TIMSS

Technical Report
Volume II

TIMSS

Third International Mathematics and Science Study

Technical Report

Volume II: Implementation and Analysis

Primary and Middle School Years
(Population 1 and Population 2)

Edited by

Michael O. Martin
Dana L. Kelly

with contributors

Raymond J. Adams
Albert E. Beaton
Pierre Foy
Eugenio J. Gonzalez
Dirk Hastedt
Greg Macaskill
Ina V.S. Mullis
Knut Schwippert
Heiko Sibberns
Teresa Smith
Margaret L. Wu

Boston College • Chestnut Hill, Massachusetts, USA

Third International Mathematics and Science Study Technical Report, Volume II: Implementation and Analysis – Primary and Middle School Years/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: 1-889938-06-8

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 internatinal coordination of TIMSS is provided by the U.S. National Center for Educational 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

Forward

The design, implementation, and analysis of the Third International Mathematics and Science Study (TIMSS) was a collaborative effort among various institutions and individuals around the world. The conduct of TIMSS was a very ambitious undertaking that required considerable resources, expertise, and the dedication of all involved. The technical documentation is a very important component of this study. The first volume in this series, the *TIMSS Technical Report, Volume I: Design and Development*, describes the design and development of the study, including the development of the achievement tests and questionnaires, the sample design and field operations procedures, and the plans for quality assurance procedures.

I am pleased to introduce the *TIMSS Technical Report, Volume II*, documenting the implementation and analysis of the assessment of students in the primary and middle school years. The publication of this volume represents a milestone for TIMSS. The pages that follow describe the activities carried out to implement this very large international study, and the analytic procedures underlying the analysis and reporting of the data. The implementation of the sample design, the calculation of sampling weights, procedures for the estimation of sampling variability, steps involved in the international data verification, the TIMSS scaling model, and the analysis of the achievement and background data, are all presented in this volume. Together with the achievement reports presenting the study results and the international database, all released to the public within the last 15 months, this volume completes the reporting of the primary and middle school assessment. The third, and final, volume in this series will describe the implementation of the TIMSS design and the analysis and reporting of results for students in the final year of secondary school.

Albert E. Beaton
TIMSS International
Study Director

Acknowledgments

TIMSS was truly a collaborative effort among hundreds of individuals around the world. Staff from the national research centers of the participating countries, the international management, advisors, and funding agencies worked closely to design and implement the most ambitious study of international comparative achievement in mathematics and science ever undertaken. The design was implemented in each country by the TIMSS national research center staff, with the cooperation and assistance of schools, and the participation of the students and teachers. This volume documents the efforts of those involved in the implementation of the very ambitious TIMSS design, and the steps undertaken to analyze and report the international results for students in the primary and middle school years (third, fourth, seventh, and eighth grades in most countries).

It is impossible to acknowledge individually everyone who contributed to the implementation and analysis of TIMSS. Chapter authors have recognized significant contributors where appropriate, and the Acknowledgments section at the end of the volume further acknowledges the National Research Coordinators and special advisors. Without the financial support provided by the National Center for Education Statistics of the U.S. Department of Education, the U.S. National Science Foundation, the Canadian government, and the IEA, the design, development, and implementation of TIMSS would not have been possible. Special acknowledgment is given to these organizations for funding the international coordination of the study.

This report would not have been possible without the efforts of many people. We would like to thank each author for her or his contribution. We also would like to thank Albert Beaton, the TIMSS International Study Director, for his constant help and support in this endeavor.

Several individuals at the TIMSS International Study Center at Boston College deserve special recognition for the production of this report. José R. Nieto coordinated the production of the report, including designing the layout and cover, scheduling production tasks, and assembling the text and tables. Cheryl Flaherty and Michelle Range showed extraordinary patience and diligence in typing and proofing the many revisions of this report. Special thanks go to Maria Sachs for editing the text.

Michael O. Martin
Dana L. Kelly
Boston College

1 Third International Mathematics and Science Study: An Overview

Michael O. Martin
Dana L. Kelly
Boston College

1.1 INTRODUCTION

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

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

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

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

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

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

Figure 1.1 Countries Participating in TIMSS*

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

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

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

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

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

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

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

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

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

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

1.2 THE CONCEPTUAL FRAMEWORK FOR TIMSS

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

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

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

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

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

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

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

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

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

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

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

The student achievement survey provides data for the study of the attained curriculum. The wide-ranging mathematics and science tests that were administered to nationally representative samples of students at three levels of the educational system provide not only a sound basis for international comparisons of student achievement, but a rich resource for the study of the attained curriculum in each country. Information about students' characteristics, and about their attitudes, beliefs, and experiences, comes from a questionnaire completed by each participating student. This information will help to identify the student characteristics associated with learning and provide a context for the study of the attained curriculum.

1.3 THE TIMSS CURRICULUM FRAMEWORKS

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

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

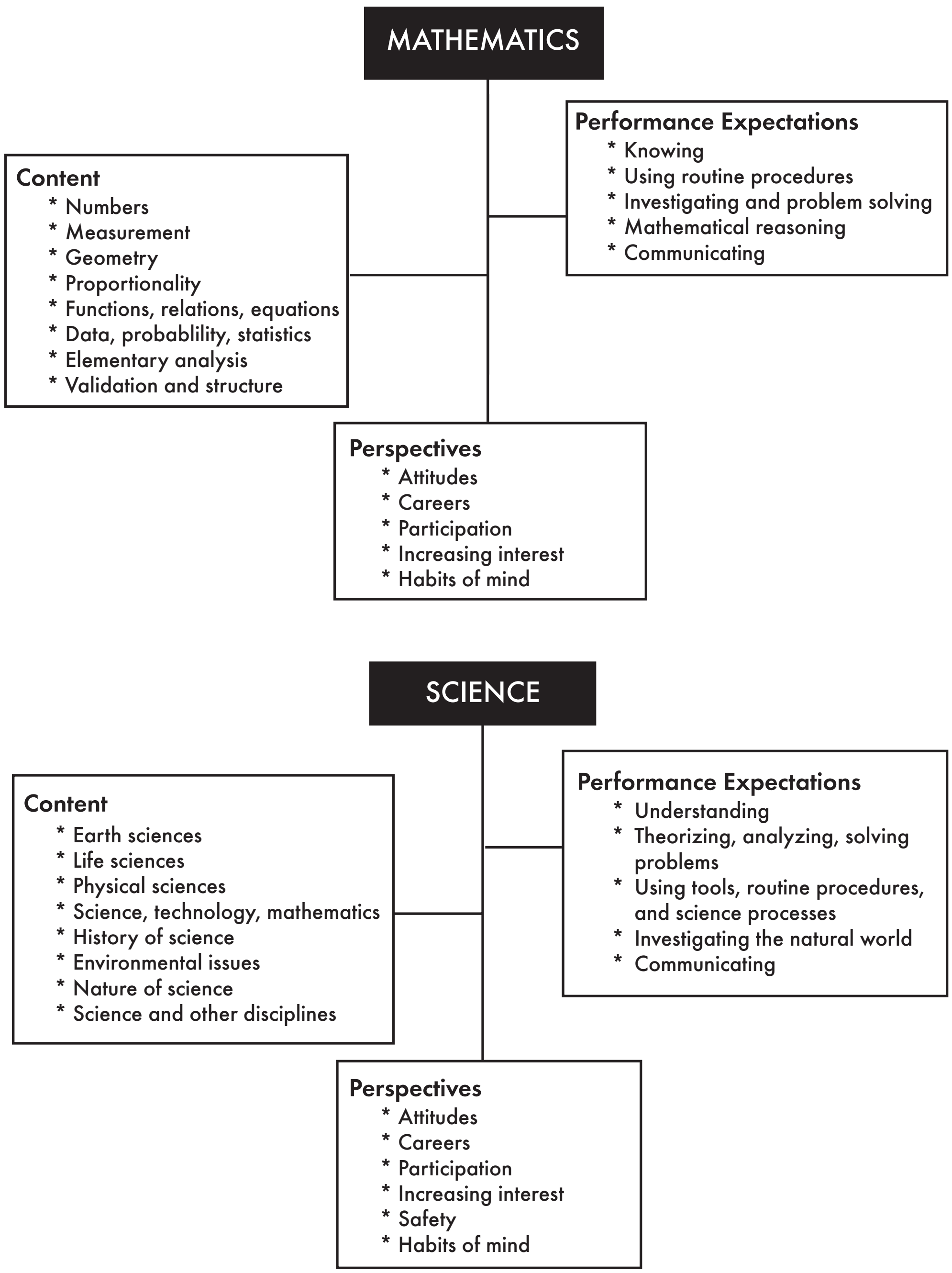
The TIMSS curriculum frameworks are an ambitious attempt to expand the concept of the content-by-cognitive-behavior grids.

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

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

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

Figure 1.2 The Major Categories of the TIMSS Curriculum Frameworks



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

1.4 THE TIMSS CURRICULUM ANALYSIS

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

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

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

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

1.5 THE STUDENT POPULATIONS

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

The student populations in TIMSS are defined below.

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

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

1.6 SURVEY ADMINISTRATION DATES FOR POPULATIONS 1 AND 2

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

1.7 THE TIMSS ACHIEVEMENT TESTS FOR POPULATIONS 1 AND 2

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

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

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

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

Despite the need for compromise in some areas, the TIMSS achievement tests have gone a long way toward meeting the ideals of their designers. They cover a wide range of subject matter, yielding, in Population 2, estimates of student proficiency in 11 areas or content area "reporting categories" of mathematics and science (6 for mathematics and 5 for science), as well as overall mathematics and science scores. In Population 1 there were ten content area reporting categories (six for mathematics and four for science), as well as overall mathematics and overall science scores. The test items include both multiple-choice and free-response items. The latter come in two varieties: "short-answer," where the student supplies a brief written response; and "extended-response," where students must provide a more extensive written answer, and sometimes explain their reasoning. The free-response items are scored using a unique 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 them were to be administered to any one student, would take almost seven hours of testing. Since the consensus among the National Research Coordinators (NRCs) was that 70 minutes was the most that could be expected for Population 1 and 90 minutes the most that could be expected for Population 2, a way of dividing the item pool among the students had to be found. Matrix sampling provided a solution by assigning subsets of items to individual students in such a way as to produce reliable estimates of the performance of the population on all the items, even though no student responded to the entire item pool. The TIMSS test design uses a variant of matrix sampling to map the mathematics and science item pool into eight student booklets each for Population 1 and Population 2 (see Adams and Gonzalez, 1996).

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

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

1.8 PERFORMANCE ASSESSMENT

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

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

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

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

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

1.9 THE BACKGROUND QUESTIONNAIRES

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

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

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

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

1.10 MANAGEMENT AND OPERATIONS

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

- Meeting with other NRCs and international project staff to plan the study and develop instruments and procedures

- Defining the school populations from which the TIMSS samples were to be drawn, selecting the sample of schools using an approved random sampling procedure, contacting the school principals and securing their agreement to participate in the study, and selecting the classes to be tested, again using an approved random sampling procedure

- Translating and adapting all of the tests, questionnaires, and administration manuals into the language of instruction of the country (and sometimes more than one language) prior to data collection

- Assembling, printing, and packaging the test booklets and questionnaires, and shipping the survey materials to the participating schools

- Ensuring that the tests and questionnaires were administered in participating schools, either by teachers in the school or by an external team of test administrators, and that the completed test protocols were returned to the TIMSS national center

- Conducting a quality assurance exercise in conjunction with the test administration, whereby some testing sessions were attended by an independent observer to confirm that all specified procedures were followed

- Recruiting and training individuals to score the free-response questions in the achievement tests, and implementing the plan for scoring the student responses, including the plan for assessing the reliability of the scoring procedure

- Recruiting and training data entry personnel for keying the responses of students, teachers, and principals into computerized data files, and conducting the data entry operation using the software provided

- Checking the accuracy and integrity of the data files prior to shipping them to the IEA Data Processing Center in Hamburg

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

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

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

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

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

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

4 See the Acknowledgments section for membership of TIMSS committees.

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

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

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

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

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

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

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

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

1.11 SUMMARY OF THIS REPORT

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

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

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

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

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

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

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

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

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

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

2 Implementation of the TIMSS Sample Design

Pierre Foy
Statistics Canada

2.1 TIMSS TARGET POPULATIONS

2.1.1 Definitions

The international desired target populations for TIMSS are defined below.[1]

> **Population 1**: All students enrolled in the two adjacent grades that contain the largest proportion of 9-year-old students at the time of testing

> **Population 2**: All students enrolled in the two adjacent grades that contain the largest proportion of 13-year-old students at the time of testing

Tables 2.1 and 2.2 summarize the grades all participating countries identified as their target populations for the TIMSS Population 1 and Population 2. These tables are those published in the TIMSS international reports (Beaton et al., 1996a; Beaton et al., 1996b; Martin et al., 1997; Mullis et al., 1997). Additional details on these definitions are provided in Appendix B. As shown in the tables, most countries tested the third and fourth grades for Population 1 and the seventh and eighth grades for Population 2. Countries that participated in the performance assessment subsampled students from the upper grade in each of these populations.

Tables 2.3 and 2.4 show the coverage of 9-year-old and 13-year-old students, respectively, across the two grades tested at each population in each country. On occasion, the selected target grades led to the sampling of students older than expected. This was the case for Colombia (Population 2), Germany (Population 2), Kuwait (Population 1 and Population 2), Romania (Population 2), Slovenia (Population 1 and Population 2), and Thailand (Population 1).

[1] A third TIMSS student population – Population 3 – consisted of students in their final year of secondary school. A technical report describing Population 3 activities is forthcoming.

Table 2.1 Information About the Grades Tested - Population 1

Country	Lower Grade		Upper Grade	
	Country's Name for Lower Grade	Years of Formal Schooling Including Lower Grade[1]	Country's Name for Upper Grade	Years of Formal Schooling Including Upper Grade[1]
[2] Australia	3 or 4	3 or 4	4 or 5	4 or 5
Austria	3	3	4	4
Canada	3	3	4	4
Cyprus	3	3	4	4
Czech Republic	3	3	4	4
England	Year 4	4	Year 5	5
Greece	3	3	4	4
Hong Kong	Primary 3	3	Primary 4	4
Hungary	3	3	4	4
Iceland	3	3	4	4
Iran, Islamic Rep.	3	3	4	4
Ireland	3rd Class	3	4th Class	4
Israel	–	–	4	4
Japan	3	3	4	4
Korea	3rd Grade	3	4th Grade	4
Kuwait	–	–	5	5
Latvia	3	3	4	4
[3] Netherlands	5	3	6	4
[4] New Zealand	Standard 2	3.5-4.5	Standard 3	4.5-5.5
Norway	2	2	3	3
Portugal	3	3	4	4
Scotland	Year 4	4	Year 5	5
Singapore	Primary 3	3	Primary 4	4
Slovenia	3	3	4	4
Thailand	Primary 3	3	Primary 4	4
United States	3	3	4	4

[1] Years of schooling based on the number of years children in the grade level have been in formal schooling, beginning with primary education (International Standard Classification of Education Level 1). Does not include preprimary education.

