TIMSS
Technical Report
Volume I

Third International Mathematics and Science Study

Technical Report

Volume I: Design and Development

Edited by
Michael O. Martin
Dana L. Kelly

with contributors
Raymond J. Adams
Leland Cogan
Pierre Foy
Robert A. Garden
Eugenio J. Gonzalez
Maryellen Harmon
Chancey Jones
Svein Lie
Beverley Maxwell
Ina V.S. Mullis
Graham Orpwood
Keith Rust
Andreas Schleicher
William H. Schmidt
Maria Teresa Siniscalco
Alan Taylor

Boston College • Chestnut Hill, Massachusetts

Third International Mathematics and Science Study Technical Report Volume I:
Design and Development/ edited by Michael O. Martin, Dana L. Kelly
Publisher: Center for the Study of Testing, Evaluation, and Educational Policy,
Boston College.

Library of Congress Catalog Card Number: 96-86397

ISBN 1-889938-00-9

For more information about TIMSS contact:

TIMSS International Study Center
Center for the Study of Testing, Evaluation, and Educational Policy
Campion Hall
School of Education
Boston College
Chestnut Hill, MA 02167
United States

This report is also available on the World Wide Web:
http://wwwcsteep.bc.edu/timss

Funding for the international coordination of TIMSS is provided by the U.S. National Center for Education Statistics, the U.S. National Science Foundation, the IEA, and the Canadian government. Each participating country provides funding for the national implementation of TIMSS.

Printed and bound in the United States.

CONTENTS

Michael O. Martin, Ina V.S. Mullis, and Dana L. Kelly

Martin, M.O. (1996) "Third International Mathematics and Science Study: An Overview" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

1. Third International Mathematics and Science Study: An Overview

Michael O. Martin

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. 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. This report documents in detail the development and implementation of the TIMSS achievement survey.

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 are planned for release in late 1996; and further international reports will be issued through 1997. 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, curriculum design, survey research, test construction, psychometrics, survey sampling, and data analysis.

An achievement survey of the scale of TIMSS not only has a long life cycle, but also passes through several distinct stages. In the development stage, attention focuses on refining the aims of the study, establishing the parameters of the survey, designing and developing the data collection instruments, and developing data collection procedures. In the operational stage, samples are drawn, survey materials are distributed, training is conducted, and data are collected, checked, scored, and entered into databases. In the analysis and reporting stage, the data are processed, summarized, and presented, first in simple descriptive reports and later in more complex analytical volumes.

Figure 1.1 Countries Participating in the TIMSS Achievement Survey

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

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* is an important part of this effort. While the details of the TIMSS procedures are described in the various procedural manuals, this report presents the technical aspects of the study design, and provides the background to and the rationale for many of the design decisions taken.

Because of the long life cycle of TIMSS, and the involvement of so many individuals at its various stages, it was desired to issue the *TIMSS Technical Report* in two volumes, each documenting a major stage of the project and produced soon after the completion of that stage. Accordingly, the first volume documents the study design and the development of TIMSS up to, but not including, the operational stage of main data collection. The second volume will describe the operational stage, which consisted mainly of collecting and processing the data, and will describe the analytic procedures underlying the analysis and reporting phase of TIMSS.

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 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 constrained 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–a necessary though not sufficient condition for student learning.

Considering the curriculum in all its aspects 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 educationalists. 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.

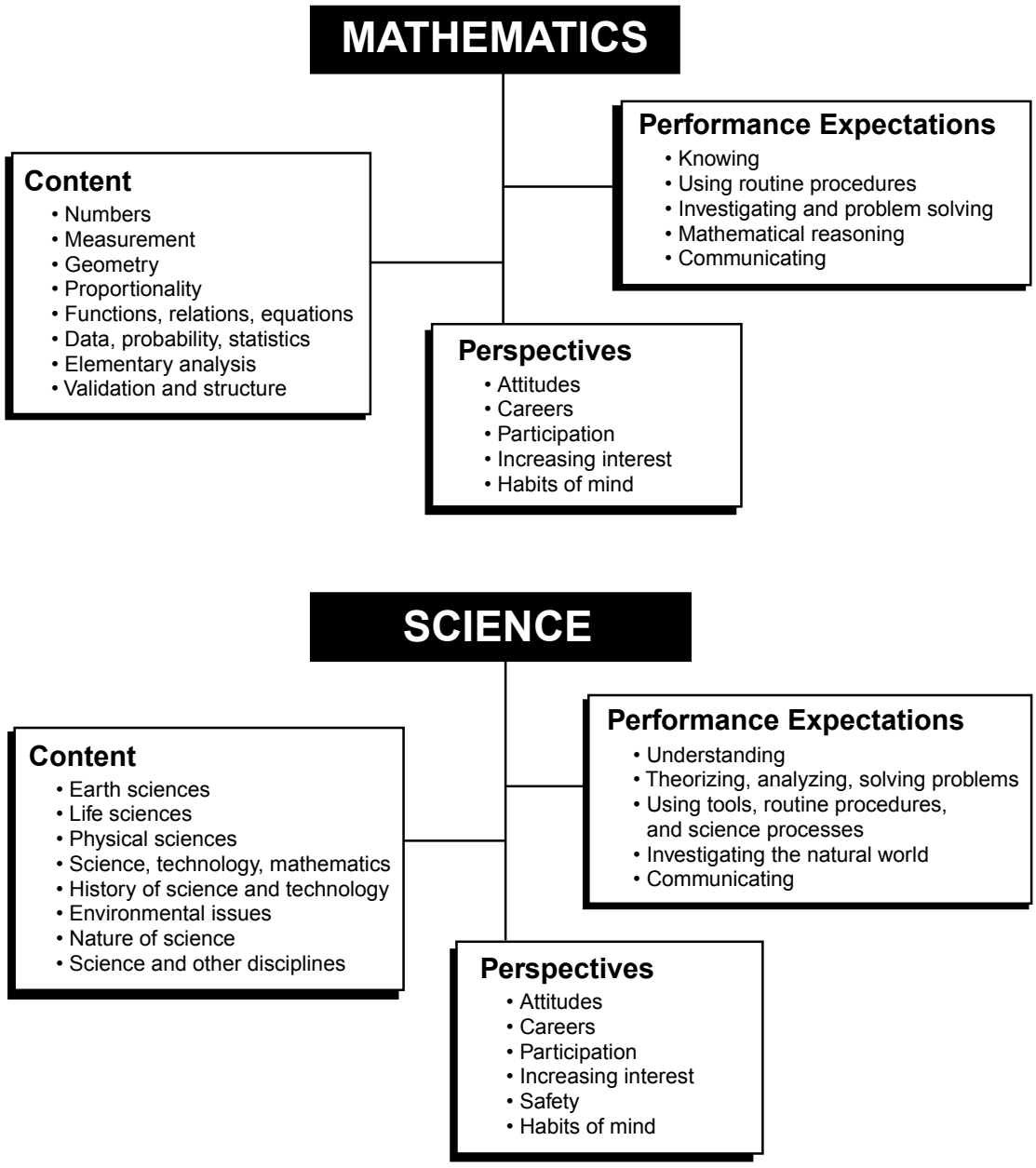
> For the purposes of TIMSS, curriculum consists of the concepts, processes, and attitudes of school mathematics and science that are intended for, implemented in, or attained during students' schooling experiences. Any piece of curriculum so conceived–whether intended, implemented, or attained, whether 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, 43).

Subject-matter content, performance expectations, and perspectives constitute the three dimensions, or aspects, of the TIMSS curriculum frameworks. *Subject-matter content* 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 the context within which the material is presented.

Each of the three aspects is partitioned into a number of categories, which are themselves 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.[1]

[1] The complete TIMSS curriculum frameworks can be found in Robitaille, D., et al. (1993). *TIMSS Monograph No. 1: Curriculum Frameworks for Mathematics and Science*. Vancouver: Pacific Educational Press.

Figure 1.2 The Major Categories of the TIMSS Curriculum Frameworks



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, finalized, 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, in press).

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 will be presented in two volumes: *Many Visions, Many Aims: A Cross National Investigation of Curricular Intentions in School Mathematics* (Schmidt, W., McKnight, C., Valverde, G., Houang, R., and Wiley, D., in press) and *Many Visions, Many Aims: A Cross National Investigation of Curricular Intentions in School Science* (Schmidt, W., Raizen, S., Britton, E., Bianchi, L., Wolfe, R., in press).

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 has 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 one-year age group (9-year-olds for Population 1, and 13-year-olds for Population 2). Most cross-country comparisons in TIMSS will be based on grade levels, since educational systems are organized around grade levels, but it will also be

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 as follows.

- 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 taking advanced courses in mathematics*
  - *Students taking advanced courses in 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).

1.6 SURVEY ADMINISTRATION DATES

Since school systems in countries in the Northern and Southern Hemispheres do not have the same school year, TIMSS had to operate two survey administration schedules. The schedules are shown in Table 1.1. These periods were chosen with the aim of testing students as late in the school year as practicable, so as to reflect the knowledge gained throughout the year.

Table 1.1 Survey Administration Dates

	Populations 1 and 2	**Population 3**
Southern Hemisphere	September-November 1994	August 1995
Northern Hemisphere	February-May 1995	February-May 1995

1.7 THE TIMSS ACHIEVEMENT TESTS

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.

Although 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), 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, 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 very 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 their everyday school experience. In particular, there was a requirement that test items make use of a variety of task types and response formats, and not be 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 in the interests of producing 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 for example, estimates of student proficiency in 11 areas or "reporting categories" of mathematics and science, as well as overall mathematics and 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 two-digit 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 it were to be administered to any one student, would take almost seven hours of testing. Since the consensus among the National Research Coordinators was that 90 minutes was the most that could be expected for this population, a way of dividing the item pool among the students had to be found. Matrix sampling provided a solution to this problem 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 has 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 for each Population 1 and Population 2, and nine booklets for Population 3.

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.

The approach to assessing achievement in mathematics and science at Population 3 was quite different from that for the younger populations. At Population 3 there are really three populations to be tested. For all students in the population (the final year of secondary schooling), TIMSS plans to report measures of mathematics and science literacy. The item pool for this domain consists of four clusters of items, assembled in two booklets distributed across the entire Population 3 sample. The other two populations to be reported on are the students at Population 3 who are taking advanced mathematics courses, and those taking physics courses. Since each group will have received advanced preparation, item pools had to be developed that treated these subjects with the appropriate depth of coverage and level of difficulty.

There are four clusters of advanced mathematics items, assembled into three booklets for students taking courses in advanced mathematics. The pool of physics items is also grouped into four clusters and assembled into three booklets distributed to the students in the sample who are taking physics courses. A ninth booklet, consisting of one cluster of literacy items, one of mathematics items, and one of physics items, is designed just for students taking courses in both advanced mathematics and physics.

1.8 PERFORMANCE ASSESSMENT

Educators have long advocated the use of using 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 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 when developing a practical test of student performance. In addition to the usual difficulties 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. It was conducted at Populations 1 and 2 only.

The performance assessment in TIMSS consists of a set of 13 tasks, 12 of which are used 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 of the 45 countries participating in Population 2, and by 11 of the 28 countries participating in Population 1.

1.9 THE CONTEXT 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. The student and school questionnaires were administered at all three populations, and the questionnaires for mathematics and science teachers at Populations 1 and 2. National Research Coordinators (NRCs) provided information about the structure of their educational 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, if any, are not included in their country's

intended curriculum. The results of this Test-Curriculum Matching Analysis will be reported in the first international reports.

The student questionnaire addresses 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. A special version of the student questionnaire was prepared for countries where physics, chemistry, and biology are taught as separate subjects. 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 received 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 TIMSS design did not include a teacher questionnaire for teachers of Population 3 students.

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 responsibilities of NRCs in this regard were the following.

- Meeting with other NRCs and international project staff to plan the study and to 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 into 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 observed 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 coding the student responses, including the plan for assessing the reliability of the coding 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.[2]

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.

- Planning, conducting and coordinating all international TIMSS activities, including meetings of the International Steering Committee, NRCs, and advisory committees
- Development, including field testing, of all data collection instruments

[2] A list of the TIMSS National Research Coordinators is provided in Appendix A.

- Development of sampling procedures for efficiently selecting representative samples of students in each country, and monitoring sampling operations to ensure that they conformed to TIMSS requirements
- Development and documentation of operational procedures to ensure efficient collection of all TIMSS data
- Design and implementation of a quality assurance program encompassing all aspects of the TIMSS data collection, including monitoring of test administration sessions in participating countries
- Supervision of the checking and cleaning of the data from the participating countries, and construction of the TIMSS international database, including the computation of sampling weights and the scaling of the achievement items
- Analysis of international data, and writing and dissemination of international reports.

The International Study Center was supported in its work by the following advisory committees.[3]

- The International Steering Committee advises on policy issues and on the general direction of the study.
- The Subject Matter Advisory Committee advises on all matters relating to mathematics and science subject matter, particularly the content of the achievement tests.
- The Technical Advisory Committee advises 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 developed the TIMSS performance assessment and advises on the analysis and reporting of the performance assessment data.
- The Free-Response Item Coding Committee developed the coding rubrics for the free-response items.
- The Quality Assurance Committee helped to develop the TIMSS quality assurance program.
- The Advisory Committee on Curriculum Analysis 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 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, and in particular has managed translation verification in the achievement test development process; and has published several monographs in the TIMSS monograph series.

[3] See Appendix A for membership of TIMSS committees.

- The IEA Data Processing Center (DPC), located in Hamburg, Germany is 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, is 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, has participated in the development of the achievement tests, has conducted psychometric analyses of field trial data, and was responsible for the development of scaling software and for scaling the achievement test data.

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 THE REPORT

In Chapter 2, Robert Garden and Graham Orpwood (subject-matter coordinators in mathematics and science, respectively) describe the long and sometimes arduous process of developing the TIMSS achievement tests. They outline the tensions between the wish for comprehensive coverage of mathematics and science curricula, the limited time available for student testing, the need to be sensitive to curricular variations from country to country, the desire to include innovative assessment methods, and the requirements of a rigorous approach to measuring student achievement. The authors describe how these tensions were resolved, the compromises that were made, and the characteristics of the final pool of achievement items. They show how the items in this pool address a wide range of subject matter in mathematics and science, and seek to evoke from the students a wide range of performance types, from exhibiting acquired knowledge to engaging in complex problem solving.

In Chapter 3, Raymond Adams and Eugenio Gonzalez present the TIMSS achievement test design. The design describes, for each student population, how the pool of achievement questions was organized into achievement booklets that were given to the students selected to take part in TIMSS. Since the entire item pool was much too large to be administered to every student in the time available, a matrix sampling approach was used in which subsets of items drawn from the total item pool were administered to random subsamples of students. This procedure provides accurate estimates of population parameters based on all items, even though not every student responds to every item.

Pierre Foy, Keith Rust, and Andreas Schleicher describe in Chapter 4 the student populations that were the target of TIMSS, and the designs that were developed to draw samples from these populations. They pay particular attention to the principles by which the target populations were defined in participating countries. This process involved specifying exactly which students were eligible for selection as a sample, and which subgroups (e.g., mentally handicapped students), if any, were to be excluded. The authors

present the sampling-precision requirements of TIMSS, and show how these were used to determine sample size in the participating countries. They go on to describe the TIMSS sampling designs, including the use of stratification and multistage sampling, and illustrate the general method used in sampling schools in TIMSS (the sampling of classrooms is described in Chapter 9 on field operations).

The background of the development of the student, teacher, and school questionnaires is the subject of Chapter 5, by William Schmidt and Leland Cogan. The difficulties devising achievement tests that can provide valid and reliable data from countries with diverse social, cultural, and educational traditions are discussed in Chapter 2; the difficulties of developing questionnaires that can elicit useful information about the educational context in an array of countries are no less formidable. Factors that are fundamental to the understanding of student achievement in one country may be much less relevant in another. Schmidt and Cogan recount the enormous efforts made to build consensus on the conceptual framework for TIMSS and to derive from it a set of data collection instruments that would be useable in all participating countries and would do justice to the aims of the project.

In measuring student achievement, TIMSS sought to ensure the validity of its tests as well as their reliability by combining traditional objective multiple-choice items with innovative and challenging free-response questions. The economic realities of a large-scale international assessment dictated that the data be collected mainly through written responses to written questions. However, TIMSS acknowledged that a comprehensive assessment of student competency in mathematics, and particularly in science, demands that students also be given realistic problems that must be answered by manipulating tools and materials. The TIMSS performance assessment, which is described by Maryellen Harmon and Dana Kelly in Chapter 6, was developed to meet that need. The performance assessment is a set of tasks that were administered to a subsample of the students selected for the main survey. Although an integral feature of the overall TIMSS design, the extra expense of conducting the performance assessment was beyond the resources of some participants. Accordingly, it was presented as an optional component.

TIMSS was committed from an early stage to measuring student achievement through a variety of item types. The main survey contains three types of items: multiple-choice, short-answer, and extended-response. All of the items in the performance assessment are either short-answer or extended-response. Unlike multiple-choice items, short-answer and extended-response items are free-response items that require a coding rubric, or set of rules, so that a code may be assigned to each response. Svein Lie, Alan Taylor, and Maryellen Harmon in Chapter 7, describe the evolution of the innovative two-digit coding rubric used with all TIMSS free-response items. This coding rubric provides for coding the "correctness" of a response to an item as a score level in the left digit, and information about the detailed nature of the response in both digits together. This coding rubric promises to provide a rich store of information about the most common student responses to the free-response items, and in particular about the most common misconceptions about mathematics and science concepts..

In order to implement the TIMSS survey in the 45 participating countries, it was necessary to translate the achievement tests, the student, teacher, and school questionnaires, and in many cases the manuals and tracking forms from English, the language in which they were developed, to the language of the country. In all, the TIMSS instruments were translated into 30 languages. Even where the language of testing was English, cultural adaptations had to be made to suit national language usage. In Chapter 8, Beverley Maxwell describes the procedures that were used to ensure that the translations and cultural adaptations made in each country produced local versions that corresponded closely in meaning to the international versions, and in particular that the items in the achievement tests were not made easier or more difficult through translation.

As a comparative sample survey of student achievement conducted simultaneously in 45 countries, TIMSS depends crucially on its data collection procedures to obtain high-quality data. In Chapter 9, Andreas Schleicher and Maria Teresa Siniscalco describe the procedures developed to ensure that the TIMSS data were collected in a timely and cost-effective manner while adhering to the highest standards of survey research. The authors outline the extensive list of procedural manuals that describe in detail all aspects of the TIMSS field operations. They describe also the software systems that were provided to participants to help them conduct their data collection activities. The development of practical and effective methods of sampling classes within schools that would be generally applicable was a particular challenge to TIMSS. The authors present these methods in some detail, and particularly emphasize the system of documentation that is an integral component of each procedure. This documentation consists of a series of tracking forms, which record how schools, classes, teachers, and students were selected to take part in the study. This documentation facilitates the operational aspects of data collection, and provides vital information for other aspects of the study (e.g., the computation of sampling weights, the provision of checks for quality assurance).

As a consequence of its emphasis on authentic and innovative assessment of achievement, much of the testing time was used to provide answers to free-response items. Approximately one-third of student time for the main survey, and all of the time for the performance assessment, was devoted to free-response items. This resulted in a huge number of student responses that had to be coded using the two-digit coding scheme described in Chapter 7. In order to code reliably and in the same way in each country, and to ensure that the performance assessment was administered consistently across countries, an extensive training program for National Research Coordinators and their staff members was conducted. In Chapter 10, Ina Mullis, Chancey Jones, and Robert Garden outline the content and format of this training program, and describe the logistics of conducting ten training meetings in locations all around the globe.

A major part of the role of the TIMSS International Study Center was to ensure that all aspects of the study were carried out to the highest standards of survey research. In Chapter 11, Michael Martin, Ina Mullis, and Dana Kelly describe the procedures and activities that comprised the TIMSS quality assurance program. The International Study

Center sought to ensure a high-quality study in the first instance by providing participants with complete and explicit documentation of all procedures and materials, supplemented by meetings with consultants and training meetings at every opportunity. An integral part of the documentation for each procedure was a set of forms that had to be completed in order to carry out the procedure. The completed forms constitute a record that can be reviewed for quality control purposes. A major component of the quality assurance activities was a program of visits to each participating country by monitors appointed by the International Study Center. These quality control monitors visited the national research centers and interviewed the NRCs about all aspects of the implementation of TIMSS. They also visited a sample of ten of the schools taking part in the study to interview the School Coordinator and Test Administrator, and to observe a test administration in one classroom.

1.12 SUMMARY

This report provides an overview of the main features of the TIMSS project, and summarizes the technical background to the development of the study. The development of the achievement tests and questionnaires, the sampling and operations procedures, the procedures for coding the free-response items, and quality assurance activities are all described in detail. The activities involved in the collection of the TIMSS data, and in analysis and reporting, will be presented in a subsequent volume.

REFERENCES

Beaton, A.E., Martin, M.O., and Mullis, I.V.S. (in press). Providing Data for Educational Policy in an International Context: The Methodologically Ambitious Nature of TIMSS. *European Journal of Psychological Assessment.*

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 & 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. (in press). *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. (in press). *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.

Garden R.A. and Orpwood, G (1996) "Development of the TIMSS Achievement Tests" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

2. Development of the TIMSS Achievement Tests

Robert A. Garden
Graham Orpwood

2.1 OVERVIEW

The task of putting together the achievement item pools for the three TIMSS student populations was immense, and took more than three years to complete. Developing the TIMSS achievement tests necessitated building international consensus among National Research Coordinators (NRCs), their national committees, mathematics and science experts, and measurement specialists. All NRCs worked to ensure that the items used in the survey were appropriate for their students and reflected their countries' curriculum, enabling students to give a good account of their knowledge and ability and ensuring that international comparisons of student achievement could be based on a "level playing field" insofar as possible. This chapter describes the steps involved in constructing the TIMSS tests, including the development of the item pool, piloting of the items, item review, and the assembly of test booklets.

2.2 ITEM TYPES

Large-scale surveys of student achievement have traditionally used, either exclusively or mainly, multiple-choice items. Well constructed tests composed of such items typically have high reliability and high validity. In addition, practical considerations make multiple-choice items popular in many applications: testing conditions can be easily standardized,

the administration costs are low, and where machine scoring is appropriate, very large samples may be processed economically and efficiently.

Multiple-choice items have served IEA studies well, and are likely to continue to do so. In previous studies, tests and subtests composed of multiple-choice items have provided teachers, curriculum developers, researchers, and policy makers with valid information about the strengths and weaknesses of system-level educational practices. Used in conjunction with information from questionnaires completed by administrators, teachers, and students, the achievement survey results have made it possible to identify and describe system- and subsystem-level strengths and weaknesses. They have also been used to suggest promising avenues for remedial action.

In the past few years, educators have become more and more aware that some important achievement outcomes are either impossible to measure, or difficult to measure well, using multiple-choice items. Constructing a proof in mathematics, for example, communicating findings in science or mathematics, or making a case for action based on scientific principles all require skills not adequately measured by multiple-choice items. It was also believed that tasks requiring complex multistep reasoning are measured with greater validity by constructed- or free-response items, which demand written responses from students. Such items, especially those that demand an extended response, also convey to the students that the ability to present a lucid written account of their reasoning is an important component of learning. It was therefore decided at the outset that the TIMSS test should employ a variety of item types for the best coverage of the outcomes of school mathematics and science education. Three types of achievement items were included in the item pools for TIMSS: multiple-choice items; free-response items (both short-answer and extended-response items); and performance tasks.

1. Multiple-Choice Items	Multiple-choice items used in TIMSS consist of a stem and either four or five answer choices, of which only one is the best or the correct answer. Neither "I don't know" nor "None of the above" is an option in any of the items. In the instructions at the front of the test booklets, students are encouraged to choose "the answer [they] think is best" when they are unsure. The instructions do not suggest or imply that students should guess where they do not know the answer.
2. Free-Response Items	For the free-response items–both short-answer and extended-response types–students write their responses, and these are coded using the two-digit coding system developed TIMSS. See Chapter 7 for a discussion of the coding system.
3. Performance Tasks	Some of the skills and abilities that mathematics and science programs are intended to transmit to students are not easily assessed by the kinds of items usually found in a written test. Only "hands-on" activities allow students to demonstrate their ability to make, record, and communicate

observations correctly; to take measurements or collect experimental data, and to present them systematically; to design and conduct a scientific investigation; or to solve certain types of problems. A set of such "hands-on" activities–referred to as performance tasks–was developed for the study and used at the Population 1 and 2 levels. This component of the study, is described in Chapter 6.

2.3 DEVELOPING THE ITEM POOLS

Candidate items for use in TIMSS were drawn from diverse sources. Achievement in TIMSS was initially intended to be linked with the results of two earlier IEA studies, the Second International Mathematics Study (SIMS) and the Second International Science Study (SISS). Items from these studies were therefore examined, and those judged to be appropriate for TIMSS' purposes were selected for piloting.[1] As is usual in IEA studies, personnel in the national centers were also asked to submit items considered suitable, and the International Coordinating Center (ICC) received a large number of multiple-choice and free-response items from these sources.

Items submitted by national centers were classified according to the content and performance expectation codes of the TIMSS curriculum frameworks (Robitaille et al., 1993). For many items more than one such code was allocated. A detailed test blueprint for content-by-performance category was developed by an iterative process, and an interim item specification framework developed in 1991 was used for initial selection of items to be piloted. This draft blueprint was in lieu of a more refined version to evolve later from data collected in the curriculum analysis component of TIMSS. The draft blueprint indicated approximate numbers of items needed for each subtopic and for each performance expectation category. Items were distributed across content areas with a weighting reflecting the emphasis national committees placed on individual topics. For purposes of assignment to categories of the blueprint, items with multiple codes were classified according to the code judged to relate to the primary content and performance categories being assessed. Inevitably, key stages of test development revealed shortages of items with particular specifications, and new items had to be written or gathered. This will be described in more detail later in the chapter.

In December 1991 an international panel of subject-matter and assessment experts met to select items from the initial collection for use in a pilot study. Although large pools of items had been assembled, a disproportionate number were found to assess computation, recall, or simple application in limited content areas. For some content areas an adequate number of potentially good items were available, but for others there were too few items of good quality. Also, because most items had been written for use within particular countries, the panel had to reject many for use in TIMSS because of cultural bias, or because translation was likely to lead to ambiguity or misunderstanding. However, items that were not too culture-bound, or specific to the curricula of too few countries, or were not too time-consuming were considered for the TIMSS item pool.

[1] Formal links between TIMSS and SISS were never realized because the target populations were not equivalent.

Preparing a pool of items for Population 1 was especially challenging. Very few countries have national assessments at this level, so there were few sources of good items. In addition, the mathematics and science taught to 9-year-olds varies more from country to country than for 13-year-olds.

To ensure that the required number of items in each content area would be available for piloting, additional items were gathered and written during the December 1991 meeting, and subsequently by ICC personnel. For content areas with a plentiful supply of items, an attempt was made to ensure that items selected for the pilot covered the range of performance categories in the TIMSS curriculum frameworks.

In May 1992, test development for the study was contracted to the Beacon Institute. The Beacon Institute conducted an international item review in which national centers were asked to have a panel of experts review candidate items. As a result, many items were discarded. At this time, too, a limited trial of extended-response items was undertaken. The newly formed Subject Matter Advisory Committee (SMAC) first met in July 1992 and, as part of its brief, began to advise on test development.

In November 1992 the ICC resumed responsibility for test development. New items were written to replace those that had been discarded after the international item review, and to accommodate some changes that had been made to the test specifications. In January 1993, the SMAC reviewed the items in the new item pools, rejected some items, and modified others. The SMAC expressed reservations about the overall quality of items, and there was concern that many items proposed for the Population 1 students would prove too difficult. Further items were written at the ICC, and pilot test booklets were distributed to national centers for the pilot held in April - June 1993.

Preparation of an adequate item pool for Population 3 piloting was delayed, partly because there was uncertainty as to which students were to be included in the target population, and partly because more emphasis had been placed on preparation of item pools for the younger populations. It became apparent that it would not be possible to gather and organize enough items of acceptable quality in time for piloting at the same time as the Populations 1 and 2 items; thus it was decided to delay the Population 3 pilot.

2.3.1 ITEM PILOT

The Populations 1 and 2 item pilots were administered to judgment samples of students in 43 countries in April and May of 1993. The design called for sample sizes of at least 100 students per item, and in most countries that target was exceeded. At the national centers, committees that included people with subject-matter, evaluation, and teaching expertise reviewed each item for its appropriateness to the national curriculum and its overall quality. Items considered to be biased were targeted, and national review committees identified those they believed should not be included in TIMSS. This information was used in conjunction with item statistics to determine which items would be retained and which discarded.

To be retained for further consideration an item had to meet the following criteria:

- Be appropriate to the curricula of more than 70 percent of countries
- Be recommended for deletion by less than 30 percent of countries
- Have p-values of greater than .20 (for five-choice items) or .25 (for four-choice items)
- Have positive point-biserial correlations for correct responses and negative point-biserial correlations for all distracters.

The number of items meeting all of the criteria was 137 (69% of those piloted) for Population 1 and 279 (81% of those piloted) for Population 2. However, acceptable items were not distributed evenly across the content or performance domains. Behavior such as recall or computation was assessed by many more items than necessary, while items assessing more complex performance were in short supply. Similarly, there was an oversupply of items in some content areas and an undersupply in others.

Several national committees leveled criticism at the item pool. The major criticisms came from a few countries in which curricular changes, or changes in forms of assessment, were in train and whose national committees believed that the TIMSS tests should reflect these changes. In particular there was a call from some quarters for more "contextualized" items. There was also a fairly common view that tests based on the items piloted would be too difficult, especially for students in Population 1 and in those countries in which children enter school only at age seven. It was believed that both the subject-matter content of the items and, especially with science items, their readability level would be too difficult for nine-year-olds.

The results of the item pilot and review did not support some of the more extreme criticism; however, general concerns about suitability of language and content, especially for Population 1, were borne out, and the shortage of items with a "real-world" context was recognized. There was clearly a need for a comprehensive overhaul of the item pools, involving extensive editing of existing items and introduction of many new items. In particular, the requests for more contextualized items needed to be met. This, in turn, meant a further round of piloting.

2.3.2 AUGMENTATION OF THE ITEM POOLS

As soon as the results of the item pilot and review had been assessed, the project management took various initiatives to remedy the perceived problems.

- Two test development coordinators were appointed to manage and oversee development of the tests: Graham Orpwood for science and Robert Garden for mathematics
- NRCs were again asked to propose items for consideration
- The Center for Assessment of Educational Progress at Educational Testing Service was contracted to produce additional items for some test blueprint areas for Populations 1 and 2, where shortages had been identified, and test booklets for the Populations 1 and 2 field trial
- The Australian Council for Educational Research (ACER) was contracted to produce additional items for Population 3 tests and the test booklets for Population 3 piloting.

The addition of so many new items meant that an additional item pilot had to be conducted. The schedule did not allow for a further round of item piloting before the field trial (originally intended to try operational procedures only). It was therefore decided that the field trials would be used to pilot the additional items being produced for Populations 1 and 2, and all Population 3 items. The Population 1 and 2 field trial was scheduled for February 1994 and the Population 3 field trial for May 1994.

In August 1993, on the initiative of the National Science Foundation, two American Educational Research Association (AERA) "Think Tank" meetings were convened in Vancouver.[2] The purpose of one of the meetings was to review the status of the Populations 1 and 2 item pools and make recommendations to enhance them. The second meeting was concerned with formulation of a rationale and plans for assessing Population 3 mathematics and science literacy.

The international group reviewing the Populations 1 and 2 item pools recommended enlisting the help of further professional testing agencies to produce supplementary items for areas of shortage. Shortly after this meeting SRI International was contracted to produce additional mathematics and science items for Populations 2 and mathematics items for Population 1 to supplement the work already under way at Educational Testing Service. For Population 3, working groups met several times to write and select items for the advanced mathematics, physics, and mathematics and science literacy tests.

New items continued to be generated from TIMSS sources, but the additional items from Educational Testing Service and SRI International ensured that the very tight deadlines for test production were met. Many of the items provided by these agencies had been piloted already, or had been used in large-scale surveys, and therefore had known properties. As a result of these activities and the inclusion of a further selection of items from SIMS, the size and quality of the pool of items from which the field trial tests were to be selected was greatly enhanced.

2.3.3 PREPARATION FOR THE FIELD TRIAL/ITEM PILOT

The development of the field trial tests from the augmented item pool involved progressive selection and development based on the following considerations:

- Matching of the item pool to the revised test blueprint
- Selection of items based on empirical considerations (item pilot and field trial)
- The professional judgments of subject matter experts
- Other considerations imposed by the test design.

In this section, the schedule of this process is shown, together with descriptions of the final blueprint development, the process of item selection by the SMAC, and the development of tests for both the 1994 field trials and the 1995 main survey.

[2] The "Think Tank" is part of a grants program, sponsored by the AERA and funded by the National Science Foundation and the National Center for Education Statistics, that is aimed at building the educational research infrastructure. The program includes a mechanism for bringing together outstanding scholars to address pressing issues in educational research, policy, or practice.

By August 1993 a number of test-related issues had been resolved. The time to be allowed for testing at each of the population levels had been determined, the desired reporting categories had been identified, a draft test design had been developed, and plans for finalizing the tests had been formulated. However, the time remaining to implement these plans was very short and there followed a period of sustained and intense activity. Table 2.1, presents the schedule of events related to the test development from 1993 to the assembly of the main survey test booklets in 1994.

Table 2.1 Schedule of Test Development From August 1993

DATE	POPULATIONS 1 AND 2	POPULATION 3
January 93	Final selection of items for international pilot	Postponement of item pilot until field trial
March 93	Pilot booklets distributed	
April-June 93	Item pilot	
June-August 93	Pilot data analyzed	
August 93	AERA Think Tank on test development	AERA Think Tank on mathematics and science literacy
August-September 93	SRI and ETS contracted to produce additional items	ACER contracted to produce additional items
August-September 93	Preselection of items for field trial	Preselection of items for field trial
September 93	Selection and editing of field trial items by the SMAC	Review of item selection; Population 3 workgroup set up
September-November 93	ETS prepared draft field trial booklets	
October 93	Blueprints for main survey tests drafted from curriculum analysis data	
November 93	NRCs approved blueprints; NRCs approved field trial items	NRCs approved delay of Population 3 field trial tests
November 93		Working group meeting on mathematics and science literacy items
December 93	ETS completed field trial booklet preparation	
December 93		Working group meetings on advanced mathematics and physics items
December 93		Selection and editing of field trial items
February 94	Field trial	
April-May 94	Analysis of field trial data; Development of coding system	Development of coding system
May 94		Field trial
June 94	Preselection of main survey items	
June-July 94		Analysis of field trial data
July 94	Selection of main survey items; final coding rubrics developed	
July-August 94	Clustering of items and booklet preparation	
August 94	Final booklet approval	
August-September 94		Preselection of main survey items
October 94		Selection of main survey items

Table 2.1 Schedule of Test Development From August 1993 (continued)

DATE	POPULATIONS 1 AND 2	POPULATION 3
October-November 94	Tests administered (Southern Hemisphere)	
November 94		Item selection approved; coding rubrics finalized
December 94		Items assembled into clusters and test booklets prepared
March-May 95	Tests administered (Northern Hemisphere)	Tests administered (Northern Hemisphere)
August-September 95		Tests administered (Southern Hemisphere)

The beginning of this stage of intensive test development activity coincided with a period of transition, in which responsibility for the overall direction of TIMSS was transferred from the ICC in Vancouver, Canada, to the International Study Center at Boston College. At the same time a number of study activities were delegated to centers around the globe, under the direction of the International Study Director. It is worth noting the extent to which aspects of test development were dispersed. The study was managed from Boston, USA. Test development coordinators Robert Garden and Graham Orpwood were located in Vancouver, Canada, and Toronto, Canada, respectively. Contractors produced additional items in California, USA, New Jersey, USA, and Melbourne, Australia. Field trial test booklets were prepared in New Jersey, Melbourne, and Boston. Field trial data from participating countries were processed at the IEA Data Processing Center in Hamburg, Germany, and further analyzed at ACER in Melbourne, Australia, before results were sent to the International Study Center at Boston College and to the test development coordinators in Vancouver and Toronto. The potential for administrative problems and delays is obvious, but through extensive use of modern communication and information transfer methods, efficient management, and excellent cooperation from all those involved, the task was accomplished smoothly.

2.4 TEST BLUEPRINT FINALIZATION

While preliminary test blueprints for the achievement tests were drafted early in the study to guide item collection and development, the blueprints were not finalized until October 1993. This reflected the desire to use data from the curriculum analysis project to confirm that the blueprints represented the best attainable fit to the curricula of the participating countries. The task of translating curriculum data into draft test blueprints was undertaken by a group of people invited to Michigan State University in East Lansing, Michigan, in October 1993. This version of the test blueprint (McKnight et al., 1993), amended very slightly by the SMAC, was approved by NRCs in November 1993.[3] In general this blueprint was closely adhered to through to the production of the final instruments,

[3] The final TIMSS test blueprints are provided in Appendix B.

although results of the field trials and additional constraints (such as the reduction of testing time in Population 1) affected the final item distribution somewhat.

The TIMSS curriculum frameworks provided a unifying system of categorization for both curriculum analysis and test development. For the purposes of test development, two dimensions of the frameworks were used–subject-matter content and performance expectations. The former denoted the mathematics or science topic being tested using any given item, and the latter characterized the type of student performance called for by the item. The item classification system used in TIMSS permitted an item to draw on multiple content areas and/or involve more than one performance expectation, so that an item could have several content and performance codes. However, for the purpose of test construction only the principal code was used on each of the two dimensions.

TIMSS was designed to permit detailed analysis of student performance in many content-by-performance expectation categories. However, because of limitations in data collection and resources, many of these detailed categories had to be combined into a few "reporting categories" for analysis and presentation in the international reports. The final set of reporting categories was based on major areas of mathematics and science content, and on the topics identified as "in-depth topics" for the curriculum analysis.

In Population 1 mathematics, the blueprint content categories 'Whole numbers: place value' and 'Whole numbers: other content' were combined to form the reporting category 'Whole numbers.' 'Decimal fractions,' 'Common fractions' and 'Proportionality' were joined to form 'Fractions and proportionality.' 'Estimation and number sense' and 'Measurement' form 'Measurement, estimation, and number sense.' 'Data analysis' and 'Probability' were combined to form 'Data representation, analysis, and probability.' The content categories 'Geometry' and 'Patterns, relations, and functions' remained as separate reporting categories.

In Population 1 science, the content categories 'Earth features' and 'Earth science: other content' were combined to form the reporting category 'Earth science,' while 'Human biology' and 'Life science: other content' were combined to form 'Life science.' 'Physical science' remains as a reporting category, while 'Environment' and 'Other content' were combined to form 'Environmental issues and the nature of science.'

In Population 2 mathematics, 'Common fractions: meaning, representation,' 'Common fractions: operations, relations, and proportions,' 'Decimal fractions' and 'Estimation and number sense' were combined into the reporting category 'Fractions and number sense.' 'Congruence and similarity' and 'Other geometry' were combined to form 'Geometry,' and 'Linear equations' and 'Other algebra' to form 'Algebra.' 'Data representation and analysis' was combined with 'Probability' to form 'Data representation, analysis, and probability.' 'Measurement' and 'Proportionality' remained as separate reporting categories.

In Population 2 science, 'Earth features' and 'Earth science: other content' were combined to form 'Earth science.' 'Life science' was composed of 'Human biology' and 'Life science: other content.' 'Energy types,' 'Light,' and 'Physics: other content' were combined

to form 'Physics,' while the content category 'Chemistry' remained a separate reporting category. 'Environment' and 'Other content' were combined to form 'Environmental issues and the nature of science.'

In Population 3, mathematics and science literacy was composed of three reporting categories: 'Mathematics literacy,' 'Science literacy,' and 'Reasoning and social utility.' 'Number sense,' 'Algebraic sense,' and 'Measurement and estimation' were combined to form 'Mathematics literacy.' 'Earth science,' 'Human biology,' 'Other life science,' 'Energy,' and 'Other physical science' were combined to form 'Science literacy.' The 'Reasoning and social utility' categories from the mathematics and science blueprints were combined to form a single reporting category 'Reasoning and social utility.'

In Population 3 advanced mathematics the reporting categories correspond to the blueprint content areas. In physics, 'Forces and motion' was renamed 'Mechanics' for reporting purposes. 'Electricity and magnetism' remained as a reporting category, while the blueprint content category 'Thermal and wave phenomena' was broken into two reporting categories: 'Heat' and 'Wave phenomena.' 'Particle evaluation' was labeled 'Particle, quantum, astrophysics, and relativity' for reporting purposes.

In Table 2.2, the reporting categories for the mathematics and science content areas are shown. Table 2.3 presents the performance expectations categories which were recommended as reporting categories.

Table 2.2 Reporting Categories for Mathematics and Science Content Areas

	Mathematics	Science
Population 1	Whole numbers Fractions and proportionality Measurement, estimation, and number sense Data representation, analysis, and probability Geometry Patterns, relations and functions	Earth science Life science Physical science Environmental issues and the nature of science
Population 2	Fractions and number sense Geometry Algebra Data representation, analysis, and probability Measurement Proportionality	Earth science Life science Physics Chemistry Environmental issues and the nature of science
Population 3	Numbers, equations, and functions Analysis (calculus) Geometry Probability and statistics Validation and structure	Mechanics Electricity and magnetism Heat Wave phenomena Particle, quantum, astrophysics, and relativity
Population 3 (literacy)	Mathematics literacy Reasoning and social utility	Science literacy Reasoning and social utility

Table 2.3 Reporting Categories for Performance Expectations

	Mathematics	Science
Populations 1, 2, and 3	Knowing Routine procedures Complex procedures Solving problems Justifying and proving Communicating	Understanding Theorizing, analyzing, and solving problems Using tools, routine procedures, and science processes Investigating the natural world

Several factors were considered in determining the distribution of items across the cells of the blueprints. A major concern was that each reporting category would be represented by sufficient items to generate a reliable scale. Other important factors are outlined below.

- *Amount of testing time*. NRCs had set the maximum testing time for students at 90 minutes (this was subsequently reduced to 70 minutes for Population 1). In order to allocate items to booklets so that optimal use was made of student time, the amount of time a student needed to complete each of the item types had to be estimated. (See Table 2.4.)

Table 2.4 Estimated Time Required by Different Populations to Complete Items of Different Types

	Multiple-Choice	Short-Answer	Extended-Response
Population 1	1 minute	1 minute	3 minutes
Population 2	1 minute	2 minutes	5 minutes
Population 3 (literacy)	1 minute	2 minutes	5 minutes
Population 3 (specialist)	3 minutes	3 minutes	5 minutes

By assembling items in 90-minute booklets distributed to the field trial sample, it was possible to include items needing a total testing time of 260 minutes at Population 1, and 396 minutes at Population 2, split equally between mathematics and science. At Population 3, the item pilot comprised 210 minutes of testing time for physics and specialist mathematics and 90 minutes of testing time for mathematics and science literacy items (combined). About twice the number of items required for the main survey were included in the field trial.

- *Coverage of subject-matter content.* At the time the blueprints were developed, preliminary data were available from about 20 countries from the modified topic trace mapping and document analyses data collected for the curriculum analyses. These data showed the proportion of each country's curriculum that was allocated to each content topic. Rough averages of these numbers provided a basis for determining the proportion of total test time to be allocated to each content topic. These were then adjusted to ensure that adequate test time was given to in-depth topics. The resulting grids were prepared for mathematics and science separately.
- *Coverage of performance expectations.* Once the total number of minutes had been allocated to a given content topic, it was distributed across performance categories using the best professional judgment of the group. It was intended that no more than 70% of the total testing time would be allocated to multiple-choice items. In the case of mathematics, the number of items by type was allocated to each cell of the grid. In the case of science, the total number of minutes per cell was allocated, leaving the specific numbers of each type of item in each cell to be determined later. This procedure gave science item selection more flexibility.

2.5 THE FIELD TRIAL

Armed with the new blueprint, the test development coordinators, assisted by selected subject-matter specialists and supported by the International Study Center, organized collections of items for the field trial to ensure that approximately twice the number of items eventually required would be tested in all countries. This preselection was based on the results of the item pilot and review described earlier and included new items drawn from the work by SRI International and Educational Testing Service (Populations 1 and 2) and Australian Council for Educational Research (Population 3). Subject to approval by the NRCs and the International Study Director, responsibility for final selection of test items for the field trial was largely in the hands of the SMAC, supplemented from time to time by selected NRCs and other subject-matter specialists.

At the September 1993 SMAC meeting, members were provided with preselected items for each subject-matter content category of the blueprint in each population. Subgroups of mathematics and science experts scrutinized items for the three TIMSS target populations.

Some items were accepted as they were, others were edited to improve substance or layout, and still others were replaced by items that were more to the liking of the committee members. SMAC members had at their disposal the p-values and discrimination indices for all items that had been used in the item pilot. Items having p-values outside the range 0.2 through 0.85, or point-biserial coefficients below 0.2 (0.3 for medium p-values), were automatically excluded, except where modifications in a piloted item were expected to improve the item significantly.

Data from the NRCs' review of items also played an important part in selection decisions. Items that had been judged unacceptable by more than a few national committees were rejected. Most "unacceptable" ratings from the NRC review reflected students' lack of opportunity to learn the content addressed by the item, perceived cultural bias, or lack of face validity. To ensure that there would be sufficient items from which to choose, the field trial item pool included twice as many items from each cell of the blueprint as were required for the final tests.

The Population 3 item pool was not considered ready for field testing. SMAC therefore suggested to the International Study Director that a further delay of the Population 3 field trial be considered and that a special working group be established to work with the ACER contractors to ensure that a high-quality item pool be available.

Following the SMAC meeting, the Center for the Assessment of Educational Progress at Educational Testing Service was contracted to prepare master copies of test booklets for the Populations 1 and 2 field trial scheduled for February 1994. As part of the process, however, the NRCs were given the opportunity to review the proposed field trial items. Educational Testing Service prepared draft field trial booklets and these were examined and commented on by NRCs from each country during the course of a meeting. Many suggestions were made, and were taken into account as far as was possible.[4]

The purpose of the field trials was to verify the properties of the items developed since the 1993 item pilot, and to try out all procedures to be used in the main survey, and so national centers were strongly encouraged to participate fully. However, timing of the trial in relation to the school year made this impossible for some countries, and others were not able to muster the necessary resources to include every population. Most were able to carry out the trial for at least one population, and this gave a good spread of countries at each level for item-piloting purposes. National centers were asked to administer the achievement tests to judgment samples of about 100 students per item. Table 2.5 lists the countries that participated in the field trial.

[4] One suggestion, for example, resulted in a complete restructuring of the booklets. The TIMSS Technical Advisory Committee had thought it desirable to concentrate all items in a given reporting category in one booklet to allow for testing of scales, and the draft booklets were so arranged. However, NRCs believed that this organization of items would distress students who had not been taught the particular topics at all and who could answer none of the questions in a booklet. As a result, the field trial booklets were reorganized so that each contained items from several content areas. The final field trial item pool was organized in 16 booklets for each population.

