: all students enrolled in the two adjacent grades that contain the largest proportion of students of age 9 years at the time of testing
- Population 2: all students enrolled in the two adjacent grades that contain the largest proportion of students of age 13 years at the time of testing
- Population 3: all students in their final year of secondary education, including students in vocational education programs; Population 3 has two optional subpopulations: students having taken advanced mathematics and students having taken physics

Population 2 was compulsory for all participating countries. Countries could choose whether or not to participate in Populations 1 and 3 (and the subpopulations of Population 3). The Population 3 implementation and analysis is addressed in the forthcoming *TIMSS Technical Report, Volume III*.

## 1.6 SURVEY ADMINISTRATION DATES FOR POPULATIONS 1 AND 2

Since school systems in countries in the Northern and the Southern Hemispheres do not have the same school year, TIMSS had to set two survey administration schedules. Countries on the Southern Hemisphere timeline administered the tests between September and November 1994. Countries on the Northern Hemisphere timeline administered the tests between February and May 1995. These periods were chosen with the aim of testing students as late in the school year as practical so as to reflect the knowledge gained throughout the year.

### 1.7 THE TIMSS ACHIEVEMENT TESTS FOR POPULATIONS 1 AND 2

The measurement of student achievement in a school subject is a challenge under any circumstances. The measurement of student achievement in two subjects at three student levels in 45 countries (through the local language of instruction), in a manner that does justice to the curriculum to which the students have been exposed and that allows the students to display the full range of their knowledge and abilities, is indeed a formidable task. This, nonetheless, is the task that TIMSS set for itself.

The IEA had conducted separate studies of student achievement in mathematics and science on two earlier occasions (mathematics in 1964 and 1980-82, and science in 1970-71 and 1983-84), but TIMSS was the first IEA study to test mathematics and science together. Since there is a limit to the amount of student testing time that may reasonably be requested from schools, assessing student achievement in two subjects simultaneously constrains the number of questions that may be asked, and therefore limits the amount of information that may be collected from any one student.

Recent IEA studies, particularly the Second International Mathematics Study (Robitaille and Garden, 1989), placed great emphasis on the role of curriculum in all its manifestations in the achievement of students. This concern with curriculum coverage, together with the desire of curriculum specialists and educators generally to ensure that both subjects be assessed as widely as possible, led to pressure for ambitious coverage in the TIMSS achievement tests. Further, there was concern that the assessment of student knowledge and abilities be as "authentic" as possible, with the questions asked and the problems posed in a form that students are used to. In particular, test items were to make use of a variety of task types and response formats, and not exclusively multiple choice.

Reconciling the demands for the form and extent of the TIMSS achievement tests was a lengthy and difficult process. It involved extensive consensus building through which the concerns of all interested parties had to be balanced so as to produce a reliable measuring instrument that could serve as a valid index of student achievement in mathematics and science in all of the participating countries. The tests that finally emerged were necessarily a compromise between what might have been attempted in an ideal world of infinite time and resources, and the real world of short timelines and limited resources.

Despite the need for compromise in some areas, the TIMSS achievement tests have gone a long way toward meeting the ideals of their designers. They cover a wide range of subject matter, yielding, in Population 2, estimates of student proficiency in 11 areas or content area "reporting categories" of mathematics and science (6 for mathematics and 5 for science), as well as overall mathematics and science scores. In Population 1 there were ten content area reporting categories (six for mathematics and four for science), as well as overall mathematics and overall science scores. The test items include both multiple-choice and free-response items. The latter come in two varieties: "short-answer," where the student supplies a brief written response; and "extended-response," where students must provide a more extensive written answer, and sometimes explain their reasoning. The free-response items are scored using a unique twodigit coding rubric that yields both a score for the response and an indication of the nature of the response. The free-response data will be a rich source of information about student understanding, and misunderstanding, of mathematics and science topics.