[2] Australia: Each state/territory has its own policy regarding age of entry to primary school. In 4 of the 8 states/territories students were sampled from grades 3 and 4; in the other four states/territories students were sampled from grades 4 and 5.

[3] In the Netherlands kindergarten is integrated with primary education. Grade-counting starts at age 4 (formerly kindergarten 1). Formal schooling in reading, writing, and arithmetic starts in grade 3, age 6.

[4] New Zealand: The majority of students begin primary school on or near their 5th birthday so the "years of formal schooling" vary.

Table 2.2 Information About the Grades Tested - Population 2

Country	Lower Grade: Country's Name for Lower Grade	Lower Grade: Years of Formal Schooling Including Lower Grade[1]	Upper Grade: Country's Name for Upper Grade	Upper Grade: Years of Formal Schooling Including Upper Grade[1]
[2] Australia	7 or 8	7 or 8	8 or 9	8 or 9
Austria	3. Klasse	7	4. Klasse	8
Belgium (Fl)	1A	7	2A & 2P	8
Belgium (Fr)	1A	7	2A & 2P	8
Bulgaria	7	7	8	8
Canada	7	7	8	8
Colombia	7	7	8	8
[3] Cyprus	7	7	8	8
Czech Republic	7	7	8	8
Denmark	6	6	7	7
England	Year 8	8	Year 9	9
France	5ème	7	4ème (90%) or 4ème Technologique (10%)	8
Germany	7	7	8	8
Greece	Secondary 1	7	Secondary 2	8
Hong Kong	Secondary 1	7	Secondary 2	8
Hungary	7	7	8	8
Iceland	7	7	8	8
Iran, Islamic Rep.	7	7	8	8
Ireland	1st Year	7	2nd Year	8
Israel	–	–	8	8
Japan	1st Grade Lower Secondary	7	2nd Grade Lower Secondary	8
Korea, Republic of	1st Grade Middle School	7	2nd Grade Middle School	8
Kuwait	–	–	9	9
Latvia	7	7	8	8
Lithuania	7	7	8	8
Netherlands	Secondary 1	7	Secondary 2	8
[3,4] New Zealand	Form 2	7.5 - 8.5	Form 3	8.5 - 9.5
[3] Norway	6	6	7	7
[3] Philippines	Grade 6 Elementary	6	1st Year High School	7
Portugal	Grade 7	7	Grade 8	8
Romania	7	7	8	8
[5] Russian Federation	7	6 or 7	8	7 or 8
Scotland	Secondary 1	8	Secondary 2	9
Singapore	Secondary 1	7	Secondary 2	8
Slovak Republic	7	7	8	8
Slovenia	7	7	8	8
Spain	7 EGB	7	8 EGB	8
[3] South Africa	Standard 5	7	Standard 6	8
[3] Sweden	6	6	7	7
[3] Switzerland				
(German)	6	6	7	7
(French and Italian)	7	7	8	8
Thailand	Secondary 1	7	Secondary 2	8
United States	7	7	8	8

[1] Years of schooling based on the number of years children in the grade level have been in formal schooling, beginning with primary education (International Standard Classification of Education Level 1). Does not include preprimary education.

[2] Australia: Each state/territory has its own policy regarding age of entry to primary school. In 4 of the 8 states/territories students were sampled from grades 7 and 8; in the other four states/territories students were sampled from grades 8 and 9.

[3] Indicates that there is a system-split between the lower and upper grades. In Cyprus, system-split occurs only in the large or city schools. In Switzerland there is a system-split in 14 of 26 cantons.

[4] New Zealand: The majority of students begin primary school on or near their 5th birthday so the "years of formal schooling" vary.

[5] Russian Federation: 70% of students in the seventh grade have had 6 years of formal schooling; 70% in the eighth grade have had 7 years of formal schooling.

Table 2.3 Coverage of 9-Year-Old Students

Country	Percent of 9-Year-Olds in Lower Grade (Third Grade*)	Percent of 9-Year-Olds in Upper Grade (Fourth Grade*)	Percent of 9-Year-Olds in Both Grades
Australia	65	29	94
Austria	72	15	87
Canada	46	48	94
Cyprus	35	63	98
Czech Republic	75	15	91
England	58	41	99
Greece	11	88	99
Hong Kong	43	50	93
Hungary	70	19	89
Iceland	15	84	99
Iran, Islamic Rep.	51	32	83
Ireland	68	23	92
Israel	-	-	-
Japan	91	9	99
Korea	67	24	91
Kuwait	-	-	-
Latvia (LSS)	55	21	76
Netherlands	63	30	93
New Zealand	50	49	99
Norway	38	62	100
Portugal	45	48	93
Scotland	23	76	99
Singapore	80	17	98
Slovenia	60	0	60
Thailand	60	11	71
United States	61	34	95

*Third and fourth grades in most countries; see Table 2.1 for more information about the grades tested in each country.
A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.
Because results are rounded to the nearest whole number some totals may appear inconsistent.

Table 2.4 Coverage of 13-Year-Old Students

Country	Percent of 13-Year-Olds in Lower Grade (Seventh Grade*)	Percent of 13-Year-Olds in Upper Grade (Eighth Grade*)	Percent of 13-Year-Olds in Both Grades
Australia	64	28	92
Austria	62	27	89
Belgium (Fl)	46	49	94
Belgium (Fr)	41	46	87
Bulgaria	58	37	95
Canada	48	43	91
Colombia	30	15	45
Cyprus	28	70	98
Czech Republic	73	17	90
Denmark	35	64	98
England	57	42	99
France	44	35	78
Germany	71	2	73
Greece	11	85	96
Hong Kong	44	46	90
Hungary	65	24	89
Iceland	16	83	100
Iran, Islamic Rep.	47	25	72
Ireland	69	17	86
Israel	-	-	-
Japan	91	9	100
Korea	70	28	98
Kuwait	-	-	-
Latvia (LSS)	60	26	86
Lithuania	64	26	90
Netherlands	59	31	90
New Zealand	52	47	99
Norway	43	57	100
Philippines	-	-	-
Portugal	44	32	76
Romania	67	9	76
Russian Federation	50	44	95
Scotland	24	75	99
Singapore	82	15	97
Slovak Republic	73	22	95
Slovenia	65	2	67
South Africa	36	20	55
Spain	46	39	85
Sweden	45	54	99
Switzerland	48	44	92
Thailand	58	20	78
United States	58	33	91

*Seventh and eighth grades in most countries; see Table 2.2 for more information about the grades tested in each country.
A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.
Because results are rounded to the nearest whole number, some totals may appear inconsistent.

2.1.2 Coverage and Exclusions

Tables 2.5 and 2.6 summarize the extent of national coverage and exclusions in the TIMSS target populations. These tables are those published in the TIMSS international reports. National coverage of the international desired target populations was generally comprehensive, with the few exceptions detailed in the tables. School-level exclusions generally consisted of schools for the disabled and very small schools; however, there were some national deviations that are documented in Appendix B. Within-school exclusions, generally consisted of disabled students and students that could not be assessed in the language of the national tests. A few countries had no within-school exclusions.

Table 2.5 Coverage of TIMSS Target Population - Population 1

The international desired population is defined as follows for Population 1:
All students enrolled in the two adjacent grades with the largest proportion of 9-year-old students at the time of testing.

Country	International Desired Population		National Desired Population		
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Australia	100%		0.1%	1.6%	1.8%
Austria	100%		2.6%	0.2%	2.8%
Canada	100%		2.5%	3.6%	6.2%
Cyprus	100%		3.1%	0.1%	3.2%
Czech Republic	100%		4.1%	0.0%	4.1%
[2] England	100%		8.6%	3.5%	12.1%
Greece	100%		1.5%	4.0%	5.4%
Hong Kong	100%		2.6%	0.0%	2.7%
Hungary	100%		3.8%	0.0%	3.8%
Iceland	100%		1.9%	4.3%	6.2%
Iran, Islamic Rep.	100%		0.3%	1.0%	1.3%
Ireland	100%		5.3%	1.6%	6.9%
[1] Israel	72%	Hebrew Public Education System	1.1%	0.1%	1.2%
Japan	100%		3.0%	0.0%	3.0%
Korea	100%		3.9%	2.6%	6.6%
Kuwait	100%		0.0%	0.0%	0.0%
[1] Latvia (LSS)	60%	Latvian-speaking schools	2.1%	0.0%	2.1%
Netherlands	100%		4.0%	0.4%	4.4%
New Zealand	100%		0.7%	0.6%	1.3%
Norway	100%		1.1%	2.0%	3.1%
Portugal	100%		6.6%	0.7%	7.3%
Scotland	100%		2.4%	4.3%	6.7%
Singapore	100%		0.0%	0.0%	0.0%
Slovenia	100%		1.9%	0.0%	1.9%
Thailand	100%		6.8%	1.5%	8.3%
United States	100%		0.4%	4.3%	4.7%

[1] National Desired Population does not cover all of International Desired Population. Because coverage falls below 65%, Latvia is annotated LSS for Latvian Speaking Schools only.

[2] National Defined Population covers less than 90 percent of National Desired Population.

Table 2.6 Coverage of TIMSS Target Population - Population 2

The international desired population is defined as follows for Population 2:
All students enrolled in the two adjacent grades with the largest proportion of 13-year-old students at the time of testing.

Country	International Desired Population		National Desired Population		
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Australia	100%		0.2%	0.7%	0.8%
Austria	100%		2.9%	0.2%	3.1%
Belgium (Fl)	100%		3.8%	0.0%	3.8%
Belgium (Fr)	100%		4.5%	0.0%	4.5%
Bulgaria	100%		0.6%	0.0%	0.6%
Canada	100%		2.4%	2.1%	4.5%
Colombia	100%		3.8%	0.0%	3.8%
Cyprus	100%		0.0%	0.0%	0.0%
Czech Republic	100%		4.9%	0.0%	4.9%
Denmark	100%		0.0%	0.0%	0.0%
[2] England	100%		8.4%	2.9%	11.3%
France	100%		2.0%	0.0%	2.0%
[1] Germany	88%	15 of 16 regions*	8.8%	0.9%	9.7%
Greece	100%		1.5%	1.3%	2.8%
Hong Kong	100%		2.0%	0.0%	2.0%
Hungary	100%		3.8%	0.0%	3.8%
Iceland	100%		1.7%	2.9%	4.5%
Iran, Islamic Rep.	100%		0.3%	0.0%	0.3%
Ireland	100%		0.0%	0.4%	0.4%
[1] Israel	74%	Hebrew Public Education System	3.1%	0.0%	3.1%
Japan	100%		0.6%	0.0%	0.6%
Korea	100%		2.2%	1.6%	3.8%
Kuwait	100%		0.0%	0.0%	0.0%
[1] Latvia (LSS)	51%	Latvian-speaking schools	2.9%	0.0%	2.9%
[1] Lithuania	84%	Lithuanian-speaking schools	6.6%	0.0%	6.6%
Netherlands	100%		1.2%	0.0%	1.2%
New Zealand	100%		1.3%	0.4%	1.7%
Norway	100%		0.3%	1.9%	2.2%
Philippines	91%	2 provinces and autonomous regions excluded	6.5%	0.0%	6.5%
Portugal	100%		0.0%	0.3%	0.3%
Romania	100%		2.8%	0.0%	2.8%
Russian Federation	100%		6.1%	0.2%	6.3%
Scotland	100%		0.3%	1.9%	2.2%
Singapore	100%		4.6%	0.0%	4.6%
Slovak Republic	100%		7.4%	0.1%	7.4%
Slovenia	100%		2.4%	0.2%	2.6%
South Africa	100%		9.6%	0.0%	9.6%
Spain	100%		6.0%	2.7%	8.7%
Sweden	100%		0.0%	0.9%	0.9%
[1] Switzerland	86%	22 of 26 cantons	4.4%	0.8%	5.3%
Thailand	100%		6.2%	0.0%	6.2%
United States	100%		0.4%	1.7%	2.1%

[1] National Desired Population does not cover all of International Desired Population. Because coverage falls below 65%, Latvia is annotated LSS for Latvian Speaking Schools only.

[2] National Defined Population covers less than 90 percent of National Desired Population

*One region (Baden-Wuerttemberg) did not participate.

For the performance assessment, in the interest of ensuring the quality of the administration, countries could exclude additional schools if the schools had fewer than nine students in the upper grade and thus could not provide a full complement of students for the performance assessment rotation or if the schools were in a geographically remote region (see Harmon and Kelly, 1996). The exclusion rate for the performance assessment sample was not to exceed 25 percent of the national desired population. Tables 2.7 and 2.8 show the main assessment school exclusion rates, the performance assessment school exclusion rates, the within-sample exclusion rates, and the overall exclusion rates for the upper grades for Populations 1 and 2, respectively.

Table 2.7 Coverage of TIMSS Target Population - Performance Assessment – Fourth Grade*

The international desired target population is defined as follows:
Fourth Grade - All students enrolled in the higher of the two adjacent grades with the largest proportion of 9-year-old students at the time of testing.

Country	International Desired Target Population		National Desired Target Population			
	Coverage	Notes on Coverage	Main Assessment School-Level Exclusions	Performance Assessment School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Australia	100%		0.1%	15.1%	1.4%	16.7%
Canada	100%		2.5%	15.4%	3.1%	21.0%
Cyprus	100%		3.1%	0.0%	0.1%	3.2%
Hong Kong	100%		2.6%	1.9%	0.0%	4.6%
Iran, Islamic Rep.	100%		0.3%	17.5%	0.9%	18.7%
[2] Israel	72%	Hebrew Public Education System	1.1%	0.0%	0.1%	1.2%
[1] New Zealand	100%		0.7%	25.8%	0.4%	27.0%
Portugal	100%		6.6%	0.0%	0.7%	7.3%
Slovenia	100%		1.9%	0.7%	0.0%	2.6%
United States	100%		0.4%	0.0%	4.3%	4.7%

* Fourth grade in most countries; see Table 2.1 for information about the grades tested in each country.
[1] School-level exclusions for performance assessment exceed 25% of the National Desired Population.
[2] National Desired Population does not cover all of International Desired Population.
Because results are rounded, some totals may appear inconsistent.

Table 2.8 Coverage of TIMSS Target Population - Performance Assessment – Eighth Grade*

The international desired target population is defined as follows:
Eighth Grade - All students enrolled in the higher of the two adjacent grades with the largest proportion of 13-year-old students at the time of testing.