Table 2.5 Participation in the TIMSS Field Trial

Population 1			
Australia	Greece	Kuwait	Portugal
Austria	Indonesia	Latvia	Singapore
Canada (British Columbia)	Iran	Netherlands	Slovak Republic
Canada (Alberta)	Ireland	Norway	Slovenia
Canada (Ontario)	Japan	Philippines	USA
England			

Population 2			
Australia	Germany	Latvia	Slovak Republic
Austria	Greece	Netherlands	Slovenia
Belgium	Indonesia	Norway	Spain
Canada (British Columbia)	Iran	Philippines	Sweden
Canada (Alberta)	Ireland	Portugal	Switzerland
Canada (Ontario)	Japan	Romania	Tunisia
Denmark	Kuwait	Singapore	USA
England			

Population 3			
Australia	Czech Republic	Mexico	Russia
Austria	Denmark	Netherlands	Sweden
Canada (Alberta)	France	Norway	Switzerland
Canada (Ontario)	Latvia	New Zealand	USA

2.6 PREPARATION FOR THE MAIN SURVEY

2.6.1 ITEM SELECTION FOR THE MAIN SURVEY

The process followed in developing the achievement instruments for the main survey was similar to that which proved successful for the field trial and which the IEA Technical Advisory Committee had judged appropriate. Preliminary analysis of the field trial achievement data was carried out at the IEA Data Processing Center in Hamburg, with further analysis at the Australian Council for Educational Research. These analyses yielded both classical and Rasch item analyses, and displays of item-by-country interactions.

As part of the field trial, national committees reviewed each item. Each item was given a rating of 1 to 4 in four carefully described areas. These can be briefly characterized as coverage (the extent to which the content of an item was judged to be taught and emphasized in a country), familiarity (with the teaching approach implied by what is being assessed), difficulty (a judgment of what proportion of students would answer correctly), and appeal (a rating of the "quality" of the item independent of whether it was appropriate to the local curriculum). Mean ratings were used to categorize items according to whether, on the basis of the national reviews, they were likely, possible, or unlikely candidates for inclusion in the main survey. National review committees also scrutinized each item for

possible cultural or other bias. A very few field trial items were excluded from consideration for the main tests on these grounds.

On the basis of the field trial results, preliminary selections of items were made by the mathematics and science coordinators with advice and assistance from other subject-matter specialists. For each cell of the TIMSS blueprint, items were chosen to meet, as nearly as possible, the specifications for the numbers of each item type required. The intention was to have items within each cell, and especially within each content line and reporting category, that elicit in a variety of ways what students have learned in these areas. The principal factors that influenced the selection of items in each cell were item statistics, item review data, and NRC comments. These were balanced against the need for varied items that sampled a range of content and performance expectations within that cell of the blueprint. With few exceptions, the selected items had mean field trial p-values between 0.3 and 0.8, discrimination indices (point-biserial correlation between item and booklet scores) above 0.3, and mean review ratings above 2.5 in each of the four review categories. However, the shortage of acceptable items in some cells meant that there were minor deviations from the Population 1 and Population 3 blueprints at this stage.

The draft selections of items were considered by the SMAC and selected NRCs at two meetings, one for Populations 1 and 2 and the other for Population 3. To facilitate item selection, each item was printed on one sheet with its summary field trial and review statistics and, for free-response items, the scoring rubric that had been used. In addition, displays of item-by-country interaction for each item were presented. The proposed selections were considered item-by-item on their merits both as individual items and as components of a scale based on subject-matter content.

Following the SMAC item-review meetings, the refined selections were formatted into booklets and presented for final review at a general meeting of all NRCs. NRCs paid particular attention to items that might cause problems in translation from English to the language of testing. NRCs proposed a number of minor change in wording and layout of items. Most of these suggestions were followed and served to improve overall test quality. At the end of the meeting the NRCs formally approved the item selections for the main survey.

2.6.2 Free-Response Item Coding and Test Development

The Free-Response Item Coding Committee (FRICC) was established to develop coding guides for the free-response items. The work of the FRICC and the principles of the coding system adopted for TIMSS are described in Chapter 7 of this report. Ideally, test items and coding rubrics would have been developed simultaneously, but a fully evolved coding scheme was not available until the test development process had been under way for some time. Nevertheless, development of the coding scheme played an important role in the selection and editing of items for the main survey.

The coding guides for the 1993 item pilot and for the 1994 field trial were designed to produce a single "correctness" score on a three- or four-point scale. There was, however,

considerable interest in obtaining more informative "diagnostic" data from the free-response items. Accordingly, following the field trial, researchers in some of the Nordic countries collaborated to prepare and trial an alternative coding system of double-digit codes that provided not only "correctness" scores for each response but also qualitative distinctions among different responses having the same score. The TIMSS codes finally developed were based on that proposal.

The more detailed system of coding suggested additional criteria for developing and refining items. Free-response items selected by the SMAC (and in some cases edited in light of results from the field trial, or suggestions from the NRCs or subject-matter specialists) were then assessed by the FRICC for applicability of the two-digit trial coding system. Evidence from small-scale trials was available. The FRICC then developed the coding rubrics for the items and in many cases proposed further editorial changes in the items. Where changes were judged very unlikely to invalidate field trial item statistics or review data, the test development coordinators approved them. Because of the close relationship between the wording of a free-response item and its coding, the FRICC and the SMAC worked closely together in the final development of both tests and codes.

2.6.3 Item Clustering and Test Booklet Preparation

Chapter 3 of this report describes the overall test design in detail. This design called for items to be grouped into "clusters," which were distributed (or "rotated") through the test booklets so as to obtain eight booklets of approximately equal difficulty and equivalent content coverage. Some items (the core cluster) appeared in all booklets, some (the focus clusters) in three or four booklets, some (the free-response clusters) in two booklets, and the remainder (the breadth clusters) in one booklet only. In addition, each booklet was designed to contain approximately equal numbers of mathematics and science items.

After the final item pool had been determined, items were assigned to clusters in several steps. First, items were allocated to clusters; second, they were sequenced within clusters; and third, the order of the response options for the multiple-choice items was checked, and where necessary reorganized to prevent undesirable patterns of correct responses.

The test design specified the numbers of multiple-choice, short-answer, and extended-response items in mathematics and science to be included in each cluster. Items were therefore selected collaboratively by the mathematics and science coordinators. The aim was to develop clusters with certain characteristics, described below.

- Clusters should be of approximately equal difficulty (based on p-values of items from the field trial)
- The test booklets should have approximately equal difficulty
- The core and focus clusters should consist of items with p-values close to the 0.5-0.6 range; discrimination indices (item-booklet point-biserial correlations) that exceeded 0.3 for correct responses and were negative for distracters; low item-by-country interactions; and a good spread of subject-matter content and performance categories

Once the draft clusters were in place, the pattern of correct responses for each multiple-choice cluster and each booklet was checked to ensure that, as far as possible, each correct response (A, B, C, etc.) occurred with equal frequency both within clusters and within booklets, and that regular patterns of such responses (e.g. A, B, A, B, . . .) were avoided. This meant either changing the sequence of items within a cluster or editing items to change the sequence of distracters. This type of editing could be done only with items whose distracters were not in a logical sequence.

Further minor resequencing of items within clusters was influenced by the need to place items on the page in such a way as to keep the overall number of booklet pages as small as possible, yet allow enough space for the translation of the items into other languages (item sequence and page layout was to be retained across all languages). A check was also made to ensure that items in a cluster or booklet did not provide clues to the answers to other items in the same cluster or booklet.

The result of the entire process was the final set of item clusters for each of the three student populations as set out in the test design. Artwork for the items, formatting of booklets, and final editing were done by International Study Center staff. The International Study Center also distributed the booklets, both electronically and in hard copy, to national centers.

2.6.4 Linking Items Across Populations

In order to link achievement areas across the TIMSS populations, items were used where possible in two adjacent populations. This means that some items were common to Populations 1 and 2, and some to Populations 2 and 3. Links to SIMS were maintained by including SIMS items at Populations 2 and 3 (See Table 2.6).

Table 2.6 Link Items

TIMSS Population 1 and TIMSS Population 2	32 items
TIMSS Population 2 and TIMSS Population 3 (literacy)	21 items
TIMSS Population 3 (literacy) and SIMS Population A	7 items
TIMSS Population 3 (advanced mathematics) and SIMS Population B	32 items

2.7 CALCULATORS AND MEASURING INSTRUMENTS

Opinions, sometimes strongly held, differed on whether the use of calculators should be allowed for TIMSS tests. The following decisions were reached after careful consideration of all the issues involved:

Population 1 — calculating devices NOT permitted

Population 2 — calculating devices NOT permitted

Population 3 — calculating devices permitted.

The fact that calculators were allowed for TIMSS Population 3 mathematics and science literacy tests but not for TIMSS Population 2 tests may call into question the comparability of achievement measures on a small number of link items between these populations; however, none of the items involved is likely to be made significantly easier by the use of a calculator. Link items between TIMSS Population 3 advanced mathematics and SIMS Population B, between TIMSS Population 2 and SIMS Population A, and between TIMSS Population 1 and TIMSS Population 2 are unaffected by the policy on calculator use.

Measuring instruments (such as graduated rulers and protractors) were NOT permitted for any of the student populations because several items call for estimation.

REFERENCES

McKnight, C.C., Schmidt, W.H., and Raizen S.A. (1993). *Test Blueprints: A Description of the TIMSS Achievement Test Content and Design* (Doc. Ref.: ICC797/NRC357). Document prepared for the Third International Mathematics and Science Study (TIMSS).

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.

Adams, R.J. and Gonzalez, E.J. (1996) "The TIMSS Test Design" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

3. The TIMSS Test Design

Raymond J. Adams
Eugenio J. Gonzalez

3.1 OVERVIEW

This chapter describes the TIMSS test design for each of the three TIMSS student populations. Underpinning the design is the recognition that the main goal of TIMSS was to estimate various characteristics of the defined populations in participating countries over a broad range of outcomes in mathematics and science. Accommodating the research demands and priorities of participants around the world, the curricular differences among educational systems, the precision necessary for the estimation of parameters, the constraints on testing time, and the need for simple administrative procedures required a test design which was flexible enough to meet the needs of individual participants yet rigid enough to meet acceptable quality standards. These competing demands imposed on the TIMSS test design can be summarized as a tension between the desire for wide subject-matter coverage and the limitations imposed by available resources.

Fortunately, modern test-scaling methods such as item response theory (Lord, 1980; Wright and Stone, 1979) and plausible value technology (Rubin, 1987; Mislevy and Sheehan, 1987) have made it possible to deal with much of the tension among practicality, coverage, and precision. Most important, it has been shown (Lord, 1962; Beaton, 1987) that when these methods are used, accurate estimates can be obtained for populations without precise measurement of individual students. That is, not all students in an assessment sample need to respond to all the test items. To derive population estimates, a multiple matrix sampling design is used in which subsets of items selected from the total item pool are administered

to random subsamples of students. The testing time for each student, and consequently the number of items administered to each student, is limited to an acceptable level, yet the population performance on a range of dimensions can be characterized.

The TIMSS tests were designed to be administered to three TIMSS student populations.

Population 1

Students in Population 1 (the two adjacent grades containing the largest proportion of 9-year-old students) were to be given tests that contained items in both mathematics and science. Proficiency estimates were required for each of ten reporting categories in mathematics (six) and science (four). A total of 199 unique items were to be distributed across eight test booklets for this population.

Population 2

Students in Population 2 (the two adjacent grades containing the largest proportion of 13-year-old students) were also to be given tests that contained items in both mathematics and science. Proficiency estimates were required for each of eleven reporting categories in mathematics (six) and science (five). A total of 286 unique items were to be distributed across eight test booklets for this population.

Population 3

Students in Population 3 (the final year of secondary schooling) were to be given tests containing items of general mathematical and scientific knowledge and on applications of mathematical and scientific principles to everyday situations. Proficiency estimates were required for each of three reporting categories: mathematics literacy, science literacy, and reasoning and social utility. A total of 76 unique items were to be distributed across two test booklets for Population 3 students.

Students in Population 3 taking advanced courses in mathematics constituted a subpopulation for which separate reporting was required. These students were to be given tests in advanced mathematics. Proficiency estimates in five reporting categories were required. A total of 65 advanced mathematics items were to be distributed across four booklets for these students. Similarly, students in Population 3 taking advanced courses in physics constituted a subpopulation. These students were to be given tests in physics, for which proficiency estimates in five reporting categories were also required. A total of 65 physics items were to be distributed across four booklets for these students.

3.2 CONSTRAINTS OF THE TIMSS TEST DESIGN

In the design of the test booklets several constraints had to be taken into account. Because of the different characteristics of the populations, the constraints for the Population 3 test booklets differed from those for Populations 1 and 2. Some of the key constraints for the Population 1 and 2 tests are detailed below.

- The total testing time for individual students in Population 1 was not to exceed 70 minutes

- The total testing time for individual students in Population 2 was not to exceed 90 minutes
- There were to be a maximum of eight different test booklets each for Populations 1 and 2, to be distributed randomly ("rotated") in each sampled classroom
- There was to be a small subset of items common to all test booklets, to allow the verification of within-school booklet rotation and testing of the equivalence of the student samples assigned each booklet
- The booklets were to be approximately parallel in content coverage, although it was acceptable for some booklets to have more mathematics than science content and vice versa
- The booklets were to be approximately parallel in difficulty
- Each booklet was to contain some mathematics and some science items.[1]
- Content coverage in each subject (mathematics and science) was to be as broad as possible. In Population 1, items from 11 mathematics content areas and 7 science content areas of the TIMSS curriculum frameworks (Robitaille et al., 1993) were to be included in the tests. In Population 2, items from 12 mathematics content areas and 10 science content areas were to be included
- It was anticipated that resources would be insufficient to provide precise estimates of item-level statistics for all items in the TIMSS item pool. Therefore, it was decided that such estimates would be required only for a subset of test items. Those items should therefore appear in as many booklets as possible. Others would appear in only one or two booklets, and would have less precise item-level statistics
- For a subset of the items, all of the item intercorrelations were required, whereas for the remaining items only a sample of item intercorrelations was estimated

Some of the key constraints for the design of tests for Population 3 are listed below.

- The total testing time for individual students in Population 3 was not to exceed 90 minutes
- There were to be test booklets for the general Population 3 students, as well as test booklets for each subpopulation (students taking advanced mathematics and students taking physics)
- Booklets for the general population were to be approximately parallel in content, with a set of common items
- Booklets for the advanced mathematics subpopulation were to be approximately parallel in content, with a subset of items common to all of the advanced mathematics booklets
- Booklets for the physics subpopulation were to be approximately parallel in content, with a subset of items common to all of the physics booklets
- There was to be a booklet containing advanced mathematics items and physics items for students taking advanced courses in both mathematics and physics. This booklet

[1] While some mathematics and some science occurs in every booklet, it was not necessary to balance the booklets by having an equal proportion of mathematics and science, since this would reduce the number of within-subject inter-item covariances that can be estimated.

was also to contain reasoning and social utility items designed for the general Population 3 students

- The mathematics and science literacy booklets were to include items from mathematics, science, and reasoning and social utility
- Items from five mathematics content areas were to be included in the booklets for the students taking advanced mathematics
- Items from five physics content areas were to be included in the booklets for the physics students.

Aside from the constraints listed above, the test design for all three populations had to cater to the use of three test-item types: multiple-choice items, short-answer items, and extended-response items (short-answer and extended-response items are collectively referred to as free response items). Multiple-choice items require the student to recognize and identify the correct response to an item on the test form. The short-answer items require the student to give a short response such as a number, phrase, sentence, or diagram. The extended-response items involve more elaborate responses, and may require that students explain their reasoning or detail the steps they took in solving a problem. While the multiple-choice and most short-answer items can be adequately scored as either correct or incorrect, the extended-response items require a more complex system of partial-credit scoring to do justice to the wider range of student responses. Some free-response items had two parts, each analyzed separately. In TIMSS, short-answer and extended-response items can receive anywhere from 0 to 4 score points, depending on the complexity and quality of the response elicited and whether it is a multi-part item, while all multiple-choice items are scored 0 or 1.

The inclusion of a variety of item types required that assumptions be made, at the time of test construction, about the average response time of students to the different types of items. The working assumption for the required response time, for each item type for each population, are presented in Table 3.1.

Table 3.1 Minutes of Response Time Required

Item Type	Population 1	Population 2	Population 3 Literacy	Population 3 Advanced
Multiple-Choice	1	1	1	3
Short-Answer	1	2	2	3
Extended-Response	3	5	5	5

3.3 A CLUSTER-BASED DESIGN

The design chosen for the TIMSS tests for all three populations calls for two stages. Items in the item pool were first assigned to one of a set of mutually exclusive groups, or "clusters." The clusters of items were then systematically assigned to test booklets. An item cluster is a small group of items that are collected and then treated as a block for the purpose of test construction. A cluster might appear in more than one test booklet;

furthermore, a cluster that appears in more than one booklet might appear in a different position within each booklet (e.g., first in one booklet, second in another, third in another). In each of Populations 1 and 2 there is one cluster of items (the core cluster) that appears in all eight booklets for that population.

Each item cluster has an identical format and layout wherever it appears. A test booklet is made up of item clusters and corresponds to the set of items that will be administered to an individual student. Because item clusters may be allocated to more than one booklet, the booklets do not contain mutually exclusive subsets of the total item pool: some test items will appear in more than one booklet.

The number of items in each cluster varies, because the cluster design is based on the estimated number of minutes it would take a typical student to answer the items, rather than on the total number of items. For Population 1, one item cluster has been designed to take 10 minutes while the remaining 25 have been designed to take 9. The Population 2 item clusters vary in length: 8 of the clusters are estimated to take 10 minutes, 8 to take 12 minutes, and 10 to take 22 minutes. This means that there is a total pool of 235 unique testing minutes in Population 1 and a total of 396 unique testing minutes in Population 2. Half of these pools were allocated to mathematics and half to science in each population. For Population 3, the test items have been grouped into 12 clusters of varying length. There are 4 core clusters, each 30 minutes in length. The remaining 8 clusters vary in length, from 30 minutes (for the literacy rotated clusters) to 60 minutes (for the advanced mathematics and physics rotated clusters).

3.4 TIMSS POPULATION 1 TEST DESIGN

All test items for Population 1 were grouped into 26 mutually exclusive item clusters, each identified by a letter of the alphabet (A through Z). There is a core cluster (cluster A) that appears in the second position in every booklet. Each cluster contains items in mathematics, science, or both. Each test booklet for Population 1 comprises up to 7 item clusters. The booklets are divided into two parts and are administered in two consecutive testing sessions. Some clusters appear in all booklets, some in four, some in three, some in two, and some in only one booklet. Of the 26 clusters at Population 1, 18 take 9 minutes and 8 take 10 minutes.

In Population 1, it is convenient to regard the clusters as being of the six types described in the following.

1. Core Cluster — One Core cluster comprising five mathematics and five science multiple-choice items was assigned to all booklets. It is labeled cluster A.

2. Focus Clusters — There were seven Focus clusters, each containing nine minutes of science and mathematics items. Four of the Focus clusters contain five mathematics and four science multiple-choice items each, and three of them contain four mathematics and five multiple-choice items.[2] These clusters were assigned to each of the first seven booklets. They are called Focus clusters because each appears in at least three booklets, so that the items in them were answered by a relatively large fraction (three-eighths) of the student sample in each country–enough to permit accurate reporting of the item statistics. The Focus clusters are labeled B through H.

3. Mathematics Breadth Clusters — There were five Mathematics Breadth clusters, each containing nine minutes of mathematics items and consisting largely, but not exclusively, of multiple-choice items. These clusters appear in only one booklet, and consequently the number of students responding to each item was small. While the items in these clusters contribute to the breadth of content coverage of the tests, the accuracy of the item statistics would be relatively low. These clusters are labeled I through M.

4. Science Breadth Clusters — There were five Science Breadth clusters, each containing nine minutes of science items and consisting largely, but not exclusively, of multiple-choice items. These clusters appear in only one booklet, and consequently the number of students responding to each item was small. While the items in these clusters contribute to the breadth of the tests, the accuracy of their item statistics would be relatively low. These clusters are labeled N through R.

5. Mathematics Free-Response Clusters — There were four Mathematics Free-Response clusters, each containing nine minutes of short-answer and extended-response mathematics items. These clusters were each assigned to two booklets, so that item statistics of reasonable accuracy would be available. These clusters are labeled S through V.

6. Science Free-Response Clusters — There were four Science Free-Response clusters, each containing nine minutes of short-answer and extended-response science items. These clusters were each assigned to two booklets, so that item statistics of reasonable accuracy would be available. These clusters are labeled W through Z.

[2] In the final design two of the clusters (E and H) were assigned one short-answer science item instead of a multiple-choice item.

Tables 3.2 and 3.3 list the clusters and the number of items of each type allocated to each cluster for mathematics and science respectively. The number of items per cluster varies because items have been allocated to clusters on the basis of the estimated number of minutes that it would take a typical student to answer them, rather than on the total number of items.

Table 3.2 Number of Mathematics Items per Cluster, by Item Type, for Population 1

Cluster	Multiple-Choice	Short-Answer	Extended-Response	Total
A	5	-	-	5
B	5	-	-	5
C	4	-	-	4
D	5	-	-	5
E	4	-	-	4
F	5	-	-	5
G	4	-	-	4
H	5	-	-	5
I	9	-	-	9
J	9	-	-	9
K	9	-	-	9
L	8	1	-	9
M	7	2	-	9
S	-	3	2	5
T	-	3	2	5
U	-	3	2	5
V	-	3	2	5
Total	**79**	**15**	**8**	**102**

Table 3.3 Number of Science Items per Cluster, by Item Type, for Population 1

Cluster	Multiple-Choice	Short-Answer	Extended-Response	Total
A	5	-	-	5
B	4	-	-	4
C	5	-	-	5
D	4	-	-	4
E	4	1	-	5
F	4	-	-	4
G	5	-	-	5
H	3	1	-	4
N	9	-	-	9
O	7	2	-	9
P	8	1	-	9
Q	7	2	-	9
R	8	1	-	9
W	-	3	2	5
X	1	2	2	5
Y	-	-	3	3
Z	-	-	3	3
Total	**74**	**13**	**10**	**97**

3.4.1 ORGANIZATION OF THE TEST BOOKLETS

In Population 1, the test design specifies eight booklets, each estimated to take a student 64 minutes to complete. Each booklet was constructed from one ten-minute Core cluster (cluster A) and six nine-minute clusters. Table 3.4 shows the assignment of clusters to booklets, as well as the position of each cluster within the booklet.

Table 3.4 Assignment of Item Clusters to Population 1 Booklets[3]

Cluster Type	Cluster Label	Booklet 1	2	3	4	5	6	7	8
Core (10 minutes)	A	2	2	2	2	2	2	2	2
Focus	B	1				5		3	1
(10 minutes)	C	3	1				5		
	D		3	1				5	
	E	5		3	1				
	F		5		3	1			
	G			5		3	1		
	H				5		3	1	
Breadth	I								5
(Mathematics)	J	6							
(9 minutes)	K			6					
	L					6			
	M							6	
Breadth	N		6						
(Science)	O				6				
(9 minutes)	P						6		
	Q								6
	R								3
Mathematics Free-Response	S	4							7
(9 minutes)	T	7		4					
	U			7		4			
	V					7		4	
Science Free-Response	W		4					7	
(9 minutes)	X		7		4				
	Y				7		4		
	Z						7		4

[3] Numbers in the cells indicate the position of the cluster within the booklet.

The order of the clusters within the Population 1 booklets is shown in Table 3.5. Cluster A is the Core cluster and has been assigned to all booklets. The rotation design used to assign clusters B through H to booklets 1 through 7 allows the estimation of all item covariances for the items in clusters A through H. Booklet 8 serves primarily to increase the content coverage of the tests. Apart from booklet 8 (which has three), each booklet has only one Breadth cluster, and each Breadth cluster appears in only one booklet.

Table 3.5 Ordering of Item Clusters Within Population 1 Booklets

	Booklet							
Cluster Order	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1st	B	C	D	E	F	G	H	B
2nd	A	A	A	A	A	A	A	A
3rd	C	D	E	F	G	H	B	R
4th	S	W	T	X	U	Y	V	Z
BREAK								
5th	E	F	G	H	B	C	D	I
6th	J	N	K	O	L	P	M	Q
7th	T	X	U	Y	V	Z	W	S

The Population 1 test design has the following features.

- The Core cluster (cluster A) appears in the second position in all test booklets.
- The Focus clusters (clusters B through H) each appear in at least three booklets, each time in a different position. They are assigned to each of the first seven booklets following a Balanced Incomplete Block design. In booklets 1 through 7, each Focus cluster appears together once with each of the remaining Focus clusters.
- Each of the Focus clusters occurs once in the first, third, and fifth positions in booklets 1 through 7.
- All test booklets contain mathematics and science items. Test booklets 1, 3, 5, and 7 have more mathematics items; booklets 2, 4, 6, and 8 have more science items.
- The test booklets are designed to be administered in two consecutive testing sessions with a 15-20-minute break in between. The first four clusters of items in the Population 1 test booklets were administered during the first testing session (37 minutes); after the break the remaining three clusters were administered (27 minutes).
- There are Free-Response clusters in Part 1 as well as in Part 2 of each test booklet (fourth and seventh cluster in each booklet).
- The design provides a total of 235 minutes, 118 for mathematics and 117 for science.

3.4.2 CONTENT OF THE TEST BOOKLETS

Test items were included in the Population 1 tests from 11 content areas in mathematics and 7 content areas in science. Some of these content areas are merged, for the purpose of scaling and reporting, into 6 reporting categories for mathematics and 4 for science. That is,

it will be possible to characterize the TIMSS Population 1 with respect to 6 mathematics and 4 science achievement dimensions.

The 6 mathematics reporting categories are:

- Whole numbers
- Fractions and proportionality
- Measurement, estimation, and number sense
- Data representation, analysis, and probability
- Geometry
- Patterns, relations, and functions.

The 4 science reporting categories are:

- Earth science
- Life science
- Physical science
- Environmental issues and the nature of science.

The TIMSS test blueprints (see Chapter 2) and the TIMSS curriculum frameworks provide more information on the composition of these reporting categories.

The Core and Focus clusters contain multiple-choice items only. The Breadth clusters include multiple-choice items and some short-answer items. Free-Response clusters consist almost exclusively of short-answer and extended-response items.

When items from both mathematics and science were included in a cluster they were grouped so that all of the mathematics items appear in a contiguous sequence, as do all of the science items. In half of these clusters the mathematics items were presented first; in the other half, the science items. Within each sequence, items were placed in order of estimated difficulty.

Table 3.6 shows the number of items by type, and the associated number of score points, for each of the content-based reporting categories for Population 1 mathematics. Table 3.7 provides the same information for Population 1 science.

Table 3.6 Number of Mathematics Items of Each Type and Score Points, by Reporting Category, Population 1

Reporting Category	Multiple-Choice	Short-Answer	Extended-Response	Total Items	Score Points
Whole numbers	19	5	1	25	28
Fractions and proportionality	15	2	4	21	28
Measurement, estimation, and number sense	16	3	1	20	21
Data representation, analysis, and probability	8	2	2	12	15
Geometry	12	2	-	14	14
Patterns, relations, and functions	9	1	-	10	10
Total	**79**	**15**	**8**	**102**	**116**

Table 3.7 Number of Science Items of Each Type and Score Points, by Reporting Category, Population 1

Reporting Category	Multiple-Choice	Short-Answer	Extended-Response	Total Items	Score Points
Earth science	13	2	2	17	19
Life science	33	5	3	41	45
Physical science	23	4	3	30	33
Environmental issues and the nature of science	5	2	2	9	11
Total	**74**	**13**	**10**	**97**	**108**

Tables 3.8 and 3.9 show the number of items from each reporting category that are included in each of the eight test booklets for mathematics and for science. Tables 3.10 and 3.11 show the maximum number of possible score points for each reporting category.

Table 3.8 Number of Mathematics Items in Each Booklet by Reporting Category, Population 1

Reporting Category	Booklet							
	1	2	3	4	5	6	7	8
Whole numbers	9	4	7	6	10	6	11	7
Fractions and proportionality	9	6	10	4	8	4	7	8
Measurement, estimation, and number sense	9	4	9	5	8	4	6	3
Data representation, analysis, and probability	5	3	3	1	4	1	4	2
Geometry	4	2	4	2	4	1	4	3
Patterns, relations, and functions	1	-	4	1	4	2	2	1
Total	**37**	**19**	**37**	**19**	**38**	**18**	**34**	**24**

Table 3.9 Number of Science Items in Each Booklet by Reporting Category, Population 1

Reporting Category	Booklet							
	1	2	3	4	5	6	7	8
Earth science	6	7	7	10	7	8	5	4
Life science	6	16	6	13	6	14	10	12
Physical science	4	10	4	8	4	9	5	11
Environmental issues and the nature of science	3	4	2	4	1	3	2	3
Total	**19**	**37**	**19**	**35**	**18**	**34**	**22**	**30**

Table 3.10 Maximum Number of Mathematics Score Points in Each Booklet by Reporting Category, Population 1

Reporting Category	Booklet							
	1	2	3	4	5	6	7	8
Whole numbers	9	4	7	6	13	6	14	7
Fractions and proportionality	13	6	15	4	11	4	8	9
Measurement, estimation, and number sense	9	4	10	5	9	4	6	3
Data representation, analysis, and probability	8	3	5	1	4	1	4	3
Geometry	4	2	4	2	4	1	4	3
Patterns, relations, and functions	1	-	4	1	4	2	2	1
Total	**44**	**19**	**45**	**19**	**45**	**18**	**38**	**26**

Table 3.11 Maximum Number of Science Score Points in Each Booklet by Reporting Category, Population 1

Reporting Category	Booklet							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Earth science	6	7	7	11	7	10	5	5
Life science	6	16	6	15	6	18	10	14
Physical science	4	12	4	9	4	10	6	12
Environmental issues and the nature of science	3	5	2	5	1	3	2	3
Total	**19**	**41**	**19**	**40**	**18**	**41**	**24**	**34**

Table 3.12 shows the rotation ratios (the number of times an item appears in a booklet) for items that belong to each cluster and the number of times that items from pairs of clusters appear together. The rotation ratios for the items in each cluster are shown in the diagonal elements of the matrix (only the diagonal and lower-triangular parts of the matrix are shown). For example, the Core cluster appears in all 8 booklets, so its rotation ratio is 8. For the Focus clusters the rotation ratio is 3 (except for cluster B, which has a rotation ratio of 4).

The nondiagonal elements of Table 3.12 give the number of times that a pair of clusters appears together in a booklet. If two clusters appear together, then the covariances between the items in them can be estimated (of course all the covariances between items within each cluster will be available). A dash indicates that the corresponding pair of clusters never appears in the same booklet. The first column of the matrix shows that the Core cluster appears at least once with each other cluster. Further, the matrix shows that few covariances between the items in different Breadth clusters, and between those in different Mathematics and Science Free-Response clusters, will be available. The design does, however, ensure that all covariances between items in Focus clusters, and most covariances between items in Focus and Free-Response clusters, will be available.

Table 3.12 Rotation Ratios for Item Clusters and Cluster Pairings, Population 1

A	8	*1. Core*
B	4 4	*2. Focus*
C	3 1 3	
D	3 1 1 3	
E	3 1 1 1 3	
F	3 1 1 1 1 3	
G	3 1 1 1 1 1 3	
H	3 1 1 1 1 1 1 3	
I	1 1 - - - - - - - 1	*3. Mathematics Breadth*
J	1 1 1 -1 - - - -1	
K	1 - - 1 1 - 1 - - - 1	
L	1 1 - - - 1 1 - - - - 1	
M	1 1 - 1 - - - 1 - - - - 1	
N	1 - 1 1 - 1 - - - - - - - - -- 1	*4. Science Breadth*
O	1 - - - 1 1 - 1 - - - - - - -- - 1	
P	1 - 1 - - - 1 1 - - - - - - - - 1	
Q	1 1 - - - - - - - 1 - - - -- - - - 1	
R	1 1 - - - - - - - 1 - - - -- - - - 1 1	
S	2 2 1 -1 - - - 1 1 - - - - - - 1 1 2	*5. Mathematics Free-Response*
T	2 1 1 1 2 -1 - -1 1 - - - - - - - 1 2	
U	2 1 - 1 1 1 2 - - - 1 1 - - - - - - - - 1 2	
V	2 2 - 1 - 1 1 1 - - - 1 1 - - - - - - - - 1 2	
W	2 1 1 2 - 1 - 1 - - - - 1 1 - - - - - - - 1 2	*6. Science Free-Response*
X	2 - 1 1 1 2 - 1 - - - - - - 1 1 - - - - - - - - 1 2	
Y	2 - 1 -1 1 1 2 - - - - -- - 1 1 - - - - - - - - 1 2	
Z	2 1 1 - - - 1 1 1 - - - -- - - 1 1 1 1 - - - - - 1 2	
	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z	**CLUSTER**

3.5 TIMSS POPULATION 2 TEST DESIGN

The test design for Population 2 is very similar to that for Population 1. As in the Population 1 design, all test items in Population 2 were grouped into 26 mutually exclusive item clusters, each identified with a letter of the alphabet (A through Z). There is a core cluster (cluster A) that appears in the second position in every booklet. Each cluster contains items in mathematics, science, or both. Each booklet comprises of up to seven item clusters. The booklets are divided into two parts and administered in two consecutive testing sessions. One cluster appears in all booklets and some clusters appear in all three, two, or only one booklet. Of the 26 clusters in Population 2, 8 take 12 minutes, 10 take 22 minutes, and 8 take 10 minutes. The design thus provides 396 unique testing minutes, 198 for science and 198 for mathematics.

In Population 2, it is convenient to regard the clusters as being of five types.

1. Core Cluster	One Core cluster comprising six mathematics and six science multiple-choice items was assigned to all booklets. It is labeled cluster A.
2. Focus Clusters	There were seven Focus clusters, each containing six mathematics and six science multiple-choice items. These Focus clusters are assigned to Booklets 1 through 7. They are called Focus clusters because each appears in at least three booklets, so that the items in them were answered by a relatively large fraction (three-eighths) of the student sample in each country–enough to permit accurate reporting of the item statistics. These clusters are labeled B through H.
3. Mathematics and Science Breadth Clusters	There were 10 Breadth clusters, each containing 11 minutes of mathematics and 11 minutes of science items. These clusters consist largely but not exclusively of multiple-choice items. They appear in only one booklet and consequently the number of students responding to each item was small. While the items in these clusters contribute to the breadth of content coverage of the tests, the accuracy of their item statistics would be relatively low. These clusters are labeled I through R.
4. Mathematics Free-Response Clusters	There were four Mathematics Free-Response clusters, each containing 10 minutes of short-answer and extended-response items. These clusters were each assigned to two booklets, so that item statistics of reasonable accuracy would be available. These items are labeled S through V.
5. Science Free-Response Clusters	There were four Science Free-Response clusters, each containing 10 minutes of short-answer and extended-response items. These clusters were each assigned to two booklets, so that item statistics of reasonable accuracy would be available. These clusters are labeled W through Z.

Tables 3.13 and 3.14 list the clusters and the number of items of each type allocated to each cluster for mathematics and science, respectively. The number of items in each cluster varies because the clusters have been designed on the basis of the estimated number of minutes that it would take a typical student to answer them, rather than on the total number of items.

Table 3.13 Number of Mathematics Items per Cluster, by Item Type, for Population 2

Cluster	Multiple-Choice	Short-Answer	Extended-Response	Total
A	6	-	-	6
B	6	-	-	6
C	6	-	-	6
D	6	-	-	6
E	6	-	-	6
F	6	-	-	6
G	6	-	-	6
H	6	-	-	6
I	7	2	-	9
J	7	2	-	9
K	7	2	-	9
L	9	1	-	10
M	7	2	-	9
N	7	2	-	9
O	7	2	-	9
P	9	1	-	10
Q	9	1	-	10
R	7	2	-	9
S	-	-	2	2
T	-	-	2	2
U	-	-	2	2
V	1	2	1	4
Total	**125**	**19**	**7**	**151**

Table 3.14 Number of Science Items per Cluster, by Item Type, for Population 2

Cluster	Multiple-Choice	Short-Answer	Extended-Response	Total
A	6	-	-	6
B	6	-	-	6
C	6	-	-	6
D	6	-	-	6
E	6	-	-	6
F	6	-	-	6
G	6	-	-	6
H	6	-	-	6
I	9	1	-	10
J	7	2	-	9
K	8	2	-	10
L	6	-	1	7
M	2	2	1	5
N	8	2	-	10
O	4	4	-	8
P	3	4	-	7
Q	5	3	-	8
R	2	2	1	5
W	-	-	2	2
X	-	-	2	2
Y	-	-	2	2
Z	-	-	2	2
Total	**102**	**22**	**11**	**135**

3.5.1 ORGANIZATION OF THE TEST BOOKLETS

As for Population 1, the Population 2 test design specifies eight booklets. While for Population 1 each booklet required 64 minutes for a student to complete, for Population 2 90 minutes were required. Except for booklet 8, each booklet was constructed from one 12-minute Core cluster (cluster A), three 12-minute Focus clusters, one 22-minute Breadth cluster, and two 10-minute Mathematics or Science Free-Response clusters. Table 3.15 shows the assignment of clusters to booklets, as well as the position of each cluster within the booklet.

Table 3.15 Assignment of Item Clusters to Population 2 Booklets[4]

Cluster Type	Cluster Label	Booklet 1	2	3	4	5	6	7	8
Core	A	2	2	2	2	2	2	2	2
(12 minutes)	B	1				5		3	1
	C	3	1				5		
Focus	D		3	1				5	
(12 minutes)	E	5		3	1				
	F		5		3	1			
	G			5		3	1		
	H				5		3	1	
Breadth	I	6							
(Mathematics and Science)	J		6						
(22 minutes)	K			6					
	L				6				
	M					6			
	N						6		
	O							6	
	P								6
	Q								3
	R								5
Mathematics Free-Response	S	4							
(10 minutes)	T	7		4					
	U			7		4			
	V					7		4	
Science Free-Response	W		4					7	
(10 minutes)	X		7		4				
	Y				7		4		
	Z						7		

[4] Numbers in the cells indicate the position of the cluster within the booklet.

The order of the clusters within the Population 2 booklets is shown in Table 3.16. Cluster A is the Core cluster and has been assigned to all booklets. The rotation design used to assign clusters B through H to booklets 1 through 7 allows the estimation of all item covariances for the items in clusters A through H.

Table 3.16 Ordering of Clusters Within Population 2 Booklets

Cluster Order	Booklet							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1st	B	C	D	E	F	G	H	B
2nd	A	A	A	A	A	A	A	A
3rd	C	D	E	F	G	H	B	Q
4th	S	W	T	X	U	Y	V	
BREAK								
5th	E	F	G	H	B	C	D	R
6th	I	J	K	L	M	N	O	P
7th	T	X	U	Y	V	Z	W	

Booklet 8 serves primarily to increase the content coverage of the tests. Apart from booklet 8 (which has three), each booklet has only one Breadth cluster, and each Breadth cluster appears in only one booklet. This means that covariances between items in different Breadth clusters cannot be directly estimated. For each item in each Breadth cluster, covariances can be directly estimated with half of the items in the Focus clusters and with all of the items in the Core cluster.

Similarly, the rotation of the Free-Response clusters restricts estimation of the covariances between items in different Free-Response clusters and between free-response and multiple-choice items. Most of the covariances between the items in the Free-Response and Focus clusters can be directly estimated, as can more than half of those between items in the mathematics Free-Response clusters. The same situation applies to science. Only a small number of covariances between items in the Mathematics and Science Free-Response clusters can be estimated.

The Population 2 test design has the following features.

- The Core cluster (cluster A) appears in the second position in all test booklets.
- The Focus clusters (clusters B through H) each appear in at least three booklets, each time in a different position. They are assigned to each of the first seven booklets following a Balanced Incomplete Block design. In booklets 1 through 7, each Focus cluster appears together once with each of the remaining Focus clusters.
- Each of the Focus clusters occurs once in the first, third, and fifth positions in test booklets 1 through 7.
- All test booklets contain mathematics and science items. Test booklets 1, 3, 5, and 7 have more mathematics items; booklets 2, 4, 6, and 8 have more science items.

- The test booklets are designed to be administered in two consecutive testing sessions with a 15-20-minute break in between. The first four clusters of items in the Population 2 test booklets were administered during the first testing session (46 minutes); after the break the remaining three clusters were administered (44 minutes).
- There are Free-Response clusters in Part 1 as well as in Part 2 of each test booklet (fourth and seventh cluster in each booklet).
- The design provides a total of 396 testing minutes, 198 for science and 198 for mathematics.

3.5.2 CONTENT OF THE TEST BOOKLETS

Test items were included in the Population 2 tests from 12 content areas in mathematics and 10 in science. Some of these content areas are merged, for the purpose of scaling and reporting, into six reporting categories for mathematics and five reporting categories for science. That is, it will be possible to characterize the TIMSS Population 2 with respect to six mathematics and five science achievement dimensions. The six mathematics reporting categories are:

- Fractions and number sense
- Geometry
- Algebra
- Data representation, analysis, and probability
- Measurement
- Proportionality.

The five science reporting categories are:

- Earth science
- Life science
- Physics
- Chemistry
- Environmental issues and the nature of science.

The TIMSS test blueprints (see Chapter 2) and the TIMSS curriculum frameworks provide more information on these reporting categories.

The Core and Focus clusters contain multiple-choice items only. The Breadth clusters include multiple-choice items and some short-answer items. Free-Response clusters consist almost exclusively of short-answer and extended-response items.

When items from both mathematics and science were included in a cluster, all the mathematics items appear in a contiguous sequence, as do all of the science items. In half of these clusters the mathematics items are presented first; in the other half, the science items. Within each sequence, items were placed in order of estimated difficulty.

Table 3.17 shows the number of items by type, and the associated maximum number of score points, for each of the content-based reporting categories for Population 2 mathematics. Table 3.18 provides the same information for Population 2 science.

Table 3.17 Number of Mathematics Items of Each Type, and Maximum Score Points, by Reporting Category, Population 2

Reporting Category	Multiple-Choice	Short-Answer	Extended-Response	Total Items	Score Points
Fractions and number sense	41	9	1	51	52
Geometry	22	1	-	23	23
Algebra	22	3	2	27	30
Data representation, analysis, and probability	19	1	1	21	23
Measurement	13	3	2	18	23
Proportionality	8	2	1	11	12
Total	**125**	**19**	**7**	**151**	**163**

Table 3.18 Number of Science Items of Each Type, and Maximum Score Points, by Reporting Category, Population 2

Reporting Category	Multiple-Choice	Short-Answer	Extended-Response	Total Items	Score Points
Earth science	17	3	2	22	25
Life science	31	5	4	40	46
Physics	28	9	3	40	44
Chemistry	15	3	1	19	21
Environmental issues and the nature of science	11	2	1	14	15
Total	**102**	**22**	**11**	**135**	**151**

Tables 3.19 and 3.20 show the number of items from each reporting category that are included in each of the eight test booklets for mathematics and science. Tables 3.21 and 3.22 show the maximum number of possible score points for each reporting category in each booklet.

Table 3.19 Number of Mathematics Items in Each Booklet by Reporting Category, Population 2

Reporting Category	Booklet							
	1	2	3	4	5	6	7	8
Fractions and number sense	11	10	11	10	10	11	11	14
Geometry	5	6	6	3	6	4	5	6
Algebra	8	5	6	8	4	6	6	9
Data representation, analysis, and probability	5	4	4	6	7	6	7	5
Measurement	5	5	6	4	6	4	4	3
Proportionality	3	3	4	3	6	2	4	4
Total	**37**	**33**	**37**	**34**	**39**	**33**	**37**	**41**

Table 3.20 Number of Science Items in Each Booklet by Reporting Category, Population 2

Reporting Category	Booklet							
	1	2	3	4	5	6	7	8
Earth science	7	7	6	6	5	5	10	7
Life science	9	11	11	13	7	9	8	6
Physics	10	9	12	11	9	10	11	13
Chemistry	2	7	2	3	5	7	4	4
Environmental issues and the nature of science	6	3	3	2	3	7	1	2
Total	**34**	**37**	**34**	**35**	**29**	**38**	**34**	**32**

Table 3.21 Maximum Number of Mathematics Score Points in Each Booklet by Reporting Category, Population 2

Reporting Category	Booklet							
	1	2	3	4	5	6	7	8
Fractions and number sense	11	10	12	10	11	11	11	14
Geometry	5	6	6	3	6	4	5	6
Algebra	11	5	8	8	4	6	6	9
Data representation, analysis, and probability	5	4	4	6	9	6	9	5
Measurement	7	5	9	4	9	4	4	3
Proportionality	4	3	5	3	6	2	4	4
Total	**43**	**33**	**44**	**34**	**45**	**33**	**39**	**41**

Table 3.22 Total Number of Science Score Points in Each Booklet by Reporting Category, Population 2

Reporting Category	Booklet							
	1	2	3	4	5	6	7	8
Earth science	7	9	6	6	5	5	13	7
Life science	9	13	11	15	9	9	8	8
Physics	10	9	12	14	9	12	11	14
Chemistry	2	7	2	3	5	9	4	4
Environmental issues and the nature of science	6	3	3	2	3	8	1	2
Total	**34**	**41**	**34**	**40**	**31**	**43**	**37**	**35**

Table 3.23 shows the rotation ratios for items that belong to each cluster and the number of times that items from pairs of clusters appear together. The rotation ratios for the items in each cluster are shown in the diagonal elements of the matrix in Table 3.23 (only the diagonal and lower-triangular parts of the matrix are shown). For example, the Core cluster appears in all 8 booklets, so its rotation ratio is 8. For the Focus clusters the rotation ratio is 3 (except for cluster B which has a rotation ratio of 4).

The nondiagonal elements of Table 3.23 give the number of times that a pair of clusters appears together in a booklet. If two clusters appear together, then the covariances between the items in them can be estimated (of course all the covariances between items within each cluster will be available). A dash indicates that the corresponding pair of clusters never appears in the same booklet. The first column of the matrix shows that the Core cluster appears at least once with each other cluster. Further, the matrix shows that few covariances between the items in different Breadth clusters, and between those in different mathematics and science Free-Response clusters, will be available. The design does, however, ensure that all covariances between items in Focus clusters and most covariances between items in Focus and Free-Response clusters, will be available.