The wide coverage and detailed reporting requirements of the achievement tests resulted in a pool of mathematics and science items in Population 2 that, if all of them were to be administered to any one student, would take almost seven hours of testing. Since the consensus among the National Research Coordinators (NRCs) was that 70 minutes was the most that could be expected for Population 1 and 90 minutes the most that could be expected for Population 2, a way of dividing the item pool among the students had to be found. Matrix sampling provided a solution by assigning subsets of items to individual students in such a way as to produce reliable estimates of the performance of the population on all the items, even though no student responded to the entire item pool. The TIMSS test design uses a variant of matrix sampling to map the mathematics and science item pool into eight student booklets each for Population 1 and Population 2 (see Adams and Gonzalez, 1996).

The TIMSS test design sought breadth of subject-matter coverage and reliable reporting of summary statistics for each of the reporting categories. However, because of the interest in the details of student performance at the item level, at least some of the items also had to be administered to enough students to permit accurate reporting of their item statistics. The TIMSS item pool for both Populations 1 and 2 was therefore divided into 26 sets, or clusters, of items. These were then arranged in various ways to make up eight test booklets, each containing seven item clusters. One cluster, the core cluster, appears in each booklet. Seven "focus" clusters appear in three of the eight booklets. The items in these eight clusters should be sufficient to permit accurate reporting of their statistics. There are also 12 "breadth" clusters, each of which appears in just one test booklet. These help ensure wide coverage, but the accuracy of their statistics may be relatively low. Finally, there are eight "free-response clusters," each of which appears in two booklets. These items are a rich source of information about the nature of student responses, and should have relatively accurate statistics.

The eight student booklets were distributed systematically in each classroom, one per student. This is efficient from a sampling viewpoint, and since there are eight substantially different booklets in use in each classroom, it reduces the likelihood of students copying answers from their neighbors.

### 1.8 PERFORMANCE ASSESSMENT

Educators have long advocated the use of practical tasks to assess student performance in mathematics and particularly in science. The inclusion of such a "performance assessment" was a design goal from the beginning of TIMSS. The performance expectations aspect of the TIMSS curriculum frameworks explicitly mentions skills such as measurement, data collection, and use of equipment, that cannot be adequately assessed with traditional paper-and-pencil tests. However, the obstacles to including a performance assessment component in a study like TIMSS are formidable. The difficulties inherent in developing a valid international measure of student achievement using just paper and pencil are greatly compounded in the development of a practical test of student performance. In addition to the usual problems of translation and adaptation, there is the question of standardization of materials and of administration procedures, and the greatly increased cost of data collection.

The TIMSS performance assessment was designed to obtain measures of students' responses to hands-on tasks in mathematics and science and to demonstrate the feasibility of including a performance assessment in a large-scale international student assessment. The students that participated were a subsample of the upper-grade students in Populations 1 and 2 that also participated in the main assessment.

The performance assessment in TIMSS consists of a set of 13 tasks, of which 12 were administered at Population 1 and 12 at Population 2. While 11 of the tasks are common to both populations, there were important differences in presentation. For the younger students (Population 1), the tasks were presented with more explicit instructions, or "scaffolding," while for the older students (Population 2) there were usually more activities to be done or additional questions to be answered.

The tasks were organized into a circuit of nine stations, with each station consisting of one long task (taking about 30 minutes to complete) or two shorter tasks (which together took about 30 minutes). An administration of the performance assessment required nine students, who were a subsample of the students selected for the main survey, and 90 minutes of testing time. Each student visited three of the stations during this time; the choice of stations and the order in which they were visited was determined by a task assignment plan.

Because of the cost and complexity of this kind of data collection endeavor, the performance assessment was an optional component of the study. The performance assessment component of TIMSS was conducted by 21 countries participating in Population 2, and by 10 countries participating in Population 1. The international results of that assessment are available in *Performance Assessment in IEA's Third International Mathematics and Science Study* (Harmon et al., 1997).

### 1.9 THE BACKGROUND QUESTIONNAIRES

To obtain information about the contexts for learning mathematics and science, TIMSS included questionnaires for the participating students, their mathematics and science teachers, and the principals of their schools. National Research Coordinators provided information about the structure of their education systems, educational decision-making processes, qualifications required for teaching, and course structures in mathematics and science. In an exercise to investigate the curricular relevance of the TIMSS achievement tests, NRCs were asked to indicate which items in the tests, if any, were not included in their country's intended curriculum. This Test-Curriculum Matching Analysis is described in Chapter 10 of this volume, and results are reported in the first international reports.