Country	International Desired Target Population		National Desired Target Population			
	Coverage	Notes on coverage	Main Assessment School-Level Exclusions	Performance Assessment School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Australia	100%		0.2%	16.3%	0.6%	17.0%
Canada	100%		2.4%	15.0%	1.8%	19.1%
Colombia	100%		3.8%	0.0%	0.0%	3.8%
Cyprus	100%		0.0%	0.0%	0.0%	0.0%
Czech Republic	100%		4.9%	0.0%	0.0%	4.9%
[2] England	100%		8.4%	16.6%	2.4%	27.3%
Hong Kong	100%		2.0%	1.0%	0.0%	3.0%
Iran, Islamic Rep.	100%		0.3%	17.0%	0.0%	17.3%
[1] Israel	74%	Hebrew Public Education System	3.1%	0.0%	0.0%	3.1%
Netherlands	100%		1.2%	0.0%	0.0%	1.2%
New Zealand	100%		1.3%	10.5%	0.4%	12.1%
Norway	100%		0.3%	22.6%	1.5%	24.4%
Portugal	100%		0.0%	0.0%	0.3%	0.3%
[3] Romania	100%		2.8%	28.5%	0.0%	31.3%
Scotland	100%		0.3%	9.3%	1.7%	11.3%
Singapore	100%		4.6%	0.0%	0.0%	4.6%
Slovenia	100%		2.4%	0.7%	0.2%	3.2%
Spain	100%		6.0%	1.7%	2.6%	10.3%
Sweden	100%		0.0%	23.5%	0.7%	24.2%
[1] Switzerland	75%	German Cantons	4.4%	8.4%	0.8%	13.6%
United States	100%		0.4%	1.3%	1.7%	3.4%

* Eighth grade in most countries; see Table 2.2 for information about the grades tested in each country.

[1] National Desired Population does not cover all of International Desired Population.

[2] National Defined Population covers less than 90 percent of National Desired Population for the main assessment (school-level plust within-sample exclusions).

[3] School-level exclusions for performance assessment exceed 25% of the National Desired Population.

Because results are rounded, some totals may appear inconsistent.

2.2 SAMPLING OF SCHOOLS AND STUDENTS

2.2.1 General Sample Design[2]

The basic sample design used in TIMSS was a two-stage stratified cluster design. The first stage consisted of a sample of schools; the second stage consisted of samples of intact mathematics classrooms from each eligible target grade in the sampled schools. The design required schools to be sampled using a probability proportional to size (PPS) systematic method, as described by Foy, Rust and Schleicher (1996), and classrooms to be sampled with equal probabilities (Schleicher and Siniscalco, 1996). The

[2] The TIMSS sample design is described in detail by Foy, Rust, and Schleicher (1996). See Schleicher and Siniscalco (1996) for TIMSS within-school sampling procedures. This chapter describes the outcome of the sampling for Population 1 (third and fourth grades in most countries) and Population 2 (seventh and eighth grades in most countries), including country-by-country descriptions of the national samples.

TIMSS sampling approach was designed to yield 150 schools for each of Populations 1 and 2, and one classroom for each grade, for a total of 3,750 students per grade per population.

The TIMSS sampling approach allowed countries to stratify the school sampling frame explicitly or implicitly or both. Explicit stratification consisted of categorizing schools according to some criterion (e.g., region of the country), and ensuring that a predetermined number of schools were selected from each explicit stratum. Implicit stratification consisted of sorting the school sampling frame according to a set of criteria prior to sampling. This produces an allocation of the school sample proportional to the implicit strata when schools are selected using a systematic PPS sampling method.

Most participants sampled 150 schools, with one classroom per grade within sampled schools and all students within sampled classrooms. There were, however, some variations in the sampling of schools and students, which are documented in Appendix B. Classrooms were generally selected with equal probabilities, unless student subsampling occurred; in that case classrooms were sampled with PPS. Any student subsampling within selected classrooms was done with equal probabilities within classrooms. Some participants chose to subsample a fixed number of students within sampled classrooms. This usually occurred in countries where large classrooms are the norm and subsampling within classrooms was a means of reducing the data collection effort. Some participating countries chose to sample two classrooms at the upper grade in each school. One country did not sample intact classrooms, but rather sampled students within schools.

For the performance assessment, TIMSS participants were to sample at least 50 schools from those already selected for the written assessment, and from each school a sample of either 9 or 18 upper-grade students already selected for the written assessment. This yielded a sample of about 450 students in each of the eighth and fourth grades in each country.

2.2.2 Target Population Sizes

Tables 2.9 and 2.10 summarize the national target population sizes based on the sampling frame counts, as well as the sample sizes for participating schools and students. From the computed sampling weights (see Chapter 4) an estimated student population size was computed, which was expected to match closely the student population size from the sampling frame. All counts are aggregates over the two grades selected, except for Israel and Kuwait where only one grade was tested. The student population size for the Russian Federation's Population 2 is an estimate based on total enrollment in their schools. The number of schools in the United States' Population 1 and Population 2 are estimates based on the number of schools in the primary sampling units that they sampled. Because of difficulties in computing sampling weights for the Philippines, the population size for its Population 2 cannot be estimated from the sample.

Table 2.9 Population and Sample Sizes - Population 1 (Third and Fourth Grades*)

Country	Population		Sample		
	Schools	Students	Schools	Students	Est. Pop.
Australia	7,588	495,803	178	11,248	483,463
Austria	3,395	184,598	133	5,171	177,434
Canada	10,388	765,653	391	16,002	760,325
Cyprus	193	19,308	147	6,684	19,736
Czech Republic	4,256	259,641	188	6,524	236,457
England	12,844	1,006,305	128	6,182	1,066,604
Greece	6,626	246,998	174	6,008	205,181
Hong Kong	882	158,391	124	8,807	173,749
Hungary	2,999	244,190	150	6,044	234,007
Iceland	153	7,784	144	3,507	7,474
Iran	59,367	3,742,497	180	6,746	2,825,173
Ireland	2,669	121,657	161	5,762	119,000
[1] Israel	1,081	70,327	87	2,351	66,967
Japan	24,676	2,929,794	142	8,612	2,827,215
Korea	4,910	1,357,238	150	5,589	1,222,011
[1] Kuwait	150	24,168	150	4,318	24,071
Latvia	632	35,434	125	4,270	34,003
Netherlands	7,873	345,600	130	5,314	344,969
New Zealand	2,121	100,591	149	4,925	100,640
Norway	2,817	101,773	139	4,476	98,933
Portugal	3,210	277,961	143	5,503	247,961
Scotland	2,004	126,007	152	6,433	118,447
Singapore	191	83,025	191	14,169	83,147
Slovenia	422	53,066	122	5,087	55,139
Thailand	31,417	1,760,339	154	5,862	1,748,290
United States	55,526	7,163,600	186	11,115	7,207,188

*Third and fourth grades in most countries; see Table 2.1 for more information about the grades tested in each country.

[1] Israel and Kuwait tested only the upper grade of the target population.

Table 2.10 Population and Sample Sizes - Population 2 (Seventh and Eighth Grades*)

Country	Population Schools	Population Students	Sample Schools	Sample Students	Sample Est. Pop.
Australia	2,341	473,731	161	12,852	469,644
Austria	1,433	180,773	125	5,786	176,332
Belgium (Fl)	770	139,192	141	5,662	139,246
Belgium (Fr)	558	108,234	120	4,883	109,167
Bulgaria	2,563	231,885	115	3,771	288,073
Canada	6,993	755,100	367	16,581	755,158
Colombia	6,803	1,072,824	140	5,304	1,146,607
Cyprus	55	19,362	55	5,852	19,380
Czech Republic	3,124	303,326	150	6,672	304,986
Denmark	2,115	109,215	144	4,370	99,153
England	3,941	993,992	122	3,579	950,737
France	7,893	1,634,436	127	6,014	1,676,167
Germany	11,234	1,378,020	134	5,763	1,468,435
Greece	1,769	293,642	156	7,921	252,134
Hong Kong	392	172,806	86	6,752	177,164
Hungary	2,999	244,190	150	5,978	231,164
Iceland	161	8,719	144	3,730	8,447
Iran	18,303	2,492,070	192	7,429	1,987,889
Ireland	752	140,670	132	6,203	136,121
[1] Israel	656	67,348	46	1,415	60,585
Japan	11,292	3,092,592	151	10,271	3,204,359
Korea	2,388	1,617,301	150	5,827	1,608,813
[1] Kuwait	69	15,085	69	1,655	13,093
Latvia	553	34,428	142	4,976	32,456
Lithuania	1,096	80,254	145	5,056	76,251
Netherlands	1,235	375,201	95	4,084	367,083
New Zealand	1,297	100,377	274	6,867	99,642
Norway	6,117	102,842	249	5,736	101,389
[2] Philippines	23,556	2,524,238	387	11,853	-
Portugal	1,009	295,088	142	6,753	284,341
Romania	7,018	636,278	163	7,471	591,881
Russian Federation	68,270	4,030,000	174	8,160	4,172,955
Scotland	445	131,715	129	5,776	126,576
Singapore	137	75,464	137	8,285	72,719
Slovak Republic	1,349	155,037	145	7,101	162,840
Slovenia	422	55,085	122	5,606	54,060
South Africa	11,742	1,384,532	227	9,792	1,415,513
Spain	11,946	1,141,065	153	7,596	1,096,145
Sweden	4,720	198,544	270	6,906	194,688
Switzerland	3,543	135,298	324	8,940	136,414
Thailand	2,128	1,158,397	147	11,695	1,342,740
United States	27,330	6,574,200	183	10,973	6,345,142

*Seventh and eighth grades in most countries; see Table 2.2 for more information about the grades tested in each country.

[1] Israel and Kuwait tested only the upper grade of the target population.

[2] Population size for the Philippines cannot be estimated.

2.2.3 Participation Rates

Weighted school, student, and overall participation rates were computed for each participating country for each grade. The procedures for computing participation (response) rates is documented by Foy, Rust, and Schleicher (1996). The level of participation of schools and students was one aspect of the national samples used to evaluate the quality of the samples and potential biases. Countries were required to obtain a school participation rate of 85%, a student participation rate of 85%, or an overall participation rate (product of school and student participation rates) of 75%. In cases where these rates were not obtained, with or without the use of replacement schools, achievement results were reported in a separate section of the tables in the international reports. Foy, Martin, and Kelly (1996) further document the procedures for evaluating the quality of the national samples and reporting the achievement results. Tables 2.11 through 2.15 present the school, student, and overall participation rates and achieved sample sizes for the Population 1 main assessment; Tables 2.16 through 2.20 show the corresponding information for the Population 2 main assessment. Tables 2.21 and 2.22 show that information for the performance assessment.

Appendix B contains further information on the characteristics of individual national samples, including target population definitions, population coverage and exclusions, use of stratification variables, and any deviations from the general TIMSS design.

Table 2.11 School Participation Rates and Sample Sizes - Population 1 Upper Grade (Fourth Grade*)

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated[1]		Total Number of Schools That Participated
						Procedural	Other	
Australia	66	69	268	268	169	9	0	178
Austria	51	72	150	150	71	31	31	133
Canada	90	90	423	420	390	0	0	390
Cyprus	97	97	150	150	146	0	0	146
Czech Republic	91	94	215	215	181	7	0	188
England	63	88	150	145	92	35	0	127
Greece	93	93	187	187	174	0	0	174
Hong Kong	84	84	156	148	124	0	0	124
Hungary	100	100	150	150	150	0	0	150
Iceland	95	95	153	151	144	0	0	144
Iran, Islamic Rep.	100	100	180	180	180	0	0	180
Ireland	94	96	175	173	161	4	0	165
Israel	40	40	100	100	40	0	47	87
Japan	93	96	150	150	137	4	0	141
Korea	100	100	150	150	150	0	0	150
Kuwait	100	100	150	150	150	0	0	150
Latvia (LSS)	74	74	169	169	125	0	0	125
Netherlands	31	62	196	196	63	67	0	130
New Zealand	80	99	150	150	120	29	0	149
Norwa	85	94	150	148	126	13	0	139
Portugal	95	95	150	150	143	0	0	143
Scotland	78	83	184	184	143	9	0	152
Singapore	100	100	191	191	191	0	0	191
Slovenia	81	81	150	150	121	0	0	121
Thailand	96	96	155	155	154	0	0	154
United States	85	85	220	213	182	0	0	182

*Fourth grade in most countries; see Table 2.1 for more information about the grades tested in each country.

[1] Replacement schools selected in accordance with the TIMSS sampling procedures are listed in the "procedural" column. Those selected using unapproved methods are listed in the "other" column and were not included in the computation of school participation rates.

Table 2.12 School Participation Rates and Sample Sizes - Population 1 Lower Grade (Third Grade*)

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated[1]		Total Number of Schools That Participated
						Proce-dural	Other	
Australia	66	69	268	264	166	9	0	175
Austria	49	70	150	149	68	29	31	128
Canada	88	88	423	418	375	0	0	375
Cyprus	98	98	150	150	147	0	0	147
Czech Republic	91	93	215	215	180	7	0	187
England	64	88	150	145	93	35	0	128
Greece	91	91	187	187	171	0	0	171
Hong Kong	84	84	156	147	123	0	0	123
Hungary	99	99	150	150	149	0	0	149
Iceland	95	95	153	152	144	0	0	144
Iran, Islamic Rep.	99	99	180	180	178	0	0	178
Ireland	94	96	175	173	160	4	0	164
Israel	-	-	-	-	-	-	-	-
Japan	93	95	150	150	137	5	0	142
Korea	100	100	150	150	150	0	0	150
Kuwait	-	-	-	-	-	-	-	-
Latvia (LSS)	73	73	169	168	123	0	0	123
Netherlands	29	62	196	195	60	69	0	129
New Zealand	80	99	150	150	120	29	0	149
Norway	83	92	150	148	124	12	0	136
Portugal	95	95	150	150	143	0	0	143
Scotland	77	81	184	184	142	8	0	150
Singapore	100	100	191	191	191	0	0	191
Slovenia	81	81	150	149	122	0	0	122
Thailand	96	96	155	154	153	0	0	153
United States	86	86	220	217	186	0	0	186

*Third grade in most countries; see Table 2.1 for more information about the grades tested in each country.

A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

[1] Replacement schools selected in accordance with the TIMSS sampling procedures are listed in the "procedural" column. Those selected using unapproved methods are listed in the "other" column and were not included in the computation of school participation rates.