Table 3.23 Rotation Ratios for Item Clusters and Cluster Pairings, Population 2

A	8	*1. Core*
B	4 4	*2. Focus*
C	3 1 3	
D	3 1 1 3	
E	3 1 1 1 3	
F	3 1 1 1 1 3	
G	3 1 1 1 1 1 3	
H	3 1 1 1 1 1 1 3	
I	1 1 1 -1 - - - 1	*3. Breadth*
J	1 - 1 1 - 1 - - -1	
K	1 - - 1 1 - 1 - - - 1	
L	1 - - -1 1 - 1 - - -1	
M	1 1 1 - - 1 1 - - - - - 1	
N	1 - - - - -1 1 - - - - -1	
O	1 1 - 1 - - - 1 - - - - - - - 1	
P	1 1 - - - - - - - - - - - - - -1	
Q	1 1 - - - - - - - - - - - - - -1 1	
R	1 1 - - - - - - - - - - - - - -1 1 1	
S	1 1 1 -1 - - - 1 - - - - - - - - - - 1	*4. Mathematics Free-Response*
T	2 1 1 1 2 - 1 - 1 - 1 - - - - - - - - 1 2	
U	2 1 - 1 1 1 2 - - - 1 - 1 - - - - - - - 1 2	
V	2 2 - 1 - 1 1 1 - - - - 1 - 1 - - - - - 1 2	
W	2 1 1 2 - 1 - 1 -1 - - - - 1 - - - - - 1 2	*5. Science Free-Response*
X	2 - 1 1 1 2 - 1 -1 -1 - - - - - - - - - 1 2	
Y	2 - 1 -1 1 1 2 - - - 1 -1 - - - - - - - - 1 2	
Z	1 - 1 - - - 1 1 - - - - - 1 - - - - - - - - - 1 1	
	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z	**CLUSTER**

3.6 TIMSS POPULATION 3 TEST DESIGN

The TIMSS design for Population 3 requires the assessment of the mathematical and scientific literacy of students in their final year of secondary schooling, and of the mathematics and physics proficiency of students within that population who are taking advanced courses in those fields. The test design therefore differs significantly from those in Populations 1 and 2.

Because the educational backgrounds of the general Population 3 differs from that of the students taking advanced mathematics and physics courses, the test design had to ensure that each group received appropriate test materials. To achieve this, the students in Population 3 in each country were stratified by educational background. Each student was dichotomously characterized as being in advanced mathematics courses (M) or not (O), and as being in physics courses (P) or not (O). This two-way classification yielded four mutually exclusive and exhaustive categories:

OO Students studying neither advanced mathematics nor physics

OP Students studying physics but not advanced mathematics

MO Students studying advanced mathematics but not physics

MP Students studying both advanced mathematics and physics.

Four types of test booklets were designed to target these student categories:

- Two literacy booklets (booklets 1A and 1B) containing mathematics and science literacy items, as well as items in reasoning and social utility
- Three physics booklets (booklets 2A, 2B and 2C) containing physics items only
- Three mathematics booklets (booklets 3A, 3B and 3C) containing advanced mathematics items only
- One mathematics/physics booklet (booklet 4) containing items in physics, advanced mathematics, and reasoning and social utility items.

The design of the TIMSS tests for Population 3 builds 12 mutually exclusive clusters of items and distributes these clusters among the four types of test booklets in a systematic fashion. The 12 clusters are labeled A through L. Each cluster could appear in more than one test booklet and, in a few cases, in different positions within the booklets. The items within a cluster always appear in the same order and position.

The four types of item clusters as classified by domain in the Population 3 tests are described below.

1. One Reasoning and Social Utility cluster containing 12 items (30 minutes of testing time), of which 6 are related to concepts in mathematics and 6 to science concepts. These items may be multiple-choice, short-answer, or extended-response. This cluster is labeled cluster A.

2. Three Mathematics and Science Literacy clusters, each containing 30 minutes of testing time. These clusters are labeled B, C and D. The Core Literacy cluster (cluster B) appears in booklets 1A and 1B, and the other two clusters each appear in one of the literacy booklets.

3. The four clusters with physics items are labeled E, F, G, and H. Cluster E is a Core physics cluster that contains 30 minutes of multiple-choice items. Cluster E is the first cluster to appear in each of the physics booklets (booklets 2A, 2B, and 2C), and the second in the advanced mathematics/physics booklet (booklet 4). The remaining clusters contain multiple-choice, short-answer and extended-response items and are rotated amongst the physics booklets, with each appearing in one booklet only.

4. There are four clusters with advanced mathematics items. These clusters are labeled I, J, K and L. Cluster I is a Core mathematics cluster that contains 30 minutes of multiple-choice items. Cluster I is the first cluster to appear in each of the advanced mathematics booklets (booklets 3A, 3B, and 3C), and is the third in the advanced mathematics/physics booklet (booklet 4). The remaining clusters contain multiple-choice, short-answer, and extended-response items and are rotated amongst the advanced mathematics booklets, with each appearing in one booklet only.

5. Another way of classifying the clusters is as either Core or Rotated clusters. Table 3.24 presents this classification of the clusters by domains tested.

Table 3.24 Classification of the Clusters by Content and Cluster Type, Population 3

Domain	Core Cluster	Rotated Cluster
Reasoning and Social Utility (RSU)	A	(none)
Mathematics and Science Literacy (MSL)	B	C, D
Physics (P)	E	F, G, H
Advanced Mathematics (MA)	I	J, K, L

Again, the number of items per cluster varies because the clusters have been designed on the basis of the estimated number of minutes it would take a typical student to answer them. All four Core clusters (A, B, E and I) and the Rotated mathematics and science literacy clusters (C and D) are each 30 minutes in length. Each of the Rotated clusters for the physics and mathematics students (F, G, H, J, K and L) are 60 minutes in length. The total testing time per cluster is shown in Table 3.25.

Table 3.25 Allocation of Testing Time to Item Clusters, Population 3

Cluster Label	Cluster Type	Time Allocated
A	Reasoning and Social Utility (RSU)	30
B	Literacy Content Core (MSL)	30
C	Literacy Content Rotated a (MSLa)	30
D	Literacy Content Rotated b (MSLb)	30
E	Physics Core (PC)	30
F	Physics Rotated a (PRa)	60
G	Physics Rotated b (PRb)	60
H	Physics Rotated c (PRc)	60
I	Advanced Mathematics Core (MAC)	30
J	Advanced Mathematics Rotated a (MARa)	60
K	Advanced Mathematics Rotated b (MARb)	60
L	Advanced Mathematics Rotated c (MARc)	60

3.6.1 ORGANIZATION OF THE TEST BOOKLETS

In Population 3, the design calls for nine booklets, each estimated to require 90 minutes to complete. Each booklet has either two or three clusters of items. Table 3.26 shows the assignment of clusters to booklets, and the position of the clusters within each booklet.

Table 3.26 Assignment of Item Clusters to Population 3 Booklets[5]

Cluster Type	Cluster Label	Booklet 1A	1B	2A	2B	2C	3A	3B	3C	4
RSU	A	1	2							1
MSL	B	2	1							
MSLa	C	3								
MSLb	D		3							
PC	E			1	1	1				2
PRa	F			2						
PRb	G				2					
PRc	H					2				
MAC	I						1	1	1	3
MARa	J						2			
MARb	K							2		
MARc	L								2	

[5] Number in cell indicates position of item cluster within the test booklet.

Table 3.27 summarizes the information in Table 3.26 by cluster order. It shows, for each booklet, the clusters assigned and the order in which they appear in the booklet.

Table 3.27 Ordering of Clusters Within Population 3 Booklets

	Booklet								
Cluster Order	**1A**	**1B**	**2A**	**2B**	**2C**	**3A**	**3B**	**3C**	**4**
1st	A	B	E	E	E	I	I	I	A
2nd	B	A	F	G	H	J	K	L	E
3rd	C	D	-	-	-	-	-	-	I

The design summarized in Tables 3.26 and 3.27 has the following features.

- Each test booklet comprises up to three item clusters, each of which can each appear in more than one booklet.
- Each of the mathematics and science literacy booklets (booklets 1A and 1B) contains the cluster with reasoning and social utility items (cluster A), the Core cluster for mathematics and science literacy (cluster B), and one of the mathematics and science literacy Rotated clusters (C or D).
- The physics booklets (booklets 2A, 2B, and 2C) contain the Core cluster for physics (cluster E), followed by one of the Rotated physics clusters (F, G, or H).
- The advanced mathematics booklets (booklets 3A, 3B and 3C) contain the Core cluster for advanced mathematics (cluster I), followed by one of the Rotated advanced mathematics clusters (clusters J, K, or L).
- The advanced mathematics/physics booklet (booklet 4) contains the reasoning and social utility cluster (cluster A), as well as the Core clusters for the physics and advanced mathematics items (clusters E and I).
- The expected completion time for reasoning and social utility (cluster A), the mathematics and science literacy (B through D), and the Core physics (E) and Core advanced mathematics (I) clusters is 30 minutes each. The expected completion time for each of the physics (F, G, H) and advanced mathematics (J, K, L) Rotated clusters is 60 minutes. As a result of the assignment of clusters to booklets, the expected completion time for each of the booklets is 90 minutes.
- Each booklet was administered in one 90-minute session with no break.

3.6.2 ASSIGNMENT OF BOOKLETS TO STUDENTS

In Populations 1 and 2, all of the test booklets were rotated through all students in the sample. In Population 3, it was necessary to specify a separate rotation scheme for each student classification: OO, MO, OP, and MP.

The booklet that a student was eligible to receive depended upon the classification of that student. Table 2.28 shows the booklets to be rotated for students of each type.

Table 3.28 Assignment of Test Booklets to Students, Population 3

	Student Grouping			
Test Booklet	**OO**	**OP**	**MO**	**MP**
1A	X	X	X	X
1B	X	X	X	X
2A		X		X
2B		X		X
2C		X		X
3A			X	X
3B			X	X
3C			X	X
4				X

The rotation ratios for items in each cluster for each student classification are shown in Table 3.29. These ratios give some indication of the relative precision of statistics that are expected at the item level. The higher the rotation ratio, the smaller the proportion of the sampled students that will respond to the item. While this figure does not address the absolute precision of item statistics, it does make it clear that items allocated to the Core clusters (RSU, MSL, MAC and PC) are likely to have more precise statistics than those assigned to the rotated clusters.

Table 3.29 Rotation Ratios for Items in Each Cluster, Assuming Uniform Rotation Within Domains, Population 3

	Student Classification			
Clusters	**OO**	**OP**	**MO**	**MP**
RSU	1	2.5	2.5	3
MSL	1	2.5	2.5	4.5
MSLa	2	5	5	9
MSLb	2	5	5	9
MAC	-	-	1.67	2.25
MARa	-	-	5	9
MARb	-	-	5	9
MARc	-	-	5	9
PC	-	1.67	-	2.25
PRa	-	5	-	9
PRb	-	5	-	9
PRc	-	5	-	9

3.6.3 CONTENT OF THE TEST BOOKLETS

Items were included in the Population 3 tests to cover several content areas in mathematics and science literacy, advanced mathematics, and physics. For the purpose of scaling and reporting, some of these content areas were merged into the reporting categories below.

The mathematics and science literacy reporting categories are:

- Mathematics literacy
- Science literacy
- Reasoning and social utility.

The physics reporting categories are:

- Mechanics
- Electricity and magnetism
- Heat
- Wave phenomena
- Particle, quantum, astrophysics and relativity.

The advanced mathematics reporting categories are:

- Numbers and equations
- Analysis (calculus)
- Geometry

- Probability and statistics
- Validation and structure.

The TIMSS test blueprints (see Chapter 2) and the curriculum frameworks describe in more detail the content areas upon which the reporting categories are based.

The Core literacy, advanced mathematics, and physics clusters contain multiple-choice items only. The Rotated clusters were composed of multiple-choice, short-answer, and extended-response items. In the literacy and reasoning and social utility clusters, the items were grouped by subject area (science and mathematics) within the cluster.

Tables 3.30 through 3.35 summarize the test design for Population 3. Tables 3.30 through 3.32 present for each scale, the number of items in each booklet for each reporting category. Tables 3.33 through 3.35 present for each scale, the maximum number of possible score points in each booklet for each reporting category.

Table 3.30 Number of Test Items per Booklet by Reporting Category, Population 3 Mathematics and Science Literacy

Reporting Category	Booklet								
	1A	1B	2A	2B	2C	3A	3B	3C	4
Mathematics literacy	26	25	-	-	-	-	-	-	-
Science literacy	21	18	-	-	-	-	-	-	-
Reasoning and social utility	12	12	-	-	-	-	-	-	12
Total	**59**	**55**	**-**	**-**	**-**	**-**	**-**	**-**	**12**

Table 3.31 Number of Test Items per Booklet by Reporting Category, Population 3 Physics

Reporting Category	Booklet								
	1A	1B	2A	2B	2C	3A	3B	3C	4
Mechanics	-	-	6	8	6	-	-	-	2
Electricity and magnetism	-	-	7	7	8	-	-	-	3
Heat	-	-	3	4	4	-	-	-	1
Wave phenomena	-	-	5	4	5	-	-	-	2
Particle, quantum, astrophysics, and relativity	-	-	6	6	6	-	-	-	2
Total	**-**	**-**	**27**	**29**	**29**	**-**	**-**	**-**	**10**

Table 3.32 Number of Test Items per Booklet by Reporting Category, Population 3 Advanced Mathematics

Reporting Category	Booklet								
	1A	1B	2A	2B	2C	3A	3B	3C	4
Numbers, equations, and functions	-	-	-	-	-	7	8	8	3
Analysis (calculus)	-	-	-	-	-	7	7	5	2
Geometry	-	-	-	-	-	10	10	9	3
Probability and statistics	-	-	-	-	-	3	2	4	1
Validation and structure	-	-	-	-	-	2	1	2	1
Total	**-**	**-**	**-**	**-**	**-**	**29**	**28**	**28**	**10**

Table 3.33 Number of Score Points per Booklet by Reporting Category, Population 3 Mathematics and Science Literacy

Reporting Category	Booklet								
	1A	1B	2A	2B	2C	3A	3B	3C	4
Mathematics literacy	28	30	-	-	-	-	-	-	-
Science literacy	24	21	-	-	-	-	-	-	-
Reasoning and social utility	21	21	-	-	-	-	-	-	21
Total	**73**	**72**	**-**	**-**	**-**	**-**	**-**	**-**	**21**

Table 3.34 Number of Score Points per Booklet by Reporting Category, Population 3 Physics

Reporting Category	Booklet								
	1A	1B	2A	2B	2C	3A	3B	3C	4
Mechanics	-	-	7	10	6	-	-	-	2
Electricity and magnetism	-	-	9	8	10	-	-	-	3
Heat	-	-	4	5	5	-	-	-	1
Wave phenomena	-	-	5	4	7	-	-	-	2
Particle, quantum, astrophysics and relativity	-	-	7	7	7	-	-	-	2
Total	**-**	**-**	**32**	**34**	**35**	**-**	**-**	**-**	**10**

Table 3.35 Number of Score Points per Booklet by Reporting Category, Population 3 Advanced Mathematics

Reporting Category	Booklet								
	1A	1B	2A	2B	2C	3A	3B	3C	4
Numbers, equations, and functions	-	-	-	-	-	7	11	10	3
Analysis (calculus)	-	-	-	-	-	9	9	5	2
Geometry	-	-	-	-	-	12	12	11	3
Probability and statistics	-	-	-	-	-	3	2	5	1
Validation and structure	-	-	-	-	-	3	1	2	1
Total	**-**	**-**	**-**	**-**	**-**	**34**	**35**	**33**	**10**

Tables 3.36 through 3.38 present the number of items in the Population 3 item pool, organized by item type and by reporting category. It also presents the maximum number of score points in each of the Population 3 reporting categories.

Table 3.36 Number of Test Items of Each Type, and Maximum Score Points, by Reporting Category, Population 3 Mathematics and Science Literacy

Reporting Category	Item Type				
	Multiple -Choice	Short-Answer	Extended-Response	Number of Items	Score Points
Mathematics literacy	31	7	-	38	42
Science literacy	16	7	3	26	43
Reasoning and social utility	5	3	4	12	21
Total	**52**	**17**	**7**	**76**	**106**

Table 3.37 Number of Test Items of Each Type, and Maximum Score Points, by Reporting Category, Population 3 Physics

Reporting Category	Item Type				
	Multiple-Choice	Short-Answer	Extended-Response	Number of Items	Score Points
Mechanics	11	4	1	16	19
Electricity and magnetism	10	3	3	16	21
Heat	6	3	-	9	12
Wave phenomena	6	3	1	10	12
Particle, quantum, astrophysics, and relativity	9	2	3	14	17
Total	**42**	**15**	**8**	**65**	**81**

Table 3.38 Number of Test Items of Each Type and Score Points, by Reporting Category, Population 3, Advanced Mathematics

Reporting Category	Item Type				
	Multiple-Choice	Short-Answer	Extended-Response	Number of Items	Score Points
Numbers, equations, and functions	13	2	2	17	22
Analysis (calculus)	12	2	1	15	19
Geometry	15	4	4	23	29
Probability and statistics	5	2	-	7	8
Validation and structure	2	-	1	3	4
Total	**47**	**10**	**8**	**65**	**82**

REFERENCES

Beaton, A.E. (1987). *Implementing the New Design: The NAEP 1983-84 Technical Report.* Report No. 15-TR-2. Princeton, NJ: Educational Testing Service.

Lord, F.M. (1962). Estimating Norms By Item-sampling. *Educational and Psychological Measurement,* 22(2), 259-267.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. and Sheehan, K. (1987). "Marginal Estimation Procedures" in A.E. Beaton (ed.), *Implementing the New Design: The NAEP 1983-84 Technical Report.* Report No: 15-TR-2. Princeton, NJ: Educational Testing Service.

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.

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

Wright, B. and Stone, M. (1979). *Best Test Design*. Chicago: MESA Press, University of Chicago.

Foy, P., Rust, K., and Schleicher, A. (1996) "Sample Design" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

4. Sample Design

Pierre Foy
Keith Rust
Andreas Schleicher

4.1 OVERVIEW

This chapter describes the procedures developed to ensure that the student populations that were the focus of the study were properly sampled in each participating country. To be acceptable for TIMSS, national sample designs had to result in probability samples which give accurate weighted estimates of population parameters, and for which estimates of sampling variance could be computed. An effort was made in designing TIMSS to strike a balance between the analytical requirements and operational constraints, while keeping the survey design simple enough for all participants to implement it. The selection of valid and efficient samples was crucial to the success of the project. The accuracy of the survey results are dependent on the quality of the sampling information available at the design stage, and particularly on the implementation of the sampling procedures.

The National Research Coordinators (NRCs) were aware that in a study as ambitious as TIMSS, the sample design and sampling procedures would be complex, and that the gathering of the required information about the national education systems would place considerable demands on resources and expertise. At the same time, those directing and coordinating the project realized that the national centers had only limited numbers of qualified sampling personnel. Simplifying the sampling procedures to the extent possible, especially the sample selection within schools, was thus a major consideration in developing

the sample design. Sometimes simplicity and practicality had to be given a higher priority than optimizing the sample design in terms of precision and cost.

NRCs were allowed to adapt the sample design for their educational system, using more sampling information and more sophisticated sample designs and procedures than the base design provided. However, these solutions had to be approved and monitored by the international project management (the International Coordinating Center at the University of British Columbia, Vancouver, until August 1993, and the International Study Center at Boston College thereafter).

The international project management provided manuals and expert advice to help NRCs to adapt the TIMSS sample design to their national system, and to guide them through the phases of sampling. The *Sampling Plan* (TIMSS, 1992) provided an overview of the sample design and described the survey design options offered. The *Sampling Manual* (TIMSS, 1994a) described how to implement the sampling plan and offered advice on initial planning, working within constraints, establishing appropriate sample selection procedures, and fieldwork. The *Survey Operations Manuals* (TIMSS, 1994d, 1994e) and *School Coordinator Manuals* (1994b, 1994c) provided information on sample selection and execution within schools, the assignment of rotated test instruments to selected students, and administration and monitoring procedures used to identify and track respondents and nonrespondents. NRCs also received software designed to automate the sometimes complex within-school sampling procedures.

NRCs also had several sources of expert support. Statistics Canada, in consultation with the TIMSS sampling referee and the TIMSS Technical Advisory Committee (TAC), reviewed and approved the national sampling plans, sampling data, and sampling frames, and the sample execution. In addition, Statistics Canada provided advice and support to NRCs at all stages of the sampling process.

4.2 TARGET POPULATIONS AND EXCLUSIONS

In IEA studies, the target population for all countries is known as the *International Desired Population*. TIMSS chose to study student achievement in three such populations in each country. The international desired populations for TIMSS were as follows:

- **Population 1.** All students enrolled in the two adjacent grades that contain the largest proportion of 9-year-olds at the time of testing.
- **Population 2.** All students enrolled in the two adjacent grades that contain the largest proportion of 13-year-olds at the time of testing.
- **Population 3.** Students enrolled in their final year of secondary education. Population 3 had two optional subpopulations:
    - — Students taking advanced courses in mathematics
    - — Students taking advanced courses in physics.

4.2.1 POPULATIONS 1 AND 2

In defining populations for international comparisons of student achievement it is usually necessary to choose between age and grade level as the basis of comparison. An age-based definition focuses on a specific age cohort, for example all 13-year-old students in an education system. A grade-based definition focuses on a specific grade, for example the eighth grade in an education system, counting from the beginning of primary schooling. Since TIMSS is mainly a survey of mathematics and science instruction, with the classrooms functioning as units of analysis as well as sampling units, a grade-based definition was chosen. It was difficult, however, to identify internationally comparable grades, for lack of standard international grade definitions. It was therefore decided to identify the target grades on the basis of an age cohort.

The Population 1 and Population 2 target populations are thus defined as the two adjacent grades that will maximize coverage of a specific age cohort (9-year-olds for Population 1, and 13-year-olds for Population 2). Two adjacent grades were chosen to ensure extensive coverage of the age cohort for most countries–thereby increasing the likelihood of producing useful age-based comparisons also. Furthermore, two grades allow the measurement of growth between grades.

4.2.2 POPULATION 3

The intention in surveying Population 3 was to try to measure what might be considered the "yield" of the elementary and secondary education systems of a country with regard to mathematics and science. Thus the definition of the population is student-oriented; it is the body of students who are in *their* last year of school. For many students, this does not represent the highest level of education, especially mathematics and science education, available in the country.

For each secondary-education track in the country, the final grade of the track was identified as being part of Population 3. This allowed substantial coverage of students in their final year of schooling. For example, grade 10 might be the final year of a vocational program, and grade 12 the final year of an academic program. Both of these grade/track combinations are considered to be part of Population 3 (but grade 10 in the academic track is not).

There are two further difficulties in defining the international desired population for Population 3. The first is that many students drop out before the final year of any track. This is addressed in the TIMSS Population 3 assessment by the calculation of a Secondary Education Coverage Index which quantifies the proportion of the general population that reaches the final year. The Secondary Education Coverage Index (SECI) was defined as follows:

$$SECI = \frac{5 * Total\ Enrollment\ in\ Population\ 3\ in\ 1995}{Total\ National\ Population\ Aged\ 15-19\ in\ 1995}$$

This definition reflected the fact that Population 3 is likely to be almost entirely a subset of the population of 15- to 19-year-olds, and that, by age 19, someone who has never

belonged to Population 3 during any of the five most recent years is very unlikely to ever belong to Population 3. The SECI represents a kind of moving average measure of the proportion of the general population that undertakes the final year of a track of the secondary education system.

The second issue is that some students repeat the final year of a track, or take the final year in more than one of the tracks at two different times. That is, some students who are in the final year of a track are not in fact completing their secondary education that year. At the time of the TIMSS testing, these students would generally not have been aware (or at least certain) whether this was to be their final year. If this occurs within a country to any great extent, sampling students from the final grade may bias the estimate of the educational "yield." On the one hand, students who in fact are not completing their education still have the potential to gain further knowledge in additional years of schooling, and thus will not have attained their full yield at the time of the TIMSS assessment. On the other hand, and of more serious concern, the presence both of students who are repeating the final track, and of those who will repeat that track can contribute a substantial downward bias to the estimated achievement of the population. Repeating students are represented twice in the population, and are likely to be lower-achieving on average than those who do not repeat. The only practical way for TIMSS to deal with this problem was to exclude students who were repeating the final year. Thus Population 3 is formally defined as those students taking the final year of one track of the secondary system for the first time.

The International Study Center tried to maximize standardization across countries for the definition of Population 3. However, the precise definitions of the mathematics and physics subpopulations was necessarily a consultative process. Each country identified the group of students that it wished to compare internationally, based on a consideration of the general contents of the tests and practical considerations in sampling and administration. The analysis of Population 3 will include for each country a measure of the proportion of the total test population who were included in the advanced mathematics subpopulation, and the proportion who were included in the physics subpopulation.

The interest in measuring mathematics and science literacy levels extended to the whole of Population 3, not just the nonspecialist students. This means that the comparability of countries with regard to the literacy assessment is not affected by how the countries chose to define their mathematics and physics subpopulations. It also means that the sample design for Population 3 had to ensure that a representative sample of the advanced course-taking students took the literacy assessment, in addition to those taking the specialist tests.

4.2.3 School and Within-Sample Exclusions

TIMSS expected all participating countries to define their national desired populations to correspond as closely as possible to its definition of the international desired populations. However, sometimes NRCs had to restrict their coverage. For example, some countries had to restrict geographical coverage by excluding remote regions; or by excluding a segment of its education system. The international reports will document any deviations from the international definition of the TIMSS target populations. Significant differences in

terms of number of students excluded would mean that the survey results will be deemed not representative of the whole national school system.

Using their national desired populations as a basis, participating countries had to operationally define their populations for sampling purposes. This operational definition, known in IEA terminology as the *National Defined Population*, is essentially the sampling frame from which the first stage of sampling takes place. The national defined populations could be subsets of the national desired populations. All schools and students from the former excluded from the latter are referred to as excluded populations.

TIMSS participants were expected to keep such exclusions to no more than 10% of the national desired populations. Exclusions could occur at the school level, within schools, or both. Because national desired populations were restricted to schools that contain the required grades, schools not containing any of the target grades were not considered as excluded. In general, practical reasons were invoked for excluding schools or students, such as increased survey costs, increased complexity in the sample design, and difficult test conditions. The size of the excluded populations were documented and serve as an index of the coverage and representativeness of the selected samples.

Participants could exclude schools from the sampling frame for the following reasons:

- They are in geographically remote regions
- They are of extremely small size
- They offer a curriculum, or school structure, that is different from the mainstream educational system(s)
- They provide instruction only to students in the exclusion categories defined under "within-school exclusions."

Within-school exclusions were limited to students who, because of some disability, were unable to take the TIMSS tests. TIMSS participants were asked to define anticipated within-school exclusions. Because these definitions can vary internationally, they were also asked to follow certain rules, adapted to their jurisdictions. In addition, they were to estimate the size of such exclusions so that their compliance with the 10% rule could be gauged.

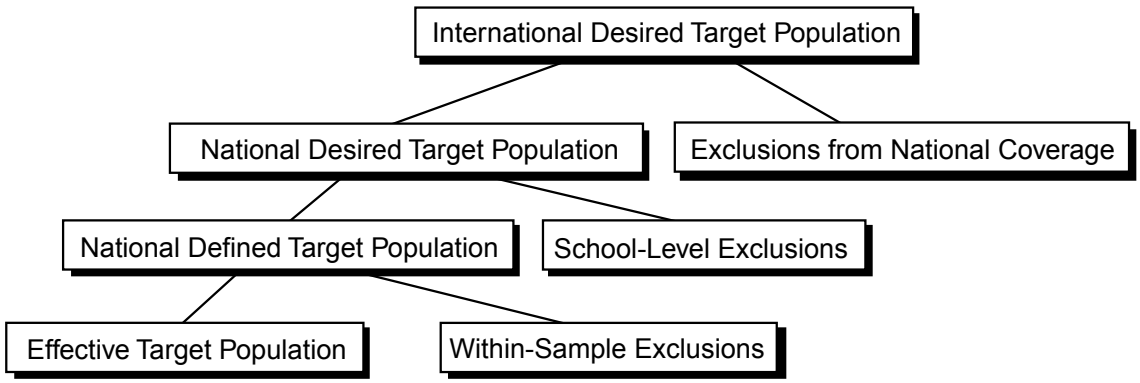
The general TIMSS rules for defining within-school exclusions are the following.

- **Educable mentally disabled students.** These are students who are considered, in the professional opinion of the school principal or other qualified staff members, to be educable mentally disabled, or who have been so diagnosed in psychological tests. This includes students who are emotionally or mentally unable to follow even the general instructions of the TIMSS test. It does not include students who merely exhibit poor academic performance or discipline problems.
- **Functionally disabled students.** These are students who are permanently physically disabled in such a way that they cannot perform in the TIMSS tests. Functionally disabled students who can perform should be included in the testing.
- **Non-native-language speakers.** These are students who cannot read or speak the language of the test and so could not overcome the language barrier of testing.

- **Non-native-language speakers.** These are students who cannot read or speak the language of the test and so could not overcome the language barrier of testing. Typically, a student who has received less than one year of instruction in the language of the test should be excluded, but this definition should be adapted in different countries.

The stated objective in TIMSS was that the effective population, the population actually sampled by TIMSS, be as close as possible to the international desired population. Figure 4.1 illustrates the relationship between the desired populations and the excluded populations. Any exclusion of eligible students from the international desired population had to be accounted for. This applies to school-level exclusions as well as within-sample exclusions.

Figure 4.1 Relationship Between the Desired Populations and Exclusions



4.3 SAMPLE DESIGN

The basic sample design proposed for TIMSS is generally referred to as a two-stage stratified cluster sample design. The first stage consists of a sample of schools[1], which may be stratified; the second stage consists of samples of classrooms from each eligible target grade in sampled schools. In some countries a third stage was added, in which students were sampled within classrooms. This design lends itself to the many analytical requirements of TIMSS. Survey estimates were required for students, teachers, classrooms, and schools.

4.3.1 UNITS OF ANALYSIS AND SAMPLING UNITS

The TIMSS analytical focus is both on the cumulative learning of students and on instructional characteristics affecting learning. The sample design, therefore, had to address both the measurement of explanatory characteristics thought to influence cumulative learning and the measurement of specific characteristics of the instructional settings. The first focus included characteristics of system organization, school organization and differentiation, national cross-grade curriculum specifications, resource allocations, national goals, and the like. The second focus included the measurement of teacher characteristics, classroom composition, teaching practices, implemented curriculum, and measurements of

[1] In some very large countries, it was necessary to include an extra preliminary stage, where school districts were sampled first, and then schools.

and students would all be potential units of analysis. They therefore had to be considered as sampling units in the sample design in order to meet specific requirements for data quality and sampling precision at all levels.

Although in the second sampling stage the sampling units were intact classrooms, the ultimate sampling elements were students, and so it was important that each student from the target grades be a member of one, and only one, of the classes in a school from which the sampled classes would be selected. Ideally, from a sampling perspective, the student should belong to the same class for both mathematics and science instruction. In most education systems, the mathematics class coincided with a student homeroom or science class, especially in Population 1. However, in some systems, mathematics and science classes did not coincide; students formed different groups for mathematics and for science instruction. In that case, participating countries were asked to define the classrooms on the basis of mathematics instruction. If not all students in the national desired population belonged to a mathematics class, then an alternative definition of the classroom was required for ensuring that the nonmathematics students had an opportunity to be selected.

The analytical objectives for Population 3 focused on the achievement of students in their final year of secondary schooling, rather than on the instructional context. In fact, there was no teacher questionnaire for Population 3, which meant that classrooms need not be a sampling unit. In practical terms, however, many education systems define classrooms by curriculum tracks. This made classrooms a useful sampling unit in those systems, especially when separate samples were selected for the advanced students. In education systems where the advanced course-taking students were not conveniently clustered in classrooms, student samples were selected at random within selected schools, using specified procedures.

4.3.2 Sampling Precision and Sample Size

Sample sizes for TIMSS had to be specified so as to meet the analytic requirements of the study. Since students were the principal units of analysis, the emphasis for data reliability was placed on the ability to produce reliable estimates of student characteristics. The TIMSS standard for sampling precision requires that all population samples have an effective sample size of at least 400 students for the main criterion variables. In other words, all population samples should yield sampling errors that are no greater than those that would be obtained from a simple random sample of 400 students.

Furthermore, since TIMSS planned to conduct analyses at the school and classroom levels, at least 150 schools were to be selected per target population. A sample of 150 schools yields 95% confidence limits for school- and classroom-level mean estimates that are precise to within ± 16% of their standard deviations. To ensure sufficient sample precision for these units of analysis, some participants had to sample more schools than they would have selected otherwise.

An effective sample size of 400 students results in the following approximate 95% confidence limits for sample estimates of population means, percentages, and correlation coefficients.

- Means: $m \pm 0.1s$ (where m is the mean estimate and s is the estimated standard deviation for students)
- Percentages: $p \pm 5.0\%$ (where p is a percentage estimate)
- Correlations: $r \pm 0.1$ (where r is a correlation estimate).

Multistage cluster sample designs are generally affected by what is called the clustering effect. A classroom as a sampling unit constitutes a cluster of students who tend to be more like each other than like other members of the population. The *intraclass correlation* is a measure of this within-class similarity. Sampling 30 students from a single classroom, when the intraclass correlation is positive, will yield less information than a random sample of 30 students spread across all classrooms in a school. Such sample designs are less efficient, in terms of sampling precision, than a simple random sample of the same size. This clustering effect was a factor to be considered in determining the overall sample size for TIMSS.

The magnitude of the clustering effect is determined by the size of the cluster (classroom) and the size of the intraclass correlation. For TIMSS the intraclass correlation for each country was estimated from past studies or national assessments. In the absence of these sources, an intraclass correlation of 0.3 was assumed.

To allow the planning of sample sizes, each participant had to specify a cluster size, known as the minimum cluster size for that country. Since most participants chose to test intact classrooms, the minimum cluster size was in fact the average classroom size. For participants who chose to subsample students from selected classrooms, the minimum cluster size was the number of students subsampled per classroom. The specification of the minimum cluster size not only affected the number of schools to sample, but also affected how small schools and small classrooms would be treated.

Sample-design tables were produced and included in the *Sampling Manual* (TIMSS, 1994a) (see Table 4.1 for an example). These tables illustrated the number of schools to sample for a range of intraclass correlations and minimum cluster size values. TIMSS participants could refer to these tables to determine how many schools they should sample given their intraclass correlation and minimum cluster size. A participant whose intraclass correlation was expected to be 0.6 and whose average classroom size was 30 would need to sample a minimum of 186 schools. Whenever the estimated number of schools to sample fell below 150, participants were asked to sample at least 150 schools.

The sample-design tables could be used also to determine sample sizes for more complex designs. For example, a stratum of small schools could be constructed where a smaller minimum cluster size could be specified, thereby avoiding the administrative complexity of defining pseudo-schools. (See section 4.4.1 Small Schools).

Table 4.1 Sample-Design Table, Populations 1 and 2

95% Confidence Limits For Means ± 0.1s/Percentages ± 5.0%										
Minimum Cluster Size		Intraclass Correlation								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	a	94	118	142	166	190	214	238	262	286
	n1	470	590	710	830	950	1,070	1,190	1,310	1,430
	n2	470	590	710	830	950	1,070	1,190	1,310	1,430
10	a	62	89	116	143	170	197	224	251	278
	n1	620	890	1,160	1,430	1,700	1,970	2,240	2,510	2,780
	n2	620	890	1,160	1,430	1,700	1,970	2,240	2,510	2,780
15	a	52	80	108	136	164	192	220	248	276
	n1	780	1,200	1,620	2,040	2,460	2,880	3,300	3,720	4,140
	n2	780	1,200	1,620	2,040	2,460	2,880	3,300	3,720	4,140
20	a	46	75	103	132	160	189	217	246	274
	n1	920	1,500	2,060	2,640	3,200	3,780	4,340	4,920	5,480
	n2	920	1,500	2,060	2,640	3,200	3,780	4,340	4,920	5,480
25	a	43	72	101	130	158	187	216	245	274
	n1	1,075	1,800	2,525	3,250	3,950	4,675	5,400	6,125	6,850
	n2	1,075	1,800	2,525	3,250	3,950	4,675	5,400	6,125	6,850
30	a	41	70	99	128	157	186	215	244	273
	n1	1,230	2,100	2,970	3,840	4,710	5,580	6,450	7,320	8,190
	n2	1,230	2,100	2,970	3,840	4,710	5,580	6,450	7,320	8,190
35	a	40	69	98	127	156	185	214	244	273
	n1	1,400	2,415	3,430	4,445	5,460	6,475	7,490	8,540	9,555
	n2	1,400	2,415	3,430	4,445	5,460	6,475	7,490	8,540	9,555
40	a	38	68	97	126	155	185	214	243	272
	n1	1,520	2,720	3,880	5,040	6,200	7,400	8,560	9,720	10,880
	n2	1,520	2,720	3,880	5,040	6,200	7,400	8,560	9,720	10,880

a= number of sampled schools

n1 = number of sampled students in upper grade

n2 = number of sampled students in lower grade

4.3.3 STRATIFICATION

Stratification is the grouping of schools according to some attribute or variable. It is generally used for the following reasons:

- To improve the efficiency of the sample design, thereby making survey estimates more reliable
- To apply different sample designs, or disproportionate sample-size allocations, to specific groups of schools (such as those within certain states or provinces)
- To ensure adequate representation in the sample of specific groups from the target population.

Examples of stratification variables for school samples are geography (such as states or provinces, school type (such as public and private schools), and level of urbanization (such as rural and urban). Stratification variables in the TIMSS sample design could be used explicitly, implicitly, or both.

Explicit stratification consists of building separate school lists, or sampling frames, according to the stratification variables under consideration. If, for example, geographic regions were an explicit stratification variable, then separate school sampling frames would be constructed for each region. Possibly different sample designs, or different sampling fractions, would then be applied to each school-sampling frame to select the sample of schools. In practice, the major reason for considering explicit stratification in the context of TIMSS was disproportionate allocation of the school sample to the strata. For example, the same number of schools might have been required from each stratum, regardless of the relative size of each stratum.

Implicit stratification makes use of a single school-sampling frame, but sorts the schools in this frame by a set of implicit stratification variables. This type of stratification is a simple way of ensuring proportional sample allocation without the complexity of explicit stratification. It can also improve the reliability of survey estimates, provided the implicit stratification variables are related to school mean student achievement in mathematics and science.

4.4 FIRST SAMPLING STAGE

The sample-selection method proposed for first-stage sampling in TIMSS makes use of a systematic probability-proportional-to-size (PPS) technique. In order to use this method it is necessary to have some measure of size (MOS) of the sampling units. Ideally this should be the number of sampling elements within the unit (e.g. number of students in the target grades in the school). If this is unavailable, some other, highly correlated measure, such as total school enrollment, may be used.

The schools in each explicit stratum are listed in order of the implicit stratification variables, together with the MOS for each school. They are further sorted by MOS within implicit stratification variable. The measures of size are accumulated from school to school, and the running total (the cumulative MOS) is listed next to each school (see Table 4.2). The total cumulative MOS is a measure of the size of the population of sampling elements.

Dividing the total cumulative MOS by the number of schools to be sampled gives the *sampling interval*.

The first school is sampled by choosing a random number in the range between 1 and the sampling interval. The school whose cumulative MOS contains the random number is the sampled school. By adding the sampling interval to that first random number, a second school is identified. This process of consistently adding the sampling interval to the previous selection number results in a PPS sample of the required size.

If an implicit stratification is in effect, then the resulting school sample will be allocated proportionately to the sizes of the implicit strata. Furthermore, if the implicit stratification variables used act to reduce sampling variance, then this sample selection method will reap that benefit, resulting in more reliable estimates than would otherwise be achieved.

Of the many benefits of this sample-selection method, the main reasons for its use in TIMSS, are that it is easy to implement, and it is easy to verify that it was implemented properly. The latter is critical since one of TIMSS' major objectives was to ensure that a sound sampling methodology be used.

Table 4.2 illustrates the PPS systematic sampling method applied to a fictitious sampling frame. The first three sampled schools are shown, as well as their preselected replacement schools should the originally selected schools not participate (see Section 4.4.3).

Table 4.2 Application of the PPS Systematic Sampling Method

Total MOS: 392,154 Sampling Interval: 23,614.3600

School Sample: 150 Random Start: 1,135.1551

School Code	School MOS	Cumulative MOS	Sample
917740	532	532	
875870	517	1049	
924942	487	1536	√
893204	461	1997	R1
952774	459	2456	R2
806290	437	2893	
161758	406	3299	
357056	385	3684	
997650	350	4034	√
778732	341	4375	R1
216873	328	4703	R2
336426	311	5014	
97015	299	5313	
486237	275	5588	
221573	266	5854	
696152	247	6101	
645538	215	6316	
540649	195	6511	√
330383	174	6685	R1
914017	152	6837	R2
76874	133	6970	
406509	121	7091	
66513	107	7198	
429291	103	7301	
88501	97	7398	

√ = Sampled School

R1, R2 = Replacement Schools

4.4.1 SMALL SCHOOLS

Small schools tend to be problematic in PPS samples because students sampled from these schools get disproportionately large sampling weights, and when the school size falls below the minimum cluster size, they reduce the overall student sample size. A school was deemed to be small for TIMSS' purposes if it could not yield an adequate sample of students per grade, as specified by the minimum cluster size. For example, if the minimum cluster size was set at 20, then a school with fewer than 20 students in each target grade was considered a small school.

In TIMSS, small schools were handled either through explicit stratification or through the use of pseudo-schools. In the first case, an explicit stratum of small schools was created for which a smaller number of students were required. The second approach consisted of creating clusters of small schools, called pseudo-schools, that would be sampled as a single unit. Any sampled cluster, or pseudo-school, would then be able to provide the required number of students.

The construction of pseudo-schools complicates data collection. Therefore, they were used only when absolutely necessary. In TIMSS, pseudo-schools were required whenever student enrollment in small schools exceeded 5% of total student enrollment. Also, participants who proposed sample designs with suitable explicit stratification for small schools were not required to construct pseudo-schools.

4.4.2 OPTIONAL PRELIMINARY SAMPLING STAGE

Some very large countries chose to introduce a preliminary sampling stage before sampling schools. This consisted of a PPS sample of geographic regions. A sample of schools was then be selected from each sampled region. This design was used mostly as a cost-reduction measure. The construction of a comprehensive list of schools would have been either impossible or prohibitively expensive. Also, this additional sampling stage reduces the dispersion of the school sample, thereby potentially reducing travel costs.

Sampling guidelines were put in place to ensure that an adequate number of sampling units would be sampled from this preliminary stage. The sampling frame had to consist of at least 100 primary sampling units, of which at least 50 had to be sampled at this stage.

4.4.3 REPLACEMENT SCHOOLS

A high participation rate among sampled schools is not always possible. To avoid sample-size losses, a mechanism was instituted to identify, a priori, replacement schools for each sampled school. For each sampled school the next school on the ordered school-sampling frame was identified as its replacement; and the one after that as a second replacement, should it be necessary.

The use of implicit stratification variables and the subsequent ordering of the school-sampling frame by size ensured that any sampled school's replacement would have similar characteristics. Although this approach was not guaranteed to avoid response bias, it

would tend to minimize the potential for bias. Furthermore, it was deemed more acceptable than oversampling to accommodate a low response rate.

4.5 SECOND SAMPLING STAGE

For Populations 1 and 2, the second sampling stage consisted of selecting classrooms within sampled schools. As a rule, one classroom per target grade was sampled, although some participants opted to sample two classrooms per grade.

Classrooms were selected either with equal probabilities or with probabilities proportional to their size. Participants who opted to test all students in selected classrooms sampled classrooms with equal probabilities. This was the method of choice for most participants. Participants who chose to subsample students within selected classrooms sampled classrooms with PPS.

4.5.1 SMALL CLASSROOMS

Generally, classrooms in an education system tend to be of roughly equal size. Frequently, however, small classrooms are devoted to special situations, such as remedial or accelerated programs. These classrooms can become problematic since they can lead to a shortfall in sample size, and thus introduce some instability in the resulting sampling weights when classrooms are selected with equal probabilities.

In order to avoid these problems, it was suggested that any classroom smaller than half the specified minimum cluster size be combined with another classroom from the same grade and school. For example, if the minimum cluster size was set at 30, then any classroom with fewer than 15 students should be combined with another. The resulting pseudo-classroom would then constitute a sampling unit. If a pseudo-classroom was sampled, then all of its component classrooms would fall in the sample.

4.5.2 POPULATION 3

For Population 3, the second sampling stage consisted either of sampling classrooms or of sampling students directly from the target grades, depending on how students taking advanced courses in mathematics or physics were organized into schools and classes. Chapter 9 describes the within-school sampling at Population 3, for systems where students could be selected in intact classes and for systems where students in each subpopulation were sampled from across the entire grade level in a school.

4.6 OPTIONAL THIRD SAMPLING STAGE

An optional third sampling stage consisted of selecting students within sampled classrooms. Generally, all students in selected classrooms were included in the TIMSS sample. Participants with particularly large classrooms in their education system could opt to subsample a fixed number of students per selected classroom. This was done using a simple random sampling method whereby all students in a sampled classroom were assigned equal selection probabilities.

4.7 RESPONSE RATES

Weighted and unweighted response rates were computed for each participant by grade, at the school level and at the student level. Specific criteria were put in place to determine acceptable response rates at each level.

4.7.1 SCHOOL-LEVEL RESPONSE RATES

The minimum acceptable school-level response rate, before the use of replacement schools, was set at 85%. This criterion was applied to the unweighted school-level response rate. School-level response rates will be computed and reported by grade weighted and unweighted, with and without replacement schools. The general formula for computing weighted school-level response rates is shown in the following equation:

$$R_{wgt}(sch) = \frac{\sum_{part} MOS_i / \pi_i}{\sum_{elig} MOS_i / \pi_i}$$

For each sampled school, the ratio of its MOS to its selection probability (π_i) is computed. The weighted school-level response rate is the sum of the ratios for all participating schools divided by the sum of the ratios for all eligible schools. The unweighted school-level response rates are computed in a similar way, where all school ratios are set to unity. This becomes simply the number of participating schools in the sample divided by the number of eligible schools in the sample. Since in most cases, in selecting the sample, the value of π_i was set proportional to MOS_i within each explicit stratum, it is generally the case that weighted and unweighted rates are similar.

4.7.2 STUDENT-LEVEL RESPONSE RATES

Like the school-level response rate, the minimum acceptable student-level response rate was set at 85%. This criterion was applied to the unweighted student-level response rate. Student-level response rates will be computed and reported by grade, weighted and unweighted. The general formula for computing student-level response rates is shown in the following equation:

$$R_{wgt}(stu) = \frac{\sum_{part} 1/p_j}{\sum_{elig} 1/p_j}$$

where p_j denotes the probability of selection of the student, incorporating all stages of selection. Thus 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. The unweighted student response rates will be computed in a similar way, but with each student contributing equal weight.

Student-level response rates in Population 3 will be calculated separately by subpopulation. There will therefore be separate student-level response rates for the general

population, and for students taking courses in advanced mathematics, and for students taking courses in physics.

4.7.3 Overall Response Rates

The minimum acceptable overall response rate was set at 75% for the upper grade. This overall response rate for each grade was calculated as the product of the weighted school-level response rate at the grade without replacement schools and the weighted student-level response rate at the grade.

Weighted overall response rates will be computed and reported by grade, both with and without replacement schools.

REFERENCES

Third International Mathematics and Science Study (TIMSS). (1992). *Sampling Plan* (Doc. Ref.: ICC438/NPC116). Prepared by Richard Wolfe and David Wiley.