The *student questionnaire* explores students' attitudes towards mathematics and science, parental expectations, and out-of-school activities. Students also were asked about their classroom activities in mathematics and the sciences, and about the courses they had taken. At Population 2, there were two versions of the student questionnaire. One was prepared for countries where physics, chemistry, and biology are taught as separate subjects (specialized version) and one for countries where science is taught as an intregrated subject (non-specialized version). Although not strictly related to the question of what students have learned in mathematics or science, characteristics of pupils can be important correlates for understanding educational processes and attainments. Therefore, students also provided general home and demographic information.

The *teacher questionnaires* had two sections. The first section covered general background information about preparation, training, and experience, and about how teachers spend their time in school. Teachers also were asked about the amount of support and resources they had in fulfilling their teaching duties. The second part of the questionnaire related to instructional practices in the classrooms selected for TIMSS testing. To obtain information about the implemented curriculum, teachers were asked how many periods the class spent on topics from the TIMSS curriculum frameworks. They also were asked about their use of textbooks in teaching mathematics and science and about the instructional strategies used in the class, including the use of calculators and computers. In optional sections of the questionnaire, teachers were asked to review selected items from the achievement tests and indicate whether their students had been exposed to the content covered by the items, and to respond to a set of questions that probed their pedagogic beliefs. At Population 2, there were separate versions of the questionnaire for mathematics teachers and science teachers.

The *school questionnaire* was designed to provide information about overall organization and resources. It asked about staffing, facilities, staff development, enrollment, course offerings, and the amount of school time for students, primarily in relation to mathematics and science instruction. School principals also were asked about the functions that schools perform in maintaining relationships with the community and students' families.

### 1.10 MANAGEMENT AND OPERATIONS

Like all previous IEA studies, TIMSS was essentially a cooperative venture among independent research centers around the world. While country representatives came together to plan the study and to agree on instruments and procedures, participants were each responsible for conducting TIMSS in their own country in accordance with the international standards. Each national center provided its own funding and contributed to the support of the international coordination of the study. A study of the scope and magnitude of TIMSS offers a tremendous operational and logistic challenge. In order to yield comparable data, the achievement survey must be replicated in each participating country in a timely and consistent manner. This was the responsibility of the NRC in each country. Among the major tasks of NRCs in this regard were the following:

- Meeting with other NRCs and international project staff to plan the study and develop instruments and procedures
- Defining the school populations from which the TIMSS samples were to be drawn, selecting the sample of schools using an approved random sampling procedure, contacting the school principals and securing their agreement to participate in the study, and selecting the classes to be tested, again using an approved random sampling procedure
- Translating and adapting all of the tests, questionnaires, and administration manuals into the language of instruction of the country (and sometimes more than one language) prior to data collection
- Assembling, printing, and packaging the test booklets and questionnaires, and shipping the survey materials to the participating schools
- Ensuring that the tests and questionnaires were administered in participating schools, either by teachers in the school or by an external team of test administrators, and that the completed test protocols were returned to the TIMSS national center
- Conducting a quality assurance exercise in conjunction with the test administration, whereby some testing sessions were attended by an independent observer to confirm that all specified procedures were followed
- Recruiting and training individuals to score the free-response questions in the achievement tests, and implementing the plan for scoring the student responses, including the plan for assessing the reliability of the scoring procedure
- Recruiting and training data entry personnel for keying the responses of students, teachers, and principals into computerized data files, and conducting the data entry operation using the software provided
- Checking the accuracy and integrity of the data files prior to shipping them to the IEA Data Processing Center in Hamburg

In addition to their role in implementing the TIMSS data collection procedures, NRCs were responsible for conducting analyses of their national data and for reporting on the results of TIMSS in their own countries.[3]

The TIMSS International Study Director was responsible for the overall direction and coordination of the project. The TIMSS International Study Center, located at Boston College in the United States, was responsible for supervising all aspects of the design and implementation of the study at the international level. This included the following:

[3] A list of the TIMSS National Research Coordinators appears in the Acknowledgments section.