Table 2.13 Student Participation Rates and Sample Sizes - Population 1 Upper Grade (Fourth Grade*)

Country	Within School Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Total Number of Students Assessed
Australia	96	6930	37	104	6789	282	6507
Austria	96	2779	12	6	2761	116	2645
Canada	96	9193	81	268	8844	436	8408
Cyprus	86	3972	4	3	3965	589	3376
Czech Republic	92	3555	7	0	3548	280	3268
England	95	3489	73	122	3294	168	3126
Greece	95	3358	6	116	3236	183	3053
Hong Kong	98	4475	0	1	4474	63	4411
Hungary	92	3272	0	0	3272	266	3006
Iceland	90	2149	23	101	2025	216	1809
Iran, Islamic Rep.	97	3521	5	36	3480	95	3385
Ireland	93	3134	14	40	3080	207	2873
Israel	94	2486	0	3	2483	132	2351
Japan	97	4453	0	0	4453	147	4306
Korea	95	2971	133	0	2838	26	2812
Kuwait	95	4578	34	0	4544	226	4318
Latvia (LSS)	93	2390	12	1	2377	161	2216
Netherlands	96	2639	0	4	2635	111	2524
New Zealand	96	2627	82	20	2525	104	2421
Norway	97	2391	16	42	2333	76	2257
Portugal	96	2994	15	16	2963	110	2853
Scotland	92	3735	0	139	3596	295	3301
Singapore	98	7274	14	0	7260	121	7139
Slovenia	94	2720	3	0	2717	151	2566
Thailand	100	3042	0	50	2992	0	2992
United States	94	8224	61	412	7751	455	7296

*Fourth grade in most countries; see Table 2.1 for more information about the grades tested in each country.

Table 2.14 Student Participation Rates and Sample Sizes - Population 1 Lower Grade (Third Grade*)

Country	Within School Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Total Number of Students Assessed
Australia	95	5138	31	92	5015	274	4741
Austria	96	2655	10	6	2639	113	2526
Canada	96	8433	77	307	8049	455	7594
Cyprus	85	3913	5	2	3906	598	3308
Czech Republic	93	3484	8	0	3476	220	3256
England	94	3468	70	158	3240	184	3056
Greece	94	3263	4	133	3126	171	2955
Hong Kong	99	4455	0	2	4453	57	4396
Hungary	94	3270	0	0	3270	232	3038
Iceland	91	2017	19	89	1909	211	1698
Iran, Islamic Rep.	98	3504	12	49	3443	82	3361
Ireland	94	3127	14	39	3074	185	2889
Israel	-	-	-	-	-	-	-
Japan	97	4433	0	0	4433	127	4306
Korea	94	2969	138	2	2829	52	2777
Kuwait	-	-	-	-	-	-	-
Latvia (LSS)	94	2218	8	0	2210	156	2054
Netherlands	96	2923	0	14	2909	119	2790
New Zealand	95	2733	91	9	2633	129	2504
Norway	97	2362	8	59	2295	76	2219
Portugal	97	2790	13	31	2746	96	2650
Scotland	90	3663	0	187	3476	344	3132
Singapore	98	7223	14	0	7209	179	7030
Slovenia	95	2659	5	0	2654	133	2521
Thailand	100	2945	0	74	2871	1	2870
United States	95	4280	40	201	4039	220	3819

*Third grade in most countries; see Table 2.1 for more information about the grades tested in each country.

A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

Table 2.15 Overall Participation Rates - Population 1 Lower and Upper Grades (Third and Fourth Grades*)

Country	Upper Grade: Overall Participation Before Replacement (Weighted Percentage)	Upper Grade: Overall Participation After Replacement (Weighted Percentage)	Lower Grade: Overall Participation Before Replacement (Weighted Percentage)	Lower Grade: Overall Participation After Replacement (Weighted Percentage)
Australia	63	66	62	65
Austria	49	69	46	67
Canada	86	86	84	84
Cyprus	83	83	83	83
Czech Republic	84	86	85	87
England	60	83	61	83
Greece	88	88	86	86
Hong Kong	83	83	83	83
Hungary	92	92	93	93
Iceland	86	86	86	86
Iran, Islamic Rep.	97	97	97	97
Ireland	88	90	88	91
Israel	38	38	-	-
Japan	90	92	90	93
Korea	95	95	94	94
Kuwait	95	95	-	-
Latvia (LSS)	69	69	69	69
Netherlands	29	59	28	60
New Zealand	77	95	76	95
Norway	82	91	81	89
Portugal	92	92	92	92
Scotland	71	76	69	73
Singapore	98	98	98	98
Slovenia	76	76	77	77
Thailand	96	96	96	96
United States	80	80	81	81

*Third and fourth grades in most countries; see Table 2.1 for information about the grades tested in each country.

A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

Table 2.16 School Participation Rates and Sample Sizes - Population 2 Upper Grade (Eighth Grade*)

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated	Total Number of Schools That Participated
Australia	75	77	214	214	158	3	161
Austria	41	84	159	159	62	62	124
Belgium (Fl)	61	94	150	150	92	49	141
Belgium (Fr)	57	79	150	150	85	34	119
Bulgaria	72	74	167	167	111	4	115
Canada	90	91	413	388	363	1	364
Colombia	91	93	150	150	136	4	140
Cyprus	100	100	55	55	55	0	55
Czech Republic	96	100	150	149	143	6	149
Denmark	93	93	158	157	144	0	144
England	56	85	150	144	80	41	121
France	86	86	151	151	127	0	127
Germany	72	93	153	150	102	32	134
Greece	87	87	180	180	156	0	156
Hong Kong	82	82	105	104	85	0	85
Hungary	100	100	150	150	150	0	150
Iceland	98	98	161	132	129	0	129
Iran, Islamic Rep.	100	100	192	191	191	0	191
Ireland	84	89	150	149	125	7	132
Israel	45	46	100	100	45	1	46
Japan	92	95	158	158	146	5	151
Korea	100	100	150	150	150	0	150
Kuwait	100	100	69	69	69	0	69
Latvia (LSS)	83	83	170	169	140	1	141
Lithuania	96	96	151	151	145	0	145
Netherlands	24	63	150	150	36	59	95
New Zealand	91	99	150	150	137	12	149
Norway	91	97	150	150	136	10	146
Philippines	96 **	97 **	200	200	192	1	193
Portugal	95	95	150	150	142	0	142
Romania	94	94	176	176	163	0	163
Russian Federation	97	100	175	175	170	4	174
Scotland	79	83	153	153	119	8	127
Singapore	100	100	137	137	137	0	137
Slovak Republic	91	97	150	150	136	9	145
Slovenia	81	81	150	150	121	0	121
South Africa	60	64	180	180	107	7	114
Spain	96	100	155	154	147	6	153
Sweden	97	97	120	120	116	0	116
Switzerland	93	95	259	258	247	3	250
Thailand	99	99	150	150	147	0	147
United States	77	85	220	217	169	14	183

* Eighth grade in most countries; see Table 2.2 for more information about the grades tested in each country.

** Participation rates for the Philippines are unweighted.

Table 2.17 School Participation Rates and Sample Sizes - Population 2 Lower Grade (Seventh Grade*)

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated	Total Number of Schools That Participated
Australia	75	76	214	213	156	3	159
Austria	43	86	159	159	63	62	125
Belgium (Fl)	61	93	150	150	91	49	140
Belgium (Fr)	57	80	150	150	85	35	120
Bulgaria	75	77	150	150	101	3	104
Canada	90	90	413	390	366	1	367
Colombia	91	93	150	150	136	4	140
Cyprus	100	100	55	55	55	0	55
Czech Republic	96	100	150	150	144	6	150
Denmark	88	88	158	154	137	0	137
England	57	85	150	145	81	41	122
France	87	87	151	151	126	0	126
Germany	70	90	153	153	101	31	132
Greece	87	87	180	180	156	0	156
Hong Kong	83	83	105	104	86	0	86
Hungary	99	99	150	150	149	0	149
Iceland	97	97	161	149	144	0	144
Iran, Islamic Rep.	100	100	192	192	192	0	192
Ireland	82	87	150	148	122	7	129
Israel	-	-	-	-	-	-	-
Japan	92	95	158	158	146	5	151
Korea	100	100	150	150	150	0	150
Kuwait	-	-	-	-	-	-	-
Latvia (LSS)	83	84	170	169	141	1	142
Lithuania	96	96	151	151	145	0	145
Netherlands	23	61	150	150	34	58	92
New Zealand	90	99	150	150	135	13	148
Norway	84	96	150	147	124	17	141
Philippines	97 **	97 **	200	200	194	0	194
Portugal	94	94	150	150	141	0	141
Romania	94	94	176	175	162	0	162
Russian Federation	97	100	175	175	170	4	174
Scotland	79	85	153	153	120	9	129
Singapore	100	100	137	137	137	0	137
Slovak Republic	91	97	150	150	136	9	145
Slovenia	81	81	150	150	122	0	122
South Africa	83	85	161	161	133	4	137
Spain	96	100	155	154	147	6	153
Sweden	96	96	160	160	154	0	154
Switzerland	90	94	217	217	200	6	206
Thailand	99	99	150	150	146	0	146
United States	77	84	220	214	165	14	179

* Seventh grade in most countries; see Table 2.2 for more information about the grades tested in each country.

** Participation rates for the Philippines are unweighted.

A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

Table 2.18 Student Participation Rates and Sample Sizes - Population 2 Upper Grade (Eighth Grade*)

Country	Within School Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Total Number of Students Assessed
Australia	92	8027	63	61	7903	650	7253
Austria	95	2969	14	4	2951	178	2773
Belgium (Fl)	97	2979	1	0	2978	84	2894
Belgium (Fr)	91	2824	0	1	2823	232	2591
Bulgaria	86	2300	0	0	2300	327	1973
Canada	93	9240	134	206	8900	538	8362
Colombia	94	2843	6	0	2837	188	2649
Cyprus	97	3045	15	0	3030	107	2923
Czech	92	3608	6	0	3602	275	3327
Denmark	93	2487	0	0	2487	190	2297
England	91	2015	37	60	1918	142	1776
France	95	3141	0	0	3141	143	2998
Germany	87	3318	0	35	3283	413	2870
Greece	97	4154	27	23	4104	114	3990
Hong Kong	98	3415	12	0	3403	64	3339
Hungary	87	3339	0	0	3339	427	2912
Iceland	90	2025	10	65	1950	177	1773
Iran, Islamic Rep.	98	3770	20	0	3750	56	3694
Ireland	91	3411	28	10	3373	297	3076
Israel	98	1453	6	0	1447	32	1415
Japan	95	5441	0	0	5441	300	5141
Korea	95	2998	31	0	2967	47	2920
Kuwait	83	1980	3	0	1977	322	1655
Latvia (LSS)	90	2705	19	0	2686	277	2409
Lithuania	87	2915	2	0	2913	388	2525
Netherlands	95	2112	14	1	2097	110	1987
New Zealand	94	4038	121	12	3905	222	3683
Norway	96	3482	26	49	3407	140	3267
Philippines	91 **	6586	93	0	6493	492	6001
Portugal	97	3589	70	13	3506	115	3391
Romania	96	3899	0	0	3899	174	3725
Russian	95	4311	42	10	4259	237	4022
Scotland	88	3289	0	46	3243	380	2863
Singapore	95	4910	18	0	4892	248	4644
Slovak Republic	95	3718	5	3	3710	209	3501
Slovenia	95	2869	15	8	2846	138	2708
South	97	4793	0	0	4793	302	4491
Spain	95	4198	27	102	4069	214	3855
Sweden	93	4483	71	28	4384	309	4075
Switzerland	98	4989	16	24	4949	94	4855
Thailand	100	5850	0	0	5850	0	5850
United States	92	8026	104	108	7814	727	7087

* Eighth grade in most countries; see Table 2.2 for more information about the grades tested in each country.

** Participation rates for the Philippines are unweighted.

Table 2.19 Student Participation Rates and Sample Sizes - Population 2 Lower Grade (Seventh Grade*)

Country	Within School Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Total Number of Students Assessed
Australia	93	6067	26	21	6020	421	5599
Austria	95	3196	22	5	3169	156	3013
Belgium (Fl)	97	2857	3	0	2854	86	2768
Belgium (Fr)	95	2418	0	1	2417	125	2292
Bulgaria	87	2080	0	0	2080	282	1798
Canada	95	8962	89	248	8625	406	8219
Colombia	93	2840	2	0	2838	183	2655
Cyprus	98	3028	17	0	3011	82	2929
Czech Republic	92	3641	11	0	3630	285	3345
Denmark	86	2408	0	0	2408	335	2073
England	92	2031	31	67	1933	130	1803
France	95	3164	0	0	3164	148	3016
Germany	87	3388	0	37	3351	458	2893
Greece	97	4166	30	78	4058	127	3931
Hong Kong	98	3507	11	0	3496	83	3413
Hungary	94	3266	0	0	3266	200	3066
Iceland	92	2243	11	72	2160	203	1957
Iran, Islamic Rep.	99	3789	18	0	3771	36	3735
Ireland	91	3480	23	17	3440	313	3127
Israel	-	-	-	-	-	-	-
Japan	96	5337	0	0	5337	207	5130
Korea	94	2996	51	0	2945	38	2907
Kuwait	-	-	-	-	-	-	-
Latvia (LSS)	91	2853	7	0	2846	279	2567
Lithuania	89	2852	3	0	2849	318	2531
Netherlands	95	2220	23	0	2197	100	2097
New Zealand	95	3471	98	17	3356	172	3184
Norway	96	2629	8	53	2568	99	2469
Philippines	93 **	6283	29	1	6253	401	5852
Portugal	96	3594	80	4	3510	148	3362
Romania	95	3938	0	0	3938	192	3746
Russian Federation	96	4408	39	11	4358	220	4138
Scotland	90	3313	0	81	3232	319	2913
Singapore	98	3744	19	0	3725	84	3641
Slovak Republic	95	3797	10	3	3784	184	3600
Slovenia	95	3058	12	4	3042	144	2898
South Africa	96	5532	0	0	5532	231	5301
Spain	95	4087	38	116	3933	192	3741
Sweden	95	3055	27	36	2992	161	2831
Switzerland	99	4199	14	44	4141	56	4085
Thailand	100	5845	0	0	5845	0	5845
United States	94	4295	42	85	4168	282	3886

* Seventh grade in most countries; see Table 2.2 for more information about the grades tested in each country.

** Participation rates for the Philippines are unweighted.

A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower-grade.