Third International Mathematics and Science Study (TIMSS). (1994a). *Sampling Manual–Version 4* (Doc. Ref.: ICC 439/NPC117). Prepared by Pierre Foy and Andreas Schleicher. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994b). *School Coordinator Manual–Populations 1 and 2* (Doc. Ref.: ICC891/NRC427). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994c). *School Coordinator Manual–Population 3* (Doc. Ref.: ICC907/NRC440). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994d). *Survey Operations Manual–Populations 1 and 2* (Doc. Ref.: ICC889/NRC425). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994e). *Survey Operations Manual–Population 3* (Doc. Ref.: ICC 906/NRC439). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Schmidt, W.H. and Cogan, L.S. (1996) "Development of the TIMSS Context Questionnaires" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

5. Development of the TIMSS Context Questionnaires

William H. Schmidt
Leland S. Cogan

5.1 OVERVIEW

The Third International Mathematics and Science Study was designed to investigate students' learning of mathematics and the sciences internationally. The IEA's Second International Mathematics Study (SIMS), recognizing the importance of curriculum in any study of student achievement, developed a tripartite model that placed the curriculum at the center of the education process. The factors that influence the education process at three different levels–system, classroom, and student–are represented in this model by three aspects of curriculum: the intended, implemented, and attained curriculum. The intended curriculum refers to the educational system's goals and the structures established to reach them. The implemented curriculum refers to the range of practices, activities, and institutional arrangements within the school and classroom that are designed to implement the visions and goals of the intended curriculum. The attained curriculum refers to the products of schooling, what students have actually gained from their educational experiences. Building on this conceptualization of the education process, TIMSS sought to assess, through context questionnaires, the factors at the system, school, teacher, and student level that are likely to influence students' learning of mathematics and the sciences.

The Survey of Mathematics and Science Opportunities (SMSO) was funded by the National Science Foundation and the U.S. National Center for Educational Statistics as a small-scale international research project. Its task was, first, to construct a model of the

educational experiences of students; and, second, to develop a comprehensive battery of survey instruments for TIMSS that could be used to study the student, teacher, and school characteristics that explain cross-national differences in student achievement in mathematics and the sciences. A team of educational researchers from six countries collaborated in the development, piloting, and revision of all aspects of the instrumentation.

The principal contributors to this effort were Richard Wolfe (Canada), Emilie Barrier (France), Toshio Sawada and Katsuhiko Shimizu (Japan), Doris Jorde and Svein Lie (Norway), Ignacio Gonzalo (Spain), Urs Moser (Switzerland), and Edward Britton, Leigh Burstein, Leland Cogan, Curtis McKnight, Senta Raizen, Gilbert Valverde, David Wiley, and William Schmidt (United States). Others made significant contributions by conducting teacher interviews and classroom observations and by participating in analytical discussions. Among these people are Daniel Robin and Josette Le Coq from France, Masao Miyake and Eizo Nagasaki from Japan, José Antonio López Varona, Reyes Hernández, Blanca Valtierra, and Icíar Eraña from Spain, Erich Ramseier from Switzerland, and Carol Crumbaugh, Pam Jakworth, Mary Kino, and Margaret Savage from the United States.

5.2 INITIAL CONCEPTUAL MODELS AND PROCESSES

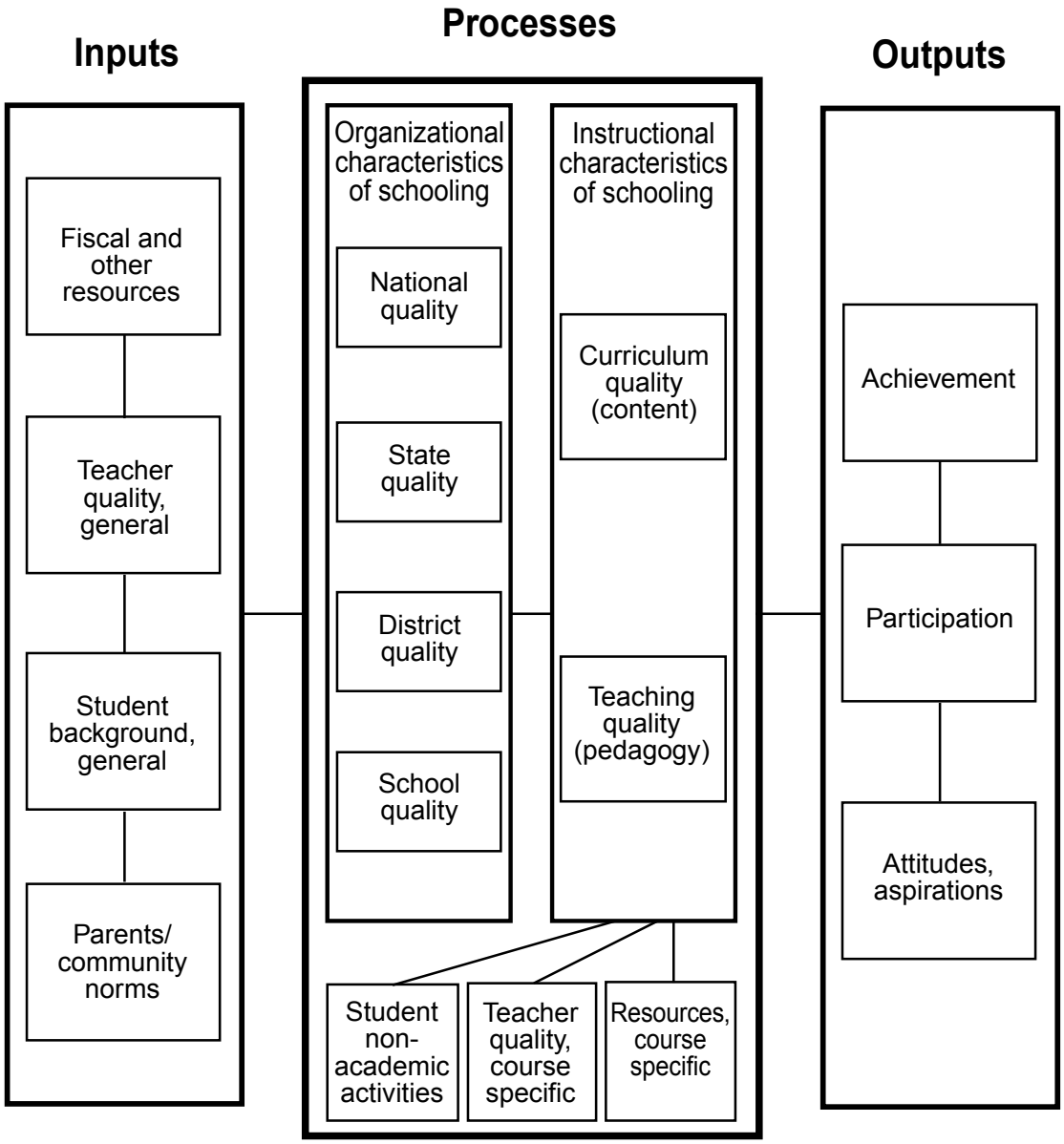
The U.S. National Center for Education Statistics (NCES) provided funding for a series of focus groups to begin to identify issues for specific data-gathering instruments. Each focus group concentrated on one of four levels of the educational system–the system; the school; the classroom and the teacher; and the student–and developed the corresponding questionnaires. The group concentrating on system-level characteristics developed the TIMSS participation questionnaire, which was used to gather some of the earliest TIMSS data. This group was chaired by David Wiley (United States) and included Manfred Lehrke (Germany), David Stevenson (United States), Ian Westbury (United States), and Timothy Wyatt (Australia). The school questionnaire focus group was chaired by Andrew Porter (United States) and consisted of Ray Adams (Australia), David Baker (United States), Ingrid Munck (Sweden), and Timothy Wyatt (Australia). The focus group for the teacher questionnaire was co-chaired by Leigh Burstein and Richard Prawat (United States) and included Ginette DeLandshere (Belgium), Jong-Ha Han (Korea), Mary Kennedy (United States), Frederick K. S. Leung (Hong Kong), Eizo Nagasaki (Japan), and Teresa Tatto (Mexico). The student questionnaire focus group was chaired by Judith Torney-Purta (United States) and included Chan Siew Eng (Singapore), Lois Peak (United States), Jack Schwille (United States), and Peter Vari (Hungary).

The development of each questionnaire began with a conceptual framework or model of the explanatory factors related to the object of the questionnaire. These models were based on the research literature and on previous IEA studies. For example, the initial identification of school-related concepts to be included in TIMSS was based on an indicator model of school processes developed by Porter (1991), shown in Figure 5.1.

The educational research literature has identified a profusion of important teacher characteristics that are related to student performance in mathematics and science. These

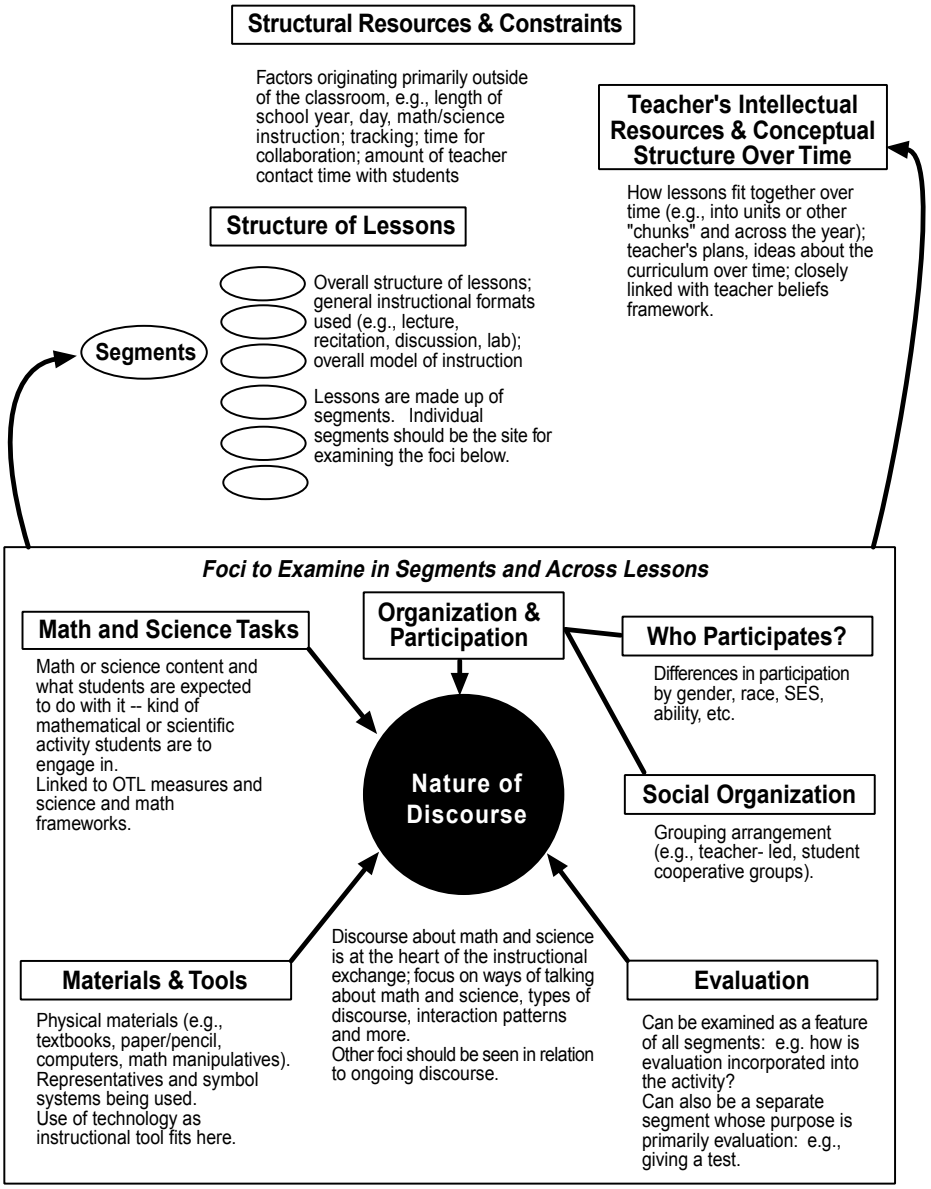
include the amount of conceptual coherence or focus that teachers build into their lessons (which reflects their own conceptual understanding), how teachers represent the subject matter, the organization and nature of instructional tasks, the patterns of classroom discourse, and the types of evaluation. In addition, the availability of technological and other material resources has proved to be significant for student learning.

Figure 5.1 An Indicator Model of School Processes



The conceptual model for instructional practices, shown in Figure 5.2, which was based upon reviews of the research literature (Prawat, 1989a, Prawat 1989b), integrated these factors for the first phase of instrument development.

Figure 5.2 Factors That Influence Instructional Practices

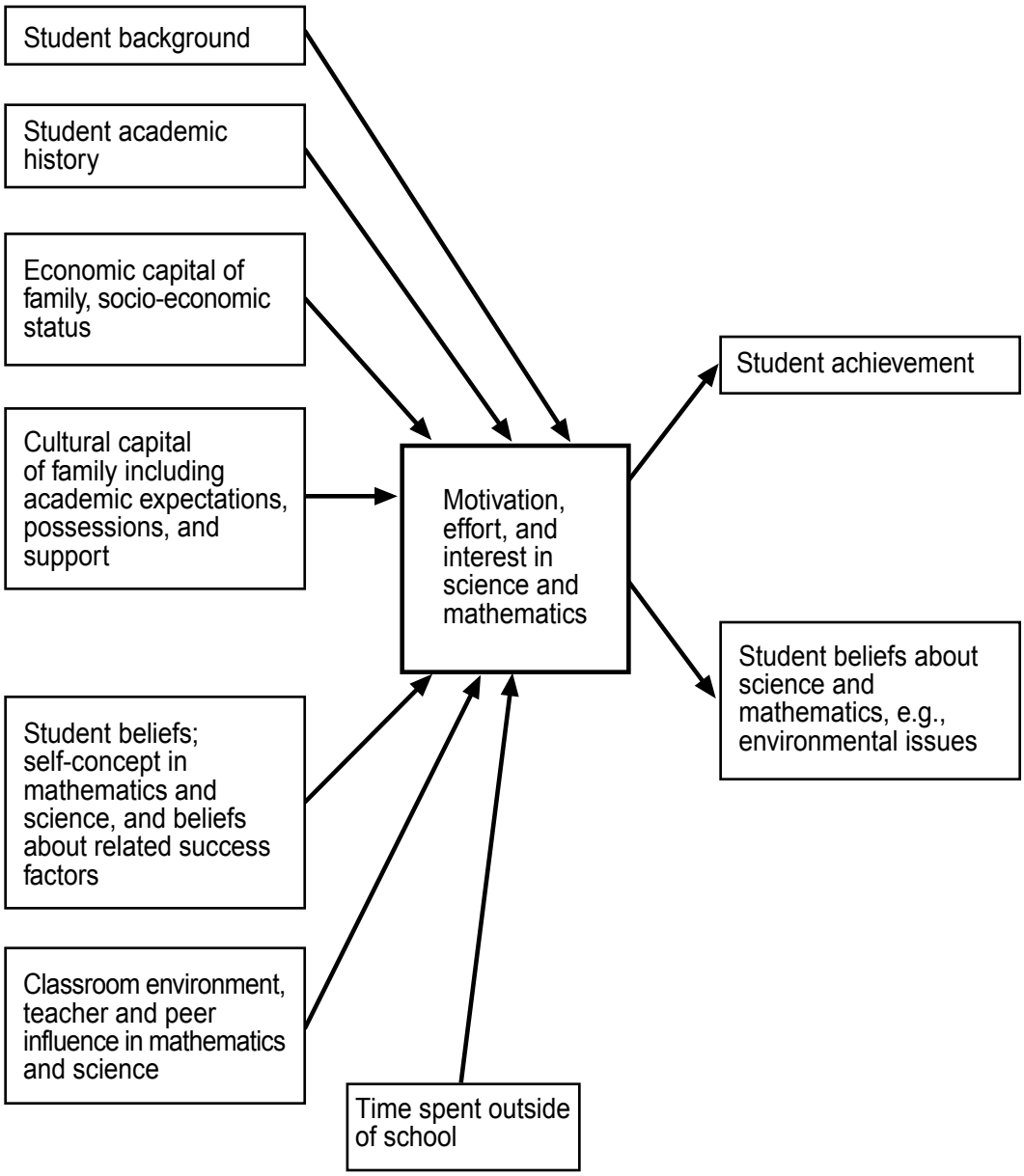


The initial list of student characteristics to be examined in TIMSS was drawn from the literature. Conceptual models of student achievement abound in the literature and most have a common set of constructs. Given the limits of a large-scale survey and the amount of student response time available, the TIMSS student focus group identified the following student constructs for consideration: demographic characteristics; home and family environment; attitudes and expectations; activities; perceptions of school context; and perceptions of classroom context.

Next, a draft student questionnaire was developed and piloted in a few countries. In addition, most countries reviewed the questionnaire, with some disappointing results. A group of Scandinavian researchers (Kjell Gisselberg, Marit Kjaernsli, Svein Lie, Borge Prien, Ingemar Wedman, Peter Weng, and Anita Wester) advanced work in this area by developing a conceptual framework that stressed the central role of motivation and effort in student achievement. That model was then integrated with the original framework. It is designed to address two questions: (1) what have students learned about science and mathematics (including ideas and beliefs about these subjects)? and (2) what student characteristics are related to student learning? The revised model is presented in Figure 5.3.

The model in Figure 5.3 suggests some of the factors that influence the motivation and interest a student has in studying science and mathematics. This motivation in turn influences student achievement, and also student beliefs about science and mathematics. Interest, motivation, and effort have been fused into one conceptual unit because of the difficulty of distinguishing among them on the basis of limited questionnaire data.

Figure 5.3 Revised Model of Student Characteristics



5.3 EDUCATIONAL OPPORTUNITY AS AN UNDERLYING THEME

The models described in the previous sections assume particular points of view, each aimed at a specific aspect of school learning. The model of Figure 5.2 represents a psycho-social view of classroom instruction consistent with the cognitive-psychology literature. The model of Figure 5.3 portrays a view of student learning influenced by theories of individual differences and motivation and sociological concepts such as family background. The school framework is based on an indicator model of school processes (Porter, 1991).

In a study of cross-national differences a more comprehensive perspective is essential–one in which instructional practices, individual student learning, and the organization of the school are all part of a larger system in which educational experiences are realized. Such a view recognizes that educational systems, schools, teachers, and the students themselves all influence the learning opportunities and experiences of individual students. From this perspective, educational opportunity can be regarded as a unifying theme of the TIMSS explanatory framework. Curriculum, instruction, and teacher characteristics are factors that both provide and delimit the educational opportunities of students to learn mathematics and sciences.

The curriculum, by specifying the learning goals at the national or regional level, emphasizes certain opportunities to learn and constrains others. For example, in a country with a mandatory national curriculum, the inclusion of a learning goal in that curriculum greatly increases the probability that classrooms will offer an opportunity to learn that topic. By the same token, the absence of a learning goal decreases the probability that educational opportunities related to that goal will be provided.

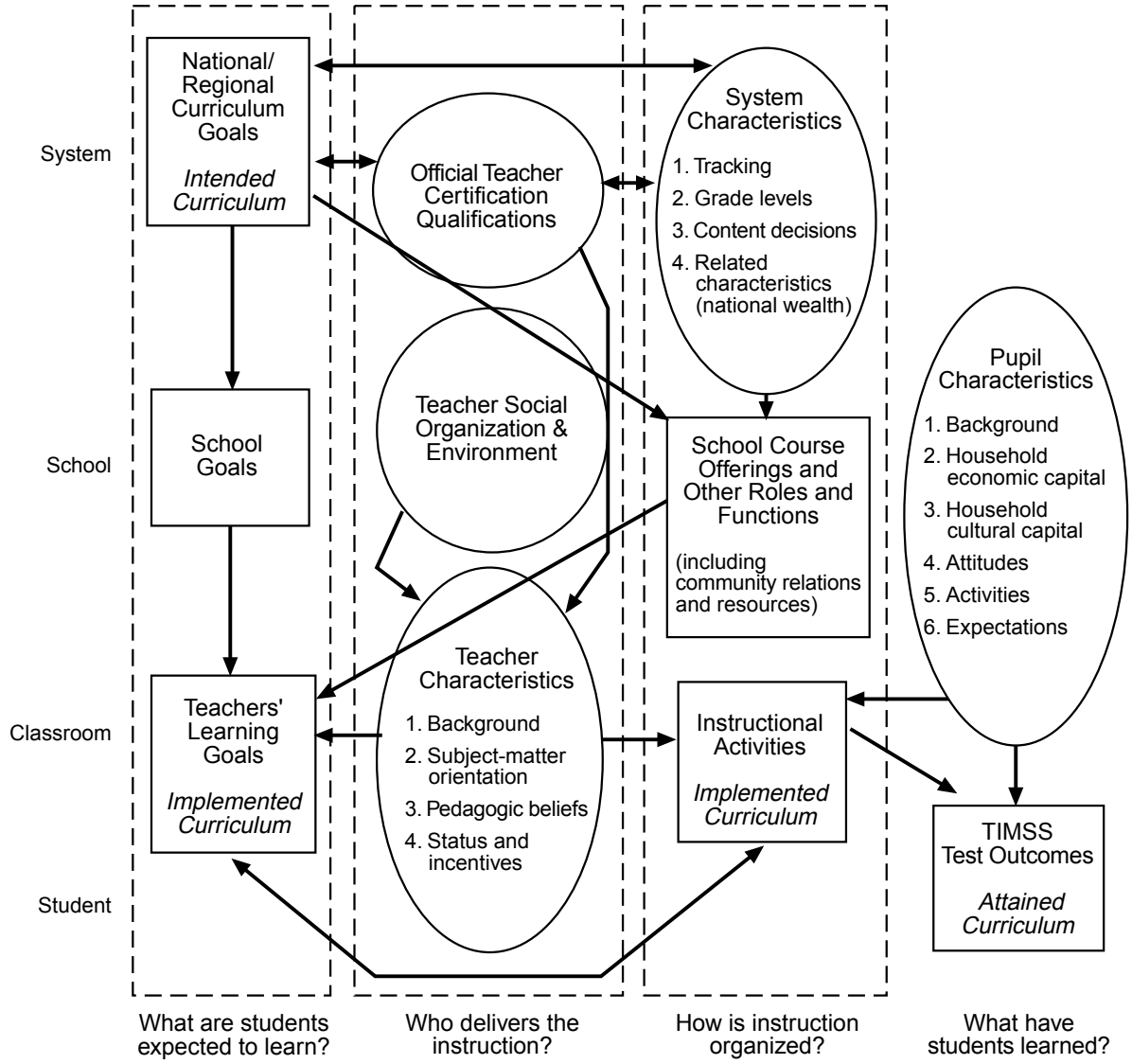
Differences across countries in the specification of learning goals, and the policies related to the learning goals, are critically important to understanding the nature of educational opportunity in those countries. The system-level specification of learning goals sets parameters by which educational opportunities are constrained in the first instance.

Schools and teachers, by their characteristics and activities, further frame educational opportunities. Both the curricular organization of the school and the qualifications and subject-matter knowledge of the teachers affect the provision and quality of educational opportunities. Teachers' instructional practices and the schools' course offerings further shape those opportunities.

To undergird the development of the data collection instruments, provision of educational opportunity was considered at the levels of the educational system, the school, and the classroom in terms of the four general research questions of TIMSS: (1) What are students expected to learn? (2) Who delivers the instruction? (3) How is instruction organized? and (4) What have students learned? This conceptual framework is presented schematically in Figure 5.4.

What are students expected to learn? There are three main levels of the educational system at which learning goals are commonly set: the national or regional level, the school level, and the classroom level. This first research question addresses not only the specification of learning goals for a system or country as a whole, but also the differentiation of such goals for divisions within the larger educational system, such as regions, tracks, school types, and grade levels. Learning goals specified at the national or regional level are, in the terminology developed within IEA for SIMS, the *intended* curriculum, whereas those specified at the school or classroom level are part of the *implemented* curriculum.

Figure 5.4 TIMSS Conceptual Framework: The Educational Experience Opportunity



Who delivers the instruction? Students' learning in school is shaped to a great extent by their teachers. The teaching force in a country may be characterized on a number of levels. At the system level are official teacher certification qualifications–including grade and subject restrictions, required education for licensing, and perhaps specific required coursework or experience. At the school level, the social organization and environment in which teachers work may influence their instructional practices. An important area here is the allocation of teacher time–the proportion of professional time spent during a school day in planning and teaching mathematics or science, and the amount of cross-grade-level teaching (Doyle, 1986; Lockhead, 1987). Collaboration among teachers in planning

instructional sequences and strategies may also greatly influence what occurs within the classroom.

At the classroom level the characteristics of the individual teacher may affect the quality of instruction and hence the quality of students' educational experiences. Such characteristics include teachers' background and beliefs (see Porter, 1991). Teacher background variables include age, gender, education, subject taught, and teaching experience. Teacher beliefs include subject-matter orientation–the views teachers have about the disciplines of mathematics and the sciences, which have been shown to affect instructional practices and student achievement (Thompson, 1992; Putnam, 1992; Peterson, 1990). Teacher beliefs also include pedagogical beliefs–their views about what is a good way to teach a particular topic.

How is the instruction organized? The organization of instruction influences the implemented curriculum and the learning experiences of students. Decision making concerning instruction is distributed across all levels of the education system. This diffusion affects many organizational aspects–the age-grade structure of education systems, the nature of the schools serving different arrays of grades, and the various curricular tracks into which students are placed. Economic resources also influence how instruction is organized, as do the qualifications of the teaching force, the instructional resources available to the teachers, and the time and material resources available to the students.

Instructional organization also subsumes course offerings and support systems for mathematics and science instruction, and the implementation of curriculum in classrooms, including textbook use, structure of lessons, instructional materials, classroom management, student evaluation, student participation, homework, and in-class grouping of students.

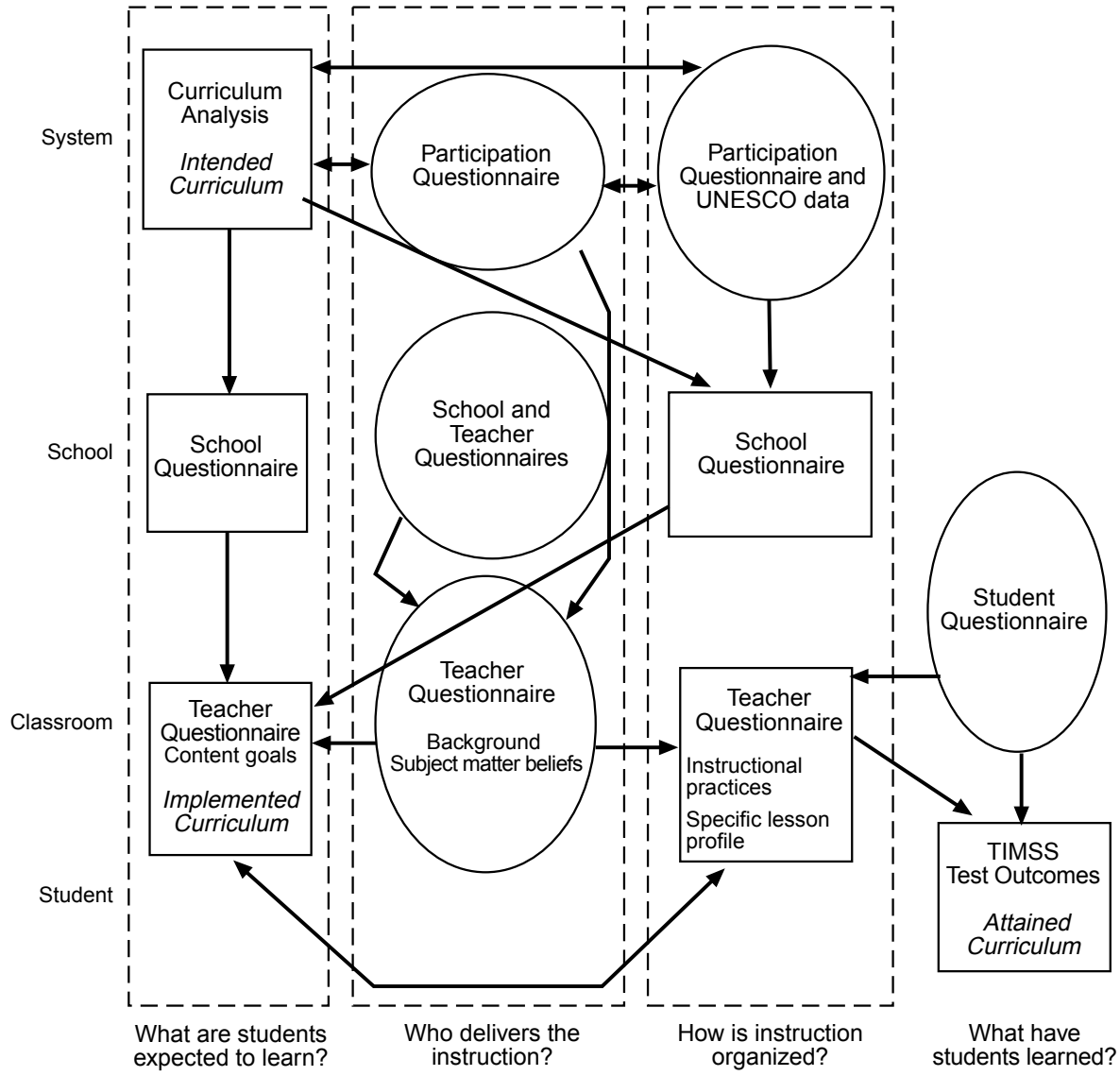
What have students learned? Comparing what students have learned in terms of their performance on the TIMSS achievement tests is a major focus of the study. However, beyond such comparisons TIMSS wanted to investigate the factors associated with student learning. Aside from curriculum goals, teachers, and instructional organization, characteristics of the students themselves influence what and how they learn. These characteristics include students' academic history, the economic and cultural capital of the family, students' self-concept, how students spend time outside school, and students' beliefs, motivation, effort, and interest in education and school subjects.

It is not possible to identify and measure every possible factor that affects student learning. However, the educational-opportunity model recognizes the connections among major components of the educational system in a very general way. This generic model can be used to describe many specific educational systems. It does not advocate a particular system but rather is intended as a template against which to study systemic variations; in this sense, it is particularly appropriate for cross-national comparisons.

The data collection instruments developed by SMSO, specifically the participation, school, teacher, and student questionnaires and the curriculum analysis, were all developed

concomitantly with the educational opportunity model to examine specific model components. These are presented schematically in Figure 5.5.

Figure 5.5 TIMSS Instruments Assessing Educational Opportunity



5.4 INSTRUMENTATION REVIEW AND REVISION

In addition to the NCES focus groups that identified the initial issues and questions for the various instruments, many others were involved in the review and revision process. National Research Coordinators (NRCs) from the countries participating in TIMSS had opportunities to review the school, teacher, and student questionnaires at various stages.

Comments from NRCs were always carefully considered in producing subsequent versions for further rounds of piloting, review, and revision.

Upon several occasions, special groups of researchers were assembled to review, revise, and reorganize the questionnaires. The SMSO, the International Coordinating Center (ICC), and the International Study Center brought together groups to work in this area. As part of the development of the questionnaires, TIMSS conducted small informal pilot studies with teachers, students, and school administrators, as well as large-scale formal pilot studies. The student questionnaire was piloted during the item pilot conducted by the ICC in most of the TIMSS countries in April and May 1993, and the teacher and school questionnaires during September and October 1993; key portions of the latter two questionnaires were also included in the field trial in April and May 1994.

For the 1993 pilot study of the teacher and school questionnaires, each participating country translated the questionnaires into the local language, obtained responses from teachers and principals, and recorded those responses in computer files. Twenty-two countries participated in this pilot in some fashion. Twenty countries–Canada (Alberta), Argentina, Australia, Czech Republic, Denmark, France, Greece, Indonesia, Iran, Ireland, Korea, Mexico, New Zealand, Portugal, Romania, Singapore, Spain, Sweden, Switzerland, and the United States–submitted data files. Table 5.1 shows the number of responses submitted and analyzed.

Table 5.1 Responses in Pilot Study of School and Teacher Questionnaires

Questionnaire		Number of Responses
Teacher Questionnaire	Population 1	488
	Population 2	296
	Population 3	290
School Questionnaire	Population 1	133
	Population 2	174
	Population 3	58

In addition to the data files, 15 countries–Canada (Alberta), Australia, Czech Republic, France, Greece, Ireland, Korea, Netherlands, New Zealand, Portugal, Singapore, Sweden, Switzerland, Tunisia, and the United States–submitted written reports on the pilot studies in their countries.

Three types of data from the pilot study were used to revise the teacher and school questionnaires. First, all comments concerning the questionnaires made in NRCs' reports or by other sources were placed into an electronic database. This was organized by item within each questionnaire. Table 5.2 shows examples of comments on two items, one from the school questionnaire and one from the teacher questionnaire.

Table 5.2 Examples of Comments on Questionnaire Items Entered into Database

Question	Country	Comment
SC1-12	CSK	Principals teach regularly, they must prepare for their lessons, some of them even work as homeroom teachers. These activities are missing in the list.
TQ1 General	NLD	Each questionnaire needs a general instruction in front of the questionnaire indicating the purpose of it (gathering information about the implemented curriculum, which is related to information about the attained and intended curriculum as well) and saying that most questions can be answered by checking one or more boxes. Note: same comments for TQ2M-Gen. and TQ2S-Gen.

The second type of data from which revisions were made came from the written responses to the "other" options that were part of many items in the piloted questionnaires. These responses were translated into English, placed into a database, and sorted by questionnaire type and item. The third type of data came from multiple-choice questionnaire items that were stored in the data files.

The written responses to the "other" options were used to expand the options for some items and to revise others. Instructions and options were rewritten to clarify the intent of some questions and to facilitate the generation of an appropriate response. The multiple-choice item data were analyzed to eliminate options for some items, rewrite some options, and confirm that some options should be retained rather than eliminated.

The pilot study gave rise to the following conclusions about the draft questionnaires.

- The questionnaires were too long and took too much time to complete
- Some of the language was too technical
- Considerable cross-country variation in item responses was evident. This variation, which makes international comparisons interesting, also makes it difficult to develop items that are meaningful and relevant within all countries
- There was a good distribution of responses across the item options. Respondents seemed to have no difficulty responding to options with three, four, or five categories
- Much of the formatting needed to be simplified. Some countries were unable to reproduce shaded areas and many respondents found the skip patterns difficult to follow

The results of the pilot study led to extensive revision of the questionnaires. In June 1994, a meeting was held in Hamburg, Germany, for the purpose of reviewing and revising the Populations 1 and 2 school and teacher questionnaires. Hosted by Neville Postlethwaite and chaired by William Schmidt, the working group included Michael Martin (International Study Center) and the following NRCs: Wendy Keys (England), Christiane Brusselmans-Dehairs (Belgium, Flemish), and Wilmad Kuiper (Netherlands). The International Study Center then made the recommended changes and disseminated the revised versions of the questionnaires to all NRCs and TIMSS committees. Simultaneously,

the student questionnaires for Populations 1 and 2 were reformatted, revised, and distributed for review. The Populations 1 and 2 context questionnaires were endorsed by the TIMSS NRCs in August 1994 and both paper and electronic versions were provided to the participating countries for translation, duplication, and administration.

In October 1994, the Population 3 school and student questionnaires were revised. In early November 1994, a group of NRCs reviewed the questionnaires and made suggestions for restructuring them. The International Study Center made the changes and distributed the revised versions to a small group of NRCs nominated by their colleagues for review before dissemination. In December 1994, the final versions of the Population 3 student and school questionnaires were disseminated to all participating countries for translation, duplication, and administration.

The model of educational opportunity guided questionnaire development, item evaluation, and revision throughout. The identification of key research questions led to the creation of a conceptual framework matrix in which various issues were assigned to specific instruments. This model links the three main areas of investigation in TIMSS: the curriculum analysis, the context questionnaires, and the student test.

5.5 THE FINAL INSTRUMENTS

The participation questionnaires gathered general information about a country's education system and its organization and structure. This information was used in the early stages of TIMSS to make decisions about sampling and about which curriculum guides and textbooks would be appropriate for the curriculum analysis. It was also used to identify issues that would need further clarification from the other instruments.

The school questionnaires at each population level sought information about the school's community, staff, students, curriculum and programs of study, and instructional resources and time. The number of years students are taught by the same teacher is addressed in the Population 1 and 2 versions but is not relevant at the Population 3 level. The school's requirements for graduation or successful completion of schooling are addressed in the Population 3 version but not in the others. Questions that address programs of study are expanded in the Population 3 version since this issue is considerably more complex at this level. The content and purpose of each item and the correspondences and differences among the three versions are detailed in Table 5.3.

The teacher questionnaires for Population 2 address four major areas: teacher's background, instructional practices, students' opportunity to learn, and teacher's pedagogic beliefs. There are separate questionnaires for teachers of mathematics and of science. Since most Population 1 teachers teach all subjects, a single teacher questionnaire at this level addresses both mathematics and science. This has constrained coverage such that only items addressing teacher's background and instructional practices are included. In general, the focus for most questions is mathematics. However, the item assessing teacher's content goals is asked about both mathematics and science, since this is the main link in the teacher

questionnaire to the TIMSS curriculum analysis. The content and purpose of each item and the similarities and differences among the three versions are detailed in Table 5.4.

In general, the structure and content of the student questionnaires are consistent across populations. A few items were not included in the Population 1 version, such as students' reports of parents' education, since responses were not considered reliable. Also, most response categories were reduced in the Population 1 version from four to three. Two versions of student questionnaires for Population 2 were developed: one for use in systems teaching general science and another for use in systems where students take courses in specific sciences such as biology, chemistry, earth science, or physics. Some items are unique to the Population 3 student questionnaire. These were developed to gather information regarding students' academic history and their plans for further education. The content and purpose of each item is detailed in Table 5.5.

Table 5.3 Contents of the School Questionnaires for Populations 1, 2, and 3

Question Number				
POPULATION			Item Content	Description
1	2	3		
1	1	1	Community	Situates the school within a community of a specific type.
2	2	2	Grade Levels	Identifies the grade levels present in the school.
3, 4, & 5	3, 4, & 5	3, 4, & 5	Staff	Describes the school's professional full- and part-time staff and the percentage of teachers at the school for 5 or more years.
6	6	6 - 9	Teaching Load	Describes percentage of time teachers teach mathematics, the sciences, and/or other subjects.
7	7	–	Students with Teacher	Indicates the number of years students typically stay with the same teacher.
8	8 & 9	–	Teacher Time	Indicates the amount of time a teacher usually has for teaching mathematics/science classes and doing related tasks.
9	10	–	Collaboration Policy	Identifies the existence of a school policy promoting teacher cooperation and collaboration.
–	–	10 & 11	University Certification	Indicates the percentage of mathematics and science teachers who have university certification in their subject matter.
10	11	13	Principal's Time	Indicates the amount of time a school's lead administrator typically spends on particular roles and functions.
11	12	14	School Decisions	Identifies for the school who has responsibility for various decisions.
12	13	15	Curriculum Decisions	Identifies the amount of influence various individuals and educational and community groups have on curriculum decisions.
13	14	16	Formal Goals Statement	Indicates the existence of school-level curriculum goals for mathematics and science.
14	15	–	Availability of Computers	Indicates the number of computers available to staff and students for specific types of use.
15	16	12	Instructional Resources	Provides a description of the material factors limiting a school's instructional activities.
16	17	19	Students	Provides enrollment and attendance data, students' enrollment in mathematics and science courses, and typical class sizes.
17	18	17	Student Behaviors	Provides a description of the frequency with which schools encounter various unacceptable student behaviors.
18	19		Instructional Time	Indicates the amount of instructional time scheduled, according to the school's academic calendar.
19	20		Instructional Periods	Indicates the existence and length of weekly instructional periods.
20 - 23	21 - 24		Remedial and Enrichment	Describes the school's provision for remedial and enrichment programs in mathematics and science.

Table 5.3 Contents of the School Questionnaires for Populations 1, 2, and 3 (continued)

Question Number				
POPULATION			Item Content	Description
1	2	3		
24 & 26	25 & 27	20-22	Programs of Study	Describes the existence of different educational tracks or programs for studying mathematics and the sciences, and the instructional time for each program.
25 & 27	26 & 28	18	Program Decision Factors	Indicates how important various factors are in assigning students to different educational programs or tracks.
-	-	21	Graduates	Describes the academic standards required of students who successfully graduate or leave the school.
INTERNATIONAL OPTIONS				
28	29		Student Demographics	Indicates the percentage of students with various backgrounds.
29	30		Admissions	Describes the basis on which students are admitted to the school.

Table 5.4 Contents of the Teacher Questionnaires for Populations 1 and 2

Question Number				
POPULATION			Item Content	Description
1	2M	2S		
SECTION A:				
1 - 2	1 - 2	1 - 2	Age and Sex	Identifies teachers' sex and age-range category.
3	3	3	Education	Describes teachers' preparation for teaching according to 8 internationally defined categories of education and teacher training. Labels for categories are country-specific with only relevant categories being used.
4 - 5	4 - 5	4 - 5	Teaching This Year	Describes at which grade levels teacher is teaching math and/or science.
6 - 8	6 - 8	6 - 8	Teaching Experience	Identifies teachers as either full- or part-time, the number of years of teaching experience, and an indication of experience in last 5 years with teaching at various grade levels.
–	9 - 11	9 - 11	Formal Teaching Responsibilities	Describes the scope and depth of the formally scheduled teaching responsibilities of teachers of mathematics and the sciences.
9	12	12	Other Teaching-Related Activities	Describes the amount of time teachers are involved in various professional responsibilities *outside* the formally scheduled school day.
10	13	13	Meet With Other Teachers	Describes the frequency that teachers' collaborate and consult with their colleagues.
–	14	14	Teachers' Influence	Describes the amount of influence that teachers' perceive they have on various instructional decisions.
11	15	15	Being Good at Maths/Science	Describes teachers' beliefs about what skills are necessary for students to be good at mathematics/science.
12	16	16	Ideas about Maths/Science	Indicates teachers' beliefs about the nature of mathematics/science and how the subject should be taught.
13	17	17	Document Familiarity	Describes teachers' knowledge of curriculum guides, teaching guides, and examination prescriptions. (country-specific options)
–	–	18	Topics Prepared to Teach	Provides an indication of teachers' perceptions of their own preparedness to teach the TIMSS in-depth topic areas.
INTERNATIONAL OPTIONS				
–	18-23	19-24	Teacher Status	Describes teacher's occupational satisfaction, perceived social status of teaching, and the number of books in the home.
SECTION B: INSTRUCTIONAL PRACTICES (Pertains to Target Class)				
14	B-1	B-1	Target Class	Identifies the number of students in the TIMSS tested class.
15	B-2	B-2	Student Achievement	Describes teacher's perception of the achievement levels of students in the TIMSS tested class compared to other students nationally.
16	B-3	B-3	Instructional Time	Identifies the number of minutes per week the class is taught.

Table 5.4 Contents of the Teacher Questionnaires for Populations 1 and 2 (continued)

Question Number				
POPULATION			Item Content	Description
1	2M	2S		
17	B-4	B-4	Textbook Used	Identifies the textbook used in the TIMSS target class.
18	B-5	B-5	Percent Textbook Used	Identifies the approximate percentage of teacher's weekly teaching that is based on the textbook.
–	B-6	B-6	Textbook Alternatives	Identifies resources that a teacher uses in addition to or in the place of a textbook.
19	–	–	Teaching Groups	Identifies the frequency with which the teacher divides the class into groups for teaching.
20	B-7	B-7	Classroom Factors	Identifies the extent to which teachers perceive that various factors limit classroom instructional activities.
1 22	B-8 B-9	B-8 B-9	Calculators	Describes the availability of calculators and how they are used in the target class.
23 24	B-10 B-11	B-10 B-11	Planning Lessons	Identifies the extent to which a teacher relies on various sources for planning lessons.
25-M 37-S	B-12	B-12	Topic Coverage	Indicates the extent of teachers' content coverage with the TARGET CLASS according to categories from the TIMSS Curriculum Frameworks.
26	B-13	B-13	Recent Class Hour	Describes the length, topic (according to the TIMSS frameworks), type (introduction, continuation, or end), and homework assigned for a recent lesson.
27	B-14	B-14	Lesson Order	Characterizes a recent lesson; the sequence of instructional activities and the amount of time devoted to each activity.
28	B-15	B-15	Asking Students Questions	Describes the type, manner, and purpose for which teachers ask students various types of questions and ask students to perform various activities during lessons.
29	B-16	B-16	Incorrect Response	Identifies the frequency with which a teacher responds to a student's incorrect response in several different ways.
30-M 36-S	B-17	B-17	Students' Work Arrangements	Describes how often students working in various group arrangements.
31 32	B-18 B-19	B-18 B-19	Amount of Homework Assigned	Describes the frequency and amount of homework assigned to target class students.
33 34	B-20 B-21	B-20 B-21	Type and Use of Homework	Describes the nature of homework assignments and how homework is used by the teacher.
35	–	–	Science	Indicates the weekly amount of science instruction and whether science is taught as a separate subject.
–	B-22 B-23	B-22 B-23	Assessment	Describes the nature and use of various forms of student assessment in the target class.
SECTION C: OPPORTUNITY TO LEARN				
–	I to XIV	I to XIII	Opportunity to Learn	Describes students opportunity to learn items from the in-depth topic areas. Items used in this section come from the TIMSS student test.
SECTION D: PEDAGOGICAL APPROACH				
	1 - 2	1 - 3	Pedagogical Beliefs	Provides an indication of teachers' instructional beliefs systems about teaching specific subject matter (i.e. mathematics or science).

Table 5.5 Contents of the Student Questionnaires for Populations 1, 2, and 3

Question Number					
POPULATION				Item Content	Description
1	2	2 (s)	3		
1 - 4	1 - 4	1 - 4	1 - 5	Student's Demographics	Provides basic demographic information to contextualize students' responses: age; sex; language of the home; if born in the country and if not how long he/she has lived in country.
5	5	5	15	Academic Activities Outside of School	Provides information on student activities that can impact their academic achievement.
6	6	6	16	Time Outside of School	Provides information on students' recreational and study habits outside of school.
7 - 8	7 - 8	7 - 8	6 - 7	People Living in the Home	Provides information about the home environment as an indicator of cultural and economic capital.
–	9	9	11	Parental Education	Provides an indicator of the home environment and data to create an indicator of socio-economic status.
9	10	10	4	Parent's Country of Birth	Provides information regarding immigrant status.
10	11	11	8	Books in Home	Provides an indicator of the cultural capital of the home environment.
11	12	12	9	Possessions in the Home List	Provides information to create an indicator of socio-economic status.
–	–	–	10	Residence While Attending School	Identifies the type of living situation students have while attending school.
–	–	–	12	Others' Ideas for Student's Future	Describes students' perceptions of what parents, teachers, and peers think student should do upon completion of school.
12	13	13	13	Mother's Values	Provides an indicator of the home environment and general academic press.
–	14	14	–	Students' Behavior in Math Class	Provides a description of typical student behavior during math lessons.
13	15	15	13	Peers' Values	Provides a description of peers' values and student's social environment.
14	16	16	13	Student's Values	Provides a description of student's values.
–	–	–	14	Student's Future Education Plans	Identifies what plans student has for further education.