- Planning, conducting and coordinating all international TIMSS activities, including meetings of the International Steering Committee, NRCs, and advisory committees
- Developing and field testing the data collection instruments
- Developing sampling procedures for efficiently selecting representative samples of students in each country, and monitoring sampling operations to ensure that they conformed to TIMSS requirements
- Designing and documenting operational procedures to ensure efficient collection of all TIMSS data
- Designing and implementing a quality assurance program encompassing all aspects of the TIMSS data collection, including monitoring of test administration sessions in participating countries
- Supervising the checking and cleaning of the data from the participating countries, the construction of the TIMSS international database, the computation of sampling weights, and the scaling of the achievement data
- Analysis of international data, and writing and disseminating the international reports

The International Study Center was supported in its work by the following advisory committees:[4]

- The International Steering Committee, which advised on policy issues and on the general direction of the study
- The Subject Matter Advisory Committee, which advised on all matters relating to mathematics and science subject matter, particularly the content of the achievement tests
- The Technical Advisory Committee, which advised on all technical issues related to the study, including study design, sampling design, achievement test construction and scaling, questionnaire design, database construction, data analysis, and reporting
- The Performance Assessment Committee, which developed the TIMSS performance assessment and advised on the analysis and reporting of the performance assessment data
- The Free-Response Item Coding Committee, which developed the coding rubrics for the free-response items

[4] See the Acknowledgments section for membership of TIMSS committees.

- The Quality Assurance Committee, which helped to develop the TIMSS quality assurance program

- The Advisory Committee on Curriculum Analysis, which advised the International Study Director on matters related to the curriculum analysis

Several important TIMSS functions, including test and questionnaire development, translation checking, sampling consultations, data processing, and data analysis, were conducted by centers around the world under the direction of the TIMSS International Study Center. In particular, the following centers have played important roles in the TIMSS project.

- The IEA Data Processing Center (DPC), located in Hamburg, Germany, was responsible for checking and processing all TIMSS data and for constructing the international database. The DPC played a major role in developing and documenting the TIMSS field operations procedures.

- Statistics Canada, located in Ottawa, Canada, was responsible for advising NRCs on their sampling plans, for monitoring progress in all aspects of sampling, and for the computation of sampling weights.

- The Australian Council for Educational Research (ACER), located in Melbourne, Australia, participated in the development of the achievement tests, conducted psychometric analyses of field trial data, and was responsible for the development of scaling software and for scaling the achievement test data.

- The International Coordinating Center (ICC) in Vancouver, Canada, was responsible for international project coordination prior to the establishment of the International Study Center in August 1993. Since then, the ICC has provided support to the International Study Center, particularly in managing translation verification in the achievement test development process, and has published several monographs in the TIMSS monograph series.

- As Sampling Referee, Keith Rust of Westat, Inc., (United States) worked with Statistics Canada and the NRCs to ensure that sampling plans met the TIMSS standards, and advised the International Study Director on all matters relating to sampling.

### 1.11 SUMMARY OF THIS REPORT

The selection of valid and efficient samples is crucial to the quality and success of an international comparative study such as TIMSS. The accuracy of the survey results depends on the quality of the available sampling information and of the sampling activities themselves. For TIMSS, NRCs worked on all phases of sampling with staff from Statistics Canada. NRCs were trained in how to select the school and student samples and how to use the sampling software. In consultation with the TIMSS sampling referee, staff from Statistics Canada reviewed the national sampling plans, sampling data, sampling frames, and sample execution. This documentation was used by the International Study Center in consultation with Statistics Canada, the sampling referee, and the Technical Advisory Committee to evaluate the quality of the samples. In Chapter 2, Pierre Foy (Statistics Canada) describes the general TIMSS sample design and the TIMSS national samples, including the grades tested, population coverage, exclusion rates, and sample sizes. Participation rates for schools and students also are documented, as is the particular design for each country (e.g. stratification variables, number of classrooms sampled).