Table 2.20 Overall Participation Rates - Population 2 Upper and Lower Grades (Seventh and Eighth Grades*)

Country	Upper Grade: Overall Participation Before Replacement (Weighted Percentage)	Upper Grade: Overall Participation After Replacement (Weighted Percentage)	Lower Grade: Overall Participation Before Replacement (Weighted Percentage)	Lower Grade: Overall Participation After Replacement (Weighted Percentage)
Australia	69	70	69	71
Austria	39	80	41	82
Belgium (Fl)	59	91	59	91
Belgium (Fr)	52	72	54	76
Bulgaria	62	63	65	67
Canada	84	84	86	86
Colombia	85	87	84	86
Cyprus	97	97	98	98
Czech Republic	89	92	88	92
Denmark	86	86	76	76
England	51	77	52	78
France	82	82	82	82
Germany	63	81	61	78
Greece	84	84	84	84
Hong Kong	81	81	81	81
Hungary	87	87	93	93
Iceland	88	88	89	89
Iran, Islamic Rep.	98	98	99	99
Ireland	76	81	75	79
Israel	44	45	-	-
Japan	87	90	88	91
Korea	95	95	94	94
Kuwait	83	83	-	-
Latvia (LSS)	75	75	75	76
Lithuania	83	83	86	86
Netherlands	23	60	22	58
New Zealand	86	94	85	94
Norway	87	93	81	92
Philippines	87 **	88 **	90 **	90 **
Portugal	92	92	90	90
Romania	89	89	89	89
Russian Federation	93	95	93	95
Scotland	69	73	71	76
Singapore	95	95	98	98
Slovak Republic	86	91	86	92
Slovenia	77	77	77	77
South Africa	58	62	79	82
Spain	91	94	91	95
Sweden	90	90	91	91
Switzerland	92	94	89	93
Thailand	99	99	99	99
United States	71	78	72	79

* Seventh and eighth grades in most countries; see Table 2.2 for information about the grades tested in each country.

** Participation rates for the Philippines are unweighted.

A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

Table 2.21 School Participation Rates and Sample Sizes - Performance Assessment Fourth Grade*

Country	School Participation Rate Before Replacement (Weighted Percentage)	School Participation Rate After Replacement (Weighted Percentage)	Within-School Student Participation Rate (Weighted Percentage)	Overall Participation Rate Before Replacement (Weighted Percentage)	Overall Participation Rate After Replacement (Weighted Percentage)
Australia	47	56	76	36	43
Canada	91	92	95	87	88
Cyprus	98	100	86	85	86
Hong Kong	61	77	95	58	73
Iran, Islamic Rep.	97	100	93	90	93
Israel	50 **	83 **	30 **	15 **	25 **
New Zealand	72	93	90	65	83
Portugal	96	96	94	91	91
Slovenia	98	100	91	89	91
United States	83	84	88	73	74

* Fourth grade in most countries; see Table 2.1 for information about the grades tested in each country.
**Unweighted participation rates.

Table 2.22 School Participation Rates and Sample Sizes - Performance Assessment Eighth Grade*

Country	School Participation Rate Before Replacement (Weighted Percentage)	School Participation Rate After Replacement (Weighted Percentage)	Within-School Student Participation Rate (Weighted Percentage)	Overall Participation Rate Before Replacement (Weighted Percentage)	Overall Participation Rate After Replacement (Weighted Percentage)
Australia	51	58	73	37	43
Canada	97	97	92	89	89
Colombia	91	91	96	88	88
Cyprus	96	96	93	88	88
Czech Republic	94	100	82	77	82
England	46	85	84	38	71
Hong Kong	44	44	77	34	34
Iran, Islamic Rep.	98	98	93	91	91
Israel	44 **	46 **	30 **	13 **	14 **
Netherlands	18	48	89	16	43
New Zealand	90	100	88	79	88
Norway	87	96	91	79	88
Portugal	96	96	91	87	87
Romania	90	90	94	84	84
Scotland	78	96	85	66	81
Singapore	90	100	87	79	87
Slovenia	98	100	93	91	93
Spain	94	100	93	87	93
Sweden	99	99	88	87	87
Switzerland	65	81	97	63	78
United States	71	77	86	61	66

* Eighth grade in most countries; see Table 2.2 for information about the grades tested in each country.
**Unweighted participation rates.

REFERENCES

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Kelly, D.L., and Smith, T.A. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College.

Beaton, A.E., Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Kelly, D.L., and Smith, T.A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College.

Foy, P., Martin, M.O., and Kelly, D.L. (1996). Sampling. In M.O. Martin and I.V.S. Mullis (Eds.), *TIMSS: Quality assurance in data collection.* Chestnut Hill, MA: Boston College.

Foy, P., Rust, K., and Schleicher, A. (1996). Sample design. In M.O. Martin and D.L. Kelly (Eds.), *TIMSS technical report, volume I: Design and development.* Chestnut Hill, MA: Boston College.

Harmon, M. and Kelly, D.L. (1996). Development and design of the TIMSS performance assessment. In M.O. Martin and D.L. Kelly (Eds.), *TIMSS technical report, volume I: Design and development.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Beaton, A.E., Gonzalez, E.J., Smith, T.A., and Kelly, D.L. (1997). *Science achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L., and Smith, T.A. (1997). *Mathematics achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College.

Schleicher, A. and Siniscalco, M.T. (1996). Field operations. In M.O. Martin and D.L. Kelly (Eds.), *TIMSS technical report, volume I: Design and development.* Chestnut Hill, MA: Boston College.

3 Data Management and Construction of the TIMSS Database

Heiko Sibberns
Dirk Hastedt
Michael Bruneforth
Knut Schwippert
IEA Data Processing Center

Eugenio J. Gonzalez
Boston College

The TIMSS data were processed through a closely cooperative procedure involving the TIMSS International Study Center at Boston College, the IEA Data Processing Center, the Australian Council for Educational Research, Statistics Canada, and the national research centers of the participating countries. Under the general direction of the International Study Center, each institution was responsible for specific aspects of the data processing.

The data processing consisted of six general tasks: data entry, creation of the international database, calculation of sampling weights, scaling of achievement data, analysis of the background data, and creation of the reporting tables. Although each task is crucial to ensuring the quality and accuracy of the results, data entry and the creation of the international database take center stage, since those tasks feed into the remaining four. The scaling of the TIMSS data are discussed in Chapters 7 and 8, the weighting procedures in Chapter 4, and the analysis and reporting in Chapters 9, 10, and 11. This chapter describes the process followed in data entry and the creation of the international database, and the steps that were undertaken to ensure the quality and accuracy of the international database. It also describes the responsibilities of each participant in the creation of the international database. In particular, this chapter outlines the flow of the data files between the different centers involved in the data processing; the structure of the data files submitted by each country for processing, and the resulting files that are part of the international database; the rules, methods, and procedures used for data verification and manipulation; the data products created during the data cleaning process and provided to the national centers; and the computer software used in this process.

The TIMSS international database for the primary and middle school years was released for public use in September 1997. It is available at the TIMSS website (http://wwwcsteep.bc.edu./timss) and through IEA Headquarters. The database is accompanied by a User's Guide (Gonzalez and Smith, 1997) and full documentation.

3.1 DATA FLOW

The data collected with the TIMSS survey instruments were entered into data files of a common international format at the national research centers of the participating countries. These data files were then submitted to the IEA Data Processing Center for cleaning and verification. The major responsibilities of the IEA Data Processing Center at this point were to check that the data files submitted matched the international standard and to make modifications where necessary, apply standard cleaning rules to the data to verify their consistency and accuracy, interact with the National Research Coordinators (NRCs) to ensure the accuracy of the data contained in the files, produce summary statistics of the background and achievement data for review by the TIMSS International Study Center, and, upon feedback from the individual countries and the TIMSS International Study Center, construct the international database. The IEA Data Processing Center also had primary responsibility for making all modifications to the data files and for distributing the national data files to each of the participating countries.

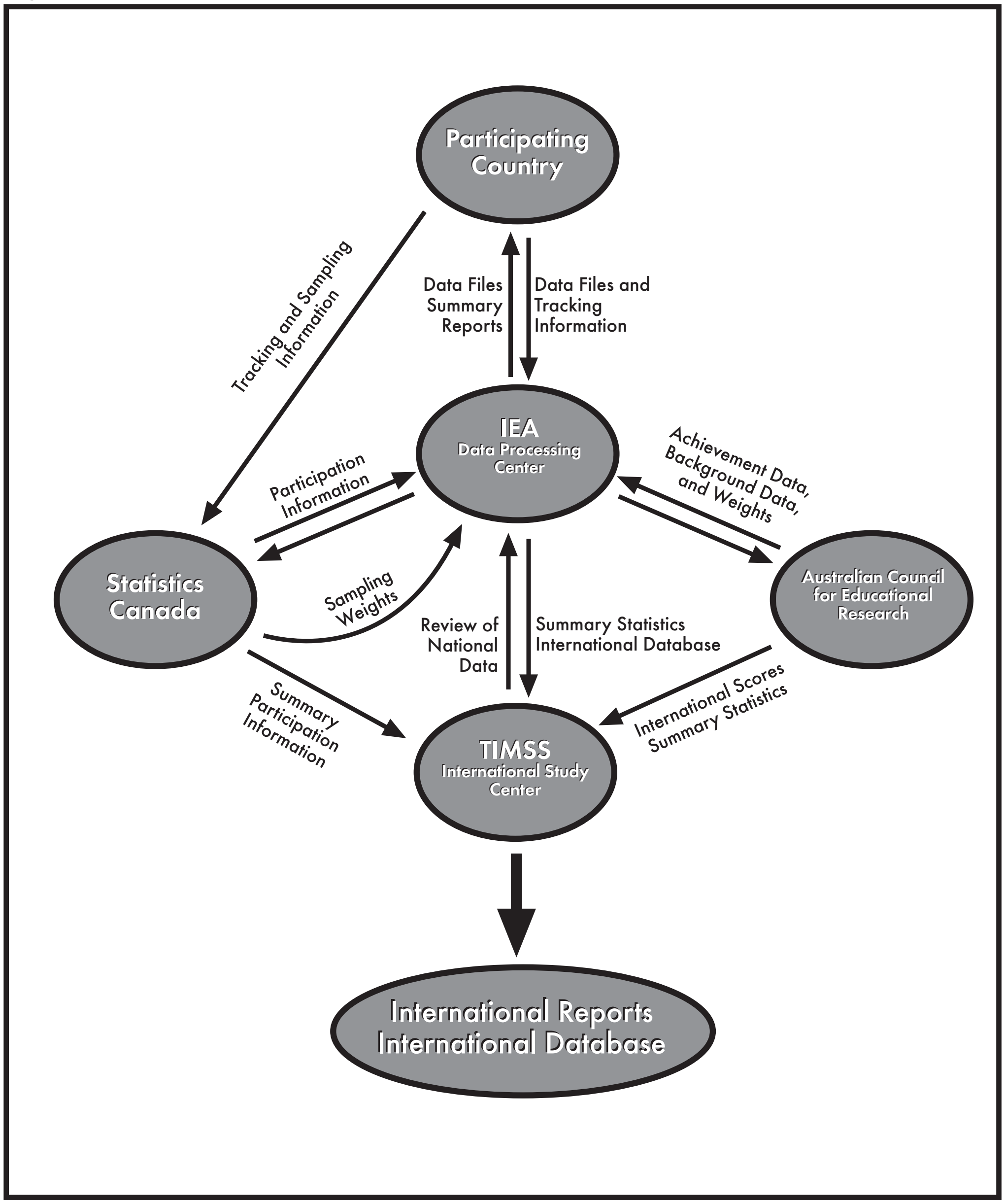
Once verified and in the international file format, the achievement data were sent to the Australian Council for Educational Research (ACER), where basic item statistics were produced for item review and an initial country-level scaling was conducted. An item review was undertaken by the staff at the TIMSS International Study Center (see Chapter 6). At the same time Statistics Canada received from the IEA Data Processing Center data files containing participation information for students in the sample. This information, together with information provided by the NRC, was used by Statistics Canada to calculate sampling weights, population coverage, and participation rates at the school and student level. The sampling weights were then sent to the TIMSS International Study Center for verification and forwarded to ACER to be used in the scaling. When the review of the item statistics was completed and the IEA Data Processing Center had updated the database accordingly, the revised data files were sent to ACER. ACER was then responsible for computing the international item difficulties and for scoring individual students on the international scales. Once the sampling weights and international scale scores were verified at the TIMSS International Study Center, they were sent to the IEA Data Processing Center for inclusion in the international database and distributed to the national research centers. The International Study Center prepared the international report tables and published the reports of the study results. A pictorial representation of the flow of the data files is presented in Figure 3.1.

A very important part of the data processing was the interaction among the staff at the TIMSS International Study Center, the staff at the IEA Data Processing Center, and the National Research Coordinators. At specific stages of the data verification, the IEA Data Processing Center returned countries' data files for checking. These data files were accompanied by computer printouts with summary statistics to be reviewed by the NRC, together with specific questions pertaining to the data.

3.2 DATA ENTRY AT THE NATIONAL RESEARCH CENTERS

Each TIMSS national research center was responsible for entering the achievement and background data into computer data files. Countries were provided with software adapted specifically for the purpose of TIMSS. The software, DATAENTRYMANAGER (DEM), was provided to each of the participating countries together with codebooks for data entry. The codebooks contained information about the variable names used for each variable in the survey instruments, and about field length, field location, labels, valid ranges, default values, and missing codes. The codebooks could be used together with DEM in the data entry process. Although this was the recommended

Figure 3.1 Flow of TIMSS Data Files



procedure, some of the participating countries elected to use a different data entry system. Data files were accepted from the countries provided they conformed to the parameters set in the international codebooks. In order to facilitate data entry, the codebooks and data files were structured to match the test instruments and questionnaires. This meant that there was a data file for each survey instrument.

Each country was responsible for submitting nine data files if participating fully in Population 1 (including performance assessment), and ten data files if participating fully in Population 2. Each of these files had its own codebook. The files for each population are listed in Table 3.1.[1] Although generally collected during the same session,

1 "Written assessment" and "achievement" are used interchangeably to refer to the items or data from the test booklets administered to students. The file name for these data is "Written Assessment."

the student background data were entered separately from the student achievement data because the tests and questionnaires were administered as separate instruments. This was done to prevent students from looking back or ahead at their work in the achievement booklet and, most important, because the open-ended achievement items had to be scored following administration. Setting the system to enter the student background data in a file separate from the achievement data allowed the data manager from each country to start entering student background data without having to wait for the scoring process to finish.

Table 3.1 Files Submitted to the IEA Data Processing Center

File	Population 1	Population 2
Student Background	x	x
Written Assessment	x	x
Written Assessment Coding Reliability	x	x
School Background	x	x
Teacher Background	x	-
Mathematics Teacher Background	-	x
Science Teacher Background	-	x
Performance Assessment Student	x	x
Performance Assessment Coding Reliability	x	x
Performance Assessment School Tracking	x	x
Performance Assessment Student Tracking	x	x

The Student Background data file contains one record for each student in the sample, whether the student participated in the testing session or not. Entries were made in this file even if the student was excluded from the testing session. This file was used to record the information given by the students in the student questionnaire and other information on identification, participation, and sampling.

The Written Assessment data file contains one record for each student who was administered a test booklet. A record also was created for any student whose booklet was lost, but not for students who did not respond to the written assessment. The necessary information for these students was found in the Student Background data file.

In order to check the reliability of the free-response item coding, the free-response items in a random sample of 10 percent of booklets were coded independently by a second coder. The Written Assessment Coding Reliability file contains one record for each student whose responses to the free-response items were coded by a second coder.

The School Background data file contains one record for each originally sampled school, whether the school participated in the survey or not. They also contain records for those schools that participated in the survey as replacement schools. This file was used to register the information from the school questionnaire and on the participation status of schools.