Table 5.5 Contents of the Student Questionnaires for Populations 1, 2, and 3 (continued)

Question Number					
POPULATION				Item Content	Item Purpose
1	2	2 (s)	3		
15	17	17	22	Competence in Math / Sciences	Provides an indication of students' self-description of their academic competence in mathematics and the sciences.
16	18	18	17	Report on Student Behaviors	Provides an indication of the existence of specific problematic student behaviors at school from the student's perspective.
17	19	19	20	Doing Well in Math	Identifies students' attributions for doing well in mathematics.
18	20	20	21	Doing Well in Science	Identifies students' attributions for doing well in the sciences
19	21	21	19	Liking Math / Sciences	Identifies how much students like specific subjects; a key component of student motivation.
20	22	22	–	Liking of Computers	Identifies how well students like working with computers, a key indicator of technology familiarity.
21	23	23	18	Interest, Importance, & Value of Mathematics	Provides a description of students' interest, importance rating, and value afforded mathematics.
–	24	24	–	Reasons to Do Well in Math	Provides the extent to which students endorse certain reasons they need to do well in mathematics.
–	–	–	23	Technology Use	Identifies the type and frequency of student's technology use.
–	–	–	24	Student's Academic Program / Track	Identifies the educational program or track in which student is enrolled .
–	–	–	25	Most Advanced Math	Identifies the most advanced math course student has taken.
–	–	–	26	Most Advanced Physics	Identifies the most advanced physics course student has taken.
–	–	–	27	Most Advanced Chemistry	Identifies the most advanced chemistry course student has taken.
–	–	–	28	Most Advanced Biology	Identifies the most advanced biology course student has taken.
–	–	–	29	Most Advanced Earth Science	Identifies the most advanced earth science course student has taken.
–	–	–	30	Math Enrollment	Identifies which math course(s) student currently take.
22	25	25	31	Classroom Practices: Math	Provides a description of students' perceptions of classroom practices in mathematics instruction.
–	26	26	–	Beginning a New Math Topic	Describes the frequency with which specific strategies are used in the classroom to introduce a new mathematics topic.

Table 5.5 Contents of the Student Questionnaires for Populations 1, 2, and 3 (continued)

Question Number					
POPULATION				Item Content	Item Purpose
1	2	2 (s)	3		
–	27	27	–	Environmental Issues	Provides an indication of students' concern and involvement in environmental issues.
–	–	28	34, 35	Sciences Enrollment	Identifies which science course(s) students are currently taking.
21	28	29, 33, 37, 41	–	Interest, Importance, & Value of the Sciences	Provides a description of students' interest, importance rating, and value afforded mathematics.
–	29	30, 34, 38, 42	–	Reasons to Do Well in the Sciences	Provides the extent to which students endorse certain reasons they need to do well in the sciences.
–	30	–	–	Science Use in a Career	Identifies preferences for sciences in careers.
23	31	31, 35, 39, 43	–	Classroom Practices: Sciences	Provides a description of students' perceptions of classroom practices in science instruction.
–	32	32, 36, 40, 44	–	Beginning a New Topic	Describes the frequency with which specific strategies are used in the classroom to introduce a new topic in the sciences.
–	–	–	32	Math Textbook	Identifies the textbook used by students in their math course.
–	–	–	33	Math Homework	Identifies the frequency with which homework is assigned in students' math course.
–	–	–	36	Classroom Practices: Physics or Other Science	Provides a description of students' perceptions of classroom instructional practices.
–	–	–	37	Physics/ Other Science Textbook	Identifies the textbook used by students in their physics or other science course.
–	–	–	38	Physics/ Other Science Homework	Identifies the frequency with which homework is assigned in students' physics or other science course.
OPTIONAL ITEMS					
24, 25	33, 34	45, 46	–	Cultural Activities	Provides a description of student's involvement in cultural events or programming such as plays and concerts.
–	–	–	39, 40	Academic Program Profile	Indicates whether students are repeating the current grade or if they have already completed any other educational program at school.

REFERENCES

Doyle, W. (1986). Classroom Organization and Management in M.C. Wittrock (ed.), *Handbook of Research on Teaching* (pp. 392-431). New York: Macmillan.

Lockhead, M. (1987). *School and Classroom Effects on Student Learning Gain.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Peterson, P.L. (1990). Doing More in the Same Amount of Time: Cathy Swift. *Educational Evaluation and Policy Analysis*, 12, 261-280.

Porter, A. C. (1991). Creating a System of School Process Indicators. *Educational Evaluation and Policy Analysis*, 13, 13-29.

Prawat, R. S. (1989a). Promoting Access to Knowledge, Strategy, and Disposition in Students: A Research Synthesis. *Review of Educational Research*, 59, 1-41.

Prawat, R. S. (1989b). Teaching for Understanding: Three Key Attributes. *Teaching and Teacher Education*, 5, 315-328.

Putnam, R. T. (1992). Teaching the "Hows" of Mathematics for Everyday Life: A Case Study of a Fifth-grade Teacher. *The Elementary School Journal*, 93, 145-152.

Thompson, A. G. (1992). Teachers' Beliefs and Conceptions: A Synthesis of the Research. in D. A. Grouws (ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 127-146). New York: Macmillan.

Harmon, M. and Kelly, D.L. (1996) "Development and Design of the TIMSS Performance Assessment" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

6. Development and Design of the TIMSS Performance Assessment

Maryellen Harmon
Dana L. Kelly

6.1 OVERVIEW

The TIMSS performance assessment was administered at Populations 1 and 2 to a sub-sample of students in the upper grades that participated in the main survey. It was designed to augment the information elicited by the multiple-choice, and free-response items presented in the TIMSS achievement booklets so that TIMSS also has measures of students' responses to hands-on, practical tasks.[1] Students that participated in the performance assessment performed a series of tasks requiring mathematics and science knowledge and performance skills. Students engaged in the tasks according to a rotation scheme whereby 12 tasks were set up across 9 stations. This chapter describes the development of the performance assessment tasks, the assessment design, procedures for sampling schools and students, and administration procedures. Scoring of the students' work is described in Chapter 7.

[1] Such forms of assessment have been called "alternative assessment" and "practical assessment." TIMSS recognizes that all achievement test items, including the multiple-choice and free-response items, assess student performance. However, "performance assessment" is used in TIMSS because it is the term most often used in the research literature for "hands-on" tasks requiring sustained, integrated strategies or routine practical procedures (TIMSS, 1994).

6.2 CONSIDERATIONS FOR THE DESIGN

The performance assessment was designed to be practical, affordable, and easily translatable to multiple languages and cultures. The main considerations that guided the development process are listed below.

- The tasks and procedures had to be replicable across administrations and countries
- The materials and resources required had to be obtainable in each country
- The number of tasks each student would complete had to be kept low, given the time required to complete just one task
- The evaluation had to be based on a *written* product rather than on observed performance, given the amount of money and time required to use observational techniques in large-scale assessment
- The number of schools and students had to be kept to a minimum, in order to minimize administration costs.

6.3 TASK DEVELOPMENT

In December 1993, the Performance Assessment Committee (PAC) was established to develop performance assessment tasks and administration procedures. The committee members are Derek Foxman (England), Robert Garden (New Zealand), Maryellen Harmon (United States), Per Morten Kind (Norway), Svein Lie (Norway), Jan Lokan (Australia), and Graham Orpwood (Canada). Edys Quellmalz (United States) also contributed to the performance assessment development. Building on prior TIMSS work, the PAC collected a number of performance assessment tasks. Some were drawn from the Assessment of Performance Unit in England (Archenhold et al., 1988; Schofield et al., 1989), the International Assessment of Educational Progress II, the IEA's Second International Science Study, and state-or province-level assessments and research studies in Australia, Canada, New Zealand, and the United States. Several tasks were written specially for TIMSS.

In addition to considerations of practicality, affordability, and feasibility, performance tasks were designed to conform to the TIMSS curriculum frameworks (Robitaille et al., 1993). While the amount of time required for the performance tasks, and hence their limited number, precluded covering all of the content areas in the frameworks, a sample was drawn from each of the main subject-matter content categories and most of the performance expectations categories.

The performance assessment tasks required students to engage in an experimental procedure or manipulation of equipment during which they responded to a number of task-related questions (hereafter referred to as "items"). Each task generally began with a statement defining a central problem or investigation, such as

> "Investigate what effect different temperatures have on the speed with which the tablet dissolves."

The items students were required to complete ranged from easy items describing approach and procedures at the beginning of the task to more difficult interpretive items,

often requiring prior concept knowledge. In some cases, particularly with Population 1, the tasks were scaffolded. That is, the items at the beginning of the task were designed to allow all students to start from the same point, regardless of prior instruction on a procedure.[2] For example, the task containing the exemplar statement shown above began with the following instructions on how to proceed with the task.

> Plan an experiment to find out what effect different temperatures have on the speed with which the tablet dissolves.
>
> Write your plan here. Your plan should include:
>
> - What you will measure
> - How many measurements you will make
> - How you will present your measurements in a table.

Between December 1993 and February 1994, 26 tasks were selected or created and piloted in six countries. In February 1994, the PAC met in London to review the available pool of tasks, adapt and revise tasks as necessary, and select those to be administered in the field trial. Of the available tasks, 22 were selected for each population for field testing. Ten of the tasks required mathematics content and performance skills, ten required science content and performance skills, and two required skills in both.

6.3.1 PERFORMANCE ASSESSMENT FIELD TRIAL

The field trial of the performance assessment took place between March and April 1994. A total of 19 countries participated.

For the field trial, a convenience sample of 64 students (eight students from each of eight schools) per population was selected in each country. In some cases, students in both Populations 1 and 2 were sampled from the same school. In each cluster of eight students, two had been identified from independent data as high achievers, four as average, and two as low achievers.

Subject matter specialists, performance assessment administrators, and National Research Coordinators (NRCs) in each country were asked to review and evaluate the tasks administered in the field trial. The Performance Assessment Committee used their ratings and comments, and data from the field trial, to select and where necessary revise tasks for the main survey.

[2] In this case, the possibility of using data from this item as confirmatory evidence about a country's curriculum was set aside in favor of enabling all students to obtain the data to respond to subsequent items in the task.

6.3.2 SELECTION OF TASKS FOR THE MAIN SURVEY

Following the field trial, the PAC reviewed the field trial results and selected tasks to be administered in the main survey. The appropriateness of tasks for 9- and 13-year-olds was assessed in terms of level of difficulty; which tasks yielded the most information about approaches to problem solving, including common errors and misconceptions; and the ratings of the tasks given by mathematics and science specialists, administrators, and NRCs in each country. The committee also tried to obtain the greatest possible coverage of the performance expectations aspect of the TIMSS curriculum frameworks given the limited time for administration and the limited pool of tasks.

Tasks were characterized by content and context; individual items were also categorized by performance expectations and the prior knowledge and concept understanding required or implied. It was recognized that no item measured only one trait, and that in many tasks both mathematics and science thinking were required. The tasks selected for the main survey are described in section 6.3.3.

The performance assessment tasks selected for the main survey have the properties described below.

- *Difficulty level*. Selected tasks had no more than 50% incorrect or missing responses on early (easier) items and no more than 70% on later (more difficult) items in the field trial. For items used in both populations, these difficulty criteria were applied with Population 2 provided Population 1 students showed some achievement on some of the items of a task.
- *Subject-matter expert ratings.* All tasks selected received ratings of 2 or higher (on a scale of 1-4, 4 = highest) from mathematics and science experts within each country, based on interest, feasibility, content quality, and congruence with curriculum and instruction within the country.
- *Administrators' ratings*. All tasks selected received ratings of 2 or above on a scale of 1-4 (4 = highest) from administrators or NRCs.
- *Balance.* Tasks were selected to maintain a balance between the number of mathematics and science tasks, and between tasks estimated to take 10-15 minutes to complete and those estimated to take 25-30 minutes to complete.
- *Framework representation.* As much as possible, the selected tasks sampled across the subject-matter content aspect of the TIMSS curriculum frameworks. Content coverage was necessarily selective, since only 12 tasks were to be administered in each population. Coverage of all age-relevant performance expectations categories in the frameworks was achieved.
- *Linkage between Populations 1 and 2.* Four complete tasks were identical for the two populations to facilitate comparisons between the two populations.
- *Professional judgment of task quality.* Tasks with lower than 50% correct or partially correct responses in the field trial were retained only if minor revisions would render them more accessible to younger students without destroying their assessment intent, and if they yielded rich information about common approaches to tasks and common errors and misconceptions.

No task was introduced into the main survey that had not been field-tested; three tasks were enhanced following the field trial by the addition of an extra question for Population 2 students.[3]

Tasks were balanced between short problem-solving tasks estimated to take 10-15 minutes to complete and longer, less structured investigation tasks estimated to take 25-30 minutes to complete. In the problem-solving tasks, the procedure to be followed was sometimes specified, while in the investigations students were to make their own decisions about what to measure, how many measurements to make, and how to present data clearly and simply. In both cases students had to describe their strategies and interpret or explain results.

In July 1994, the TIMSS Subject Matter Advisory Committee (SMAC) reviewed and endorsed the PAC's selection of tasks for the main survey. In August 1994, the TIMSS National Research Coordinators approved that selection. The final international versions of the tasks were prepared by the International Study Center and distributed to participants in paper and electronic format.

6.3.3 TASKS SELECTED FOR THE MAIN SURVEY

Science Tasks

Pulse: Students monitor the change in their pulse rate during repeated stepping up on and down from a low bench. They construct their own data table, interpret the results, and invoke prior conceptual knowledge about work, energy, and the circulatory system to develop explanations.

Magnets: Given two magnets of different magnetic strengths, students are asked to develop and describe tests to find out which one is stronger. A variety of magnetic and nonmagnetic objects are available for performing their tests.

Batteries: Students are given a flashlight and four batteries, two of them newly charged and two dead. They develop a strategy for determining which batteries are new and which are worn out, and justify their results by showing understanding of complete circuits and direct current.

Rubber Band: A rubber band with a hook on its lower end is fixed to hang vertically from a clip on a clipboard. Students measure the change in the length of the rubber band as they attach an increasing number of weights to the hook. Students record and tabulate their observations and then interpret them.

Solutions (Population 2 only): Students design an experiment to measure the effect of water temperature on the rate at which tablets dissolve, organize their data in an appropriate format, draw conclusions, evaluate the quality of the experiment, and use concepts and knowledge to hypothesize causes for their findings.

[3] These items were submitted to a limited field trial in the United States.

Containers (Population 1 only): Students use a thermometer to measure the rates of cooling when hot water is poured into containers of different material. They are expected to organize their data, report their conclusions, explain the results in terms of heat transfer, and apply their findings to a problem of "keeping ice cream cold."

Mathematics Tasks

Dice: Students investigate probability by repeatedly rolling a die, applying a computational algorithm, and recording the results. They observe patterns in the data and propose explanations in terms of probability for patterns that emerge.

Calculator: Students perform a set of multiplications with a calculator and observe and record patterns of results. These data allow students to predict the results of further multiplications beyond the scope of the calculator. Population 2 students also had to find factors for a given 3-digit number after first explaining why suggested factors could not be correct.

Folding and Cutting: Students are shown pictures of rectangular shapes with pieces cut out of them. They try to make similar shapes by folding and cutting rectangles of paper. These are evaluated for accuracy of shapes and recognition of the axes of symmetry. Population 2 students completed a fourth item requiring that they predict the axes of symmetry (fold lines) without the use of manipulatives.

Around the Bend: Students use a simulated section of hallway corridor made of cardboard, thin wood, or plastic to determine the dimensions of furniture that can be moved around a bend in the corridor. The furniture is represented by rectangles of varying dimensions cut out of cardboard. The students manipulate the rectangles in an attempt to determine rules about the maximum dimensions and the relationship between the length and width of the furniture that affects whether they will "go around the bend." The task involves understanding scale conversions and right triangle relationships.

Packaging: Students design boxes for packaging four balls by experimenting with drawing boxes of various shapes and their nets.[4] The students then construct the net of a box of actual size to hold the set of four balls.

Combined Science and Mathematics Tasks:

Shadows: A flashlight is attached to the top of a box and directed toward a wall or projection screen from a distance of about 50 cm. A 5 x 5 cm card is on a stand between the wall and the torch, perpendicular to the beam of light and parallel to the wall. Students experiment with the effect of distance on casting shadows by moving the card and measuring the different-sized shadows. They then find positions where the

[4] A net is the two-dimensional pattern of a three-dimensional figure, so drawn that when folded up it will form a box or other three-dimensional figure. The "cover" of the box may be omitted.

shadow is twice the size of the card and construct a general rule to predict when this will be true. The task samples science concepts of light and shadow formation and mathematics concepts of similar triangles and proportion.

Plasticine: Students are given a 20 g standard weight, a 50 g standard weight, and plasticine (modeling clay). Using a simple balance, they devise methods for measuring different amounts of plasticine, record their procedures, and save and label their plasticine samples so that their weight can be verified. In describing their strategies students may use concepts of proportionality or knowledge of alternative number combinations to achieve the desired masses.

The four tasks identical for both populations are Batteries, Dice, Folding and Cutting and Packaging. Tables 6.1, 6.2, and 6.3 present, for each task, the relevant content and performance expectations categories based on the TIMSS curriculum frameworks.

Table 6.1 Science Tasks[5]

Task Name and Label	Content	Performance Expectations
Pulse (S1)	1.2.2 Life processes and systems ■ energy handling 1.2.5 Human biology and health	2.2.3 Applying scientific principles to develop explanations 2.3.3 Data gathering 2.3.4 Organizing and representing data 2.4.4 Interpreting investigational data
Magnets (S2)	1.3.3 Energy and physical processes ■ magnetism	2.3.3 Data gathering 2.4.5 Drawing conclusions from investigational data
Batteries (S3)	1.3.3 Energy and physical processes ■ electricity	2.2.2 Applying scientific principles to solve quantitative problems 2.2.3 Applying scientific principles to develop explanations 2.4.4 Interpreting investigational data 2.4.5 Drawing conclusions from investigational data
Rubber Band (S4)	1.3.1 Matter ■ physical properties of matter: elasticity	2.2.3 Applying scientific principles to develop explanations 2.3.3 Data gathering 2.3.4 Organizing and representing data 2.3.5 Interpreting data (extrapolating) 2.4.4 Interpreting investigational data
Solutions (S5)	1.31. Matter ■ physical properties of matter: solubility 1.3.2 Structure of matter ■ atoms, ions, molecules 1.3.3 Energy and physical processes ■ heat and temperature 1.3.4 Physical transformations ■ physical changes ■ explanations of physical changes	2.3.1 Using equipment 2.2.3 Applying scientific principles to develop explanations 2.3.4 Organizing and representing data 2.4.2 Designing investigations 2.4.3 Conducting investigations 2.4.5 Formulating conclusions from investigational data
Containers (S6)	1.3.1 Matter ■ physical properties of matter: specific heat 1.3.3 Energy and physical processes ■ heat and temperature	2.2.3 Applying scientific principles to develop explanations 2.3.4 Organizing and representing data 2.4.3 Conducting investigations 2.4.4 Interpreting investigational data 2.4.5 Formulating conclusions from investigational data 2.5.2 Sharing scientific information

[5] Number codes refer to the TIMSS curriculum framework. Content subcategories are also shown. See Robitaille, D.F., et al., *Curriculum Frameworks for Mathematics and Science: TIMSS Monograph No. 1*. Vancouver, BC: Pacific Educational Press, 1993.

Table 6.2 Mathematics Tasks[6]

Task Name and Label	Content	Performance Expectations
Dice (M1)	1.1.1 Whole numbers ■ operations 1.7.1 Data representation and analysis 1.7.2 Uncertainty and probability	2.2.2 Performing routine procedures 2.2.3 Performing more complex procedures 2.4.4 Conjecturing 2.5.3 Describing and discussing
Calculator (M2)	1.1.1 Whole numbers ■ meaning ■ operations 1.7.1 Data representation and analysis	2.1.3 Recalling mathematics objects and properties 2.2.1 Use of equipment 2.3.3 Problem solving 2.3.4 Predicting 2.4.5 Justifying 2.5.3 Describing and discussing
Folding & Cutting (M3)	1.4 Geometry: symmetry, congruence and similarity ■ Transformations	2.3.3 Problem solving 2.3.4 Predicting
Around the Bend (M4)	1.2 Measurement ■ units 1.3 Geometry: position, visualization, shape ■ two-dimensional geometry: polygons and circles ■ three-dimensional geometry 1.5 Proportionality ■ problems	2.2.2 Performing routine procedures 2.2.3 Using complex procedures 2.3.3 Problem solving 2.4.3 Generalizing 2.4.4 Conjecturing
Packaging (M5)	1.2. Measurement ■ units 1.3 Geometry: position, visualization, shape ■ three-dimensional geometry	2.1.1 Representing 2.3.3 Problem solving

[6] Number codes refer to the TIMSS curriculum framework. Content subcategories are also shown. See Robitaille, D.F., et al., *Curriculum Frameworks for Mathematics and Science: TIMSS Monograph No. 1*. Vancouver, BC: Pacific Educational Press, 1993.

Table 6.3 Combined Science and Mathematics Tasks[7]

Task Name and Label	Content	Performance Expectations
Shadows (SM1)	**Science categories** 1.3.3 Energy and physical processes ■ light **Mathematics categories** 1.2 Measurement ■ units 1.3.3 Two-dimensional geometry: polygons and circles 1.4 Geometry: symmetry, congruence and similarity 1.5.2 Proportionality problems	**Science categories** 2.2.2 Applying scientific principles to solve quantitative problems 2.3.4 Organizing and representing data 2.4.3 Conducting investigations 2.4.4 Interpreting investigational data 2.4.5 Formulating conclusions from investigational data 2.5.2 Sharing information **Mathematics categories** 2.2.3 Performing complex procedures 2.3.3 Problem solving 2.4.3 Generalizing 2.4.4 Conjecturing 2.5.3 Describing and discussing
Plasticine (SM2)	**Science categories** 1.3.1 Matter ■ physical properties of matter **Mathematics categories** 1.2 Measurement ■ units 1.5 Proportionality ■ concepts ■ problems	**Science categories** 2.2.2 Applying scientific principles to solve quantitative problems 2.3.2 Conducting routine experimental operations 2.5.2 Sharing information **Mathematics categories** 2.2.2 Performing routine procedures 2.3.2 Developing strategy 2.3.3 Problem solving 2.5.3 Describing and discussing

[7] Number codes refer to the TIMSS curriculum framework. Content subcategories are also shown. See Robitaille, D.F., et al., *Curriculum Frameworks for Mathematics and Science: TIMSS Monograph No. 1*. Vancouver, BC: Pacific Educational Press, 1993.

6.4 PERFORMANCE ASSESSMENT DESIGN

The 12 tasks administered during each performance assessment session were presented at 9 different stations. Table 6.4 specifies which tasks students performed at each station.

Table 6.4 Assignment of Tasks to Stations

Station	Task	
A	S1	Pulse
	M1	Dice
B	S2	Magnets
	M2	Calculator
C	SM1	Shadows
D	S3	Batteries
	M3	Folding and Cutting
E	S4	Rubber Band
F	M5	Packaging
G	S5 or S6	Solutions (Population 2) Containers (Population 1)
H	M4	Around the Bend
I	SM2	Plasticine

The assignment of tasks to stations results in three stations with one "short" science and one "short" mathematics task each, two stations with one "long" science task each, two stations with one "long" mathematics task each, and two stations with one combined science/mathematics task each. Each station required about 30 minutes working time. Each student was assigned to three stations, for a total testing time of 90 minutes. Because the complete circuit of nine stations occupies nine students, students for the performance assessment were selected in sets of nine. However, the complete rotation of students required two sets of nine, or eighteen students, to assure that each task was paired with each other task at least once.

6.4.1 SAMPLING SCHOOLS[8]

All TIMSS participants involved in the performance assessment were to sample at least 50 schools from those already selected for the main survey, and a sample of either of 9 or 18 upper-grade students per selected school (lower-grade students were not included in the sample). This yielded a minimum sample of 450 students in the upper grade in each country.

[8] The procedures for sampling schools and students and for assigning schools and students to rotation schemes was developed by Pierre Foy (Statistics Canada) in consultation with the TIMSS Technical Advisory Committee. The procedures are fully explained in the Performance Assessment Administration Manual (TIMSS, 1994).

Schools in the main TIMSS sample could be excluded from subsampling on the basis of the following three criteria.

- Lower-grade schools. The school has students enrolled in the lower grade but not in the upper grade. By design, these schools were not part of the target population for the performance assessment and should be omitted from the school-sampling frame.
- Small schools. The school has fewer than nine students in the upper grade. These schools were excluded because they could not provide a full complement of students for the test sessions.
- Remote schools. The school is in a remote region where it would have been prohibitively expensive to send a fully trained test administrator. Such exclusions were kept to a minimum.

Despite the potential for bias, the TIMSS International Study Center allowed exclusion of schools containing up to 25% of students in the target grade (through reasons of small school size or remoteness), in the interests of improving the quality of administration.

In each country a random subsample of at least 50 schools was drawn from the schools participating in the main survey that were also eligible for the performance assessment. The procedure for selecting the schools for the performance assessment consisted of three steps. The first step was to make a list of all schools selected for the main survey. In the second step those schools to be excluded from the performance assessment were eliminated from the list. In the third step, every *n*th (usually third) eligible school was selected, beginning with a random start.

6.4.2 SAMPLING STUDENTS

Students for the performance assessment were sampled in groups of nine within schools. NRCs could choose to sample one group of nine students from every school, two groups of nine students from every school, or some combination of these (e.g., one group from smaller schools and two groups from larger schools).

In the TIMSS main survey, test booklets 1 to 8 were assigned to students in a manner designed to ensure a uniform and effectively random distribution across the sample. The booklets were assigned within a class at random; it was thus possible to select the students for the performance assessment on the basis of booklet assignment, which would result in a random sample. With this system, students were allocated to the performance assessment sample on the basis of their previously assigned booklet number. If a sample of nine students from a classroom was required, the first student on the Student Tracking Form with booklet 1 was selected first, then the next student with booklet 1, until all booklet 1 students were selected; then the first student with booklet 2, and so on, until 9 (or 18) students were selected (as required).

The Performance Assessment Tracking Form (shown in Figure 6.1) was used to record students selected for the performance assessment. In schools with one group of nine students, the identification information for nine students with the lowest booklet numbers

(beginning with booklet 1) on the Student Tracking Form was transcribed to the Performance Assessment Tracking Form. This group constituted the performance assessment sample for that school. Where two groups of nine were selected, the identification information for the first 18 students with the lowest booklet numbers (beginning with booklet 1) was transcribed to the Performance Assessment Tracking Form and thus constituted the sample for that school. If two classrooms per grade per school were selected for the main survey, then the Student Tracking Forms for both selected upper-grade classrooms were combined and students were selected from both forms. In the example shown in Figure 6.1, two additional students per school were also selected as replacements. These two were simply the next two eligible students on the Student Tracking Form.

6.4.3 ASSIGNING STUDENTS TO STATIONS

Since each student had enough time to visit only three of the nine available stations, a scheme had to be devised for assigning sampled students to stations. The scheme adopted by TIMSS is based on a combination of two partial balanced incomplete block designs[9] (see Table 6.5). It ensures that each task is paired with every other task at least once (but not uniquely), that each station is assigned to approximately the same number of students, and that the order in which students visit stations varies.

After the schools and students were sampled for the performance assessment, each group of students was assigned to either Rotation 1 or Rotation 2, and each student was given a sequence number. The rotation scheme and the sequence number determined which stations each student would attend and in what order. Given that either one or two groups of students could be sampled in a school, the assignment to a rotation scheme and of a sequence number could be done in one of the three ways described below.

- **Two groups per school:** In this case, each selected school provided two groups of nine students. In each selected school, the selected students were alternately assigned to Rotation 1 and Rotation 2. The performance assessment sequence numbers were assigned sequentially from 1 to 9 within each group.
- **One group per school:** Where a single group of nine students was selected per school, the rotation schemes were assigned to alternating schools. Schools were numbered 1 or 2 and rotation schemes were assigned accordingly. Students in the odd-numbered schools were assigned to Rotation 1 and students in the even-numbered schools to Rotation 2. The performance assessment sequence numbers were assigned to the sampled students sequentially from 1 to 9.
- **Combination:** In some countries, some schools provided two groups of nine students and others provided one. Rotation schemes and performance assessment sequence numbers were assigned in one of two ways, depending on the situation in the school, as described above.

In cases where replacement students were necessary, they assumed the sequence numbers of the absent students.

[9] This design was suggested by Edward Haertel of Stanford University.

Table 6.5 shows the stations each student visited according to his/her sequence number. Taken together, Tables 6.4 and 6.5 show the stations each student visited and the tasks completed according to the rotation assignment and sequence number. For example, the student with sequence number 3 participating in Rotation 2 went to stations C, A, and D, in that order. Referring to Table 6.4, you will see that the student completed Tasks SM1 (Shadows); M1 (Dice) and S1 (Pulse); and S3 (Batteries) and M3 (Folding and Cutting).

Table 6.5 Assignment of Students to Stations

Student Sequence Number	Rotation 1 Stations	Rotation 2 Stations
1	A, B, C	A, B, E
2	B, E, D	B, D, G
3	C, F, E	C, A, D
4	D, G, H	D, E, F
5	E, A, G	E, I, H
6	F, H, B	F, H, A
7	G, I, F	G, F, I
8	H, C, I	H, G, C
9	I, D, A	I, C, B

Figure 6.1 presents a completed Performance Assessment Tracking Form. In this example, there is one group of nine students in the school. The school has been assigned to Rotation 1 (column 4) and each student has been assigned a sequence number (column 5). During the performance assessment session, the administrator indicated in column 6 whether or not each student completed each station (A-I). Note also that in this example one student was absent on the day of the performance assessment and was replaced by one of the preselected substitutes. The name of the absent student is crossed out.

Figure 6.1 Example Performance Assessment Tracking Form

Performance Assessment Tracking Form

TIMSS Participant: Germany **Population:** 2 **Stratum:** Hamburg
School Name: Schiller Gymnasium

[a] School ID	[b] Class ID	[c] Class Name	[d] Grade	[e] No. of Students for Perf. Assessment
133	13301	8a	8	9

1]	[2]	[3]	[4]	[5]	[6] Participation Status								
Student Name or number	Student ID	Booklet	Rotation Scheme	Sequence number	A	B	C	D	E	F	G	H	I
DICKMANN G	1330105	1	1	1	√	√	√						
MANN Karl	1330113	1	1	2		√		√	√				
TIMM Bernd	1330122	1	1	3			√		√	√			
ECKHART Mike	1330106	2	1	4				√			√	√	
PECHSTEIN M	1330115	2	1	5	√				√		√		
~~TREUR Jörg~~	1330123	2	1	6									
FRANZKI M	1330107	3	1	7						√	√		√
PELKA Horst	1330116	3	1	8			√					√	√
WOSEGEN B	1330124	3	1	9	√			√					√
GLOCK Michael	1330108	4	1	6		√				√		√	
ROEHL Gisela	1330117	4	1										

Table 6.6, below, summarizes the station assignments, and the tasks at each station for each student sequence number in each rotation plan.

Table 6.6 Assignment of Students to Tasks

	Rotation 1		Rotation 2	
Student Sequence#	**Station**	**Task**	**Station**	**Task**
1	A	S1, M1	A	S1, M1
	B	S2, M2	B	S2, M2
	C	SM1	E	S4
2	B	S2, M2	B	S2, M2
	E	S4	D	S3, M3
	D	S3, M3	G	S5 (Pop 2) or S6 (Pop 1)
3	C	SM1	C	SM1
	F	M5	A	S1, M1
	E	S4	D	S3, M3
4	D	S3, M3	D	S3, M3
	G	S5 (Pop 2) or S6 (Pop 1)	E	S4
	H	M4	F	M5
5	E	S4	E	S4
	A	S1, M1	I	SM2
	G	S5 (Pop 2) or S6 (Pop 1)	H	M4
6	F	M5	F	M5
	H	M4	H	M4
	B	S2, M2	A	S1, M1
7	G	S5 (Pop 2) or S6 (Pop 1)	G	S5 (Pop 2) or S6 (Pop 1)
	I	SM2	F	M5
	F	M5	I	SM2
8	H	M4	H	M4
	C	SM1	G	S5 (Pop 1) or S6(Pop 2)
	I	SM2	C	SM1
9	I	SM2	I	SM2
	D	S3, M3	C	SM1
	A	S1, M1	B	S2, M2

6.4.4 SUMMARY OF SAMPLING OF SCHOOLS AND STUDENTS

The sampling of schools and students, and the rotation procedures ensure that:

- At least 450 students at the upper grade of each population were given the performance assessment in at least 50 schools
- The overall sample size for each population was kept to a minimum
- The testing time for any student did not exceed 90 minutes
- Each task was attempted by at least 150 students in each country

- These students were a subsample of those previously selected for the achievement test component
- There was random allocation of students to tasks
- Each of the 12 tasks was paired with each of the other tasks at least once (that is, completed by the same student)
- Tasks were assigned in such a way as to minimize task interaction effects
- Links can be made to the achievement booklet data.

6.5 ADMINISTRATION PROCEDURES

Specific procedures were established to ensure that the performance assessment was administered in as standardized a manner as possible across countries and schools. The National Research Coordinator in each participating country was responsible for collecting the equipment and materials required for each of the performance assessment tasks, and for assembling a set of materials for each school (in some countries, a set of materials was used for more than one administration). The *Performance Assessment Administration Manual* (TIMSS, 1994) specified the equipment for each task. The tasks were designed to require materials that were easy to obtain and inexpensive. Many of the pieces of "equipment" could be homemade; for example, one task (SM1, Plasticine) required a balance that could be made out of a coat hanger, plastic cups, and string. The *Performance Assessment Administration Manual* provided explicit instructions for setting up the equipment, described which tasks required servicing during administration, and contained instructions for recording information about the materials used that coders could refer to when scoring. In addition, each NRC was invited to a training session on administration of the performance assessment (see Chapter 10) where the materials were demonstrated.

The design for administering the performance assessment required students to move from station to station around a room according to their rotation and sequence numbers to perform the tasks assigned to them. The administrator was responsible for overseeing the activities, keeping time, directing students to their stations, maintaining and replenishing equipment as necessary, and collecting the students' work. The administrator also provided advance instruction on the use of a stopwatch, pointed out any peculiarities about the ruler or thermometer provided, and showed students how to find their pulse. The advance instruction was given only for tasks where the use of the equipment was not what was being measured. Administrators were instructed to provide no instruction on other procedures and to answer no other questions related to the activities required for the tasks.

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

At each station, students performed each assigned task. This involved performing the designated activities, answering questions, and documenting their work in booklets (one booklet per task per student). Students had 30 minutes to work at each station. When students had finished their work at a station (or when time was up), they handed their completed booklets to the administrator. Throughout the administration, the administrator kept track of the time, announced when students should move to the next station on their list, reminded them to hand in their booklets, and made students aware that they had to perform two tasks at some stations.

6.6 CONCLUSION

This chapter has described the development of the TIMSS performance assessment from the development of the tasks to the administration procedures. Given the nature of performance assessment, special attention was paid to developing tasks and procedures that were replicable across administrations and countries and that required materials and resources easily obtainable in each country, while still providing estimates of students' abilities to perform practical hands-on tasks in science and mathematics.

REFERENCES

Archenhold, F., et al. (1988). *Science at Age 15*: *A Review of APU (Assessment of Performance Unit) Survey Findings, 1980-1984*. London: Her Majesty's Stationery Office.

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.

Schofield, B., et al. (1989). *Science at Age 13: A Review of APU (Assessment of Performance Unit) Survey Findings, 1980-1984*. London: Her Majesty's Stationery Office.

Third International Mathematics and Science Study (TIMSS). (1994). *TIMSS Performance Assessment Administration Manual* (Doc. Ref.: ICC 884/NRC 421). Chestnut Hill, MA: Boston College.

Lie, S., Taylor, A., and Harmon, M. (1996) "Scoring Techniques and Criteria" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

7. Scoring Techniques and Criteria

Svein Lie
Alan Taylor
Maryellen Harmon

7.1 OVERVIEW

Free-response items play an important role in the TIMSS test design for measuring student achievement in mathematics and science. While many multiple-choice items effectively measure content and process outcomes over a range of cognitive behavior levels, they give little information about the procedures and thought processes students use in solving problems in mathematics and science. Free-response items are thus intended to supplement multiple-choice items, in an attempt to reflect the complex and multistage processes involved in mathematical and scientific thinking.

Analysis of student responses to free-response achievement items can provide valuable insights into the nature of student knowledge and understanding. The case for including free-response items in the international item pool was made by Taylor (1993a), who also noted the implications for coding reliability and the need for resources. He states:

> The inclusion of free-response items in the international item pool provides an opportunity to collect a rich source of data related not only to levels of student achievement but also to the method used by students in approaching a problem, and to the misconceptions or error types which may be demonstrated by them. Inherent in the collection of these data, however, are issues of reliability and need for additional resources in the coding process (p.1).

The TIMSS tests employed several item formats. These included multiple-choice, short-answer, and extended-response items, as well as performance tasks. The first three item types were included in the written assessment administered to all sampled students; performance tasks, were administered to subsamples of students in Populations 1 and 2. Test blueprints for the written tests allocated approximately 30 percent of testing time to free-response items for each of the three student populations. The distinction between the two free-response categories, short-answer and extended-response, was made mainly for reasons of time; the two did not differ sharply in rationale or information sought. The performance assessment component of TIMSS also required students to provide written answers to the test items in the tasks they completed. These were also considered to be free-response items, and were coded accordingly.

If student responses to free-response items were scored for correctness only, it would be sufficient for the purpose of aggregating results with corresponding multiple-choice items to develop test and subtest scores. But that would yield no information on how students approached problems. TIMSS therefore developed a special coding system that provides diagnostic information in addition to information about the correctness of student responses.

This chapter presents an overview of the development of the TIMSS coding systems for scoring free-response items. The development of the TIMSS two-digit coding system is discussed first, followed by exemplar coding rubrics for several free-response items in the TIMSS tests and for a performance task.

7.2 DEVELOPMENT OF THE TIMSS CODING SYSTEM

The TIMSS coding system was developed over several years, beginning with the early stages of the study when it was recognized that a coding system for both correctness and students' approaches and misconceptions was desirable.

7.2.1 THE 1991 PRE-PILOT STAGE

The first opportunity to code free-response items occurred with the 1991 Pre-Pilot test. The *Pre-Pilot Manual for National Research Coordinators* (Marshall et al., 1991) included directions for the translation of the items, administration of the instruments, data processing, and coding. Items were intended only as samples for the purpose of exploring the free-response scoring methodology. Codes used for free-response items at this stage of the project dealt with three types of information each for mathematics and science.

For mathematics, the three information categories were "answer," "implementation strategy," and "problem-solving strategy." The options within each category were the same across all items. For example, the "answer" category included four options: blank, correct, incorrect, and undetermined. "Implementation strategy," on the other hand, contained the following: no work shown, complete, incomplete, misinterpreted, and not clear. There were

11 options for "problem-solving strategy" (inspired by Polya's well known classification of such strategies), including typical approaches used by students such as systematic list, guess and check, work backwards, and the like.

Rubrics for science, on the other hand, were both generic and specific to each item. They included the number of ways students approached a question, what type of logic they demonstrated, how the approach was specific to the question, and the extent to which an answer was complete. The first two and the last of these were generic, whereas the second was specific to each question. For example, among the generic options under the type of logic were the following: no response, logical and appropriate, logical but not appropriate, not logical, and ambiguous.

In a review of their results from the pre-pilot, the Scandinavian TIMSS groups (Brekke et al., 1992) identified several issues stemming from the free-response codes proposed at that time. Among their recommendations for free-response item coding were the following.

- As a criterion for selecting of free-response items, the time required to code an item should be related to the value of information obtained
- The set of codes for an item should be based on empirical evidence
- Codes should relate to specific answers or strategies rather than generic types or categories of response
- Diagnostic information should in some cases be included in the rubrics themselves, to help coders understand how students might have reasoned
- The coding guide should in many cases give precise examples of responses that belong to a certain code

Information from this review helped to provide direction for the further development of coding rubrics in TIMSS.

7.2.2 THE 1992 PILOT

As a preparation for the subsequent item pilot, a group of countries voluntarily reviewed and rated a number of extended-response items. The selected items were piloted in early 1992 in these countries. Each country then grouped the responses to each task into a response "typology," or category, for classification of the most common responses. Thus, the different countries' typologies could be compared. A rather complicated meta-analysis was undertaken based on these data (Wiley, 1992), with the goal of obtaining codes that could provide rich information on student thinking throughout the world. However, the construction of such codes seemed to be a very complex task, and development along these lines was discontinued.

7.2.3 THE 1993 ITEM PILOT AND REVIEW

As a preparation for the 1993 item pilot, Taylor (1993a) proposed that information gathered from free-response items focus on three aspects of student response: degree of correctness, method or approach, and misconception or error type. The rubric for correctness ranged from zero to the number of score points for an item. Numbers for the other rubrics corresponded to major approaches or error types identified for each question. An illustration of these rubrics is shown below in Table 7.1.

Table 7.1 Coding Rubrics in the 1993 Item Pilot

Degree of Correctness	Method or Approach	Misconception/Error Type
0 - no work leading to answer	0 - no work	0 - no error type
1 - step 1 toward answer	1 - approach 1	1 - error type 1
2 - step 2 toward answer	2 - approach 2	2 - error type 2
x - correct answer	y - approach y	z - error type z
	y+1 - other	z+1 - other

The number of points for each rubric varied by item since each was unique in terms of answer, approach, and types of misconceptions generated. Although these aspects are similar to those used at the final stage of data collection for TIMSS, the design at this point suggested a separate set of codes for each aspect. Taylor proposed that the descriptions of codes within each rubric be based on student responses from the item pilot. He also made suggestions for the composition of coding committees, for training to improve inter-rater reliability, and for feedback on item responses.

Following acceptance by NRCs of the direction proposed by Taylor (1993a), a manual for coding free-response items was developed for use in the 1993 item pilot (Taylor 1993b). The manual included directions for establishing coding committees, training procedures, and coding reports. Instruments included exemplar rubrics for coding, item review forms, mark allocation forms, and correctness rubrics for all free-response items. The national centers coded student responses collected in the item pilot according to the correctness rubrics. In addition, some countries volunteered to report for each free-response item the most common responses, approaches and/or error types. Further, sample student papers were used to develop the coding rubrics.

7.2.4 THE 1994 FIELD TRIAL AND THE NORWEGIAN INITIATIVE

Plans for the 1994 field trial did not include construction of coding rubrics for more than just correctness. Due to a shortage of time and resources, student responses from the field trial were coded by score points only. At this time the Norwegian national center initiated an effort to develop a set of richer coding rubrics for the main survey, in line with the earlier work. Individuals at the TIMSS International Coordinating Center (Alan Taylor, Ed Robeck,

Ann Travers, and Beverley Maxwell) were involved in many discussions that led to the Norwegian proposal.

A series of discussion papers on free-response coding was prepared by Carl Angell and Truls Kobberstad (Angell, 1993; Kobberstad, 1993; Angell & Kobberstad, 1993). The first two papers were prepared for a meeting of the TIMSS Subject Matter Advisory Committee in September 1993, and the third for the October 1993 meeting of the TIMSS NRCs. In these papers it was proposed that a system of two-digit coding be employed for all free-response items. The first digit, ranging between 1 and 3, would be used for a correctness score, and the second digit would relate to the approach used by the student. Numbers between 70 and 79 would be assigned for different categories of incorrect response attempts, while 90 would be used if the student did not respond. The papers also presented a number of exemplar mathematics (Population 2) and physics (Population 3) items. The rubrics were described and applied on student responses. Further, some promising results were reported on inter-rater reliability using this method of coding.

The Subject Matter Advisory Committee supported the proposal for two-digit coding and recommended that an international coding committee be established to develop final versions of coding rubrics prior to final administration of the instruments. Subsequently, in 1994 the International Study Director established the Free-Response Item Coding Committee (FRICC), the purpose of which was to develop coding rubrics for the free-response items in the TIMSS tests. The FRICC included, representatives from 11 countries: Jan Lokan, (Australia); Alan Taylor, (Canada); Peter Weng, (Denmark); Josette Le Coq, (France); Nancy Law, (Hong Kong); Algirdas Zabulionis, (Lithuania); Svein Lie (chair), (Norway); Galina Kovalyova, (Russian Federation); Vladimir Burjan, (Slovak Republic); Kjell Gisselberg, (Sweden); Maryellen Harmon, (USA); Curtis McKnight, (USA); and Senta Raizen, (USA). In addition to the formal members, the following individuals made substantial contributions to the FRICC activities: Truls Kobberstad, Carl Angell, Marit Kjaernsli, and Gard Brekke from Norway, and Anna Hofslagare from Sweden.

7.3 DEVELOPMENT OF THE CODING RUBRICS FOR FREE-RESPONSE ITEMS

In order to capture the richness of the intended information efficiently and reliably, the FRICC established a set of criteria to which the coding rubrics should adhere. The TIMSS rubrics should do the following.

- Permit scoring for correctness and capture the analytical information embedded in student responses.
- Be clear, distinct, readily interpretable, and based on empirical data (student responses obtained from pilot or field trials) so as to account for the most common correct responses, typical errors, and misconceptions.
- Be capable of encoding the adequacy of an explanation, justification, or strategy as well as the frequency with which it is used.

- Be simple, in order to get high reliability and not to impose unreasonable time or resource burdens.
- As far as possible, allow for nuances of language and idiosyncratic features of various countries but avoid being so complex that coders are overwhelmed and tend to limit themselves to a few stereotypical codes.
- Have a number of codes that is not excessive, but sufficient to reduce coding ambiguity to a minimum.

The task of the FRICC was to develop scoring rubrics that could be efficiently and consistently applied, and that were based on empirical evidence in a number of countries. The Norwegian team, on the basis of a detailed review of student responses to items from the field trial in Norway, developed a draft set of rubrics for consideration by the FRICC (Angell et al., 1994). Committee members began their work by analyzing Population 1 and 2 results from the field trial in each of their countries, applying the draft rubrics prepared by researchers at the Norwegian national center (Kjaernsli, Kobberstad & Lie, 1994). In July 1994, the committee arrived at descriptors by response category for each rubric for the items in those tests. The criteria used in the development of the codes and the draft coding rubrics were presented to and approved by the NRCs in August 1994. A number of achievement items were modified following the 1994 field trial, and some of these were administered to a convenience sample in the Scandinavian countries in August 1994. Student responses were then used in the development of coding rubrics for those items.

After the coding rubrics had been developed, the International Study Center assembled the coding manuals for distribution to the participating countries (TIMSS 1995a, 1995b). The manuals included the coding rubrics developed by the FRICC and, for many items, example student responses corresponding to the appropriate codes.