To ensure the availability of comparable, high-quality data for analysis, TIMSS engaged in a set of rigorous quality control steps to create the international database. TIMSS prepared manuals and software for countries to use in entering their data so that the information would be in a standardized international format before it was forwarded to the IEA Data Processing Center in Hamburg for creation of the international database. Upon arrival at the IEA Data Processing Center, the data from each country underwent an exhaustive cleaning process. That process involved several iterative steps and procedures designed to identify, document, and correct deviations from the international instruments, file structures, and coding schemes. The process also emphasized consistency of information within national data sets and appropriate linking among the many student, teacher, and school data files. Following the data cleaning and file restructuring by the DPC, Statistics Canada computed the sampling weights and the Australian Council for Educational Research computed the item statistics and scale scores. These additional data were merged into the database by the DPC. Throughout, the International Study Center reviewed the data and managed the data flow. In Chapter 3, Heiko Sibberns, Dirk Hastedt, Michael Bruneforth, Knut Schwippert, and Eugenio Gonzalez describe the TIMSS data management, including procedures for cleaning and verifying the data and the links across files, restructuring of the national data files to the standard international format, the various data reports produced throughout the cleaning process, and the computer systems used to undertake the data cleaning and construction of the database.

Within countries, TIMSS used a two-stage sample design for Populations 1 and 2. The first stage involved selecting 150 public and private schools within each country. Within each school, the basic approach required countries to use random procedures to select one mathematics class at each grade (third and fourth or seventh and eighth, depending on the population). All of the students in those two classes were to participate in the TIMSS testing. This approach was designed to yield a representative sample of 7,500 students per country per population, with approximately 3,750 students at each grade. The complex sampling approach required the use of sampling weights to account for the differential probabilities of selection and to adjust for nonresponse in order to ensure the computation of proper survey estimates. Statistics Canada was responsible for computing the sampling weights for the TIMSS countries. In Chapter 4, Pierre Foy describes the derivation of TIMSS school, classroom, and student weights.

Because the statistics presented in the TIMSS reports are estimates of national performance based on samples of students, rather than the values that could be calculated if every student in every country had answered every question, it is important to have measures of the degree of uncertainty of the estimates. The complex sampling approach that TIMSS used had implications for estimating sampling variability. Because of the effects of cluster selection (classrooms within schools, students within classrooms, and any other front-end stratification) and because of the effects of certain adjustments to the sampling weights, procedures derived from simple random sampling assumptions for estimating the variability of sample statistics are inappropriate. TIMSS used the jackknife procedure to estimate the standard errors associated with each statistic presented in the international reports. In Chapter 5, Eugenio Gonzalez and Pierre Foy describe the jackknife technique and its application to the TIMSS data in estimating the variability of the sample statistics.

Prior to scaling, the TIMSS cognitive data were thoroughly checked by the IEA Data Processing Center, the International Study Center, and the national centers. The national centers were contacted regularly and given multiple opportunities to review the data for their countries. In conjunction with the Australian Council for Educational Research, the International Study Center conducted a review of item statistics for each of the mathematics and science items in each of the countries to identify poorly performing items. In Chapter 6, Ina Mullis and Michael Martin describe the procedures used to ensure that the cognitive data included in the scaling and the international database are comparable across countries.

The complexity of the TIMSS test design and the desire to compare countries' performance on a common scale led TIMSS to use item response theory in the analysis of the achievement results. For both populations, TIMSS reported overall mathematics and science scale scores (by grade) based on a variant of the Rasch item response model. The model, developed by Adams, Wilson, and Wang (1997), included refinements that enable reliable scores to be produced even though individual students responded to relatively small subsets of the total mathematics and science item pools. An item response model was preferred for developing comparable estimates of performance for all students, since students answered different test items depending on which of the eight test booklets they received. In Chapter 7, Ray Adams, Margaret Wu, and Greg Macaskill describe the scaling methodology and procedures used to produce the TIMSS achievement scores, including the estimation of international item parameters, and the deriviation and use of plausible values to provide estimates of performance.