The Teacher Background data file contains one record for each teacher listed as a teacher of a sampled student, even if the teacher was not administered a survey instrument. These files contain the information reported in the teacher questionnaires. The teachers for the third and fourth graders (Population 1) all received the same questionnaires with questions pertaining to the teaching of both mathematics and science. The teachers of the seventh and eighth graders (Population 2) received one of two questionnaires with questions regarding the teaching of either mathematics or science. The data for the mathematics teachers were recorded in a different file from the data for the science teachers.

The Performance Assessment Student data file created in each country contains one record for each performance assessment task that was assigned to a student, even if the student did not complete or attempt the task. Participating students are each represented with up to six entries in this data file depending on the number of tasks they were assigned to take.

The Performance Assessment School Tracking file contains one record for each school sampled for the performance assessment. This data file also contains the information recorded in the performance assessment Post Administration Form.

The Performance Assessment Student Tracking file contains one record for each entry in the Performance Assessment Tracking Form, so that each student is represented only once. This data file was meant to simplify the entering of the tracking information, which is of extreme importance for the linkage to the written assessment data. This file contains information about the specific tasks and task sequence assigned to each student.

The Performance Assessment Coding Reliability data file contains one record for each task that was coded by a second coder for reliability purposes.

Table 3.2 presents the total number of files and records of each type received from all the participating countries.

Table 3.2 Data Files Received by the IEA Data Processing Center

File	Population 1		Population 2	
	Files	Observations	Files	Observations
Written Assessment	27	206,662	43	338,908
Student Background	27	206,662	43	338,908
Written Assessment Coding Reliability	17	13,432	29	20,376
School Background	27	5,337	43	7,808
Teacher Background	27	10,757	-	-
Mathematics Teacher Background	-	-	41	13,885
Science Teacher Background	-	-	41	23,139
Performance Assessment	10	11,746	22	25,501
Performance Assessment Coding Reliability	4	641	14	1,852

In addition to submitting the data files, countries were also required to submit supporting documentation of their field procedures and copies of their national instruments (translated tests and questionnaires). The documentation included a report of their survey activities, a series of data management forms with clear indications of any changes made in the survey instruments or the structure of the database, and copies of all sampling tracking forms. Copies of these materials were archived at the IEA Data Processing Center and kept for reference purposes during data processing.

Each country was provided with a program called LINKCHK that was to be used to carry out checks on the data files prior to submitting them to the IEA Data Processing Center. The program was designed to help NRCs perform an initial check of the system of student, teacher, and school identification numbers after data entry, both within and between files.

LINKCHK performed checks for:

- Duplicate occurrences of identification numbers
- Inconsistencies in the identification numbering system
- Mismatches between different student files
- Mismatches between different teacher files
- Mismatches in the student-teacher linkage

Generally, two types of checks were made:

- Checks within the school, teacher, or student files
- Checks across linked files

The reports produced by the LINKCHK program allowed countries to correct problems in the identification system before transferring the data to the IEA Data Processing Center.

3.3 DATA CLEANING AT THE IEA DATA PROCESSING CENTER

Once the data were entered into data files at the national research center, the data files were submitted to the IEA Data Processing Center for checking and input into the international database. This process is generally referred to as data cleaning. The goals of the TIMSS data cleaning were to identify, document, and, where necessary and possible, correct deviations from the international file structure, and to correct key punch errors, systematic deviations from the international data formats, problems in linking observations between files, inconsistent tracking information between and within files, and inconsistencies within and across observations. The main objective of the process was to ensure that the data adhered to international formats and reflected accurately and consistently the information collected within each country.

Data cleaning involved several steps. Some of these were repeated in an iterative fashion until satisfactory results were achieved. During the first step of data cleaning, all incoming data files were checked and reformatted if necessary so that their file structure conformed to the international format. As a second step, all problems with identification variables, linkage across files, codes used for different groups of variables, and participation status were detected and corrected. The distribution for each variable was examined with particular attention to those variables that presented implausible or inconsistent distributions based on the information from the country involved.

During this stage, a series of data summary reports was generated for each country. The reports contained listings of codes used for each variable and pointed to outliers and changes in the structure of the data file. They also contained univariate statistics. The reports were sent to each participating country, and the NRC was asked to review the data and provide advice on how to best resolve inconsistencies in the data. In many cases the NRC was obliged to go back to the original booklets from which the data had been entered initially.

During the data cleaning process two main procedures were used to make necessary changes in the data. Errors due to incorrect data entry were usually corrected by keying the correct value directly. Inconsistencies in the hierarchical identification variables, whenever possible, were corrected by means of computer programs. In either case, all changes made in the data after they were received by the IEA Data Processing Center were documented. A database was created in which each change made in the data was recorded, and it was possible to reconstruct the original database received from a country.

In the following section each of the steps mentioned above is described in more detail.

3.3.1 Standardization of National File Structure

The first step in the data processing at the international level was to verify the compatibility of the national datasets with the international file structure as defined in the TIMSS international codebook. This was necessary before the standard cleaning with the Data Processing Center cleaning software could be performed.

Although the TIMSS international codebooks distributed with the data entry software gave clear and detailed instructions about the structure and format of the files each country was to submit to the IEA Data Processing Center, some countries opted to enter and submit their data files in other formats, using structures different from the international standard. For the most part, these differences were due to specific national circumstances.

The *TIMSS Guide to Checking, Coding, and Entering TIMSS Data* (TIMSS, 1995) asked countries to prepare and send their data files using the DEM software, which produces an extended dBase format. Some data files were also received in ASCII fixed format (raw data), SPSS format, and SAS format.

After the national files were converted into the extended dBase format, the structure of the files was inspected and deviations from the international file structure were identified. A standard software tool automatically scanned the file structure of the country files and reported the following deviations:

- International variables dropped
- National variables added
- Different variable length or number of decimal positions
- Different coding schemes or out of range values
- Specific national variables
- Gang-punched variables

Together with the inspection of the national data files, the data management and tracking forms submitted by each NRC were reviewed. As a result of this initial review, the Data Processing Center outlined and implemented necessary changes in the national data to make the files compatible with the international format. In most cases programs had to be prepared to fit the file structures and specificities of each country.

During this process some of the files were merged (for example, the Student Background and the Written Assessment data files). The structure of some of the files was also changed significantly, since direct correspondence to the instruments was no longer necessary. Some variables created during data entry for verification purposes only were not copied to the transformed data files. The changes made in the files during the cleaning process are described below. In general, variables used during data entry for verification were dropped from all files and new variables were added (e.g., reporting variables, derived variables, sampling weights, and achievement scores). What follows is a brief description of the changes performed in the files received from the countries.

3.3.1.1 Student Background File

Several new variables were added to the beginning of each record to represent students' participation status in the two testing sessions and in completing the student background questionnaire. The student's age computed from the date of testing and the date of birth were also added to the files, as were sampling weights and several achievement scores for both mathematics and science.

For Population 2, two versions of the student background questionnaire were available for administration. Each had its own data file and codebook. Although most countries chose to use one version of the questionnaire, some countries opted to use both versions. One version was tailored for educational systems where science is taught as an integrated subject (non-specialized version). The second version was tailored for educational systems where the sciences (biology, earth sciences, physics, chemistry) are taught separately (specialized version). Although a separate data file was created for data entry for each questionnaire version, these were then merged into one file with the same structure. This new file contained all variables from the version for non-specialized science teaching in the order in which they appear in the questionnaire, followed by all variables from the version for specialized science teaching that do not appear in the non-specialized version. For students who received the non-specialized version of the questionnaire, all questions that were given only in the specialized version were coded as "not administered." For students who were assigned the specialized version of the questionnaire, all questions that were asked only in the non-specialized version were coded as "not administered." The international structure of the Student Background data file is shown in Figure 3.2. In Population 1 there was only one version of the student questionnaire.

Figure 3.2 Revised Structure of the Student Background File

IDs	Tracking Information & Teacher Linkage	Background Information from the Non-Specialized Version	Background Information from the Specialized Version and not in the Non-Specialized Version	Weights	Scores	Derived Variables

3.3.1.2 Written Assessment File

The structure of the Written Assessment files prepared for data entry focused on the structure of the booklets (eight each for Populations 1 and 2). During data entry, once the version of the booklet was indicated, the data software displayed only the variables representing the items in that particular booklet. A variable was created for each item in a booklet, and the order of these variables reflected the order of the items within a booklet. This kept data entry and programming of the data entry software to a simple and rectangular structure. However, it also meant that a lot of redundant variables were created during data entry, since an item administered in more than one booklet was coded as a different variable for each booklet in which the item occurred. A useful feature of the redundancy is that it allowed the booklet administered to the student to be identified easily even if there was a key-punch error when the identification of the test booklet was entered.

After final cleaning, the Written Assessment files were restructured so that each item appeared in just one location in the student records, regardless of the test booklet it came from. This new structure reflects the item clusters used to assemble the booklets (Adams and Gonzalez, 1996) and not the booklet layout. The variables for the items that were not administered to the student were coded as "not administered." The structure of the Written Assessment file is presented in Figure 3.3.

Figure 3.3 Revised Structure of the Written Assessment File

IDs	Tracking	Achievement Item Cluster					
		A	B	C	-------------------------------	Y	Z

3.3.1.3 Written Assessment Coding Reliability File

The structure of the Written Assessment Coding Reliability file prepared for data entry also mirrored the structure of the eight test booklets. Again, a variable was created for each free-response item in a booklet, and the order reflected the order of appearance of the items within the booklets. In the final international data file the variables were rearranged so that each item was represented by only one variable regardless of the booklet in which it appears. All other variables representing items not included in the booklet administered to the student were coded as "not administered."

The final international version of the Coding Reliability file includes both the data from the 10 percent sample of students selected for reliability coding and the original data for these students. This enables the user of the Coding Reliability file to compare the codes without having to merge any files.

A third set of variables was included in the final international version of the file to reflect the agreement between the two codes assigned to the answers to the free-response items.

3.3.1.4 Teacher Background File

The structure of the Teacher Background files is similar to the that of the original data file used for data entry. For Population 2, two files were used for data entry, one corresponding to the mathematics teacher background questionnaire and one corresponding to the science teacher background questionnaire.

In some cases, a teacher taught more than one sampled class or course or, in the case of Population 2, both subjects to the same class or course. Although it would have been desirable to assign a questionnaire to a teacher for each class taught, in most countries the resulting burden to teachers was considered unacceptable. However, much of the information obtained from the questionnaires was not related to the specific class or course taught, but to background characteristics of the teacher (e.g., sex and age, teaching experience). This information was asked only once from the teachers.

Each teacher was assigned a unique identification number (Teacher ID) and a Teacher Link Number specific to each class taught by the teacher. The Teacher ID and Teacher Link Number combination identified a teacher teaching one specific class. For example, students linked to teachers identified by the same Teacher ID but different Teacher Link Numbers were taught by the same teacher but in different classes. If students were linked to a teacher observation identified by a combination of Teacher ID and

Teacher Link Number for which no data were obtained, but there was an observation in the teacher file with the same Teacher ID and a different Teacher Link Number with data available, all personal data for the teacher were transcribed to the missing observation. Thus, whether or not a teacher completed a questionnaire pertaining to a specific course, background information was sometimes available.

During data processing, teacher-related information was transcribed from other observations of the same teacher to teacher observations for which a questionnaire was not administered (or not returned). In some countries, more than two questionnaires per teacher were administered, but only one contained the personal information. In these cases, a similar transcription was made. Table 3.3 gives two examples of how teacher data have been transcribed.

Table 3.3 Examples of Teacher Data Transcribed to Files

Obs.	Teacher ID	Link No.	Class ID	Participation	Sections A & D Teacher-related Data	Sections B & C Class-related Data
1	22201	1	22203	Questionnaire completed	Data	Data
2	22201	3	22205	No questionnaire assigned	Data transcribed from Obs. 1	No data
:	:	:	:	:	:	:
10	33302	4	33301	Questionnaire completed	Data	Data
11	33302	5	33302	Only class-related part completed	Data transcribed from Obs. 10	Data

3.3.1.5 School Background File

The file structure of the cleaned school data sets in the international database is identical to the structure used for data entry. No major changes were made. The file includes a School Identification number (ID) block and the variables in order of their appearance in the school questionnaire.

3.3.1.6 Performance Assessment Files

The structure of the Performance Assessment data files submitted by the national centers to the IEA Data Processing Center mirrored the structure of the instruments and tracking forms. To make the files suitable for further analysis, the Performance Assessment Student file was rearranged from a multi-record structure (i.e., multiple records for each student – one for each task taken by the student) to a single-record structure (i.e., one record per student). In addition, information from the Performance Assessment Student file and the Performance Assessment Tracking file were combined into one file, together with particular variables from the Student Background file (age, gender, achievement scores, etc.) and sampling weights computed by Statistics Canada.

This revised file was called the Performance Assessment Combined file. The Performance Assessment Coding Reliability files were kept separate and processed in the same manner as the Written Assessment Coding Reliability files

3.3.2 Standard Cleaning

After the data received from the countries were transformed into the international format, a set of standard cleaning rules were applied to each of the data files received from each country. These rules were applied using software the IEA Data Processing Center had developed to report and in many cases to correct inconsistencies in the data. Some inconsistencies could not be solved automatically but had to be reviewed carefully and appropriate corrections devised, where possible.

In particular, the following problems were sought and corrected whenever possible (for further details, please refer to Jungclaus and Bruneforth (1996)):

- Problems with identification, tracking, and other indicator variables
- Problems with split variables, i.e. variables where respondents were allowed to check more than one option
- Problems with the variable indicating the achievement booklet administered to the student
- Problems with filter and dependent questions

After as many problems as possible were solved at the IEA Data Processing Center (by reviewing the instruments and national documentation or by applying the cleaning rules), the Data Processing Center cleaning software was used a second time to create a report of remaining data problems. These reports were summarized and sent to the NRCs with specific questions and, in some cases, suggestions as to how the problem could be solved.

For the Performance Assessment files, the tracking data regarding the performance assessment rotation scheme, sequence number, and station participation status were compared with the tasks that students performed and for which data had been recorded in the Performance Assessment Student file. The tracking information also included the number of the written test booklet which each student had completed, thus enabling a double linkage check (in addition to the student ID) to the written assessment.

3.3.3 Item Cleaning

After applying the cleaning rules described above, the achievement data underwent a careful and detailed review.

For this purpose, an item analysis was performed using the item analysis software QUEST developed by ACER (Adams and Khoo, 1993). National scores in mathematics and science based on the Rasch model were calculated and several reports were generated with these data. Some data problems, such as items with changes in the coding scheme or switched response options, were detected and corrected at this point. Reports with summary item statistics were sent to the NRCs for their review.