This process was repeated for the Population 3 items in November-December 1994. For this effort, the work of the FRICC was based on draft codes prepared by Vladimir Burjan (advanced mathematics), Carl Angell (physics), Truls Kobberstad (mathematics literacy), and Kjell Gisselberg (science literacy). Again, additional piloting was carried out for modified items in order to ensure that the coding rubrics would represent common student responses, approaches, and misconceptions.

7.4 DEVELOPMENT OF THE CODING RUBRICS FOR THE PERFORMANCE ASSESSMENT TASKS

While the FRICC established the TIMSS two-digit coding system and developed coding rubrics for the free-response items, the Performance Assessment Committee (PAC) collaborated to develop the coding guides for the performance assessment tasks. Led by Maryellen Harmon (United States) and Per Morten Kind (Norway), in 1994 the PAC developed the initial coding rubrics for the performance assessment tasks administered in the performance assessment field trial. Countries participating in the field trial coded the student responses to the items within each of the tasks. The ensuing data were used to

evaluate the tasks for suitability for use in the main survey. These initial coding rubrics served as the basis for the codes developed for the main performance assessment study.

Using the responses from the field trial and from additional piloting of the tasks in Norway and the United States, Harmon and Kind, with the assistance of the PAC, developed rubrics for each item within the tasks selected for the main survey. Like the codes for the free-response items, the codes for the performance assessment tasks were developed to include the common correct responses and the common misconceptions of students. An additional feature of the performance assessment coding rubrics was a set of criteria for what a correct response should include, as well as additional information to aid the coder in evaluating the response.

Following the development of the codes, the International Study Center assembled the *Coding Guide for Performance Assessment* (TIMSS, 1994a), which included the possible codes and examples of the most common responses to each item in all tasks. To facilitate the coding effort in the participating countries, the International Study Center also prepared the *Supplement to the Coding Guide for Performance Assessment* (TIMSS, 1995c). This included a full set of example student responses to all tasks.

7.5 THE NATURE OF FREE-RESPONSE ITEM CODING RUBRICS

The TIMSS coding system is demonstrated in Table 7.2 with a generic example of the coding scheme for a free-response item worth one score point. Actual coding rubrics for actual items are presented later, as is an example of a performance task.

Table 7.2 TIMSS Two-Digit Coding Scheme

Code	Text
10	correct response, answer category/method #1
11	correct response, answer category/method #2
12	correct response, answer category/method #3
19	correct response, some other method used
70	incorrect response, common misconception/error #1
71	incorrect response, common misconception/error #2
76	incorrect response, information in stem repeated
79	incorrect response, some other error made
90	crossed out/erased, illegible, or impossible to interpret
99	blank

Student responses coded as 10, 11, 12, or 19 were correct and earn one score point. The type of response in terms of the approach used or explanation provided is denoted by the second digit. A response coded as 10 demonstrates a correct response of answer type #1 or

method #1. For items worth more than one score point, rubrics were developed to allow partial credit and to describe the approach used or explanation provided.

Student responses coded as 70, 71, 76, or 79 were incorrect and earned zero score points. The second digit in the code represents the type of misconception displayed, incorrect strategy used, or incomplete explanation given. A code of 76 was assigned to an incorrect response in which the student merely repeated information from the item stem. In addition, countries had the option of assigning country-specific codes for correct and incorrect responses in cases where the international rubrics failed to allow for common responses. For the international analyses, the country-specific codes were recoded to 19 and interpreted as "other correct" for items worth one point (to 29 and 39 for items worth two and three points respectively), or to 79 and interpreted as "other incorrect" response.

Student responses coded as 90 or 99 also earned zero score points. A 90 indicates that a student attempted the item but did not provide a coherent response. A 99 indicates that the student did not attempt the item. The differentiation between 90 and 99 allows for the identification of a series of totally blank items towards the end of the test (deemed "not reached") versus items a student has attempted but failed to answer.

The three examples of free-response items shown below illustrate how these rubrics corresponded to specific items and provided diagnostic information on item-specific features. The fourth example demonstrates the application to an item in a performance task.

Figure 7.1 presents a science item and its coding guides. Because this item was administered to all three student populations, the coding rubrics were developed to accommodate a wide range of responses. Correct responses were coded 10, 11, 12, or 13 depending on the type or response or method employed. The most common misconceptions are covered by codes 70 (drinking makes us cool down), 71 (you dry out, particularly in your throat), and 72 (you drink to get energy).

The coding guide for this item allow a detailed study of students' conceptions, at different ages and in different countries, of water balance and temperature regulation of the human body.

Figure 7.1 Exemplar Coding Guides — Thirsty on a Hot Day

O16. Write down the reason why we get thirsty on a hot day and have to drink a lot.

Code	Response
Correct Response	
10	Refers to perspiration and its cooling effect and the need to replace lost water.
11	Refers to perspiration and only replacement of lost water. *Example: Because when we are hot, out body opens the pores on our skin and we lose a lot of salt and liquid.*
12	Refers to perspiration and only its cooling effect.
13	Refers to perspiration only. *Examples: We are sweating.* *Your body gives away much water.* *We are sweating and get drier.*
19	Other acceptable explanation.
Incorrect Response	
70	Refers to body temperature (being too hot) but does not answer why we get thirsty. *Example: You cool down by drinking something cold.*
71	Refers only to drying of the body. *Examples: Your throat/mouth gets dry.* *You get drier.* *The heat dries everything.*
72	Refers to getting more energy by drinking more water. *Example: You get exhausted.*
76	Merely repeats the information in the stem. *Examples: Because it is hot.* *You need water.*
79	Other incorrect: *Example: You loose salt.*
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret
99	BLANK

The mathematics item displayed in Figure 7.2 was administered to students in Population 2. It is a simple equation, the solution of which is straightforward for those who know the algorithm. Calculation errors are covered by codes 70 (clear indication of student confusing addition and subtraction) and 71 (other calculation errors), whereas code 72 covers responses that reach no numeric solution for x.

With this set of codes, in spite of its simplicity, one can analyze not only students' knowledge of and ability to apply the algorithm for solving a linear equation, but also the frequency of the most common errors.

Figure 7.2 Exemplar Coding Guides — Solve for X

L16. Find x if $10x - 15 = x + 20$

Answer: _____

Code	Response
Correct Response	
10	7
Incorrect Response	
70	1 OR 2.33 OR 3
71	Other incorrect numeric answers.
72	Any expression or equation containing *x*.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret
99	BLANK

Figure 7.3 presents the set of codes for a science item that deals with the conservation of mass during melting. This item was administered to students in Population 2. One score point is given for a correct answer (no change of mass), and two score points are given for an explanation that refers to the principle of conservation of mass. Among the incorrect responses there are codes for increase of mass (70 and 71) and decrease of mass (72 and 73). Further, the codes allow us to compare countries on how many responses include any explanation, and to determine whether a response is fully correct (code 20) or not (codes 10, 70, or 72).

Figure 7.3 Exemplar Coding Guides — Melting Ice Cubes

A glass of water with ice cubes in it has a mass of 300 grams. What will the mass be immediately after the ice has melted? Explain your answer.

Note: For this question do not distinguish if the student substitutes kg for g: that is, accept 300 kg as the same as 300 g.

Code	Response
Correct Response	
20	300 g with a good explanation. *Examples: 300 g. The ice changes into the same amount of water.* *The same. The ice only melts.* *The same weight. Nothing disappears.*
Partial Response	
10	300 g. Explanation is inadequate.
11	300 g. No explanation.
Incorrect Response	
70	More than 300 grams with explanation. *Examples: More. Water has higher density.* *More. Water is heavier than ice.*
71	More than 300 g. No explanation.
72	Less than 300 g. With explanation. *Examples: Less. Ice is heavier than water.* *Less. There will be water only.*
73	Less than 300 g. No explanation.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret
99	BLANK

Figure 7.4 presents the possible codes for one item in the performance assessment task Solutions, administered to Population 2 students. The TIMSS two-digit coding system is the basis for the coding rubric. Coders were also provided with a list of criteria for a complete response. The solutions task required students to investigate what effect different temperatures have on the speed with which a soluble tablet dissolves. Students were required to develop and record their plan for an experiment to investigate this, carry out their proposed tests on the tablets and record their measurements in a table. They were asked to explain what effect different temperatures have on the speed with which a soluble tablet dissolves, according to their investigation. The following coding guide was used to score students responses to this.

Figure 7.4 Exemplar Coding Guides – Item 3, Solutions Performance Assessment Task

Q3.	According to your investigation, what effect do different temperatures have on the speed with which a tablet dissolves?

Criteria for a complete response:

i) Conclusion must be consistent with data table or other presentation of data (graph or text).
ii) Conclusion must describe the relationship presented in the data.

NOTE:

- A wrong direction in the data has been coded for in Q2. However, if an anomaly has occurred and the student recognizes it as such and identifies it as such s/he should receive credit in this question.
- If the student says that temperature has no effect on rate of solution, and if this conclusion is consistent with the student's data, code 29.

Code	Response
Complete Response	
20	Correctly describes trend in the data. *Example: As the temperature increases the tablet dissolve faster.*
21	Describes explicitly only what happens in hot or in cold water but not both. *Examples: In hot water the tablet dissolves faster.* *In cold water the tablet dissolves more slowly.* *In hot water the tablet dissolves twice as fast.*
29	Other complete summaries of data.
Partially Correct Response	
10	Describes trend in the data but the temperatures are not in a reasonable range and student fails to recognize or account for this.
19	Other partially correct.
Incorrect Response	
70	Conclusion not consistent with student's data, and no explanation of the inconsistency offered.
71	Mentions that temperature has an effect but does not describe the effect. *Example: The temperature has a big effect.*
72	Conclusion erroneous: that is, temperature "does not affect rate of solution." *Example: All temperatures have the same effect.*
76	Repeats data but does not draw a conclusion or generalization.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

In performance assessment almost no item measures a single trait. For example, in order to answer many of the questions, students had to recall and synthesize two or more concepts and use a number of different skills. In addition, items within a task were interdependent both in the sense of being clustered around a common investigatory question (although calling for various skills) and in the sense that some responses depended on data collected and analyzed in previous responses. This richness within a single task or scenario is characteristic of "authentic" problems, and was intentionally structured into the tasks, even though it would render interpretation of results complex and difficult. The revised coding system attempts to reduce these levels of complexity to quantifiable, interpretable data.

In the coding example above, compromises had to be made between expanding the number of codes to capture additional alternative approaches and/or misconceptions, and limiting the coding time per item. Therefore not all possible responses were included in the codes. Decisions about which codes to include were based on empirical data: an alternative approach or an error had to have been made in at least 5% of the field trial responses to be included in the final set of codes.

The similarities in approach and application of the coding systems for performance assessment and free-response items does not imply that the two genres are equal in difficulty for coders. In fact, because of the complexity of measuring several entangled traits simultaneously, it is essential that those who code performance assessment tasks have actually done the tasks themselves or observed students doing these tasks. This is necessary for coders to understand fully what the task is intended to measure, the functioning of equipment, and possible "alternative" perceptions of the tasks by students.

7.6 SUMMARY

In this chapter we have discussed the importance of free-response items and explained how the TIMSS coding guides are used to collect information on how students respond to these items. To illustrate the application of the rubrics to actual TIMSS items, and to demonstrate the potential for analysis, some specific examples were given.

To provide a context for the rubrics, a historical overview of their conception and development was included. Given the interest and expectation from the early stages of the study, it was desired that the information gathered via the free-response items not be limited to correctness only. As a result, coding rubrics were designed to measure three aspects of student response: correctness, method or approach or type of explanation/example given, and misconception or error-type. Through use of a two-digit system it was possible to collect information on all of these aspects.

The analysis of data collected for free-response items will answer several questions of interest, in addition to their contribution to the correctness score. Analyses of students'

approaches and conceptions around the world will be of great interest to researchers in mathematics and science education. Furthermore, such data can provide valuable diagnostic information for mathematics and science teachers. We hope that not only the data themselves, but also the methods of analysis that have been briefly described here, will turn out to be useful tools for a better understanding of student thinking in science and mathematics.

REFERENCES

Angell, C. and Kobberstad, T. (1993). *Coding Rubrics for Free-Response Items* (Doc. Ref.: ICC800/NRC360). Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Angell, C. (1993). *Coding Rubrics for Short-answer and Extended-response Items*. Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Angell, C., Brekke, G., Gjortz, T., Kjaernsli, M., Kobberstad, T., and Lie S. (1994). *Experience with Coding Rubrics for Free-Response Items* (Doc. Ref.: ICC867). Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Brekke, G., Kjaernsli, M., Lie, S., Gisselberg, K., Wester-Wedman, A., Prien, B., and Weng P. (1992). *The TIMSS Pre-Pilot Test: Experience, Critical Comments and Recommendations from Norway, Sweden, and Denmark* (Doc. Ref.: ICC310/NPC083). Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Kjaernsli, M., Kobberstad, T., and Lie S. (1994). *Draft Free-response Coding Rubrics–Populations 1 and 2* (Doc. Ref.: ICC864). Document prepared for the Third International Mathematics and Science Study (TIMSS).

Kobberstad, T. (1993). Discussion Paper for the Third International Mathematics and Science Study Subject Matter Advisory Committee Meeting, September 1993.

Marshall, M., Koe, C., and Donn, S. (1991). *Pre-Pilot Manual for National Project Coordinators* (Doc. Ref.: ICC158/NPC032). Prepared for the Third International Mathematics and Science Study (TIMSS).

Taylor, A. (1993a). *Coding Rubrics for Free-Response Items* (Doc. Ref.: ICC648/NPC249). Discussion Paper for the Third International Mathematics and Science Study National Project Coordinators Meeting, March 1993.

Taylor, A. (1993b). *Manual for Coding Free-Response (Open-Ended) Items for Achievement Review* (Doc. Ref.: ICC662/NPC260). Document prepared for the Third International Mathematics and Science Study (TIMSS).

Third International Mathematics and Science Study (TIMSS). (1994a). *Coding Guide for Performance Assessment* (Doc. Ref.: ICC885/NRC422). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994b). *Coding Rubrics for Free-Response Items.* Paper prepared by the TIMSS Free-Response Item Coding Committee for the National Research Coordinators Meeting, August, 1994.

Third International Mathematics and Science Study (TIMSS). (1994c). *Manual for Coding Free-Response Items–Population 3 Field Trial* (Doc. Ref.: ICC844/NRC397). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995a). *Coding Guide for Free-Response Items–Populations 1 and 2* (Doc. Ref.: ICC897/NRC433). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995b). *Coding Guide for Free-Response Items–Population 3* (Doc. Ref.: ICC913/NRC446). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995c). *Supplement to the Coding Guide for Performance Assessment* (Doc. Ref.: ICC933/NRC456). Chestnut Hill: Boston College.

Wiley, D. (1992). *Response Typologies From the First TIMSS Achievement Pilot: Implications for Scoring Extended-Response Tasks*. Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Maxwell, B. (1996) "Translation and Cultural Adaptation of the Survey Instruments" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

8. Translation and Cultural Adaptation of the Survey Instruments

Beverley Maxwell

8.1 OVERVIEW

Although the international set of TIMSS instruments was prepared and distributed in English, the 45 countries participating in TIMSS represent 31 different languages, and the achievement booklets had to be translated into each of those languages. Because the inherent risk of error or inequity in the translations was obvious, translation validity was an issue from the very beginnings of TIMSS. Detailed guidelines for producing translations, the enlistment of qualified translators, and careful procedures for checking translations have resulted in high-quality instruments for all countries.

The first stage of ensuring high-quality of translations was to identify the most appropriate translation procedures for the study. In 1992, TIMSS commissioned Ronald K. Hambleton (University of Massachusetts, Amherst) to write a paper on translation procedures for international achievement instruments. That paper, *Translating Achievement Tests for Use in Cross-National Studies* (Hambleton, 1992), was the basis for the translation guidelines provided to the National Research Coordinators (NRCs) for the 1993 item pilot, 1994 field trial, and main survey in 1994-1995. In his paper, Hambleton acknowledges that the process in a study such as TIMSS is not merely test translation, but test adaptation:

> Some researchers prefer the term test adaptation to test translation because the former term seems to more accurately reflect the process that often takes place: Producing an equivalent test in a second language or culture often involves not only a translation that preserves the original test meaning, but also additional changes such as those affecting item format and testing procedures. [Such

> changes] may be necessary to insure the equivalence of the versions of the test in multiple languages or cultures (pp. 3-4).

The TIMSS procedures addressed all aspects of test adaptation and applied to all participating national centers. Many of the adaptations that were necessary to produce equitable items, such as changing proper nouns and units of measure, were important to all countries, regardless of the ultimate language of the test. Throughout the study, the English-speaking countries followed the same translation guidelines and verification procedures as countries producing instruments in other languages. In all of the operations manuals, the section entitled "Guidelines for Translation and Cultural Adaptation," specifies that the guidelines include English-speaking countries.

Hambleton emphasizes the need for care in translation and for ensuring equivalence:

> Unless the translation work is done well, and evidence is compiled to establish, in some sense, the equivalence of the two versions of the test, questions about the validity of the translated tests will arise. Also, the validity of comparisons among countries where different versions of the test have been administered will be in doubt until questions about the equivalence of the versions are resolved (p. 3).

Each national center was responsible for producing the instruments used in that country. The national centers had access to the best translators for the task and were familiar with the resources available within their educational systems. They could select personnel whose first language was the language into which the test was to be translated, and who had a good knowledge of the subject matter and age-appropriate language. These translations were then reviewed centrally through the International Coordinating Center (ICC) by independent certified translators who were not involved in any other TIMSS activities. Their evaluations were provided to the respective national centers to allow them to make any necessary corrections before the booklets were administered.

8.2 TRANSLATING THE TIMSS ACHIEVEMENT TESTS

Depending on their participation in pilot and field trial activities, national centers produced up to three translations of test booklets:

1. Booklets for the Population 1 and 2 item pilot in 1993
2. Booklets for the Population 1, 2, and 3 field trial in 1994
3. Booklets for the Population 1, 2, and 3 main survey in 1994-95.

Each translation was verified item by item by an independent agency. (The verification procedure is described in detail in a section below.) In the 1993 item pilot and the 1994

field trial, verification followed the administration of the instruments, because there was insufficient time between translation and administration.[1]

In the main survey, verification preceded administration. Once the booklets for a country were translated, a master copy was sent to the ICC. All items were checked by certified translators, as were the Population 2 performance assessment student worksheets.[2] For each country submission, the items and booklets were reviewed, and the results (including suggested changes in the translations) were returned to the country. Errors or deviations discovered during translation verification were, in most cases, communicated to NRCs in time for them to make corrections before test administration. This was especially effective in minimizing errors, and was certainly an important element in ensuring and confirming good translations. In cases where the national center submitted the student test booklets for translation verification only after they were administered, the report could still be used to help resolve data anomalies, and as a *post hoc* confirmation of the quality of the booklets.

As a final confirmation, each country's Quality Control Monitor (an independent reviewer of the implementation of the TIMSS procedures, described in Chapter 11) received a copy of the translation verification report and checked that the appropriate changes had been made.

8.3 TRANSLATION PROCEDURES AT THE NATIONAL CENTERS

The recommended translation procedures applied to the item pilot, field trial, and main survey, and called for multiple-forward translations of all the test items. NRCs were asked to have the instruments translated by more than one translator and to compare the results. The expectation was that a pair of independent translations would be the same for most items. For any item where the two translations differed, the differences would be discussed, and the best translation of the item selected for the test instrument. The guidelines identified five characteristics of an appropriate translator:

- A good knowledge of English
- An excellent knowledge of the target language
- Experience in both languages and cultures
- Experience with students of the target populations
- Skills in test development

---

[1] In the 1993 item pilot, the translators reviewed the test items after the tests had been administered, the results analyzed, and the desirable items identified. Only the "surviving" items were reviewed. Each participating NRC received a report on the quality of the translation of each of those items. Many of those items would be included in the 1994 field trial, so the NRC had an opportunity to improve or correct the translation before it was administered again. In the 1994 field trial, reviewing all items for all countries would have been too costly, and a compromise position was taken. If a country's item pilot translation had been good, only a 25% sample of items was reviewed, to check that the same quality existed in the field trial instruments. If the item pilot translation had not been good, or if the country had not submitted item pilot instruments, all of the items were reviewed. The results of these reviews contributed to the quality of the main survey instruments.

[2] Population 1 tasks were not verified due to limited resources. In fact, most items were very similar to those used for Population 2, and the results for the Population 2 review could be applied to Population 1 tasks.

Although multiple translations are an excellent safeguard for validity, time and budget constraints precluded these in some countries. However, by choosing competent translators, all countries were able to produce good translations. Subsequent detailed independent evaluations of the translations confirmed the quality of the instruments. The guidelines for the translators emphasized the following operations:

- Identifying and minimizing cultural differences
- Finding equivalent words and phrases
- Making sure the reading level is the same in the target language as in the original English version
- Making sure the essential meaning does not change
- Making sure the difficulty level of achievement items does not change
- Being aware of changes in layout due to translation.

In the item development stage of TIMSS, extensive review and selection went on to ensure that the items did not introduce a cultural bias to the tests. For example, an item that requires knowledge of the rules of baseball is not acceptable in an international test. Notwithstanding this scrutiny, many changes in the questions were required from country to country because of differences in culture. Thus, measurement units, seasons, names of people, places, animals, plants, currencies, and the like were adapted to be equally familiar to all students, insofar as this is possible.

Concepts and conventions that were not common to all cultures and were not related to the substance of the question were changed. For example, a graph that showed winter clothing sales by month of the year, with increasing sales in November-December and declining sales in July-August, was sensible in the Northern Hemisphere. In the Southern Hemisphere, such a graph makes less sense. To adapt the question, one country changed "coats and sweaters" to "shorts and tee-shirts," whereas another changed the month names on the graph. Both were good adaptations for a Southern Hemisphere climate. The meaning of the item, rather than the exact wording, was translated.

Changes in proper nouns were a necessary adaptation for many countries. These included the names of people, cities, and official titles. Changes in common nouns were also necessary, to ensure that children were equally familiar with the vocabulary and topic of the question. For example, in a question about vertebrates and invertebrates, a land-locked country replaced "clam" with "snail," and a seaside country replaced "crayfish" with "shrimp." In another item, where students were asked to interpret a diagram of a food web, various small animals were selected to replace the skunk shown in the English-language version.

Questions involving money were adapted in three ways. In the source instruments, the currency used was dollars. Where it was sensible to do so, the notation was directly translated into the local currency, without changing the value or the context of the item (for

example, a $20 train ticket could be changed to a £20 ticket). In some cases, simply changing the currency resulted in inappropriate values, so the item being purchased was also changed. Finally, for some questions, some translators retained the dollar currency because there was no easy adaptation. This was limited to a few countries where students were generally familiar with the dollar as a foreign currency and where it would not affect the difficulty level of the question.

Most of the adaptations in mathematics and science notation and units of measure are generally accepted. The most straightforward adaptations were in the form of decimal notation, place value notation, and time (use of the colon or period, and use of the 12- or 24-hour clock). The test consistently used metric units of measure; however, if the context allowed, imperial measure could be substituted. This was acceptable only when the values did not also need to be changed. For example, it is acceptable to change "six bags of flour, each weighing 10 kg," to "six bags of flour, each weighing 10 lb," but not to "six bags of flour, each weighing 22 lb."

In addition to their own experience and good sense, translators had two resources to inform their decisions. First, the guidelines for translators were explicit about the objectives of cultural adaptations, and provided examples of good and poor changes. Table 8.1 displays the actual examples of appropriate adaptations provided to translators. The second resource was the team of subject-matter and evaluation specialists coordinated by the International Study Center. Translators could refer uncertainties about translation to the International Study Center or the ICC. In such cases, the query was directed to the appropriate person, and a recommendation returned quickly.

Table 8.1 Examples of Acceptable Cultural Adaptations

Class of Change	Specific Change from	Specific Change to
Punctuation or Notation	decimal point place value comma	decimal comma space
Units	centimeters liters ml	inches quarts mL
Proper nouns	Ottawa Mary	Oslo Maria
Common nouns	robin elevator	kiwi lift
Spelling	center	centre
Verbs (not related to content)	skiing	sailing
Usage	Bunsen burner	hot plate

The layout of some questions also needed to be adapted in some countries. If the translated text differed in length from the English original, additional lines of text were inserted without changing the pagination of the items. This was possible because the English layout left substantial space between items. If additional space was required for an

item, NRCs could use the white space between items on that page or reduce the spacing between the lines of text.

In some items, especially questions in the negative, a word was emphasized by using all capital letters. For example, in "Which animal is NOT a mammal?", the negative is emphasized by capitalizing the word "not." In languages where this format would not be possible or meaningful to children, it was recommended that the word be emphasized in some other way. This was usually boldface, underlining, or italics. The objective was to ensure that the student not overlook important words or the negative form of a question.

The translators maintained records of each adaptation made in translating the achievement instruments. Before completing the translation, this information was forwarded to the International Study Center, which obtained a ruling on the appropriateness of each adaptation from a subject-matter specialist. Upon completion of the instruments, NRCs were instructed to compare them item by item with the English originals. The guidelines directed NRCs to check that the following conditions were met:

- All items were present in the correct order
- There were no misplaced graphics, incomplete texts, or incorrect options
- The translations were inserted precisely (correct spelling, no missing words)
- All variable names were correct and in order
- The graphics were printed correctly, especially those containing shading that was significant to the solution of the item.

After this comparison, the instruments were submitted for independent translation verification.

8.4 VERIFYING THE TRANSLATIONS

There are four types of procedures for verifying translations: multiple-forward translation, back-translation, translation review by bilingual judges, and statistical review. In TIMSS, at least two and usually three of these procedures were used.

- Multiple-forward translation. This form of verification was carried out in the individual national centers. As mentioned above, NRCs were asked to obtain multiple independent translations of the instruments, followed by an item by item comparison.
- Back-translation. Back-translation is a three-step procedure. The test is translated from English into the target language; a different translator translates that version back into English, and finally an English-speaking person compares the original test with the back-translation. This procedure was not used in TIMSS for a number of reasons. First, it would have exceeded the resources of most national centers. Additionally, the procedure can obscure significant flaws in the translated instrument.[3] Finally, "the back translator [may be] able to do a good translation even though the original translation

[3] For example, in the question, "What does a carnivore eat?", the word "carnivore" would read "meat-eater" in many translations, making the questions very much easier. But if in the back-translation "meat-eater" was translated back to "carnivore," one would not know about the flawed original translation.

was poorly done and resulted in a non-equivalent target language version of the test " (Hambleton, 1992, p. 14).

- Translation review by bilingual judges. This may be considered as a variation of back-translation; however, unlike that procedure, this focuses on both the target and the source language. This procedure was used in all stages of TIMSS. It was favored because in addition to checking the accuracy of the translations per se, it allowed checking cultural adaptations and comparison of the levels of reading difficulty.
- Statistical review. In both the item pilot and the field trial, NRCs were provided with item statistics for their sampled populations. Anomalies in the results were flagged, so that NRCs could check suspect items for translating or printing errors.

The ICC enlisted a professional translation agency in Vancouver, Canada, to select the personnel for verifying the translated TIMSS achievement instruments. The criteria for translators included:

- Formal credentials as a translator into the target language
- First-language experience in the target language
- Excellent knowledge of English
- Experience living and working in an English-language environment
- Familiarity with the culture associated with the target language.

Because of Canada's multicultural history, it was usually possible to engage translators who had immigrated to Canada from countries involved in TIMSS and who had experience in both cultures. Most of the translators lived in the Greater Vancouver area; the rest were located in other Canadian cities.

For verification of the main survey translation, each of these "verifiers" was provided with a package containing the following materials.

- A two-page introduction summarizing the TIMSS project, the instruments, and the translation goals, as background information
- A set of the translated instruments (as either assembled booklets or item clusters)
- A set of the international versions of the instruments
- A copy of "Guidelines for Translation and Cultural Adaptation" (an excerpt from the *Survey Operations Manual* (TIMSS, 1994a, 1994b), containing the original instructions for translating the instruments; this allowed the verifier to know what instructions were given to the original translator)
- Instructions for verifying the general layout (checking that the message to students appeared at the beginning of the book, the questions appeared in the correct order, the illustrations were in the right place, all labels were translated, and page breaks were the same as in the international versions)
- Instructions for verifying the message to students (a list of points that the message must have clearly communicated)

- Instructions for item-by-item checking (including the procedures for coding observations to indicate the type and severity of the error)
- An example of a verified translation, including an annotated verifier's report.

After checking the general layout and the message to students, the verifiers compared each item with its international version. If the translated item was judged equivalent to the international version, no observation was made in the verifier's report. If it differed in any way from the original, an observation was made, composed of a *severity* code, a *type* code, and an explanation.

The severity code ranged from 1 (serious error) to 4 (acceptable adaptation).

1 – Major Change or Error: This could affect the results. Examples include incorrect ordering of choices in a multiple-choice item; omission of a graph that is essential to a solution; an incorrect translation of text such that the answer is indicated by the question.

2 – Minor Change or Error: This should be corrected if possible, but will not affect the results. Examples include spelling errors that do not affect comprehension; misalignment of margins or tabs; incorrect font or font size.

3 – Suggestions for Alternative: The translation may be adequate, but the verifier suggests a different wording for the item. The NRC would be asked to review such suggestions and decide whether to make the suggested changes.

4 – Acceptable Changes: The verifier identifies changes that are acceptable and appropriate adaptations. This is done to provide information and requires no action from the NRC. An example is where a reference to winter is changed from January to July for the Southern Hemisphere.

Type codes allowed the verifier to use a "shorthand" for indicating the type of adaptation in addition to its severity. Codes A through J were used for text, and K through N for graphics and layout. For example, an appropriate change in vocabulary (coyote to dingo) would be coded as 4-C. An inappropriate change (gravity to weight) would be coded as 1-C. In cases where the verifier was unsure about the coding, a question mark was used in place of a code, and the uncertainty was elaborated in the explanation.

The type codes are:

A. Spelling
B. Grammar
C. Vocabulary
D. Incorrect number or value
E. Error in equation or numeric notation
F. Missing or additional text

G. Change in meaning
H. Change in level of reading difficulty
I. Tabs, alignment, or text layout
J. Other problem with the text
K. Labels are missing
L. Wrong picture or picture is missing
M. Picture has been modified
N. Labels have been modified.

The verifiers' reports consisted of an overall statement of the quality of the translation, followed by a list of observations associated with individual items. The reports were sent to the ICC, where they were reviewed and subsequently forwarded to the International Study Center and the appropriate NRC. It became apparent that two features of the coding greatly facilitated the review of the reports. First, for a report with numerous observations, the frequency of each severity code provided a quick indication of which and how many items required immediate attention. This was useful for the NRC, and for the quality control monitor responsible for checking that the report recommendations had been followed. Second, in some cases the observation consisted of an alternative translation without explanation. For English-speaking reviewers at the ICC and International Study Center, the severity and type codes were necessary for understanding the nature of the observation.

Finally, the translation verification reports contributed to understanding the initial analyses of the achievement data. The IEA Data Processing Center in Hamburg, Germany, received each country's data files following administration and data entry. As those files were cleaned, several routines were performed to check for anomalous data. During this process, the translation verification reports were consulted for possible explanations for the anomalies. And, as the International Study Center staff reviewed the item statistics the translation reports were reviewed.

The procedures for verifying translations in the TIMSS study were highly effective. In most cases they confirmed that the national centers had produced high quality translations; in other cases they alerted the centers to flaws in translations in time to make changes. As a serendipitous outcome of the procedures, the careful documentation of acceptable and unacceptable adaptations will be a useful resource for researchers developing guidelines and procedures in subsequent studies.

REFERENCES

Hambleton, R. (1992). *Translating Achievement Tests for Use in Cross-National Studies* (Doc. Ref.: ICC454/NRC127). Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Third International Mathematics and Science Study (TIMSS). (1994). *Survey Operations Manual–Populations 1 and 2* (Doc. Ref.: ICC889/NRC425). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994). *Survey Operations Manual–Population 3* (Doc. Ref.: ICC907/NRC440). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Schleicher, A. and Siniscalco, M.T. (1996) "Field Operations" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

9. Field Operations

Andreas Schleicher
Maria Teresa Siniscalco

9.1 OVERVIEW

The TIMSS field operations comprised the activities and responsibilities of National Research Coordinators (NRCs), School Coordinators, and Test Administrators for the execution of the study in their countries. In particular, the NRC was responsible for:

- Translating and preparing the test instruments
- Selecting the sample of schools
- Obtaining cooperation from the sampled schools

- Identifying and instructing School Coordinators and Test Administrators
- Sampling and tracking classes and students within schools and identifying their teachers
- Assigning test instruments and questionnaires to students and teachers
- Preparing, packaging, and sending assessment materials to the schools
- Scoring the free-response items
- Arranging for data entry and data management, and verifying the data and the data collection materials
- Returning the verified data files to the IEA Data Processing Center along with a report on the survey activities.

Under the direction of the NRC, the School Coordinators and Test Administrators were responsible for:

- Organizing the testing sessions and preparing the testing materials for administration
- Returning the assessment materials to the national center after data collection.

The TIMSS field operations were designed by the IEA Data Processing Center (DPC) in Hamburg in cooperation with Statistics Canada and the International Study Center, to ensure that high-quality, comparable data would be available for the analyses. The TIMSS field operations were based on procedures used successfully in previous IEA studies and enhanced on the basis of experiences in the TIMSS field trial.

To facilitate the national field operations, NRCs were provided with software systems for within-school sampling, preparation of survey tracking forms, and data entry and verification. Quality control measures were implemented throughout all phases of the main study.

This chapter briefly describes the procedures for sampling schools, outlines the rationale for the field operations, describes the procedures for sampling and tracking classes, teachers, and students, and summarizes the other main steps undertaken between the first contact with the selected schools and the verification of data files and materials after data collection. The procedures for the translation and adaptation of the instruments are described in Chapter 8, and those for scoring free-response items are described in Chapter 7.

9.2 DOCUMENTATION

NRCs were provided with the following manuals detailing the procedures for carrying out the TIMSS field operations.

- The *Survey Operations Manuals* (TIMSS, 1994g, 1994h) describes the activities and responsibilities of NRCs from the moment the testing materials arrived at the national center to the moment the cleaned data files and accompanying documentation were submitted to the IEA Data Processing Center. It includes instructions for using the computer programs provided by the DPC.
- The *Sampling Manual* (TIMSS, 1994d) defines the TIMSS target populations and sampling goals and describes the procedures for the sampling of schools.

- The *School Coordinator Manuals* (TIMSS, 1994e, 1994f) describes the activities of the School Coordinator from the time the survey tracking forms and testing materials arrived at the school to the time the completed testing materials were returned to the national center.
- The *Test Administrator Manual* (TIMSS, 1994i) covers the procedures from the beginning of testing to the return of the testing materials and completed Student Tracking Forms to the School Coordinator.
- The *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995c) provides information to enable the coding and data entry personnel in the national centers to code, enter, and verify the data.
- The *International Codebooks* (TIMSS, 1994b, 1995d) define the variables and file formats in the data files and assisted NRCs in the analysis of their data.
- The *Coding Guide for Free-Response Items* (TIMSS, 1995a, 1995b) contains instructions for scoring the short-answer and extended-response test items.
- The *Performance Assessment Administration Manual* (TIMSS, 1994c) provides instructions for sampling and administering the TIMSS performance assessment.
- The *Coding Guide for Performance Assessment* (TIMSS, 1994a) contains instructions and coding guides for scoring the performance assessment tasks.

Two software packages were supplied by the IEA Data Processing Center to assist NRCs in the main study:

- The DATAENTRYMANAGER, a computer program for data entry and data verification
- The field operations software, designed to help NRCs select the within-school sample, prepare the Survey Tracking Forms, and assign the test booklets to students.

In addition to the manuals and software, NRCs received hands-on training in the procedures and use of the software from staff from the International Study Center, the IEA Data Processing Center, and Statistics Canada.

9.3 SELECTING THE SCHOOL SAMPLE

The procedure for sampling schools is presented in Chapter 4. This chapter describes the within-school sampling procedures. To avoid nonresponse bias as much as possible, it was important to get maximum cooperation from the schools in the sample. After the sample of schools was drawn, the schools were contacted, with the permission of the relevant authorities. They received a letter describing the goals of TIMSS, requesting their cooperation, outlining what the school's participation would involve and the benefits of participation, and arranging for the appointment of a School Coordinator and the necessary further contacts. The procedures are described in detail in the *Survey Operations Manuals* (TIMSS, 1994g, 1994h).

9.4 IMPLICATIONS OF THE TIMSS DESIGN FOR WITHIN-SCHOOL FIELD OPERATIONS

The within-school sampling procedures were based on the TIMSS study design and took into account the structure of national education systems as well as various national

administrative and economic constraints. The features of the TIMSS study design that influenced within-school sampling and tracking of students, classes, and teachers are described below.

- For Populations 1 and 2, the study design anticipated relational analyses between student achievement and teacher-level data at the class level. For field operations, this meant that intact classes had to be sampled, and that for each sampled class the mathematics and science teachers had to be tracked and linked to their students. The linkage between teachers and students was established with two tracking forms, the Teacher-Student Linkage Form and the Teacher Tracking Form. Each teacher was linked to the class(es) he or she taught by a seven-digit identification code composed of the three-digit School ID, plus a two-digit sequential identification of the teacher within the school, plus a two-digit number, referred to as the "Teacher Link Number," which uniquely identified each occurrence of a teacher in the Teacher Tracking Form. Each occurrence represented a different linkage of a teacher with a class.
- For Population 3, no teacher-level data were obtained, and therefore no teacher-student linkage was required. In terms of procedures, this meant that the most efficient approach to within-school sampling was to take a simple random sample of students; fewer forms were needed to sample and track students within schools; and no teacher-level questionnaires were administered.
- For purposes of parameter estimation, Population 3 should be thought of as consisting of three populations: the general population (i.e., all students at this level), and two subpopulations–the students at this level taking advanced courses in mathematics, and those taking courses in physics. The two subpopulations form overlapping subsets of the general population. However, for the purpose of sampling operations, it was necessary to separate Population 3 students into four nonoverlapping groups or subpopulations: those taking advanced courses in mathematics and courses in physics, those taking advanced courses in mathematics but not physics, those taking courses in physics but not advanced mathematics, and those not taking either advanced mathematics or physics. Schools were asked to assign each student to one of the four subpopulations, and students were sampled within subpopulations.
- Because of the great differences in academic preparation among the subpopulations of Population 3, separate sets of test booklets had to be developed for each group. This in turn necessitated a separate booklet rotation scheme within each subpopulation. Thus, separate Student Tracking Forms had to be prepared for each subpopulation within a school to ensure correct booklet assignment and to facilitate test administration.

The field operations comprising within-school sampling, preparing the Survey Tracking Forms, and assigning test instruments to students, teachers, and school principals could be carried out either manually, using the preprinted forms included in the *Survey Operations Manuals*, or electronically, using the field operations software designed by the IEA Data Processing Center.

9.5 WITHIN-SCHOOL SAMPLING PROCEDURES FOR POPULATIONS 1 AND 2

Since the within-school sampling for Populations 1 and 2 was based on intact classes, a key step in the TIMSS field operations was the construction within each school of an exhaustive and mutually exclusive sampling framework–that is, a list of classes in which each student in the target grades was assigned to one and only one class. In education

systems in which mathematics and science classes did not form identical partitions (that is, where the students who attended a mathematics class did not all also attend the same science class), classes were defined on the basis of mathematics instruction for sampling purposes. The TIMSS instruments then had to be administered to students from the sampled mathematics classes and to the mathematics and science teachers associated with these students.

Three variations on the procedure for within-school sampling were developed, reflecting different classroom organization within the participating countries. Each procedure ensured that every student in the target grades would be assigned to one and only one class so that all students had a known chance of being included in the sample. The NRC chose the procedure most suited to the country, in consultation with the sampling and field operations consultants.

The General Procedure was designed to apply in all possible circumstances, but it had the disadvantage of being operationally very complex and demanding for School Coordinators. Nevertheless, it had to be used in cases where the mathematics classes in the target grades were not exhaustive and mutually exclusive with respect to the target population and no other class with these properties could be identified.

As a simplified alternative, Procedure B could be used if the population of mathematics classes was exhaustive and mutually exclusive; that is, if each student in the target population belonged to one and only one mathematics class, and all students in the class were taught by the same mathematics teacher(s). With Procedure B, however, students in the mathematics class might be taught by different science teachers and belong to different science classes.

As a further simplification, Procedure A could be used if each student in the target grades belonged to one and only one mathematics class and all students in the class were taught by the same mathematics *and* science teachers.

9.5.1 SURVEY TRACKING FORMS IN POPULATIONS 1 AND 2

Survey Tracking Forms were provided for sampling classes and students; for tracking schools, classes, teachers, and students; for linking students and teachers; and for recording information during test administration. They are described below. Copies of the forms are provided in Appendix C.

- The School Tracking Form keeps track of the sampled schools and their replacement schools.
- The Class Tracking Form lists all mathematics classes in the target grades of a selected school and is used in all three procedures for sampling classes within schools.
- The Teacher-Student Linkage Form is a matrix linking teachers (in the columns) and students (in the rows). This is essential in all cases where the composition of a class changes from one subject and/or teacher to another. It is not necessary with Procedure A. In Procedure B, this form is prepared only for the selected classes. In the General Procedure, it is prepared so that all students in the target grades in the selected schools

are listed, so as to identify, for sampling purposes, groups of students that are exhaustive and mutually exclusive.

- The Student Tracking Form lists all students in the sampled classes. It is used in all three procedures, but there are differences (with respect to whether the national center or the School Coordinator fills in certain information) between Procedure A on the one hand and Procedure B and the General Procedure on the other hand. When students are subsampled within classes, an additional worksheet, the Student Subsampling Form, is used for the random selection of students before the assignment of booklets, so that booklets are assigned only to the selected students and not to all the students listed in the Student Tracking Form.
- The Teacher Tracking Form lists the teachers who taught the students in the sampled classes in mathematics and/or science. It is essential when entering the data for linking students in the sampled classes to teachers.
- The Test Administration Form is used by Test Administrators to record information about the administration and timing of the testing sessions.
- The Student Response Rate Form is used by School Coordinators to calculate students' response rates in the regular and makeup sessions.

Although the general functions of these forms are the same in the three procedures for within-school class sampling, each procedure has its own version of the forms, tailored to its specific needs. Not all forms are needed in all procedures, and the order in which forms are used differs across procedures.

The following sections present the steps taken by national centers and schools in each procedure.

9.6 THE GENERAL PROCEDURE FOR WITHIN-SCHOOL SAMPLING

In the General Procedure, the list of mathematics teachers provided by the School Coordinators was used to prepare the Teacher-Student Linkage Forms for *all* classes in the target grades in the selected schools. This was necessary if there were some students in the target grades who attended more than one of the classes, and/or others who attended none of the classes. From these forms, the School Coordinator filled in the students' names and the linkages between students and mathematics teachers, and the national center prepared the Class Tracking Forms. The Teacher-Student Linkage Forms for the selected classes were sent back to the School Coordinators, who added the science teachers' names and their linkages with students. The information from these forms was then used to prepare the Student Tracking Forms and the Teacher Tracking Forms.

Table 9.1 shows the steps taken in the General Procedure. (Steps 5b and 6 could be omitted if the information on the science teachers for all the classes in the target grades was collected in step 1.)

Table 9.1 Sequence of Steps in the General Procedure, Populations 1 and 2

National Centers	Schools
1. Ask the sampled schools for a list of mathematics *teachers* who teach students in the target grades	
	2. Send the list of mathematics teachers who teach students in target grades to the national center
3. Prepare Teacher-Student Linkage Forms for each target grade with header and mathematics teacher columns completed (based on information in the list of math teachers provided by schools)	
	4. Enter name, sex, birth date, and exclusion status for each student in the target grade on the Teacher-Student Linkage Forms and indicate which mathematics teacher teaches which students. Then return this form to the national center
5a. Identify mathematics classes 5b. Prepare Class Tracking Forms for sampling classes and select the sample of mathematics classes 5c. Prepare Teacher-Student Linkage Forms for the sampled classes with information on students and on mathematics teachers and send these to schools	
	6. Enter the information for the science teachers on the Teacher-Student Linkage Forms and return them to the national center
7. Prepare Student Tracking Forms and Teacher Tracking Forms (filled out with all information obtained from Teacher-Student Linkage Forms) and send copies to the schools with the testing materials	
	8. After the test administration return Student Tracking Forms and Teacher Tracking Forms (with participation status filled in) and testing materials to the national center

9.7 PROCEDURE A FOR WITHIN-SCHOOL SAMPLING

In Procedure A, the Class Tracking Forms for sampling classes were prepared, based on the list of the mathematics classes and their mathematics and science teachers that each sampled school prepared. There was no need for School Coordinators to prepare Teacher-Student Linkage Forms. Student Tracking Forms and Teacher Tracking Forms needed to be prepared only for the selected classes.

The main steps taken in Procedure A are summarized in the following table.

Table 9.2 Sequence of Steps in Sampling Procedure A, Populations 1 and 2

National Centers	Schools
1. Ask the sampled schools for a list of mathematics classes in the target grades with the names of their mathematics and science teachers	
	2. Send the list of mathematics classes in the target grades with the names of their mathematics and science teachers to the national center
3a. Prepare Class Tracking Forms for sampling classes and select the sample of mathematics classes 3b. Prepare Student Tracking Forms for the sampled classes and send them to the schools	
	4. Enter name, sex, birth date, and exclusion status for each student in the sampled classes on Student Tracking Forms and return them to the national center
5. Complete the Student Tracking Forms (with Student IDs and booklet assignments) and Teacher Tracking Forms (with Teacher IDs, and teacher questionnaire assignments) and send copies to the schools along with the testing materials	
	6. After the test administration, return the Student Tracking Forms and Teacher Tracking Forms (with participation status filled in) along with completed testing materials to the national center

9.8 PROCEDURE B FOR WITHIN-SCHOOL SAMPLING

In Procedure B, the Class Tracking Forms for sampling classes were also prepared based on the list of the mathematics classes and their mathematics and science teachers that each sampled school prepared, because, as in Procedure A, the population of mathematics classes was exhaustive and mutually exclusive. However, in this case, the Teacher-Student Linkage Forms were necessary, because the entire group of students forming the mathematics class was not taught by a single science teacher. The Student Tracking Forms and Teacher Tracking Forms were produced with the information from the Teacher-Student Linkage Forms.

Table 9.3 shows the steps taken in Procedure B.