TIMSS reported achievement scale scores for mathematics and science overall from a number of perspectives. Mean achievement and selected percentiles were reported by country for each grade. Significant differences between countries (adjusted for multiple comparisons) also were reported for each grade. TIMSS presented mean achievement for girls and boys separately, with indications of significant differences between the genders. Although the TIMSS design was based on adjacent grades, rather than age, TIMSS was able to report median mathematics and science achievement for 9-year-olds and 13-year-olds. To show the "growth" in achievement between the primary and middle school years, TIMSS also reported achievement of the younger students on the scale constructed for the older population. In Chapter 8, Eugenio Gonzalez describes the analyses undertaken to report the achievement scale scores in these various ways in the international reports.

While achievement results for mathematics and science overall were estimated using item response theory, achievement results for the mathematics and science content areas and for individual items were analyzed using average percent correct technology. In Chapter 9, Albert Beaton and Eugenio Gonzalez describe how this technology was adapted to handle the TIMSS data and used to report achievement in the content areas and for individual items.

TIMSS developed international tests of mathematics and science that reflect as far as possible the various curricula of the participating countries. The tests were developed through a consensus-building process involving representatives from the participating countries and approved for use by each country. Despite efforts to create a test that was as comprehensive as possible and was appropriate for all countries, there were likely some items that are not addressed by the curriculum in each country. To investigate the extent to which this was the case and the impact this might have on the results, TIMSS developed and conducted the Test-Curriculum Matching Analysis. The purpose and procedures for this analysis are described by Albert Beaton and Eugenio Gonzalez in Chapter 10.

TIMSS collected a vast amount of contextual data from student, teachers, and school principals, as well as information about the education systems. Deciding what to report in terms of background data, and how to best report these data, was a difficult task. In Chapter 11, Dana Kelly, Ina Mullis, and Teresa Smith describe the analysis and reporting of the background data in the international reports, including the development of the international report outlines, the consensus and review procedures undertaken to ensure that the perspectives of many people were incorporated into the reporting, the development of analysis plans for the report tables, and special issues in reporting, including response rates and reporting teacher data.

## REFERENCES


Adams, R. J. and Gonzalez, E. J. (1996). The TIMSS test design. In Martin, M.O. and Kelly, D.L., (Eds.), *TIMSS technical report, volume I: Design and development*. Chestnut Hill, MA: Boston College.

Beaton, A.E., Martin, M.O., and Mullis, I.V.S. (1997). Providing data for educational policy in an international context: The Third International Mathematics and Science Study. *European Journal of Psychological Assessment,* 13 (1), pp. 49-58.

Martin, M.O. and Kelly, D.L., Eds. (1996). *TIMSS technical report, volume I: Design and development*. Chestnut Hill, MA: Boston College.

Martin, M.O. and Mullis, I.V.S., Eds. (1996) *TIMSS: Quality assurance in data collection*. Chestnut Hill, MA: Boston College.

Robitaille, D.F. and Garden, R.A. (1996). Design of the study. In D.F. Robitaille and R.A. Garden (Eds.), *TIMSS monograph No. 2: Research questions and study design*. Vancouver, Canada: Pacific Educational Press.

Robitaille, D.F. and Garden, R.A. (1989). *The IEA study of mathematics II: Contexts and outcomes of school mathematics.* Oxford: Pergamon Press.

Robitaille, D.F., Schmidt, W.H., Raizen, S.A., McKnight, C.C., Britton, E., and Nicol, C. (1993). *TIMSS monograph no. 1: Curriculum frameworks for mathematics and science*. Vancouver, Canada: Pacific Educational Press.

Schmidt, W.H., McKnight, C.C., Valverde, G.A., Houang, R.T., and Wiley, D.E. (1996). *Many visions, many aims: A cross-national investigation of curricular intentions in school mathematics*. Norwell, MA: Kluwer Academic Press.

Schmidt, W.H., Raizen, S.A., Britton, E.D., Bianchi, L.J., and Wolfe, R.G. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in school science*. Norwell, MA: Kluwer Academic Press.

Travers, K.J. and Westbury, I. (1989). *The IEA study of mathematics I: Analysis of mathematics curricula*. Oxford: Pergamon Press.