The Coding Reliability data were compared with the Written Assessment data. For this purpose, the percentage of agreement between the codes assigned by the two coders was calculated on two levels: agreement between the number of score points assigned to an item and agreement on the two-digit diagnostic code.

After this initial review by the IEA Data Processing Center, reports were generated with item statistics. The TIMSS International Study Center used these reports to conduct a thorough review of the achievement item data. Details of this process are presented in Chapter 6 of this report.

3.3.4 Country-Specific Cleaning

Some of the anomalies detected by the checking procedure had to be solved case by case. During this process, it was important to find individual solutions that followed general guidelines, so that uniform solutions could be applied to similar problems in other countries.

The corrections made in this cleaning step were based on the NRCs' review of the preliminary statistics from the IEA Data Processing Center, the NRC field operations reports and instruments sent with the data, and the NRCs' comments on the data almanacs produced by the TIMSS International Study Center. In particular, the following steps were performed on a country-by-country basis to correct the data:

- Correcting switched options/categories in categorical background variables
- Deleting data entered for questions that were not included in the international versions of the questionnaires
- Deleting data entered in error
- Collapsing categories to match the international coding scheme
- Deleting data made incompatible by translation problems
- Copying data from one observation to another if the information requested was identical for both observations
- Adding dummy records to the files to ensure correct linkage across files

None of these steps were performed without the cooperation of the NRCs. They had to confirm or reject the suggested data changes; more important, in many cases they had to give detailed advice about the changes to be performed to the coding scheme.

3.3.5 Other General Cleaning

After transforming the data files into the international format, performing the standard cleaning on them, and reviewing the achievement data, two other kinds of checks were made: statistical checks and consistency checks.

3.3.5.1 Statistical Checks

Statistical checks were designed to find outliers for continuous variables, variables with very high percentages of missing values, and categorical variables with different numbers of options from the international version of the instruments. Statistical checks were performed separately for each country. For such checks, several preparatory steps were necessary. In particular, descriptive statistics were computed for each variable within each country and these statistics were stored in a database. The information compiled in this way was used as outlined below.

Outlier Detection

In order to check variables for extreme values, an outlier was defined as a value in a variable that is over 5 standard deviations above the mean for that variable, or with a value twice as large as the 90th percentile for the variable. Any such variables detected were carefully examined.

For some of the variables found by this procedure (e.g., number of students in a school), additional information was used to judge the plausibility of the detected outlying values. If the file contained obvious miskeys, the variable was coded to "Invalid" for the detected cases. Cases that could not be resolved at the Data Processing Center were reported to NRCs and treated according to their suggestions.

High Percentages of Missing Observations

Variables were flagged for investigation if more than 99 percent of the cases had missing values. If such a variable was detected, the corresponding question in the questionnaire was examined. Often in such cases the question was not completed by the respondents because it was not applicable. For example, teachers were asked a question about teaching the theory of relativity. Many teachers did not respond since relativity theory was not part of the curriculum in their country. Thus, the variables related to these questions show high missing rates. Another example would be that a question was not asked, but data entry errors gave the corresponding variable(s) inconsistent missing codes. In that case, the missing codes were made consistent.

Additional Response Options for Categorical Variables

The observed values for categorical variables were compared with the valid codes specified by the international codebook. If additional codes were found, the corresponding question in the questionnaire was examined. It was possible that the additional code was due to key-punch error during data entry. Where it was determined that this was the case, the corresponding categories were recoded to "Invalid." If, on the other hand, the question that was asked allowed additional categories, the NRCs were asked to help find a way to make the new code internationally comparable. If recoding was possible, the original value for the variable was kept in a separate country-specific variable. If it was not possible to recode to meet the international coding scheme, the original data were kept in a separate variable and the international variable was coded to an explicit missing code.

Response Options with a Frequency of Zero in Categorical Variables

If a frequency of zero was detected for an option of a categorical variable, the corresponding question in the questionnaire was checked as a precaution. If a category in the original version of the question was missing, the NRC was contacted to verify that the correct categories were retained. However, if the category was not missing in the questionnaire but was not checked by any respondent, the data were not changed. Quite often, variables belonging to groups of questions had zero frequencies for one or more of the categories. For example, the school questionnaire asked for the frequency of different types of student behavior in schools. Some forms of behavior did not happen often; thus the corresponding categories had a frequency of zero.

3.3.5.2 Consistency Checks

Consistency checks dealt with problems that were discovered in the first phase of the cleaning process, but not corrected at that time because information about the problems across countries was needed to decide on the rules to be applied. The following sections describe the checks applied to all countries and the inconsistencies that were corrected.

Student's Gender, Date of Birth, Age, and Date of Testing

If a student's sex as reported in the background questionnaire differed from that in the tracking information, the tracking version was replaced by the background questionnaire version in Population 2. In Population 1 the replacement was the other way around. The same substitution procedure was followed with regard to students' dates of birth. Changes in the date of birth were made provided that the value to be used in the substitution resulted in a valid age for the student. For students whose estimated age was less than ten years of age in Population 2, or less than six years of age in Population 1, the estimated age was coded as invalid.

If the date of testing was missing, it was replaced by the modal value of the student's class when available.

Teacher File

In the Teacher Files, two lists in the Population 2 questionnaires were considered and corrected separately: a list of subjects taught during a school week and a list of tasks that must be performed during a school week. If no zeroes were used, more than four variables in a list were coded differently from "Not administered," or values greater than zero could be found, then "Omit" codes were recoded to zero.

School File

The questions concerning the same course of instruction were checked for consistent answers. If all students followed the same course of instruction (filter = Yes) and the majority of answers was consistent with the filter, all answers in the "No" list were recoded to "Not applicable." If, on the other hand, valid answers could be found in the

"No" list and only missing values could be found in the "Yes" list, the filter was changed to "No." Uncertain cases were reported and recoded directly if possible. Sometimes the appropriate response could be deduced from the answering pattern found in the data.

3.3.6 Performance Assessment Cleaning Routines

The Performance Assessment file cleaning routines were based on the data checks created for Written Assessment files, although some routines were modified to fit the structure of the Performance Assessment files. In addition, due to the design of the Performance Assessment and the linkages among the various files, it was necessary to develop special cleaning routines. These cleaning programs were of two types. One type of cleaning program flagged inconsistencies between the Performance Assessment Tracking file data and the Performance Assessment Student file data. The second type of cleaning program flagged problems associated with the Performance Assessment Combined file.

Performance Assessment cleaning problems could not be resolved automatically, but rather had to be solved case by case. It would have been very difficult to create general cleaning rules which could cover the complexity of the Performance Assessment design. The structure of Performance Assessment required case-by-case cleaning especially to resolve inconsistencies between the Performance Assessment Tracking file and Performance Assessment Student file. Problems were resolved by reviewing error report printouts and data, and through dialogue with the participating countries. All corrections were undertaken by editing the data files.

Similar to the written assessment items, the performance assessment item responses were analyzed with the QUEST program. Both Rasch statistics and classical item statistics were calculated, printed, and reviewed, as described in Section 3.3.4. The only difference to the procedure for the written assessment items is that all performance assessment item responses were scored using the two-digit coding scheme, like the open-ended items of Written Assessment.

The Performance Assessment Reliability Coding data were processed and statistics were produced for review in a similar manner as those for the written assessment.

After the Data Processing Center had reviewed all item statistics, they were sent to the participating countries and the International Study Center. Country-specific item statistics enabled NRCs to review their data, and international item statistics were sent to the International Study Center for an international review of all items for all countries.

3.4 DATA PRODUCTS

3.4.1 Data Almanacs

Together with their data files, each country received data almanacs produced by the TIMSS International Study Center that contained weighted summary statistics by grade, for each participating country, on each variable included in the survey instruments. There were two types of display. The display for *categorical variables* included an estimate of the size of the student population, the sample size, the weighted percentage of students who were not administered the question, the percentage of students choosing each of the options on the question, and the percentage of students who did not choose any of the valid options. The percentage of students to whom the question did not apply was also presented in the almanac. For *continuous variables* the display included an estimate of the size of the student population, the sample size, the weighted percentage of students who were not administered the question, the percentage who did not respond, the percentage to whom the question did not apply, the mean, mode, minimum, maximum, and the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. An example of such data displays is presented in Figures 3.4 and 3.5. These data almanacs were sent to each of the participating countries for review. When necessary, they were accompanied by specific questions about the data presented in them. These almanacs also were used by the TIMSS International Study Center during the data review and in the production of the reporting tables.

Figure 3.4 Example Data Almanac Display for Categorical Variable

1Third International Mathematics and Science Study
4:17 Sunday, September 21, 1997 1
Report on Student Background Variables - Population 2
Preliminary results: DO NOT CITE OR CIRCULATE

Question: Are you a boy or a girl? (BSBGSEX)
Location: SQ2-2

****
1.SEVENTH GRADE
GEN\STUDENT'S SEX

Country	Population	Sample	%NA	1.GIRL %	2.BOY %	Other %
Australia	238294	5599	1.3	52.0	48.0	1.3
Austria	89593	3013	3.3	53.2	46.8	3.6
Belgium (Fl)	64177	2768	0.3	49.4	50.6	0.3
Belgium (Fr)	49898	2292	1.3	53.2	46.8	1.8
Bulgaria	140979	1798	0.5	54.0	46.0	0.5
Canada	377732	8219	0.8	49.4	50.6	2.2
Colombia	619462	2655	0.9	49.9	50.1	1.1
Cyprus	10033	2929	0.2	48.9	51.1	0.3
Czech Republic	152492	3345	0.2	50.6	49.4	0.2
Denmark	44980	2073	5.1	51.2	48.8	5.1
England	465457	1803	1.8	45.7	54.3	1.8
France	860657	3016	3.4	49.6	50.4	3.5
Germany	742346	2893	0.8	50.9	49.1	1.5
Greece	130222	3931	0.2	48.2	51.8	0.4
Hong Kong	88591	3413	0.5	44.3	55.7	0.6
Hungary	118727	3066	2.0	50.4	49.6	2.4
Iceland	4212	1957	0.7	49.0	51.0	0.7
Iran, Islamic Rep.	1052795	3735	1.7	43.2	56.8	1.7
Ireland	68477	3127	1.1	54.0	46.0	1.1
Israel	.	.	.	.	.	.
Japan	1562418	5130	0.0	48.4	51.6	.
Korea	798409	2907	0.2	42.4	57.6	0.2
Kuwait	.	.	.	.	.	.
Latvia (LSS)	17041	2567	1.8	51.5	48.5	1.9
Lithuania	36551	2531	0.7	49.9	50.1	0.7
Netherlands	175419	2097	2.7	50.5	49.5	2.8
New Zealand	48508	3184	0.9	46.7	53.3	0.9
Norway	51165	2469	0.5	48.6	51.4	0.5
Portugal	146882	3362	0.6	51.6	48.4	0.6
Romania	295348	3746	0.5	51.8	48.2	0.5
Russian Federation	2168163	4138	0.2	51.0	49.0	0.2
Scotland	61938	2913	4.1	49.2	50.8	4.1
Singapore	36181	3641	0.4	49.9	50.1	0.4
Slovak Republic	83074	3600	0.0	50.9	49.1	0.0
Slovenia	28049	2898	0.2	51.3	48.7	0.2
South Africa	649180	5301	0.6	53.9	46.1	1.7
Spain	549032	3741	0.5	49.7	50.3	0.5
Sweden	96494	2831	0.5	48.8	51.2	0.6
Switzerland	66681	4085	0.6	49.7	50.3	0.6
Thailand	680225	5810	0.0	58.2	41.8	1.1
United States	3156847	3886	1.7	50.3	49.7	1.7

Figure 3.5 Example Data Almanac Display for Continuous Variable

```
1Third International Mathematics and Science Study                    14:17 Sunday, September 21, 1997   2
 Report on Student Background Variables - Population 2
 Preliminary results: DO NOT CITE OR CIRCULATE

 Question: Student's Age in Years (BSDAGE)
 Location: DERIVED

 *GRADE*=1.SEVENTH GRADE
```