Table 9.3 Sequence of Steps in Sampling Procedure B, Populations 1 and 2

National Centers	Schools
1. Ask the sampled schools for a list of mathematics classes in the target grades along with the names of their mathematics (and science) teachers	
	2. Send the list of mathematics classes in the target grades and the names of their mathematics (and science) teachers to the national center
3a. Prepare Class Tracking Forms for sampling classes and select the sample of mathematics classes 3b. Prepare Teacher-Student Linkage Forms for the sampled classes, indicating names of mathematics (and science) teachers and classes they teach in the headers (based on information in list of mathematics classes provided by the schools) and send copies to the schools	
	4. Enter name, sex, birth date, and exclusion status for each student in the sampled classes on the Teacher-Student Linkage Forms and indicate which teachers teach which students. Then return this form to the national center
5. Prepare Student Tracking Forms (with Student IDs and assigned booklets) and Teacher Tracking Forms (with Teacher IDs, teacher link numbers, and assigned questionnaires) and send copies to the schools with the test instruments	
	6. After the test administration, return Student Tracking Forms and Teacher Tracking Forms (with participation status filled in) and testing materials to the national center

9.9 EXCLUDING STUDENTS FROM TESTING

The target population included all students enrolled in the target grades. However, certain students were, for various reasons, ineligible for the TIMSS testing. In some education systems, these students were in special schools or special classes, and it was possible to exclude these schools or classes from the sample. However, when that was not

possible, these students had to be excluded by the school principal or other qualified staff member; detailed instructions therefore had to be given to the School Coordinators.

The following guidelines define general categories for the exclusion of students within schools. NRCs were asked to follow these guidelines carefully within the context of each educational system.

- *Educable mentally disabled students*. These are students who are considered, in the professional opinion of the school principal or other qualified staff members, to be educable mentally disabled, or who have been so diagnosed in psychological tests. This includes students who are emotionally or mentally unable to follow even the general instructions of the TIMSS test. It does not include students solely because of poor academic performance or discipline problems.
- *Functionally disabled students.* These are students who are permanently physically disabled in such a way that they cannot perform in the TIMSS tests. Functionally disabled students who can perform should be included in the testing.
- *Non-native-language speakers.* These are students who cannot read or speak the language of the test and so could not overcome the language barrier of testing. Typically, a student who has received less than one year of instruction in the language of the test should be excluded, but this definition should be adapted in different countries.

9.10 CLASS, STUDENT, AND TEACHER ID AND TEACHER LINK NUMBER

Within each school, a Class ID was assigned to each class in the target grades listed on the Class Tracking Form. The Class ID consisted of the three-digit School ID plus a two-digit identification number for the class within the school. To allow each class and each student to be uniquely identified later, it was essential that the same Class ID not be assigned to two classes in different target grades in the same school. Therefore, although the two-digit ID could be simply a sequential number, it was preferable to add further classifying information, such as the grade.

Each student listed on the Student Tracking Form was assigned a Student ID: a seven-digit number consisting of the five-digit Class ID plus the two-digit sequential number of the student within the class (corresponding to his or her entry in the Student Tracking Form). All students listed on the Student Tracking Form, including those marked for exclusion, had to be assigned a Student ID.

All mathematics and science teachers of the selected classes (those listed on the Teacher Tracking Form) were assigned a Teacher ID. This consisted of the three-digit School ID plus a two-digit sequential number of the teacher within the school. The teacher questionnaire included sections on professional and academic background and on teaching practices and implemented curriculum that were specific to each class/subject the teacher taught. Thus one entry in the Teacher Tracking Form had to be made for each teacher/class and teacher/subject combination. Teachers teaching more than one selected class (e.g., two classes in different target grades), or, in the case of Population 2, teaching both mathematics and science, were listed more than once on the Teacher Tracking Form. So that multiple entries for a teacher could be distinguished and that teacher linked to the correct group of

students, the five digits of the Teacher ID were followed by a two-digit number, the Teacher Link Number. The five-digit Teacher ID referred to the teacher as an individual and was therefore identical for multiple entries for the same teacher. The two-digit Teacher Link Number allowed each occurrence of a teacher in the Teacher Tracking Form to be identified uniquely.

Careful implementation of these procedures was necessary so that later each class could be linked to a teacher and student outcomes could be analyzed in relation to teacher-level variables. A teacher might teach more than one class and therefore be represented by more than one entry in the Teacher Tracking Form, but complete only one questionnaire on one of his/her courses. In that case the combination of Teacher ID and Teacher Link Number made it possible to link personal information about the teacher (like age or sex) to all related students, without falsely linking course-related information to other courses.

To cater for students who joined the class after the Student Tracking Form was created, NRCs were asked to add three further entries to the list of students in the Student Tracking Form. These entries had to be assigned Student IDs and booklets, so that Test Administrators could easily use them when required.

9.10.1 Assigning Testing Materials to Students

Before assigning test booklets and questionnaires to students, NRCs and School Coordinators had to decide whether the test should be administered to the students marked for exclusion or not.

Eight booklets were rotated within classrooms in both Populations 1 and 2. The first eligible student on the Student Tracking Form was assigned a booklet (by selecting a random number from the table of random numbers, multiplying it by 8, adding 1 to the result and using the integer part of the resulting number to identify the booklet). For subsequent eligible students, booklet numbers were assigned sequentially, continuing from the first class to all subsequent ones.

9.10.2 Assigning Questionnaires to Teachers

Each teacher listed on the Teacher Tracking Form was assigned at least one teacher questionnaire. NRCs decided whether to assign multiple questionnaires to teachers teaching more than one of the selected classes or teaching both mathematics and science. If teachers represented by multiple entries were assigned only one questionnaire, it was recommended that teachers of both subjects be linked to all students; but it was more important to cover both subjects at the upper grade than to cover both grades. Detailed instructions were provided to assist in the assignment of questionnaires.

9.11 WITHIN-SCHOOL SAMPLING PROCEDURES FOR POPULATION 3

The most cost-effective within-school sampling procedures for Population 3 involved selection of a simple random sample of students within schools. Given that students had to be sampled within subpopulations, assigning students to subpopulations was a key step.

Before they could be sampled, each eligible student had to be identified and assigned to one and only one subpopulation.

There were two procedures for within-school sampling of individual students, depending on the structure of the education system at the Population 3 level.

Procedure 1 applied to tracked education systems with only one subpopulation per school type so that student assignment to a subpopulation was known before within-school sampling. This procedure was appropriate, for example, for a two-track system composed of academic schools, where only students taking courses in both advanced mathematics and physics could be found, and vocational schools, whose students do not take such courses.

Procedure 2 applied to education systems in which more than one track or stream was found in the same school, such that students had to be assigned to subpopulations during within-school sampling. This procedure was appropriate for an education system where students taking advanced courses in mathematics or physics could be found in any school. In this case, each student in the target population in each sampled school would be assigned to the relevant subpopulation and an equal number of students would be sampled from each available subpopulation in a given school.

Some education systems contain both schools in which all the students belong to one subpopulation, and schools that contain students from more than one subpopulation. In these mixed systems, it was permissible to use different procedures for different schools.

Although the Population 3 sampling design required the sampling of individual students, in some countries it was necessary for administrative reasons to sample intact classes. Variants of Procedures 1 and 2, known as Procedure 1C, Procedure 2C, and Procedure 3C, were developed for that purpose. Procedure 1C could be used in tracked education systems in which all classes in a given school type were composed of students from the same subpopulation. Procedure 2C could be used in systems in which different classes in the same school were composed of students from different subpopulations (but students within a class belonged to the same subpopulation), so that the assignment of classes to subpopulations took place during within-school sampling. Finally, Procedure 3C could be used if students from different subpopulations were in the same class.

However, countries were warned that sampling classes (Procedure 1C, 2C, and 3C) instead of students could become rather expensive; a sample of N students from a given class is likely to have a higher intraclass correlation than a sample of N students selected from the whole eligible population within a school. Countries in which sampling classes was the only feasible way to proceed had to factor this clustering effect into calculation of their sample size.

9.11.1 Survey Tracking Forms in Population 3

The following Survey Tracking Forms were provided for sampling students (or classes); for tracking schools and students; and for recording information during test administration. Copies of the forms are provided in Appendix C.

- The School Tracking Form kept track of the selected schools and their replacement schools. It was used to assign replacement schools, establish the school identification codes, and track the sampled schools throughout the study.
- The Student Listing Form provided a list of all students in the target population of a selected school and was used in all procedures when individual students were sampled within schools (Procedures 1 and 2).
- The Class Listing Form provided a list of all classes in the target population of a selected school and was used in all procedures when intact classes were sampled within schools (Procedures 1C, 2C, and 3C).
- The Student Tracking Form listed all sampled students and was used when either individual students or intact classes were sampled. Depending on whether the national center or the School Coordinator filled in certain pieces of information, there were differences between Procedures 1 and 2 on the one hand and Procedures 1C, 2C, and 3C on the other.
- The Test Administration Form was used by Test Administrators to record information about the administration and timing of the testing sessions.
- The Student Response Rate Form was used by School Coordinators to calculate students' response rates in the regular and makeup sessions.

Again, the general functions of the forms were the same across procedures for within-school sampling, but each procedure has its own version of the forms, tailored to its specific needs.

9.11.2 PROCEDURE 1

In schools where Procedure 1 was appropriate (i.e., in tracked systems with only one student subpopulation per school where student assignment to subpopulation was known before sampling), NRCs did not need to ask School Coordinators to assign students to subpopulations.

With this procedure, the sampled schools listed all students in the target population on the Student Listing Form. The required number of students was then sampled and a Student Tracking Form, including Student IDs and booklet assignments, was prepared for the selected students at the national center. Only one such form needed to be prepared since there is only one subpopulation in each sampled school. Schools then completed the Student Tracking Form by entering the sex and birth date of each student and his or her participation status in the testing sessions.

The main steps taken in Procedure 1 are summarized in Table 9.4.

Table 9.4 Sequence of Steps in Sampling Procedure 1, Population 3

National Center	Schools
1. Ask the sampled schools to fill out the Student Listing Form with the names of all students in the target population (and their class name/location and exclusion status, if necessary)	
	2. Send the Student Listing Form (completed with the indications of the class names/locations and exclusion status) to the national center
3a. Sample the appropriate number of students from the Student Listing Form 3b. Prepare a Student Tracking Form for the sampled students (with Student IDs and booklet assignments) and send it to the school along with the assessment materials	
	4a. Enter sex and birth date of each sampled student on the Student Tracking Form 4b. After test administration, return the Student Tracking Form (with the participation status filled in for each student) and the completed testing materials to the national center

9.11.3 PROCEDURE 2

In many education systems, schools offer more than one option for mathematics and physics courses, so students cannot be grouped into the four subpopulations in advance. In such systems, students must be assigned to subpopulations during within-school sampling. Procedure 2 was developed for this situation. NRCs provided School Coordinators with a clear operational definition of the four subpopulations and asked them to assign each Population 3 student to one subpopulation.

After School Coordinators had listed the students in their schools for each subpopulation, NRCs sampled the required number of students from each subpopulation. They then prepared a Student Tracking Form for each subpopulation and assigned Student IDs and test booklets to students. The School Coordinator/Test Administrator entered, for each student, sex and birth date and participation status in the testing sessions.

The main steps taken in Procedure 2 are summarized in Table 9.5.

Table 9.5 Sequence of Steps in Sampling Procedure 2, Population 3

National Center	Schools
1a. Provide the sampled schools with an operational definition of the four subpopulations 1b. Ask them to fill out the Student Listing Form with the names of all students in the target population, assigning each of them to the correct subpopulation (and indicating the class name/ location and exclusion status, if necessary)	
	2. Send the Student Listing Form with the student assignment to subpopulations (and the indication of the class names or locations and exclusion status) to the national center
3a. For each subpopulation present in a school, sample the appropriate number of students 3b. Prepare separate Student Tracking Forms (with Student IDs and booklet assignments) for the sampled students in each subpopulation and send them to the school along with the assessment materials	
	4a. Enter sex and birth date of each sampled student on the Student Tracking Forms 4b. After test administration, return the Student Tracking Forms (with the participation status filled in for each student) and the completed testing materials to the national center

9.11.4 PROCEDURE 1C

In tracked education systems with only one subpopulation per school, it was sometimes necessary to sample intact classes rather than individual students. Procedure 1C was developed to meet this situation.

With this procedure, only one class per school was sampled. A Student Tracking Form was prepared for the sampled class, and the School Coordinator entered the name of each student in that class, along with birth date, sex, and exclusion status, if necessary. The form was then returned to the national center, where Student IDs and booklets were assigned. Finally, the Student Tracking Form was sent with the testing materials to the schools for test administration.

The steps taken in Procedure 1C are summarized in Table 9.6.

Table 9.6 Sequence of Steps in Sampling Procedure 1C, Population 3

National Center	Schools
1. Ask the sampled schools to fill out the Class Listing Form with all classes in the target population	
	2. Return the Class Listing Form with all classes in the target population to the national center
3a. Sample one class from the Class Listing Form 3b. Prepare a Student Tracking Form for the sampled class and ask the School Coordinator to list all students in that class, along with the sex, birth date, and exclusion status of each	
	4. Return the Student Tracking Form to the national center with the required information filled in
5. Assign Student IDs and booklets to students and send the Student Tracking Form back to the school along with the assessment materials	
	6. Return the Student Tracking Form (with participation status filled in) and testing materials to the national center

9.11.5 PROCEDURE 2C

In education systems where a group of classes exists within the school such that a) each student in the school belongs to exactly one class, and b) each class contains students from a single subpopulation (multiple subpopulations may be represented within each school), it was also sometimes necessary to sample intact classes rather than use Procedure 2 to select individual students. Procedure 2C was designed to handle this situation. In this case, the list of classes in the target population sent by each school included the subpopulation to which each class belonged. NRCs then sampled one class for each subpopulation in each school.

A separate Student Tracking Form was prepared for each sampled class, and the School Coordinator again filled in the names of all students in each Student Tracking Form along with birth date, sex, and exclusion status. The forms were then returned to the national center, where Student IDs and booklets were assigned according to subpopulation. Last, the Student Tracking Forms were sent with the testing materials back to the school for test administration.

The steps taken in Procedure 2C are summarized in Table 9.7.

Table 9.7 Sequence of Steps in Sampling Procedure 2C, Population 3

National Center	Schools
1. Ask the sampled schools to fill out the Class Listing Form with all classes in the target population along with an indication of the subpopulation to which each of them belongs	
	2. Return the Class Listing Form for all classes in the target population, with their subpopulations, to the national center
3a. Sample one class for each subpopulation listed on the Class Listing Form 3b. Prepare Student Tracking Forms for the sampled classes and ask the School Coordinator to list all students in each of those classes, along with the sex, birth date, and exclusion status	
	4. Return the Student Tracking Forms to the national center with the required information filled in
5. Assign Student IDs and booklets to students on each Student Tracking Form and send these forms to the school along with the assessment materials	
	6. Return the Student Tracking Forms (with participation status filled in) and testing materials to the national center

9.11.6 PROCEDURE 3C

A different procedure was used when intact classes had to be sampled (so that Procedure 2 could not be applied) but those classes contained students from more than one subpopulation; for example, where in the same class there were students taking advanced courses in mathematics and physics and students taking advanced courses in mathematics only. In that case, the School Coordinators indicated on the Student Tracking Form for the sampled class the subpopulation to which each student should be assigned. NRCs left the students from the largest subpopulation on the original Student Tracking Form and prepared new forms for those from other subpopulations.

The steps taken in Procedure 3C are summarized in Table 9.8.

Table 9.8 Sequence of Steps in Sampling Procedure 3C, Population 3

National Center	Schools
1. Ask the sampled schools to fill out the Class Listing Form with all classes in the target population	
	2. Return the Class Listing Form with all classes in the target population to the national center
3a. Sample the class(es) from the Class Listing Form 3b. Prepare Student Tracking Form(s) for the sampled class(es) and ask the School Coordinator to list all students in each of those classes, indicating for each student the subpopulation to which he or she belongs, sex, birth date and exclusion status, if necessary	
	4. Return the Student Tracking Forms to the national center with the required information filled in
5a. For each class, keep the largest subpopulation on the original Student Tracking Form and prepare new Student Tracking Form(s) for the other subpopulation(s) in the class. 5b. Assign Student IDs and booklets to students on each Student Tracking Form and send these forms to the schools along with the assessment materials	
	6. Return the Student Tracking Forms (with participation status filled in) and testing materials to the national center

9.11.7 EXCLUDING STUDENTS FROM TESTING

The guidelines for excluding students from testing in Populations 1 and 2 also applied in Population 3.

9.11.8 STUDENT ID NUMBERS

All sampled students listed on the Student Tracking Form are assigned a unique Student ID. This was a seven-digit number consisting of the three-digit School ID plus a two-digit code for the subpopulation, plus a two-digit sequential number for the student within the school. The two-digit codes for the subpopulation are assigned as follows:

- "00" for the OO group (students not taking advanced courses in mathematics or physics — see Chapter 3 on the TIMSS test design)
- "10" for the MO group (students taking advanced courses in mathematics but not physics)
- "01" for the OP group (students taking physics, but not advanced courses in mathematics)
- "11" for the MP group (students taking advanced courses in mathematics and physics).

9.11.9 ASSIGNING TESTING MATERIALS TO STUDENTS

The number of booklets to be rotated varied by subpopulation:

- 2 booklets for the OO group (Booklets 1A and 1B)
- 5 booklets for the MO group (Booklets 1A, 1B, 3A, 3B and 3C)
- 5 booklets for the OP group (Booklets 1A, 1B, 2A, 2B and 2C)
- 9 booklets for the MP group (Booklets 1A, 1B, 2A, 2B, 2C, 3A, 3B, 3C and 4).

Booklets were rotated within subpopulations as follows. The first eligible student listed in the Student Tracking Form was assigned a booklet at random (by selecting a random number in a table of random numbers, multiplying it by the number of booklets for the subpopulation, adding 1 to the result and using the integer part of the obtained number to identify the booklet).

For subsequent eligible students, booklet numbers were assigned sequentially, continuing the sequential assignment from the first tracking form to all subsequent ones in the same subpopulation. In each Student Tracking Form, booklets (as well as Students IDs) were also assigned to three extra entries in order to facilitate the work of the Test Administrator in case of loss or damage of testing materials, or (when sampling intact classes) in case there were students in the class at the time of testing who were not included in the Student Tracking Form.

9.12 RESPONSIBILITIES OF SCHOOL COORDINATORS AND TEST ADMINISTRATORS

The School Coordinator is the person in the school responsible for administering the TIMSS tests. This could be the principal, the principal's designee, or an outsider appointed by the NRC with the approval of the principal. The NRC was responsible for ensuring that the School Coordinators were familiar with their responsibilities.

The following were the major responsibilities of School Coordinators, as presented in the *School Coordinator Manuals* (TIMSS, 1994e, 1994f):

- Providing lists of mathematics classes in the target grades in the school and assisting the national center in completing the tracking forms
- Helping the national center determine the dates of test administration

- Selecting (in cooperation with the school principal) the Test Administrators who would conduct testing sessions for the school
- Making the necessary arrangements for testing
- Verifying that the testing materials received from the national center were complete and ensuring that they were kept in a secure place
- Acquainting the Test Administrators with the TIMSS study and providing them with the testing materials and instructions for quality assurance
- Distributing parental permission forms (if applicable) and ensuring that they were signed and returned on time
- Working with the school principal, the Test Administrators, and the teachers of the selected classes regarding logistics on the days of testing, such as room locations, classes involved, distribution of materials, availability of pens and pencils
- Distributing the teacher questionnaires to the selected teachers, ensuring that they were returned completed, and recording the participation information in the Teacher Tracking Form (Populations 1 and 2 only)
- Ensuring that Test Administrators returned all testing materials after the testing session (including the completed Student Tracking Form, the Test Administration Form, and any unused materials), and that they filled in all the required information in the Student Tracking Form
- Calculating the student response rate and arranging for makeup sessions if it was below 85%
- Preparing a report for the national center providing essential information about the test administration in the school
- Returning the completed and unused test instruments, the Student and (for Populations 1 and 2) Teacher Tracking Forms, and the report to the national center.

The following were the responsibilities of Test Administrators, as described in the *Test Administrator Manual* (TIMSS, 1994i):

- Ensuring that each student received the correct testing materials
- Administering the test according to the instructions provided in the *Test Administrator Manual*
- Ensuring the correct timing of the testing sessions using a stopwatch and recording on the Test Administration Form the time when the various sessions started and ended
- Recording student participation on the Student Tracking Form.

9.13 PACKAGING AND SENDING MATERIALS

Two packages were prepared at the national center for each sampled class in Populations 1 and 2 and for each subpopulation within a school in Population 3. One package contained the test booklets and the other the student questionnaires for all students listed in the Student Tracking Form and for the three extra entries.

For each participating school, the packages for all sampled classes or subpopulations were assembled and packed together with the Teacher Tracking Form and teacher

questionnaires (for Populations 1 and 2), the school questionnaire, and the materials prepared for briefing School Coordinators and Test Administrators. A set of labels and prepaid envelopes addressed to the national center were included to facilitate the return of testing materials.

9.14 CODING, DATA ENTRY, DATA VERIFICATION, AND SUBMISSION OF DATA FILES AND MATERIALS

Following the administration of the TIMSS tests, the NRC was responsible for:

- Retrieving the materials from the schools
- Training coders to code the free-response items
- Scoring the free-response items and the 10% reliability sample
- Entering the data from the achievement tests and background questionnaires
- Preparing a report on survey activities
- Submitting the data files and materials to the IEA Data Processing Center.

When the testing materials were received back from the schools, NRCs were to do the following:

- Check that the appropriate testing materials were received for every student listed on the Student Tracking Form
- Verify all identification codes on all instruments that were not precoded at the national center
- Check that the participation status recorded on the tracking forms matched the information on the test instruments
- Follow up on schools that did not return the testing materials or for which forms were missing, incomplete, or inconsistent.

NRCs then organized the instruments for scoring and data entry. The procedures involved were designed to maintain identification information that linked students to schools and teachers, minimize the time and effort spent handling the booklets, ensure reliability in the free-response coding, and document the reliability of the coding.

9.15 CODING THE FREE-RESPONSE ITEMS

The substantial number of free-response items in the TIMSS achievement booklets and the number of students tested in each country resulted in tens of thousands of student responses to be scored in each country. This required clear procedures for scoring, training sessions for the coders, and instructions for conducting coding sessions. The development of the coding guides is described in detail in Chapter 7 and the international training effort in Chapter 10. The following section describes the procedures used during the coding sessions, including within-country training, coding of the reliability sample, organization of the coding sessions, and monitoring of the coding (backreading).

9.15.1 WITHIN-COUNTRY TRAINING

As described in Chapter 10, the international training sessions for the coding of the free-response items were organized to train the person in each country responsible for the coding. The sessions covered the use of the TIMSS coding guides and presented a model for the national training of coders. The person who attended the international training session conducted the within-country training and was responsible for conducting the coding sessions.

The materials required for within-country training included:

- The *Coding Guide for Free-Response Items* (TIMSS, 1995a, 1995b)
- Selected examples of responses for items for which the coding guides are not straightforward
- Practice sets for the more complicated guides with responses illustrating a range of responses the coders could expect.

Coders were led through the items and their coding guides to become familiar with their content. They then completed the assembled practice papers and discussed each paper until a consensus on the appropriate use of the coding guides was achieved. The session leader monitored the application of the coding guides to ensure that the coders were using them correctly.

9.15.2 CODING THE 10% RELIABILITY SAMPLE

Since free-response items are a vital part of the TIMSS achievement tests, it was critical to document the reliability of coding. To gather information about the agreement among coders, TIMSS had each country arrange for 10% of the test booklets to be coded independently by two coders. Coders assigned to Coding Set A were to code every tenth booklet in Coding Set B, and vice versa. When coding the 10% reliability sample, coders recorded their codes on separate sheets of paper to ensure that the double coding was "blind"; that is, that one coder did not know the code given by the other. The coding of the reliability sample was done throughout the coding sessions to accurately reflect the reliability of the entire process.

9.15.3 PROCEDURES FOR MONITORING THE CODING (BACKREADING)

TIMSS recommended that coders be organized into teams of about six, headed by a team leader who monitored progress and the reliable use of the codes. It was suggested that the team leaders continually check and reread the responses coded in their team, systematically covering the daily work of each coder. In cases where a coder seemed to be having difficulty, this backreading was to be intensified. Any errors in the application of the coding guides were to be brought to the attention of the coder responsible and corrected immediately.

9.16 DATA ENTRY

Entry of the achievement and background data was facilitated by the *International Codebooks* (TIMSS, 1994b, 1995d), which define the variables and file formats in the data files; the DATAENTRYMANAGER program; and the *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995c).

The *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995c) outlines:

- The resources required for data entry, including estimates of time required for data entry and of data file sizes
- Guidelines on which records have to be entered into which data files
- Options to simplify entry of tracking information
- Instructions for entering some special variables
- Suggestions to facilitate data entry.

National centers were requested to verify data in order to ensure that:

- The data files were structured as specified in the *International Codebooks* (TIMSS, 1994b, 1995d)
- The data values conformed to the range validation criteria specified in the *International Codebooks* (TIMSS, 1994b, 1995d)
- There were no duplicate records in the data files
- The components of the identification codes for each record were internally consistent
- The data variables were consistent with the corresponding control and indicator variables
- There were no errors in the student-student and student-teacher linkages.

The background questionnaires were sorted by School ID and by Class ID within school. They were stored with their tracking forms so that the data entry staff could control the number of records to enter and transcribe the necessary information during data entry. NRCs were asked to arrange for double entry of a random sample of at least 5% of the test instruments and questionnaires. An error rate of 1% was considered acceptable.

9.16.1 THE NRC REPORT ON SURVEY ACTIVITIES

Following coding and data entry, the NRC was asked to prepare a report containing:

- A description of the procedures used and problems encountered in the preparation of the test instruments
- An indication of which of the within-school sampling procedures was used in which population
- The national definition of mathematics classes, mathematics and science teachers, and streams (or tracks) used for within-school sampling
- The criteria and definitions used for excluding students from testing

- A description of any problems encountered in the use of the survey tracking forms and of deviations from the recommended procedures
- An indication of the position of School Coordinators and Test Administrators in the schools and the training they received
- A description of any deviations from the prescribed timing of the testing sessions
- An indication of the procedures used for quality control and a summary of the findings from the national quality control observers
- A description of data entry and data verification problems encountered, including an indication of error rate found during the verification of double-entered data
- A description of any modifications in the international coding schemes.

9.16.2 SUBMISSION OF DATA FILES AND MATERIALS TO THE IEA DATA PROCESSING CENTER

The national centers were to send their data files and documentation to the IEA Data Processing Center three months after the last date of testing in their country. The materials they submitted were:

- Data files (and their structure files) in DATAENTRYMANAGER format
- Codebook structure files (including all changes made in the structure) for each data file, with an explanatory letter
- Copies of the national test booklets and questionnaires
- Copies all School Tracking Forms, Class Tracking Forms, and Teacher Tracking Forms
- Completed Data Management Forms indicating the names of the data files and number of diskettes submitted, the number of records in each file, the coding schemes used for optional identification variables, any country-specific changes in the questions, changes in the default validation ranges, any modification of the international coding scheme
- The report on survey activities.

The data files (and their structure files), the codebook structure file, copies of the Data Management Forms, and the report on survey activities were also submitted to the International Study Center.

9.17 CONCLUSION

This chapter presents the design of the field operations from the first contact with the sampled schools to the return of cleaned data files to the IEA Data Processing Center. The implementation of the designed procedures, major problems, deviations, and recommendations for future studies will be discussed in a later volume of the Technical Report.

REFERENCES

Third International Mathematics and Science Study (TIMSS). (1994a). *Coding Guide for Performance Assessment* (Doc. Ref.: ICC885/NRC422). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994b). *International Codebook–Populations 1 and 2* (Doc. Ref.: ICC892-893/NRC428-429). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994c). *Performance Assessment Administration Manual for the Main Survey* (Doc. Ref.: ICC884/NRC421). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994d). *Sampling Manual–Version 4* (Doc. Ref.: ICC 439/NPC117). Prepared by Pierre Foy and Andreas Schleicher. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994e). *School Coordinators Manual–Populations 1 and 2* (Doc. Ref.: ICC891/NRC427). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994f). *School Coordinators Manual–Population 3* (Doc. Ref.: ICC907/NRC440). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994g). *Survey Operations Manual–Populations 1 and 2* (Doc. Ref.: ICC889/NRC425). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994h). *Survey Operations Manual–Population 3* (Doc. Ref.: ICC 906/NRC439). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994i). *Test Administrator Manual–Populations 1 and 2* (Doc. Ref.: ICC890/NRC426). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995a). *Coding Guide for Free-Response Items–Populations 1 and 2* (Doc. Ref.: ICC897/NRC433). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995b). *Coding Guide for Free-Response Items–Population 3* (Doc. Ref.: ICC913/NRC446). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995c). *Guide to Checking, Coding, and Entering the TIMSS Data* (Doc. Ref.: ICC918/NRC449). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995d). *International Codebook–Population 3* (Doc. Ref.: ICC912/NRC445). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995e). *Supplement to the Coding Guide for Performance Assessment* (Doc. Ref.: ICC933/NRC456). Chestnut Hill: Boston College.

Mullis, I.V.S., Jones, C., and Garden, R.A. (1996) "Training Sessions for Free-Response Scoring and Administration of Performance Assessment" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

10. Training Sessions for Free-Response Scoring and Administration of Performance Assessment

Ina V.S. Mullis
Chancey Jones
Robert A. Garden

10.1 OVERVIEW

For the TIMSS main survey, about one-third of the written test time is devoted to free-response items, both short-answer and extended-response. This includes the five TIMSS tests: Population 1, Population 2, Population 3 mathematics and science literacy, Population 3 physics, and Population 3 advanced mathematics. Additionally, for Populations 1 and 2, subsamples of students in approximately 20 countries participated in a performance assessment consisting of hands-on tasks for which students were expected to record results or show other products from their activities (see Chapter 6). Across the five main surveys and the performance assessment, TIMSS included approximately 300 free-response questions and tasks.

With large within-country samples of students responding to the tests, and those student samples representing widely diverse cultures from countries spanning the world's continents, ensuring reliability of scoring was a major concern for TIMSS. The scope of the effort was enormous, with 27 countries participating for Population 1, 46 countries for Population 2, and 21 countries for Population 3. The sample size was approximately 5,000

to 7,500 per population for the main survey. Although the samples for the performance assessment component were smaller (approximately 450 students per country for each of Population 1 and Population 2), the performance assessment entailed setting up equipment and conducting testing sessions involving 12 different hands-on investigations in science and mathematics.

Because of the scope of TIMSS, the training sessions were designed to assist representatives of national centers who would then be responsible for training personnel in their countries to apply the two-digit scoring codes reliably. A four-day training session was developed in which attendees were introduced to the coding system and given practice in coding example papers. In the most effective schedule for the sessions, the first three days were devoted to Populations 1, 2, and 3, respectively, and the fourth day to the administration and coding of the performance assessment. Considering that English is not the native language of many participants and that free-response scoring is a very challenging undertaking requiring subtle distinctions, four days is about the maximum length for any such training session without driving people to total exhaustion.

The four-day training period was demanding, intense, and appropriate for most participants. However, for any future study of the scope of TIMSS, more time needs to be spent on training. For example, one day each could easily have been devoted to training for the advanced mathematics and physics for Population 3. Also, even without discussing administration procedures, a full day could easily have been spent training for coding on the performance assessment. Training for administering the performance assessment ideally would include a separate training session for administering the tests.

Training sessions were conducted in seven regions to provide easy access for participants and smaller groups for the TIMSS trainers to manage. Consistency across sessions was provided by using essentially the same training team and training materials across all the sessions. All in all, this model of "training the trainers" appears to have worked relatively successfully.

10.2 THE TIMSS FREE-RESPONSE CODING TRAINING TEAM

The members of the training team embodied considerable knowledge of the TIMSS tests and of procedures used in training coders to achieve high reliability. The team members are briefly described below.

Mr. Chancey Jones, United States. Mr. Jones was heavily involved in developing the mathematics instruments for Populations 1 and 2. As part of his work in managing development of mathematics tests at Educational Testing Service (United States), he has had extensive experience in establishing scoring criteria, training personnel in scoring procedures, and managing large-scale mathematics scoring sessions for the College Board's Advanced Placement Program and for the U.S. National Assessment of Educational Progress. Mr. Jones was also responsible for reviewing the TIMSS mathematics training materials and conducting training for scoring

mathematics items. He also assisted the International Study Center by serving as the team leader for several training sessions.

Mr. Robert Garden, New Zealand. Mr. Garden coordinated the development of the TIMSS mathematics instruments. As International Coordinator of the IEA's Second International Mathematics Study, he is experienced in conducting international studies. Mr. Garden was director of research and statistics at the Education Ministry in New Zealand, and more recently became a private consultant. He too was responsible for reviewing mathematics training materials and conducting training for scoring mathematics items.

Dr. Graham Orpwood, Canada. Dr. Orpwood coordinated the development of the TIMSS science instruments. He is a professor of science education at the Faculty of Education of York University in Ontario. Dr. Orpwood had responsibility for reviewing the science training materials and conducting training for scoring the science items. He also was involved in developing the TIMSS performance assessment and had responsibility for training related to the administration of the performance assessment tasks.

Dr. Jan Lokan, Australia. Dr. Lokan is the National Research Coordinator for TIMSS in Australia. A senior researcher at the Australian Council for Educational Research, Dr. Lokan contributed substantially to developing the coding guides for the science items and TIMSS performance assessment. She shared responsibility for conducting training for scoring the science items. She also had a central role in training related to the administration of the performance assessment tasks.

Dr. Ina Mullis, United States. Dr. Mullis, codeputy director of the TIMSS International Study Center, coordinated the activities of the training team. Before joining TIMSS, she was director of the National Assessment of Education Progress in the United States, where she gained extensive experience in the evaluation of students' answers to free-response questions in large-scale assessments. She coordinated preparation of the TIMSS manuals containing the coding guides and example responses, and of the materials used at the training sessions.

10.3 THE SCHEDULE OF THE REGIONAL TRAINING SESSIONS

As shown in Table 10.1, the regional training sessions for free-response coding and administering the performance assessment were held across a one-year period from October 1994 through September 1995. This time period was established to accommodate the different school schedules in the countries in terms of the TIMSS schedule. For example, the school schedule for Southern Hemisphere countries is such that the TIMSS tests for Populations 1 and 2, including the performance assessment, were administered in late 1994, while the Population 3 instruments were given in mid- to late 1995. The countries in South America and South Africa administered their tests for multiple populations on a schedule similar to that for Population 3 in the Southern Hemisphere. One training session, focusing

solely on administering the performance assessment, was held in Slovenia in December 1994 for countries that were doing the assessment before their scheduled regional training session.

In general, resources for TIMSS, both within countries and overall, precluded having training sessions devoted only to test administration. Yet separating training for administration and scoring activities would benefit future international assessments. First, it would enable more rigor in training for test administration and could include procedures for the main survey as well as for special components like the performance assessment. Perhaps even more important, the training for scoring could be conducted at a time that would improve those procedures. It is best to conduct scoring training after data collection has begun. The training materials are thus based on responses to the final test items incorporating all of the revisions. Also, training closer in time to the actual scoring process means that the information is fresh in the minds of the scorers.

TABLE 10.1 TIMSS Training Sessions: Free-Response Item Coding and Performance Assessment Administration

Location	Dates
Wellington, New Zealand (Populations 1 and 2)	October 10-12, 1994
Ljubljana, Slovenia (Only PA Administration)	December 18-19, 1994
Hong Kong	January 18-21. 1995
Boston, United States	January 25-28, 1995
Enschede, Netherlands	March 7-10, 1995
Budapest, Hungary	March 13-16, 1995
Pretoria, South Africa	July 18-19, 1995
Miami, United States	July 17-18, 1995
Wellington, New Zealand (Population 3)	September 6, 1995
Melbourne, Australia (Population 3)	September 28-29, 1995

10.4 DESCRIPTION OF EACH TRAINING SESSION

Wellington, New Zealand. This first session was attended by 11 representatives, from Australia (3), Korea (1), New Zealand (6), and Singapore (1). The training team included Ina Mullis, Robert Garden, and Graham Orpwood. The session was designed for countries on a school schedule necessitating the administration of Population 1 and 2 instruments in late 1994 with Population 3 administration to follow in 1995. Therefore, it did not include training for Population 3 items and was three days long rather than four. These countries either had administered the Population 1 and 2 tests, including the performance assessment, or were about to do so. The exception was Korea, which did not participate in the performance assessment.

Because the coding schemes had not yet been applied in countries, participants at the New Zealand training session were able to make an important contribution to determining how they were organized. All representatives had participated extensively in the TIMSS field tests and were familiar with the materials, approaches to free-response coding, and how to administer the performance assessment tasks.

For this initial session, training materials were prepared for all but the most straightforward items. That is, for each item, participants were given a coding guide and from about 10 to 20 example student responses, depending on the complexity of the question and the number of codes involved. The papers had been given preliminary codes to insure a range of example answers. The participants at the New Zealand session worked through the guide for each item and scored the example responses. In striving for reliable coding for all the guides, they made many clarifications and refinements in the guides for Populations 1 and 2, including both the main survey and the performance assessment.

Robin Caygill from New Zealand presented the performance assessment equipment being used in New Zealand and the group reviewed the *Performance Assessment Administration Manual for the Main Survey* (TIMSS, 1994b). Because the group was so familiar with the performance assessment materials, there was no real need to "train" participants in administrative procedures. However, their review of the materials was enormously productive. The TIMSS International Study Center is very grateful for the thoughtful work accomplished at the New Zealand session.

Ljubljana, Slovenia. The session in Slovenia dealt only with training for administering the performance assessment. The session was attended by 11 representatives, from Norway (1), Austria (1), Iceland (1), Czech Republic (2), and Slovenia (6). It was designed particularly for countries that were beginning performance assessment administration before the main survey. Graham Orpwood served as the trainer, and the representatives from Norway and Slovenia both had their performance assessment equipment available for the group to use. The participants at this session, especially those not involved in developing and field testing the performance assessment tasks, found two days of discussion about these complex administration procedures to be very helpful.

Hong Kong. Designed for countries in the Asian region, the Hong Kong session was attended by 16 representatives, from Hong Kong (12), Japan (1), the Philippines (1), and Thailand (2). The training team included Chancey Jones, Robert Garden, Graham Orpwood, and Jan Lokan.

The session began with an orientation covering the importance of coding the free-response questions and performance tasks. Topics included the need to maintain high reliability in coding, the importance of conducting similar training in the participants' own countries, and the necessity of finding exemplars within their countries to use in the training process. The remainder of the first day was devoted to the performance assessment. The Australian materials and equipment for each of the performance assessment tasks were set up for demonstration and discussion purposes. Jan Lokan described the equipment necessary for each task and gave advice about how to conduct the administration. Also, the training team worked with participants on coding approaches and practiced coding for several of the performance assessment tasks.

The second day began with a review of questions raised by the participants concerning coding procedures in general, and the significance of the first and second digits used to code

the free-response questions and performance tasks. The rest of the second day was devoted to training on Population 1 mathematics and science free-response items. Day 3 was spent primarily on training for the Population 2 free-response items, although at the conclusion of the day there was a discussion of the procedures to be followed for planning, organizing, and implementing a successful coding endeavor. This session covered the crucial nature of training materials, including exemplar student responses, the importance of subject-matter expertise in coding the Population 3 specialist items, and effective ways to organize staff to do the scoring (including information about table leaders and backreading procedures). Procedures for implementing the within-country reliability studies were discussed, and the vital need to maintain high reliability was again emphasized. Participants were told that the most important factor in coding student responses is that codes be applied accurately and consistently. Although speed is desirable, accuracy and consistency should not be sacrificed. Coders must be encouraged to follow the manual at all times. The fourth day was dedicated to training for Population 3, although Hong Kong was the only country at the training session with plans to participate in Population 3 testing.

Boston, United States. The session in Boston was attended by 12 representatives, from the United States (5), Canada (4), Mexico (1), Norway (1), and Kuwait (1). The Boston session tended to parallel that in Hong Kong. However, it was decided that beginning with this session, it was preferable to devote the last rather than the first day to the performance assessment. All countries needed to participate in the training for Population 2, but only some in the training for Populations 1 and 3 and the performance assessment. In an attempt to arrange the most convenient schedule for the most countries, the performance assessment had been placed first. This had been convenient, but it was a difficult initiation into TIMSS scoring procedures. Therefore, it was decided to begin with Population 1, follow with Populations 2 and 3, and conclude with the performance assessment on Day 4.

Chancey Jones opened the session by providing an orientation to the TIMSS scoring approach and the training session itself. During the next three days, he and Robert Garden conducted training for the mathematics items, and Graham Orpwood for the science items; and Ina Mullis discussed procedures for doing the actual coding (as described in the *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995c). In general, Day 1 was devoted to Population 1, Day 2 to Population 2, and Day 3 to Population 3. The performance assessment training took place on the fourth day using the equipment and materials from the United States. Maryellen Harmon, who coordinated development of the performance assessment tasks for the International Study Center, presented and discussed techniques for administering the tasks. Graham Orpwood and Robert Garden provided training on the science and mathematics performance tasks, respectively.

Enschede, Netherlands. With 28 participants, the session in Enschede was the largest. It was attended by representatives from Belgium (Flemish) (1), Denmark (1), England (1), France (2), Germany (1), Greece (1), Indonesia (2), Iran (1), Ireland (1), the Netherlands (4), Portugal (2), Scotland (1), Spain (2), Sweden (3), and Switzerland (5). The complete

training team was in attendance: Chancey Jones, Robert Garden, Graham Orpwood, Jan Lokan, and Ina Mullis.

Since beginning with the free-response scoring for Populations 1, 2, and 3 and then moving to the performance assessment training worked well during the Boston session, this order was followed also in the Enschede and Budapest sessions. Thus, the Enschede session began with an orientation to the TIMSS approach to coding the free-response items and the importance of coding reliably. This was followed by training for Population 1. The second day was devoted to Population 2 training and some discussion of procedures for coding and conducting the within-country reliability study. Day 3 was dedicated to training for Population 3, both the literacy and specialist components. On Day 4, Jan Lokan led a demonstration on administering the performance assessment tasks using the Australian equipment. This was followed by training in free-response coding for the performance assessment.

Budapest, Hungary. Representatives from the following 16 countries took part in the training session held in Budapest: Austria (1), Bulgaria (1), Canada (2), Cyprus (1), Czech Republic (2), Hungary (3), Iceland (1), Israel (1), Latvia (1), Lithuania (1), Norway (1), Romania (1), Russia (1), Slovak Republic (2), Slovenia (1), and the Ukraine (1). The training for the 21 participants was conducted by Chancey Jones, Robert Garden, Graham Orpwood, and Jan Lokan.

The first day followed the agenda of the Boston and Enschede sessions. After a brief orientation to free-response coding for TIMSS, the team reviewed the goals of the training session: to instruct the participants in the nature and volume of coding, to model procedures for training staff to apply the free-response codes reliably and efficiently, and to discuss staff requirements and facilities needed for successful free-response coding. The greater part of Day 1 was spent in training for Population 1.

On Day 2, it was decided to include the coding of practice examples of both mathematics literacy and science literacy for Population 3. This provided time on Day 3 to cope with the complexity of Population 3 coding for the advanced mathematics and physics items. Since some items are part of both Population 2 testing and the literacy assessment for Population 3, this change in schedule worked well at the Budapest session, where most countries were participating in both Population 2 and 3 testing. The science training for Population 2 and Population 3 literacy was followed by the mathematics training for Population 2 and Population 3 literacy. These were followed (as in earlier sessions) by the discussion of guidelines for successful coding within countries. Day 3 was devoted to training for the advanced mathematics and physics items. The extra time gained permitted discussion of additional mathematics questions that were not part of the subset used for practice coding during the training. In response to requests from the participants, training for the performance assessment was begun in an early morning session on Day 3 and concluded on Day 4.

Pretoria, South Africa. South Africa participated in TIMSS on the schedule for Southern Hemisphere countries testing Population 3, but also tested Population 2. Therefore, a special training session was held to provide training for Population 2 testing and for the Population 3 literacy free-response items. (South Africa did not participate in the specialist testing for Population 3.) There were 24 participants, all from South Africa. Robert Garden led the training session, which covered most of the mathematics and science free-response items for Population 2 and the literacy portion for Population 3. Because the Population 3 specialist tests did not need to be covered, there was additional time for covering the items relevant to South Africa. South Africa provided financial support for this training session.

Miami, United States. Like the session in South Africa, this training session was for the South American countries–Colombia and Argentina (2 and 3 representatives respectively)–that also participated on the Southern Hemisphere schedule for Population 3. Both of these countries participated only at Population 2, but for both the main survey and the performance assessment. Ina Mullis and Eugenio Gonzalez from the TIMSS International Study Center led the training session. One day was devoted to coding training for Population 2 mathematics and science for the main survey, and the second day to the performance assessment. Although some discussion was held about administering the latter, both countries had participated in the pilot, already had arranged for their equipment, and felt comfortable about administration procedures. Thus, training on the second day focused mainly on procedures for coding the performance assessment responses.

Wellington, New Zealand, and Melbourne, Australia. These two training sessions were for the two Southern Hemisphere countries — Australia and New Zealand — testing Population 3. Both sessions were led by Robert Garden. Because New Zealand participated only in the literacy testing for Population 3, that training took only one day. It was held on September 6, 1995. As Australia administered both the literacy and specialist tests, that training was held across two days with the assistance of Dr. Jan Lokan and Dr. John Lindsey, both of the Australian Council for Educational Research. It was held September 28-29, 1995. During the two days, the time for coding training was divided about equally across physics, advanced mathematics, and literacy. In contrast to the usual approach, for both the New Zealand and Australian sessions the training was held for the actual coders.

10.5 THE TRAINING MATERIALS

Each participant in the training sessions needed a considerable amount of material, including the relevant manuals and packets of example papers for practice. The participants were asked to bring their own copies of the following manuals as pertinent to their participation status:

- *Coding Guide for Performance Assessment* (TIMSS, 1994a)

- *Coding Guide for Free-Response Items–Populations 1 and 2* (TIMSS, 1995a)
- *Coding Guide for Free-Response Items–Population 3 (in three sections: Literacy Guide, Physics Guide, Mathematics Guide)* (TIMSS, 1995b)
- *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995c)
- *Performance Assessment Administration Manual* (TIMSS, 1994b).

Each coding guide contained the rubrics developed for each of the TIMSS free-response items. For the main survey, each coding category within a rubric also contained some example student responses–as part of the rubric itself, or by following the rubric with some actual student responses, or both. For the performance assessment, a separate document containing examples of coded student responses, entitled the *Supplement to the Performance Assessment Coding Guide with Student Examples* (TIMSS, 1995d), was sent to the countries after training, but before the actual coding effort began.

For the initial training session in New Zealand, the training materials were by necessity based on field-test materials. For the remaining sessions, however, the training materials for Populations 1 and 2 were based on actual test papers from the Southern Hemisphere countries that administered the tests in English: Australia, Hong Kong, and New Zealand. For the literacy and specialist tests for Population 3, again by necessity, the training materials were based on field-test materials. This problem was somewhat alleviated because several countries held a late second round of field testing of revisions to the specialist materials. Still, everything considered, trying to assemble training materials before actual testing was an enormous undertaking and is not recommended. It is better to train for scoring after testing has begun, so that the training materials can be based on actual test papers reflecting the final wording of the test items.

Training materials were prepared for the subset of items shown in Tables 10.2 and 10.3 for mathematics and science, respectively. The purpose was not to conduct the actual training for the coders, but to present a model for use in each country and an opportunity to practice with the most difficult items.

Table 10.2 Mathematics Items For Free-Response Training Sessions

Population 1 Mathematics	
S1	Graph of Numbers of Boys and Girls
T4	Girl Boy Ratio
V4	Game with Cards
Population 2 Mathematics	
T1	Apples in Box
U1	Estimate Time Songs (also, Population 3 literacy)
U2	Draw Rectangle, Explain Ratio
Population 3 Mathematics Literacy	
A12	Which Apartment Cheaper (also, Population 2)
A8	Graph of CD's
Population 3 Mathematics Specialist	
J19	Quadrilateral - Prove E Midpoint
K12	Coordinates of B'
K13	Bacteria in Colony
L15	Crickets (Template)
L16	Real Values of X Satisfy Equation
Performance Task	
M2	Calculator

Table 10.3 Science Items for Free-Response Training Sessions

Population 1 Science	
Q4	Glass Jar Over Lighted Candle (also, Population 2)
R1	Watering Can (also, Population 2)
W5	Reducing Air Pollution
X1	Soup Cooling
X3	Oil Spills
Y1	Sun and Moon (also, Population 2)
Z3	Weights of Blocks
Population 2 Science	
L18	Juanita's Experiment
K10	How Air Exists
O16	Thirsty on a Hot Day (also, Populations 1 and 3 literacy)
O17	Jose's Influenza (also, Population 3 literacy)
P2	Flashlight on the Wall
R4	Ozone Level
W2	Rain from Another Place (also, Population 3 literacy)
Population 3 Science Literacy	
A7	High Heels
A11	Painting the Bridge (also, Population 2)
Population 3 Physics Specialist	
F17	Value of Gravity and Uncertainty
G12	Collision Railway Trucks
G15	Acceleration Arrows Bouncing Ball
G18	Alpha Particles through Gold Sheet
H16	Expression Speed of Electron
Performance Tasks (Population 2 Version)	
SM1	Shadows
S1	Pulse

For each item selected for training, a packet of materials was prepared for each participant in the training session. This packet began with coded responses illustrating each of the categories in the rubric or guide for that item. These served as a basis of discussion to familiarize the participants with the rubric. The trainers presented the reasons for each of the assigned codes and answered any questions.

The packet also contained about 15 to 20 precoded student responses, with the codes known to the training team but not to the session participants. The trainer for the item would first invite participants to code five or six of these student responses. After the

coding had been completed, the trainer would read the scores and answer any questions from the group. This procedure was iterated until the group had scored all the responses. For variety, sometimes the participants took turns in reading out their scores. Although generally there was insufficient time at the training sessions to achieve a high degree of agreement on all items, the procedure provided some practice for participants and an example for how training might be conducted in each of their countries. The trainers emphasized the need for each country to prepare training materials for each item rather than for only a sample of items, and pointed out that for more difficult items more responses might be needed to help coders reach a high degree of reliability. The trainers also recommended that the training materials used in each country be based largely on student responses from that country.

10.6 CONCLUDING REMARKS

The participants in the training sessions exhibited enthusiasm, patience, understanding, and humor in successfully completing the intense and demanding training. Most of them took part in the training for two populations, while those involved in all three populations and the performance assessment attended all the sessions during the four-day training. The training activity highlighted the complexity of the TIMSS coding process, especially for Population 3 and for the performance assessment tasks. In general, future studies should consider a more rigorous process both for deciding which codes to apply internationally and for assigning the codes to the example responses used in the coding guides and training materials. More specifically, the coding guides should be developed as an integral part of item development and modified as necessary throughout the process, particularly in light of actual student responses. The example student responses should be considered to be part of each coding guide. Particular attention should be paid to the suitability of an item for such elaborate coding.

Although demanding and intense, the four-day training period was appropriate for most participants. The difficulty was trying to fit so much material into the four days. Considering the many aspects of TIMSS, perhaps extra sessions should have been held for participants who were to be trained in how to code responses to the advanced mathematics and physics items. Or perhaps other configurations of the training sessions might have helped to ease the burden for countries participating in all aspects of TIMSS.

All in all, however, the participants in the sessions, the host countries, the staff at the International Study Center at Boston College, and the training team are to be commended. Their planning coordination, good will, patience, and support were instrumental to the success of the TIMSS training endeavor.

REFERENCES

Third International Mathematics and Science Study (TIMSS). (1994a). *Coding Guide for Performance Assessment* (Doc. Ref.: ICC885/NRC422). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994b). *Performance Assessment Administration Manual for the Main Survey* (Doc. Ref.: ICC884/NRC421). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995a). *Coding Guide for Free-Response Items–Populations 1 and 2* (Doc. Ref.: ICC897/NRC433). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995b). *Coding Guide for Free-Response Items–Population 3* (Doc. Ref.: ICC913/NRC446). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995c). *Guide to Checking, Coding, and Entering the TIMSS Data* (Doc. Ref.: ICC918/NRC449). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995d). *Supplement to the Performance Assessment Coding Guide with Student Examples* (Doc. Ref.: ICC933/NRC456). Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., and Kelly, D.L. (1996) "Quality Assurance Procedures" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

11. Quality Assurance Procedures

Michael O. Martin
Ina V.S. Mullis
Dana L. Kelly

11.1 OVERVIEW

A study as ambitious as TIMSS, and one that involves the collaboration of as many individuals, requires particular attention to all aspects of quality assurance to ensure that the design is properly implemented and that the data collected are comparable across all countries. As documented in previous chapters of this report, TIMSS has expended considerable effort in developing standardized materials and procedures so that the data collected in all countries are comparable to the greatest possible extent. In addition to setting high standards for data quality, the International Study Center has tried to ensure the overall quality of the study through a dual strategy of support to the national centers and quality control checks.

This chapter describes the procedures used to ensure high-quality data across all countries, and the support afforded to the national centers by the International Study Center in the form of standardized manuals, software aids, practical training, and technical assistance. The chapter describes also the development and implementation of an important aspect of the TIMSS quality assurance efforts–the program of site visits by trained Quality Control Monitors. The data collected during these visits are presented in Martin and Mullis (1996), together with extensive documentation of the quality of

translation activities, population sampling, free-response coding, and data checking and database construction.

11.2 STANDARDIZATION OF THE TIMSS PROCEDURES

One of the main ways in which TIMSS sought to achieve uniform project implementation was by providing clear and explicit documentation of all operational procedures. Such documentation was primarily in the form of operations manuals, supported where possible by computer software systems that implement the specified procedures. Forms accompanying some of the manuals serve to document the implementation of the procedures in each country. The manuals are described below. Bibliographic references can be found at the end of the chapter.

Sampling Manual: Defines the operational definitions of the school sample and details the procedures for selecting it for Populations 1 and 2. The forms provided in the *Sampling Manual* ensure that vital information at key stages is collected and recorded in a uniform manner. Target population definitions, choice of stratifying variables, definition of excluded populations, construction of school sampling frames, and selection of school samples are clearly documented.

Sampling Guide for Population 3: This outlines the school sampling procedures for Population 3.

Survey Operations Manual: The *Survey Operations Manuals* (one for Populations 1 and 2 and one for Population 3) was prepared by the IEA Data Processing Center for the National Research Coordinators (NRCs) and their colleagues who were responsible for implementing the TIMSS procedures. It describes the activities and responsibilities of NRCs from the moment the international testing materials arrive at the national center to the moment the cleaned data sets and accompanying documentation are submitted to the IEA Data Processing Center. The manual includes:

- Procedures for translating and assembling the test instruments and questionnaires
- Descriptions of the approved procedures for within-school sampling and guidelines for selecting the appropriate procedure
- Instructions for obtaining cooperation from the selected schools
- Instructions for installing and using the IEA software to prepare the sampling and tracking forms
- Explicit procedures for packing and sending materials to the schools
- Preparations for test administration and instructions for data entry and verification.
- An important feature of the *Survey Operations Manuals* is the detailed instructions for completing the various forms that are required to implement and document the within-school sampling procedure. The forms assist NRCs in their work by making each step

explicit, and also provide an audit trail that facilitates the International Study Center's evaluation of the implementation of the procedures.

School Coordinator Manual: Describes the steps the School Coordinator follows from the moment the testing materials arrive at the school to the moment they are returned to the NRC.

Test Administrator Manual: Covers the procedures from the beginning of testing to the return of the completed tests, questionnaires, and tracking forms to the School Coordinator. Included in this manual is an administration script to be read by the Test Administrator.

Guide to Checking, Coding and Entering the TIMSS Data: Provides further instructions on the procedures for coding, entering, and verifying the TIMSS data.

Performance Assessment Administration Manual: Provides instructions for selecting the sample of students, collecting the equipment for the tasks, and administering the TIMSS performance assessment.

Coding Guide for Performance Assessment: Together with the *Supplement to the Coding Guide for Performance Assessment*, contains the coding rubrics for the performance assessment items and exemplar coded student responses.

Coding Guides for Free-Response Items: Contain the coding rubrics for the free-response items and exemplar coded student responses.

Field Operations Software: This software was provided to assist the NRCs in selecting classes and students and preparing the booklet labels with the student identification. The sampling software targeted primarily within-school sampling activities, although it also provided for sampling of schools. It automatically generated all of the required documentation forms. Training in the use of the software was provided to NRCs, and Statistics Canada gave technical support.

International Codebooks: Contain the necessary information to code, enter, and verify the data from the tests and questionnaires. They are accompanied by data entry software (DATAENTRYMANAGER, DEM), which contains the codebooks.

Data Entry Software: Study participants received software specially developed to facilitate within-school sampling activities and data entry and management (DEM, see later section). Training in the use of these software products and technical support were also provided.

Throughout TIMSS, small-group training sessions during the semi-annual meetings of the National Research Coordinators (NRCs) dealt with desktop publishing, the use of the data entry software, and the use of the sampling software. Individual consultations between NRCs and staff members from the International Study Center, the IEA Data Processing

Center, and Statistics Canada provided further training in the TIMSS procedures. Presentations at NRC meetings, and progress reports disseminated via e-mail, fax, and mail, keep NRCs up to date on the status of TIMSS and their current and future tasks.

11.3 PROCEDURES FOR TRANSLATION AND ASSEMBLY OF THE ASSESSMENT INSTRUMENTS

In any comparative study of student achievement that takes place in more than one language it is crucial that procedures for ensuring comparable translations are followed. With the administration of the TIMSS survey instruments in languages, this was especially important. Furthermore, following translation, the NRCs had to assemble the test booklets according to a complicated booklet assembly plan. This step in the preparation of the instruments introduced another area of concern–uniformity of the test booklet and questionnaire layout. In order to ensure that all instruments administered in all languages and countries were equivalent, TIMSS established a series of procedures which were documented through manuals and supplementary material.

11.3.1 TRANSLATION OF THE TIMSS INSTRUMENTS

TIMSS participants were provided with a set of procedures to help them obtain reliable and high-quality translations. The *Survey Operations Manual* (TIMSS, 1994f, 1994g) contains the following guidelines for translators:

- Identify and minimize cultural differences
- Find equivalent words and phrases
- Make sure the reading level is the same in the target language as in the original English version
- Make sure the essential meaning does not change
- Make sure the difficulty level of achievement items does not change
- Be aware of changes in layout due to translation.

Also included were guidelines for decisions about vocabulary, meaning, and item and booklet layout, and guidelines for making cultural adaptations. Translators were also cautioned to ensure that another possible correct answer for a test item was not introduced. The translations were verified by an independent translation agency (this was coordinated by the International Coordinating Center in Vancouver). The independent translators prepared a translation verification report documenting the quality of the translations and corrections to be made to the booklets. A series of statistical checks were also conducted to identify problematic translations. The TIMSS translation procedures and verification process are further described by Beverley Maxwell in Chapter 8 of this report and in Mullis, Kelly, and Haley (1996).

11.3.2 ASSEMBLY OF THE TIMSS INSTRUMENTS

For the main survey, the International Study Center provided paper and electronic versions of the achievement test and questionnaire items, and paper versions of the completed test booklets and questionnaires for NRCs to use when assembling their national versions of the instruments. Instructions for the layout, printing, and assembly of the booklets also were provided in the *Survey Operations Manuals* and in the supplementary *Instructions for the Preparation of the Instruments at the National Centers*. These materials included directions for the layout of the item clusters, with special warnings related to editing and formatting; for verifying the translation; for printing the clusters from the electronic files; and for assembling the test booklets. In addition, the questionnaires were accompanied by notes on their adaptation by the national centers.

11.4 SCORING THE OPEN-ENDED RESPONSES

Because of the heavy reliance on the use of free-response questions, ensuring reliability of scoring was a major concern for TIMSS. As one step towards this goal, the International Study Center prepared coding guides for Populations 1 and 2 free-response items, for Population 3 mathematics and science literacy, physics, and advanced mathematics items, and for the performance assessment tasks. These contain the scoring rubrics for each item, and each of these is accompanied by exemplar coded student responses to illustrate how the codes are to be applied. In addition, the *Guide for Coding, Checking, and Entering the TIMSS Data* (TIMSS, 1995c) contained specific instructions related to coding. These instructions pertained to the following.

- Arranging for staff and facilities
- Distributing booklets to coders
- Procedures for coding the 10% reliability sample
- Procedures for monitoring the coding
- Preparing materials to train the coders
- Training the coders
- The roles and responsibilities of the coders
- The roles and responsibilities of the group leaders in coding.

Furthermore, an extensive training program was established in which representatives from each country were trained in the coding procedures (see Chapter 10).

In order to document the reliability of free-response coding (i.e., the degree of agreement between coders) in each country, two coders independently coded a random sample of 10% of the student responses (or, for main survey samples larger than 7,500 students, a random sample of 100 booklets from each booklet type). To help with this process, the International Study Center developed a procedure that separated the booklets into two equivalent samples as part of receipt control (by odd- and even-numbered school identifications). The scorers were also designated as belonging to one of two equivalent groups. First, scorers in

one group coded every tenth booklet on a separate coding sheet. These data constitute the 10% reliability sample. Then, scorers in the second group scored all the booklets and record codes in the booklets for data entry. This procedure ensures that the two coders do not know each other's codes, that each booklet is coded by two different scorers, and that the reliability scoring is distributed relatively equally among scorers.

In addition, the International Study Center conducted an international coding reliability study to obtain information about the degree of agreement among coders from different countries. A comprehensive study of inter-coder agreement across countries was beyond the resources of TIMSS. However, a limited study was designed and implemented in which 39 English-speaking coders from 21 countries coded a sample of booklets from 7 countries that tested in English. The results of the reliability studies are reported in Mullis and Smith (1996).

11.5 NATIONAL QUALITY CONTROL PROGRAM

As part of the national quality control efforts, NRCs were requested to arrange a program of unannounced visits by quality control observers to the schools on the day of testing. The main purpose of these visits was to ensure the proper implementation of the TIMSS policies and procedures in the schools and during test administration. The *Survey Operations Manuals* describe the steps to be taken to arrange for the quality control observation component and contains a list of the tasks of the quality control observer.

The International Study Center made available the manual and accompanying forms developed for the international quality assurance program. NRCs were encouraged to use the international materials to conduct their national quality control programs.

11.6 TIMSS QUALITY CONTROL MONITORS

As a major part of the TIMSS quality assurance efforts, a program of site by TIMSS Quality Control Monitors hired by the International Study Center was established. The purpose of this program was to observe the administration of the achievement tests in a sample of classrooms in participating classrooms, and document the degree of compliance with prescribed procedures.

In December 1994, the TIMSS International Study Center contracted Goodison Associates (United States) to help with the hiring, training, and overseeing of a team of Quality Control Monitors. In January 1995, NRCs were asked to nominate a person, such as a retired school teacher, to serve in that capacity for their country. The International Study Center reviewed the nominations and in almost all cases selected the NRC's first suggestion for a Quality Control Monitor. The monitors were trained centrally before returning to their countries to interview the NRC and to observe classroom testing sessions.

The TIMSS Quality Control Monitors (QC Monitors) were trained in a two-day session in which they were briefed on the design and purpose of TIMSS, the responsibilities of the NRC in conducting the study in each country, and their own roles and responsibilities. In

total, five training sessions were held for QC Monitors. Most of the monitors were trained during the three originally scheduled sessions: February 1995, London; March 1995, Enschede; April 1995, Paris. Two additional training sessions were held to train the remaining QC monitors, from Argentina (August 1995, Philadelphia) and Australia and New Zealand (July 1995, New Zealand).

The *Manual for the TIMSS Quality Control Monitors* (TIMSS, 1995e) was developed by the International Study Center with the assistance of Goodison Associates and was used as the basis for the training sessions. The manual included:

- An introduction to TIMSS, outlining the purpose of the study, study schedule, management arrangements, the major components of TIMSS (populations, sampling design, test and questionnaire design), and the purpose of the quality assurance program
- An overview of the roles and responsibilities of the TIMSS Quality Control Monitor
- An overview of the major tasks of the NRC
- Instructions for visiting the national center, interviewing the NRC, collecting the required materials from the NRC, and using the translation verification report to check the implementation of the suggestions made in the international review of the translations
- A report on the interview with the NRC
- Step-by-step procedures for selecting the schools for classroom observation
- Instructions for visiting these schools: arranging the visit, observing the testing sessions, completing the Classroom Observation Record, and interviewing the School Coordinator
- A copy of the Classroom Observation Record
- Instructions for returning materials to the International Study Center.

In addition to the *Manual for Quality Control Monitors*, each QC Monitor received copies of the *Survey Operations Manuals*, the *Test Administrator Manual*, the *School Coordinator Manuals*, and the *Guide to Checking, Coding and Entering the TIMSS Data*, which describe the procedures required for the implementation of TIMSS in each country. Although QC Monitors did not need to know every TIMSS policy and procedure in detail, they were encouraged to read through all the manuals in order to become familiar with the work of NRCs and the procedures to be followed in each country participating in TIMSS.

During each training session a staff member from the International Study Center explained the structure and major components of the study, emphasizing the NRC's tasks, especially as they related to the QC Monitor's duties. Goodison Associates reviewed the roles and responsibilities of the QC Monitor, and led QC Monitors through the Interview with the National Research Coordinator and the Classroom Observation Record. QC monitors also took part in an exercise to help them select the schools for classroom observation.

11.6.1 INTERVIEW WITH THE NATIONAL RESEARCH COORDINATOR

The QC Monitor's visit to the national center included an interview with the NRC and the selection of schools for classroom observation. The structured interview dealt with the NRCs' ten major responsibilities.

- Selecting the sample of students to be tested
- Working with the School Coordinators
- Translating the test instruments
- Assembling and printing the test booklets
- Packing and shipping the necessary materials to the designated School Coordinators
- Arranging for the return of materials from the school sites
- Arranging for coding the free-response and performance assessment questions
- Entering into data files the testing results and information from students, teachers, and principals
- Conducting on-site quality assurance observations for a 10% sample of schools
- Preparing the NRC report on survey activities.

The QC Monitor recorded the NRC's responses to questions about the implementation of these responsibilities, and any additional comments made regarding the TIMSS procedures. The interview questions were designed to ascertain the degree to which the procedures and policies described in the *Survey Operations Manuals*, the *Sampling Manual*, the *Guide to Coding, Checking, and Entering the TIMSS Data,* and other documents were followed. The results of the interviews with the NRCs are summarized in Martin, Hoyle, and Gregory (1996a).

Following the interview with the NRC, the QC Monitor and the NRC worked together to select ten schools for classroom observation, plus three extra schools as potential replacements. Using the School Tracking Form, the QC Monitor and NRC selected the schools by a random selection process (albeit one subject to a number of practical constraints). The schools selected for classroom observation had to be within easy traveling distance of the QC Monitor's home so that travel and observation could be done in one working day; the NRC or QC Monitor had to be able to contact the school to ascertain the date and time of testing and to arrange the visit; the school could not be taking part in the NRC's own national quality control observation program; and the testing could not yet have taken place in that school. After the schools, the classrooms for observation were selected. Where possible, the class chosen was the upper-grade class. The school name and classroom selected for observation were recorded on the Classroom Observation Tracking Form.

At the end of the visit to the national center, the QC Monitor collected the following materials from the NRC:

- *Test Administrator Manual*

- *School Coordinator Manual*
- Test booklets (for each population assessed)
- Performance assessment tasks (for each population assessed, if participating)
- School questionnaires (for each population assessed)
- Student questionnaires (for each population assessed)
- Teacher questionnaires (for each population assessed)
- Translation Verification Report (if this was not given to the QC Monitor at the training session)
- Student Tracking Forms for each class selected for observation
- Class Tracking Forms for each school selected for observation.

QC Monitors received the Translation Verification Report either from the International Study Center during training or from the NRC on their visit to the national center. The QC Monitor checked that any deviations in translation or booklet layout were corrected before test administration, recorded that information, and submitted it to the International Study Center together with the instruments and manuals collected from the NRC.

11.7 THE QUALITY CONTROL MONITOR'S VISIT TO THE SCHOOLS

The QC Monitor was given instructions for arranging the visits to the schools selected for observation, including guidelines for telephoning the School Coordinator and discussing the objectives of the QC Monitor.

To document the activities during the testing session in each school selected for a site observation, QC Monitor used the Classroom Observation Record, which documents the following:

- Activities preliminary to the testing session, including security of the test booklets, level of preparation of the Test Administrator, and adequacy of supplies and testing environment
- Activities during the testing session, including distribution of the test booklets to the appropriate students (using the Student Tracking Form), timing of the testing and breaks, and the Test Administrator's accuracy in reading the test administration script
- The QC Monitor's general impressions of the testing session, including the orderliness of the students, the Test Administrator's answering of students' questions, documentation of any cheating, handling of defective test booklets (if any), and handling of late students (if any).

11.7.1 INTERVIEW WITH THE SCHOOL COORDINATOR

Following the observation of the testing session, the QC Monitor met with the School Coordinator to conduct a brief interview covering the School Coordinator's evaluation of the TIMSS testing and suggestions for improvement, and any additional background

information. The QC Monitor documented responses to specific questions on the Classroom Observation Record. The questions focus on:

- The School Coordinator's impression of the success of the test session
- The attitude of school staff members toward the TIMSS testing
- The shipment of testing materials from the national center
- The level of communication with the national center
- The administration of the teacher questionnaires
- The security of the testing materials before the test date
- The accommodations for testing
- The use of make-up sessions
- The training of the Test Administrators
- Feedback on the sampling procedures used to select students in the school
- Any motivation talks, special instructions, or incentives provided to students to prepare them for the assessment
- Any use of practice questions to prepare the students for the assessment
- Suggestions for improving the *School Coordinator Manuals* (1994d, 1994e).

Finally, the QC Monitor checked the Class Tracking Form for that school with the School Coordinator to ensure that the information is accurate. The QC monitor verified with the School Coordinator:

- Whether the list of mathematics classes in the grade was complete
- Whether there were any students in the grade level who were not in any of the mathematics classes on the Class Tracking Form
- Whether there were any students in the grade level who were in more than one of the mathematics classes on the Class Tracking Form.

The information collected in the Classroom Observation Record is summarized in Martin, Hoyle, and Gregory (1996b).

REFERENCES

Martin, M.O. and Mullis, I.V.S. (1996). *Third International Mathematics and Science Study: Quality Assurance in Data Collection*. Chestnut Hill, MA: Boston College.

Martin, M.O., Hoyle, C., and Gregory, K. (1996a). "Monitoring the TIMSS Data Collection" in M.O. Martin and I.V.S. Mullis (eds.), *Third International Mathematics and Science Study: Quality Assurance in Data Collection*. Chestnut Hill, MA: Boston College.

Martin, M.O., Hoyle, C., and Gregory, K. (1996b). "Observing the TIMSS Testing Sessions" in M.O. Martin and I.V.S. Mullis (eds.), *Third International Mathematics and Science Study: Quality Assurance in Data Collection*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Kelly, D.L., and Haley, K. (1996). "Translation Verification" in M.O. Martin and I.V.S. Mullis (eds.), *Third International Mathematics and Science Study: Quality Assurance in Data Collection.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S. and Smith, T.A. (1996). "Quality Control Steps for Free-Response Coding" in M.O. Martin and I.V.S. (eds.), *Third International Mathematics and Science Study: Quality Assurance in Data Collection*. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994a). *International Codebooks–Populations 1 and 2 (*Doc. Ref.: ICC892-893/NRC428-428). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994b). *Performance Assessment Administration Manual for the Main Survey* (Doc. Ref.: ICC884/NRC421). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994c). *Sampling Manual–Version 4* (Doc. Ref.: ICC 439/NPC117). Prepared by Pierre Foy and Andreas Schleicher. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994d). *School Coordinator Manual–Populations 1 and 2* (Doc. Ref.: ICC891/NRC427). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994e). *School Coordinator Manual–Population 3* (Doc. Ref.: ICC907/NRC440). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994f). *Survey Operations Manual– Populations 1 and 2* (Doc. Ref.: ICC889/NRC425). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994g). *Survey Operations Manual–Population 3* (Doc. Ref.: ICC 906/NRC439). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994h). *Test Administrator Manual* (Doc. Ref.: ICC890/NRC426). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995a). *Coding Guide for Free-Response Items–Populations 1 and 2* (Doc. Ref.: ICC897/NRC433). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995b). *Coding Guide for Free-Response Items–Population 3* (Doc. Ref.: ICC 913/NRC446). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995c). *Guide to Checking, Coding, and Entering the TIMSS Data* (Doc Ref.: ICC918/NRC449). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995d). *International Codebook–Population 3* (Doc. Ref.: ICC912/NRC445). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995e). *Manual for the TIMSS Quality Control Monitors* (Doc. Ref.: ICC920/NRC450). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995f). *Supplement to the Coding Guide for Performance Assessment* (Doc. Ref.: ICC933/NRC456). Chestnut Hill, MA: Boston College.

APPENDIX A - ACKNOWLEDGMENTS

TIMSS was truly a collaborative effort among hundreds of individuals around the world. Staff from the national research centers, the international management, advisors, and funding agencies worked closely to design and implement the most ambitious study of international comparative achievement ever undertaken. TIMSS would not have been possible without the tireless efforts of all involved. Below, the individuals and organizations are acknowledged for their contributions. Given that implementing TIMSS has spanned more than seven years and involved so many people and organizations, this list may not pay heed to all who contributed throughout the life of the project. Any omission is inadvertent. TIMSS also acknowledges the students, teachers, and school principals who contributed their time and effort to the study. This report would not be possible without them.

MANAGEMENT AND OPERATIONS

Since 1993, TIMSS has been directed by the International Study Center at Boston College in the United States. Prior to this, the study was coordinated by the International Coordinating Center at the University of British Columbia in Canada. Although the study was directed centrally by the International Study Center and its staff members implemented various parts of TIMSS, important activities also were carried out in centers around the world. The data were processed centrally by the IEA Data Processing Center in Hamburg, Germany. Statistics Canada was responsible for collecting and evaluating the sampling documentation from each country and for calculating the sampling weights. The Australian Council for Educational Research conducted the scaling of the achievement data.

International Study Center (1993-)

Albert E. Beaton, International Study Director
Michael O. Martin, Deputy International Study Director
Ina V.S. Mullis, Codeputy International Study Director
Eugenio J. Gonzalez, Director of Operations and Data Analysis
Dana L. Kelly, Research Associate
Teresa A. Smith, Research Associate
Maryellen Harmon, Performance Assessment Coordinator
Robert Jin, Computer Programmer
William J. Crowley, Fiscal Administrator
Thomas M. Hoffmann, Art Director
Debora Galanti, Art Director (former)
Jonathan R. English, Systems Manager
José Rafael Nieto, Senior Production Specialist
Ann G.A. Tan, Conference Coordinator
Mary C. Howard, Office Supervisor
Cheryl L. Flaherty, Secretary
Diane Joyce, Secretary
Leanne Teixeira, Secretary (former)
Kelvin D. Gregory, Graduate Assistant
Kathleen A. Haley, Graduate Assistant
Craig D. Hoyle, Graduate Assistant

International Coordinating Center (1991-93)

David F. Robitaille, International Coordinator
Robert A. Garden, Deputy International Coordinator
Barry Anderson, Director of Operations
Beverley Maxwell, Director of Data Management

Statistics Canada

Pierre Foy, Senior Methodologist
Suzelle Giroux, Senior Methodologist
Jean Dumais, Senior Methodologist
Nancy Darcovich, Senior Methodologist
Marc Joncas, Senior Methodologist
Laurie Reedman, Junior Methodologist
Claudio Perez, Junior Methodologist

IEA Data Processing Center

Michael Bruneforth, Senior Researcher
Jedidiah Harris, Research Assistant
Dirk Hastedt, Senior Researcher
Heiko Jungclaus, Senior Researcher
Svenja Moeller, Research Assistant
Knut Schwippert, Senior Researcher
Jockel Wolff, Research Assistant

Australian Council for Educational Research

Raymond J. Adams, Principal Research Fellow
Margaret Wu, Research Fellow
Nikolai Volodin, Research Fellow
David Roberts, Research Officer
Greg Macaskill, Research Officer

FUNDING AGENCIES

Funding for the International Study Center was provided by the National Center for Education Statistics of the U.S. Department of Education, the U.S. National Science Foundation, and the International Association for the Evaluation for Educational Achievement. Eugene Owen and Lois Peak of the National Center for Education Statistics and Larry Suter of the National Science Foundation each played a crucial role in making TIMSS possible and for ensuring the quality of the study. Funding for the International Coordinating Center was provided by the Applied Research Branch of the Strategic Policy Group of the Canadian Ministry of Human Resources Development. This initial source of funding was vital to initiate the TIMSS project. Tjeerd Plomp, Chair of the IEA and of the TIMSS Steering Committee, has been a constant source of support throughout TIMSS. It should be noted that each country provided its own funding for the implementation of the study at the national level.

NATIONAL RESEARCH COORDINATORS

The TIMSS National Research Coordinators and their staff had the enormous task of implementing the TIMSS design in their countries. This required obtaining funding for the project; participating in the development of the instruments and procedures; conducting field tests; participating in and conducting training sessions; translating the instruments and procedural manuals into the local language; selecting the sample of schools and students; working with the schools to arrange for the testing; arranging for data collection, coding, and data entry; preparing the data files for submission to the IEA Data Processing Center; contributing to the development of the international reports; and preparing national reports. The way in which the national centers operated and the resources that were available varied considerably across the TIMSS countries. In some countries, the tasks were conducted centrally, while in others, various components were subcontracted to other organizations. In some countries, resources were more than adequate, while in others, the national centers were operating with limited resources. Of course, across the life of the project, some NRCs have changed. This list attempts to include all past NRCs who served for a significant period of time as well as all the present NRCs. All of the TIMSS National Research Coordinators and their staff members are to be commended for their professionalism and their dedication in conducting all aspects of TIMSS.

Argentina
Carlos Mansilla
Universidad del Chaco
Av. Italia 350
3500 Resistencia
Chaco, Argentina

Australia
Jan Lokan
Raymond Adams*
Australian Council for Educational Research
19 Prospect Hill
Private Bag 55
Camberwell, Victoria 3124
Australia

Austria
Guenter Haider
Austrian IEA Research Centre
Universität Salzburg
Akademiestraße 26/2
A-5020 Salzburg, Austria

Belgium (Flemish)
Christiane Brusselmans-Dehairs
Rijksuniversiteit Ghent
Vakgroep Onderwijskunde &
The Ministry of Education
Henri Dunantlaan 2
B-9000 Ghent, Belgium

Belgium (French)
Georges Henry
Christian Monseur
Université de Liège
B32 Sart-Tilman
4000 Liège 1, Belgium

Bulgaria
Kiril Bankov
Foundation for Research, Communication,
Education and Informatics
Tzarigradsko Shausse 125, Bl. 5
1113 Sofia, Bulgaria

Canada
Alan Taylor
Applied Research & Evaluation Services
University of British Columbia
2125 Main Mall
Vancouver, B.C. V6T 1Z4
Canada

Colombia
Carlos Jairo Diaz
Universidad del Valle
Facultad de Ciencias
Multitaller de Materiales Didacticos
Ciudad Universitaria Meléndez
Apartado Aereo 25360
Cali, Colombia

*Past National Research Coordinator

Cyprus
Constantinos Papanastasiou
Department of Education
University of Cyprus
Kallipoleos 75
P.O. Box 537
Nicosia 133, Cyprus

Czech Republic
Jana Strakova
Vladislav Tomasek
Institute for Information on Education
Senovazne Nam. 26
111 21 Praha 1, Czech Republic

Denmark
Peter Weng
Peter Allerup
Borge Prien*
The Danish National Institute for
Educational Research
28 Hermodsgade
Dk-2200 Copenhagen N, Denmark

England
Wendy Keys
Derek Foxman*
National Foundation for Educational Research
The Mere, Upton Park
Slough, Berkshire SL1 2DQ
England

France
Anne Servant
Ministère de l'Education
Nationale 142, rue du Bac
75007 Paris, France

Josette Le Coq*
Centre International d'Etudes Pédagogiques
1 Avenue Léon Journault
93211 Sèvres, France

Germany
Rainer Lehmann
Humboldt-Universitaet zu Berlin
Institut Fuer Allgemeine Erziehungswissenschaft
Geschwister-Scholl-Str. 6
10099 Berlin, Germany

Juergen Baumert
Max-Planck Institute for Human
Development and Education
Lentzeallee 94
14191 Berlin, Germany

Manfred Lehrke
Universität Kiel
IPN Olshausen Str. 62
24098 Kiel, Germany

Greece
Georgia Kontongiannopoulou-Polydorides
Joseph Solomon
University of Athens
Department of Education
Ippokratous Str. 35
106 80 Athens, Greece

Hong Kong
Frederick Leung
Nancy Law
The University of Hong Kong
Department of Curriculum Studies
Pokfulam Road, Hong Kong

Hungary
Péter Vari
National Institute of Public Education
Centre for Evaluation Studies
Dorottya U. 8, P.O. Box 120
1051 Budapest, Hungary

Iceland
Einar Gudmundsson
Institute for Educational Research
Department of Educational Testing
and Measurement
Surdgata 39
101 Reykjavik, Iceland

Indonesia
Jahja Umar
Ministry of Education and Culture
Examination Development Center
Jalan Gunung Sahari - 4
Jakarta 10000, Indonesia

Ireland
Deirdre Stuart
Michael Martin*
Educational Research Centre
St. Patrick's College
Drumcondra
Dublin 9, Ireland

Iran, Islamic Republic
Ali Reza Kiamanesh
Ministry of Education
Center for Educational Research
Iranshahr Shomali Avenue
Teheran 15875, Iran

Israel
Pinchas Tamir
The Hebrew University
Israel Science Teaching Center
Jerusalem 91904, Israel

Italy
Anna Maria Caputo
Ministerio della Pubblica Istruzione
Centro Europeo dell Educazione
Villa Falconieri
00044 Frascati, Italy

Japan
Masao Miyake
Eizo Nagasaki
National Institute for Educational Research
6-5-22 Shimomeguro
Meguro-Ku, Tokyo 153, Japan

Korea
Jingyu Kim
Hyung Im*
National Board of Educational Evaluation
Evaluation Research Division
Chungdam-2 Dong 15-1, Kangnam-Ku
Seoul 135-102, Korea

Kuwait
Mansour Hussein
Ministry of Education
P. O. Box 7
Safat 13001, Kuwait

Latvia
Andrejs Geske
University of Latvia
Faculty of Education & Psychology
Jurmalas Gatve 74/76, Rm. 204a
Riga, Lv-1083, Latvia

Lithuania
Algirdas Zabulionis
University of Vilnius
Faculty of Mathematics
Naugarduko 24
2006 Vilnius, Lithuania

Mexico
Fernando Córdova Calderón
Director de Evaluación de Politicas y
Sistemas Educativos
Netzahualcoyotl #127 2ndo Piso
Colonia Centro
Mexico 1, D.F., Mexico

Netherlands
Wilmad Kuiper
Anja Knuver
Klaas Bos
University of Twente
Faculty of Educational Science
and Technology
Department of Curriculum
P.O. Box 217
7500 AE Enschede, Netherlands

New Zealand
Hans Wagemaker
Steve May
Ministry of Education
Research Section
45-47 Pipitea Street
Wellington, New Zealand

Norway
Svein Lie
University of Oslo
SLS Postboks 1099
Blindern 0316
Oslo 3, Norway

Gard Brekke
Alf Andersensv 13
3670 Notodden, Norway

Philippines
Milagros Ibe
University of the Philippines
Institute for Science and Mathematics
Education Development
Diliman, Quezon City
Philippines

Ester Ogena
Science Education Institute
Department of Science and Technology
Bicutan, Taquig
Metro Manila 1604, Philippines

Portugal
Gertrudes Amaro
Ministerio da Educacao
Instituto de Inovação Educacional
Rua Artilharia Um 105
1070 Lisboa, Portugal

Romania
Gabriela Noveanu
Institute for Educational Sciences
Evaluation and Forecasting Division
Str. Stirbei Voda 37
70732-Bucharest, Romania

Russian Federation
Galina Kovalyova
The Russian Academy of Education
Institute of General Secondary School
Ul. Pogodinskaya 8
Moscow 119905, Russian Federation

Scotland
Brian Semple
Scottish Office,
Education & Industry Department
Victoria Quay
Edinburgh, E86 6qq
Scotland

Singapore
Chan Siew Eng
Research and Evaluation Branch
Block A Belvedere Building
Ministry of Education
Kay Siang Road
Singapore 248922

Slovak Republic
Maria Berova
Vladimir Burjan*
SPU-National Institute for Education
Pluhova 8
P.O. Box 26
830 00 Bratislava
Slovak Republic

Slovenia
Marjan Setinc
Pedagoski Institut Pri Univerzi v Ljubljana
Gerbiceva 62, P.O. Box 76
61111 Ljubljana, Slovenia

South Africa
Derek Gray
Human Sciences Research Council
134 Pretorius Street
Private Bag X41
Pretoria 0001, South Africa

Spain
José Antonio Lopez Varona
Instituto Nacional de Calidad y Evaluación
C/San Fernando del Jarama No. 14
28071 Madrid, Spain

Sweden
Ingemar Wedman
Anna Hofslagare
Kjell Gisselberg*
Umeå University
Department of Educational Measurement
S-901 87 Umeå, Sweden

Switzerland
Erich Ramseier
Amt Für Bildungsforschung der
Erziehungsdirektion des Kantons Bern Sulgeneck
Straße 70
Ch-3005 Bern, Switzerland

Thailand
Suwaporn Semheng
Institute for the Promotion of Teaching
Science and Technology
924 Sukhumvit Road
Bangkok 10110, Thailand

United States
William Schmidt
Michigan State University
Department of Educational Psychology
463 Erikson Hall
East Lansing, MI 48824-1034
United States

TIMSS ADVISORY COMMITTEES

The International Study Center was supported in its work by several advisory committees. The International Steering Committee provided guidance to the International Study Director on policy issues and general direction of the study. The TIMSS Technical Advisory Committee provided guidance on issues related to design, sampling, instrument construction, analysis, and reporting, ensuring that the TIMSS methodologies and procedures were technically sound. The Subject Matter Advisory Committee ensured that current thinking in math and science education were addressed by TIMSS, and was instrumental in the development of the TIMSS tests. The Free-Response Item Coding Committee developed the coding rubrics for the free-response items. The Performance Assessment Committee worked with the Performance Assessment Coordinator to develop the TIMSS performance assessment. The Quality Assurance Committee helped to develop the quality assurance program.

International Steering Committee

Tjeerd Plomp (Chair), the Netherlands
Lars Ingelstam, Sweden
Daniel Levine, United States
Senta Raizen, United States
David Robitaille, Canada
Toshio Sawada, Japan
Benny Suprapto Brotosiswojo, Indonesia
William Schmidt, United States

Technical Advisory Committee

Raymond Adams, Australia
Pierre Foy, Canada
Andreas Schleicher, Germany
William Schmidt, United States
Trevor Williams, United States

Sampling Referee

Keith Rust, United States

Subject Area Coordinators

Robert Garden, New Zealand (Mathematics)
Graham Orpwood, Canada (Science)

Special Mathematics Consultant

Chancey Jones

Subject Matter Advisory Committee

Svein Lie, (Chair), Norway
Antoine Bodin, France
Peter Fensham, Australia
Robert Garden, New Zealand
Geoffrey Howson, England
Curtis McKnight, United States
Graham Orpwood, Canada
Senta Raizen, United States
David Robitaille, Canada
Pinchas Tamir, Israel
Alan Taylor, Canada
Ken Travers, United States
Theo Wubbels, the Netherlands

Free-Response Item Coding Committee

Svein Lie (Chair), Norway
Vladimir Burjan, Slovak Republic
Kjell Gisselberg, Sweden
Galina Kovalyova, Russian Federation
Nancy Law, Hong Kong
Josette Le Coq, France
Jan Lokan, Australia
Curtis McKnight, United States
Graham Orpwood, Canada
Senta Raizen, United States
Alan Taylor, Canada
Peter Weng, Denmark
Algirdas Zabulionis, Lithuania

Performance Assessment Committee

Derek Foxman, England
Robert Garden, New Zealand
Per Morten Kind, Norway
Svein Lie, Norway
Jan Lokan, Australia
Graham Orpwood, Canada

Quality Control Committee

Jules Goodison, United States
Hans Pelgrum, the Netherlands
Ken Ross, Australia

Editorial Committee

David Robitaille, Chair, Canada
Albert Beaton, International Study Director
Paul Black, England
Svein Lie, Norway
Rev. Ben Nebres, the Philippines
Judith Torney-Purta, United States
Ken Travers, United States
Theo Wubbels, the Netherlands

APPENDIX B - TIMSS TEST BLUEPRINTS

Table B.1 Population 1 Mathematics Blueprint: Time Allocations in Minutes by Content Grouping and Performance Category

	Performance Category						
Content grouping	Knowing	Routine procedures	Complex procedures	Solving problems	Justifying and proving	Commun-icating	Total minutes
Whole numbers: place value	6	2	2	3	3	0	16
Whole numbers: other content	4	5	2	0	0	0	11
Decimal fractions: meaning, representation and operations	3	2	0	3	0	0	8
Common fractions: meaning, representation and operations	4	0	1	1	0	3	9
Proportionality	1	0	4	1	0	6	12
Estimation and number sense	2	0	4	1	0	0	7
Measurement	5	6	0	4	0	0	15
Data analysis	3	0	8	3	0	0	14
Probability	0	0	1	1	0	0	2
Geometry	9	0	4	0	0	1	14
Patterns, relations, and functions	5	1	0	4	0	0	10
Total minutes	**42**	**16**	**26**	**21**	**3**	**10**	**118**

Table B.2 Population 1 Science Blueprint: Time Allocations in Minutes by Content Grouping and Performance Category

	Performance Category				
Content grouping	Understanding	Theorizing, analyzing, and solving problems	Using tools, routine procedures, and science processes	Investigating the natural world	Total minutes
Earth features	6	5	0	1	12
Earth science: other content	5	3	1	0	9
Human biology	12	2	0	0	14
Life science: other content	31	1	1	0	33
Physical science	19	13	3	1	36
Environment	9	0	0	0	9
Other content	2	0	1	1	4
Total minutes	**84**	**24**	**6**	**3**	**117**

Table B.3 Population 2 Mathematics Blueprint: Time Allocations in Minutes by Content Grouping and Performance Category

	Performance Category						
Content grouping	Knowing	Routine procedures	Complex procedures	Solving problems	Justifying and proving	Communicating	Total minutes
Common fractions: meaning, representation	7	0	3	0	0	0	10
Common fractions: operations, relations and proportions	1	5	2	9	0	0	17
Decimal fractions	3	7	2	5	0	0	17
Estimation and number sense	2	4	6	8	0	0	20
Congruence and similarity	2	2	1	1	0	0	6
Other geometry	3	4	6	5	0	0	18
Linear equations	3	2	1	8	0	0	14
Other algebra	5	10	0	4	5	0	24
Data representation and analysis	2	2	8	2	0	5	19
Probability	1	0	1	5	0	0	7
Measurement	6	3	6	9	0	5	29
Proportionality	0	5	0	7	5	0	17
Total minutes	**35**	**44**	**36**	**63**	**10**	**10**	**198**

Table B.4 Population 2 Science Blueprint: Time Allocations in Minutes by Content Grouping and Performance Category

	Performance Category				
Content grouping	**Understanding**	**Theorizing, analyzing, and solving problems**	**Using tools, routine procedures, and science processes**	**Investigating the natural world**	**Total minutes**
Earth features	10	5	1	0	16
Earth science: other content	6	9	2	0	17
Human biology	12	3	0	5	20
Life science: other content	32	7	1	1	41
Energy types etc.	5	10	0	0	15
Light	7	7	0	0	14
Physics: other content	16	13	1	2	32
Chemistry	16	9	1	0	26
Environment	5	5	0	0	10
Other content	3	3	2	2	10
Total minutes	**112**	**71**	**8**	**10**	**201**

Table B.5 Population 3 Mathematics Literacy Blueprint: Time Allocation in Minutes by Content Grouping and Performance Category

	Performance Category				
Content grouping	Knowing	Routine procedures	Complex procedures	Solving problems	Total minutes
Number sense	1	10	1	5	17
Algebraic sense	5	1	0	3	9
Measurement and estimation	3	1	5	10	19
Reasoning and social utility	0	0	2	13	15
Total minutes	**9**	**12**	**8**	**31**	**60**

Table B.6 Population 3 Science Literacy Blueprint: Time Allocation in Minutes by Content Grouping and Performance Category

	Performance Category				
Content grouping	Understanding	Theorizing, analyzing, and solving problems	Using tools, routine procedures, and science processes	Investigating the natural world	Total minutes
Earth science	1	8	0	0	9
Human biology	9	2	0	0	11
Other life science	3	0	2	1	6
Energy	3	5	0	0	8
Other physical science	5	6	0	0	11
Reasoning and social utility	4	7	5	0	16
Total minutes	**25**	**28**	**7**	**1**	**61**

Table B.7 Population 3 Advanced Mathematics Blueprint: Time Allocation in Minutes by Content Grouping and Performance Category

	Performance Category						
Content grouping	Knowing	Routine procedures	Complex procedures	Solving problems	Justifying and proving	Communicating	Total minutes
Numbers, equations and functions	3	18	6	25	0	3	55
Analysis (calculus)	6	24	0	14	0	3	47
Geometry	15	18	9	20	10	5	77
Probability and statistics	3	6	3	9	0	0	21
Validation and structure	0	3	0	0	8	0	11
Total minutes	**27**	**69**	**18**	**68**	**18**	**11**	**211**

Table B.8 Population 3 Physics Blueprint: Time Allocation in Minutes by Content Grouping and Performance Category

	Performance Category				
Content grouping	Understanding	Theorizing, analyzing, and solving problems	Using tools, routine procedures, and science processes	Investigating the natural world	Total minutes
Forces and motion	3	30	12	5	50
Electricity and magnetism	12	39	3	0	54
Thermal and wave phenomena	12	12	3	0	27
Particle physics and relativity	12	15	0	5	32
Energy	12	27	6	3	48
Total minutes	**51**	**123**	**24**	**13**	**211**