Country	Population	Cases	%Not Ad.	%Omit	%Not Ap.	Mean	Mode	Min	P5	P10	Q1	Median	Q3	P90	P95	Max
Australia	238294	5599	0.2	0.0	0.0	13.2	13.3	11.8	12.5	12.7	12.9	13.3	13.5	13.8	14.1	16.3
Austria	89593	3013	3.9	0.0	0.0	13.3	13.3	12.3	12.7	12.8	12.9	13.3	13.6	14.0	14.4	16.3
Belgium (Fl)	64177	2768	0.1	0.0	0.0	13.0	12.9	11.5	12.4	12.5	12.8	13.0	13.3	13.7	14.0	16.3
Belgium (Fr)	49898	2292	2.9	0.0	0.0	13.2	12.6	11.4	12.4	12.5	12.7	13.0	13.5	14.2	14.5	17.8
Bulgaria	140979	1798	0.5	0.0	0.0	13.1	13.2	11.2	12.6	12.7	12.9	13.1	13.3	13.4	13.8	14.5
Canada	377732	8219	0.2	0.0	0.0	13.1	13.1	10.2	12.4	12.5	12.8	13.0	13.3	13.8	14.3	18.0
Colombia	619462	2655	1.8	0.3	0.0	14.5	13.0	10.0	12.3	12.5	13.0	14.0	15.1	16.8	18.6	49.8
Cyprus	10033	2929	0.6	0.0	0.0	12.8	12.9	11.5	12.3	12.3	12.5	12.8	13.0	13.2	13.4	15.3
Czech Republic	152492	3345	0.0	0.0	0.0	13.4	13.3	10.8	12.8	12.8	13.1	13.3	13.7	13.9	14.3	16.3
Denmark	44980	2073	3.3	0.0	0.0	12.9	12.9	11.5	12.3	12.4	12.6	12.8	13.1	13.3	13.4	15.2
England	465457	1803	0.0	0.0	0.0	13.1	13.5	12.2	12.6	12.7	12.8	13.1	13.3	13.5	13.5	14.5
France	860657	3016	6.9	0.0	0.0	13.3	13.0	10.6	12.4	12.5	12.8	13.2	13.8	14.3	14.8	17.4
Germany	742346	2893	2.4	0.0	0.0	13.8	13.8	11.1	13.1	13.2	13.3	13.7	14.0	14.5	14.9	29.7
Greece	130222	3931	0.5	0.0	0.0	12.6	12.6	10.0	12.0	12.1	12.3	12.6	12.8	13.1	13.7	17.5
Hong Kong	88591	3413	0.5	0.2	0.0	13.2	13.3	10.3	12.5	12.5	12.8	13.0	13.3	14.0	14.6	18.6
Hungary	118727	3066	3.5	0.0	0.0	13.4	13.5	11.0	12.7	12.8	13.0	13.3	13.6	14.0	14.5	17.3
Iceland	4212	1957	0.1	0.0	0.0	12.6	12.8	10.4	12.2	12.3	12.4	12.7	12.8	13.0	13.1	15.1
Iran, Islamic Rep.	1052795	3735	10.9	0.0	0.0	13.6	12.7	11.2	12.7	12.7	12.9	13.4	14.1	15.0	15.7	18.3
Ireland	68477	3127	0.0	0.0	0.0	13.4	13.3	12.2	12.7	12.8	13.0	13.4	13.8	14.1	14.3	16.5
Japan	1562418	5130	0.5	0.0	0.0	13.4	13.4	12.2	12.9	13.0	13.2	13.4	13.7	13.8	13.8	15.7
Korea	798409	2907	0.0	0.0	0.0	13.2	12.8	10.5	12.8	12.8	12.9	13.2	13.5	13.7	13.7	14.9
Latvia (LSS)	17041	2567	0.2	0.1	0.0	13.3	13.0	12.0	12.7	12.8	12.9	13.2	13.6	14.0	14.4	16.4
Lithuania	36551	2531	0.0	0.0	0.0	13.4	13.1	10.1	12.8	12.8	13.0	13.3	13.6	14.0	14.4	16.2
Netherlands	175419	2097	0.4	0.0	0.0	13.2	12.9	10.3	12.6	12.7	12.8	13.2	13.5	13.9	14.3	16.5
New Zealand	48508	3184	0.0	0.0	0.0	13.0	13.1	11.2	12.4	12.5	12.8	13.0	13.3	13.5	13.6	15.1
Norway	51165	2469	0.0	0.0	0.0	12.9	13.0	11.2	12.3	12.4	12.6	12.8	13.1	13.3	13.3	14.8
Portugal	146882	3362	0.1	0.0	0.0	13.4	13.0	11.3	12.5	12.6	12.8	13.2	13.8	14.8	15.3	19.5
Romania	295348	3746	0.1	0.0	0.0	13.7	13.6	11.6	12.9	13.0	13.3	13.6	13.9	14.3	14.6	17.5
Russian Federation	2168163	4138	0.0	0.0	0.0	13.0	12.8	11.5	12.5	12.6	12.8	13.0	13.3	13.5	13.9	16.8
Scotland	61938	2913	0.5	0.0	0.0	12.7	13.0	11.3	12.2	12.3	12.4	12.7	13.0	13.1	13.2	14.3
Singapore	36181	3641	0.0	0.0	0.0	13.3	13.1	12.8	12.8	12.9	13.0	13.3	13.5	13.8	13.8	17.5
Slovak Republic	83074	3600	0.0	0.0	0.0	13.3	13.0	12.4	12.8	12.8	13.0	13.3	13.5	13.7	13.9	16.2
Slovenia	28049	2898	0.2	0.0	0.0	13.8	13.8	12.3	13.3	13.3	13.5	13.8	14.1	14.3	14.6	15.8
South Africa	649180	5301	5.4	0.0	0.0	13.9	13.0	10.0	12.0	12.1	13.0	13.4	14.9	16.1	17.3	24.8
Spain	549032	3741	0.0	0.0	0.0	13.2	13.0	12.4	12.4	12.5	12.8	13.1	13.4	14.3	14.8	16.3
Sweden	96494	2831	0.4	0.0	0.0	12.9	13.2	10.3	12.4	12.5	12.7	12.9	13.2	13.3	13.3	14.5
Switzerland	66681	4085	0.5	0.0	0.0	13.1	13.0	10.2	12.3	12.4	12.7	13.0	13.3	13.8	14.2	16.7
Thailand	680225	5810	1.8	0.0	0.0	13.5	13.6	11.6	12.5	12.8	13.2	13.5	13.8	14.1	14.6	15.1
United States	3156847	3886	0.0	0.0	0.0	13.2	13.1	10.2	12.5	12.7	12.8	13.2	13.5	13.9	14.3	16.2

3.4.2 Versions of the National Data Files

Building the international database was an iterative process. The IEA Data Processing Center provided NRCs with a new version of their countries' data files whenever a major step in data processing was completed. This also guaranteed that the NRCs had a chance to review their data and run their own checks to validate the data files.

Three versions of the data files were sent out to each of the countries before the TIMSS international database was made available. Each country received its own data only. The first version of the data files was sent to the NRC as soon as that country's data had been cleaned. These files contained national Rasch scores calculated by the Data Processing Center. Documentation, with a list of the cleaning checks and all corrections applied to the data, was included to enable the NRC to review the cleaning process. Univariate statistics for the background data and item statistics were also provided for statistical review of the data. A second version of the data files was sent to the NRCs when the weights and the international achievement scores were available and had been merged with the files. A third version of the data was sent together with the data almanacs after final updates had been made in the data files, to enable the NRCs to validate the results presented in the first international reports.

For the performance assessment, participating countries were provided with their performance assessment data as soon as they were cleaned and restructured. The data were distributed along with national item statistics and a codebook describing the new structure of the data.

When international weights and scores were available, each country received a new version of its performance assessment data and the International Study Center received the data for all countries.

3.4.3 Reports

Several reports were produced during data processing at the IEA Data Processing Center to inform and assist the NRCs, the TIMSS International Study Center, and other institutions involved in TIMSS. The NRCs were provided with diagnostic reports and univariate statistics to help them in checking their data. The TIMSS International Study Center and ACER were provided with international item statistics. The International Study Center also received international coding reliability statistics and international univariate statistics. A report was made to the TIMSS International Study Center and the TIMSS Technical Advisory Committee about each country's deviations and cleaning status as well as the major problems encountered during its data cleaning. The report also included general statistics about the number of observations per file and subpopulation and student response rates.

3.5 COMPUTER SOFTWARE

dBase was used as a standard database program for handling the incoming data. Tools for pre-cleaning and programs such as LINKCHCK (described earlier), and MANCORR and CLEAN (described below) were developed using CLIPPER for manipulating data and some data processing. Statistical analyses (e.g., univariate statistics) for data cleaning and review were carried out with SAS. The final data sets were also created using SAS. For item statistics, the Data Processing Center used the QUEST software (Adams and Khoo, 1993).

The main programs that were developed by the Data Processing Center for TIMSS are described below. Most of the programs that were written for country-specific cleaning needs are not listed. Most of the programming resources in the main cleaning process were spent developing this set of programs.

3.5.1 MANCORR

The most time-consuming and error-prone part of data cleaning is the direct or "manual" editing of errors uncovered by the review process. Based on the Data Processing Center's experience in the IEA Reading Literacy Study and the pilot phases of TIMSS, the data editing program MANCORR was developed. It is easy to use and generates automatic reports of all data manipulation. Its main advantage compared with other editors is that all changes in the data are documented in a log database, from which reports can be generated. As updated data were received from countries, the time-intensive manual changes could be automatically repeated. An "Undo" function allowed the restoration of original values that had been modified with the MANCORR program. The report on which changes were made in the data, by whom, and when was important for internal quality control and review. The MANCORR program was designed using CLIPPER in order to manipulate DATAENTRYMANAGER files.

3.5.2 CLEAN

The central program for data cleaning in TIMSS was the diagnostic program CLEAN, developed with CLIPPER. This program was based on the programs used in the IEA Reading Literacy Study and the TIMSS field tests. It checked all the TIMSS files separately, but also checked the linkages across files and made between-file comparisons. Then corrections were performed according to the rules described above (see Section 3.3.2 and, for a more detailed explanation, Jungclaus and Bruneforth, 1996). An important feature of the program is that it can be used on a data set as often as necessary. It could first be used to make automatic corrections, and subsequently for creating a report only, without performing corrections. Thus it was possible to run a check on the files at all stages of work until the file format was changed to the SAS format. This meant that the program was used not only for initial checks but also to check the work done at the Data Processing Center.

A feature of the TIMSS data cleaning tools is that all deviations are reported to a database, so that reports could be generated by type of problem or by record. Reports previously generated by the program could be compared automatically with newer reports to see which problems had been solved, and even more important, to see whether additional deviations were introduced during manual correction. Last, the databases (which included all reported deviations) were used to generate the final reports to be sent to the countries. These reports showed which deviations were initially in the data, which were solved automatically, which were solved manually, and which remained unchanged.

3.5.3 Programs Creating Meta Databases

Using SAS, several programs were developed by the Data Processing Center for reviewing and analyzing both the background data and the test items. For the background data, a meta database containing information provided by the initial analysis and by the international codebook was created. A meta database containing the relevant item parameters was also created for the achievement test items. Later, all statistical checks and reports used these meta databases instead of running the statistics over all data sets again and again. If the data for one country were changed, then statistics had to be recalculated only for this country; the tabulation program, which accessed only the meta database, could then be applied, since the other countries' values remained unchanged. This reduced the computing time for certain procedures from hours to a few minutes. Both databases are the base sources of several reports produced at both the national and international levels (e.g., for the univariate and item analysis reports).

The univariates and item statistics were prepared on a variable-by-country or country-by-variable basis to allow review at the national level and international comparison of individual variables.

3.5.4 Export programs

As mentioned above, SAS was the main program for analyzing the data. Using SAS, export programs were developed and tested to create output data sets for data distribution that are readable either by SAS or SPSS.

3.6 CONCLUSION

The structures and processes designed for the data processing of TIMSS, the largest international empirical educational study ever conducted, met the tremendous challenge provided. In planning for TIMSS data processing, the major problems were anticipated and provision for dealing with them incorporated into the data processing system. Even the most complicated school systems were handled adequately by the admittedly complex record identification system. This system had been criticized during the planning phase as too complicated, but it proved to be just complex enough to unambiguously identify observations and allow the linkage of files in every education system.

The Data Processing Center was closely involved in the planning phase of the study. This allowed the study to benefit from the Center's knowledge and experience in data processing. For example, it was anticipated that national adaptations and country-specific options would create problems not only during data processing but also in later analysis. Accordingly, international definitions were established that minimized such problems. Most of the problems encountered during data processing arose because countries sometimes modified the internationally-agreed procedures without notifying the Data Processing Center. The adaptation of record identification systems by some countries (because they felt the international system was too complex) created a lot of unexpected work.

Minor modifications, such as adding new categories to questions, switching the order of options, leaving out international response categories, or changing open-ended questions to multiple-choice questions, were easy to recode to match the international definitions unless countries completely restructured the questionnaires, resulting in the need for additional resources and energy to check and reorganize the data. This shows how important it is in any international study to verify translations of the national questionnaires and to ensure internationally comparable data.

Some problems arose due to communications difficulties. Early and continuous involvement of the data processing staff helped minimize the amount of time and work required, by the countries, the International Study Center, and the Data Processing Center, to produce clean data. It was very important that the data processing staff was easily accessible for the participating countries so that they could get help whenever they had problems. Modern technology, such as the capability to send facsimiles, as well as the Internet, makes the will to communicate, and not the distance between the participants, the most important factor in a successful study. TIMSS demonstrated this with the successful communication between the Data Processing Center in Hamburg, the TIMSS International Study Center at Boston College, Statistics Canada in Ottawa, and the Australian Council of Educational Research in Melbourne. The idea of a decentralized study proved feasible and workable. The time difference between the institutions involved occasionally even helped speed up the work: TIMSS was worked on around the clock.

REFERENCES

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 technical report, Volume I: Design and development*. Chestnut Hill, MA: Boston College.

Adams, R.J. and Khoo, S. (1993). *Quest: The interactive test analysis system*. Melbourne: Australian Council for Educational Research.

Gonzalez, E.J. and Smith, T.A., Eds. (1997). *User Guide for the TIMSS international database: Primary and middle school years – 1995 assessment.* Chestnut Hill, MA: Boston College.

Jungclaus, H. and Bruneforth, M. (1996). Data consistency checking across countries. 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.

Third International Mathematics and Science Study (TIMSS). (1995). *TIMSS guide to checking, coding, and entering the TIMSS data* (Doc. Ref.: ICC918/NRC449). Chestnut Hill, MA: Boston College.

4 Calculation of Sampling Weights

Pierre Foy
Statistics Canada

4.1 OVERVIEW

The basic sample design used in TIMSS Populations 1 and 2 was a two-stage stratified cluster design.[1] The first stage consisted of a sample of schools; the second stage consisted of samples of one intact mathematics classroom from each eligible target grade in the sampled schools. The design required schools to be sampled using a probability proportional to size (PPS) systematic method, as described by Foy, Rust, and Schleicher (1996), and classrooms to be sampled with equal probabilities (Schleicher and Siniscalco, 1996). While TIMSS had a basic design for how the national representative samples of students in Populations 1 and 2 were to be drawn, aspects of the design were adapted to national conditions and analytical needs. For example, many countries stratified the school sampling frame by variables of national interest. As another example, some countries chose to sample two classrooms from the upper grade of the target population. Chapter 2 of this report documents in detail the national samples for TIMSS Populations 1 and 2.

While a multi-stage stratified cluster design greatly enhances the feasibility of data collection, it results in differential probabilities of selection; consequently, each student in the assessment does not necessarily represent the same number of students in the population, as would be the case if a simple random sampling approach were employed. To account for differential probabilities of selection due to the nature of the design and to ensure accurate survey estimates, TIMSS computed a sampling weight for each student that participated in the assessment. This chapter documents the calculation of the sampling weights for students sampled for the Populations 1 and 2 main assessment and for those students subsampled to also take part in the performance assessment.[2]

4.2 WEIGHTING PROCEDURES

The general weighting procedure for TIMSS required three steps. The first step for all target populations consisted of calculating a school weight. The school weight also incorporates weighting factors from any additional front-end sampling stages required

[1] The target populations are defined as follows:

Population 1: Students enrolled in the two adjacent grades where most 9-year-old students are found at the time of testing (third and fourth grades in many countries)

Population 2: Students enrolled in the two adjacent grades where most 13-year-old students are found at the time of testing (seventh and eighth grades in many countries).

[2] See Harmon and Kelly (1996) for details of the sampling procedures for the performance assessment.

by some TIMSS participants.[3] A school-level nonresponse adjustment was applied to the school weight; it was calculated independently for each design domain or explicit stratum.

The second step consisted of calculating a classroom weight. A classroom-level nonresponse adjustment was not necessary since in most cases a single classroom was selected per school at each grade level. When only one of the sampled classrooms in a school participated, a grade-specific school-level response adjustment was used. When one of two selected classrooms in a school (when a country chose to sample two classrooms per grade) did not participate, the classroom weight was calculated as though a single classroom had been selected in the first place. The classroom weight was calculated independently for each school and grade.

The final step consisted of calculating a student weight. A student-level nonresponse adjustment was applied to the student weight. The student weight was calculated independently for each sampled classroom.

The overall sampling weight attached to each student record is the product of the three intermediate weights: the first stage (school) weight, the second stage (classroom) weight, and the third stage (student) weight.

The overall sampling weight attached to each student in the performance assessment sub-sample is the product of the first stage weight adjusted for the subsampling of schools required, the second stage weight, and the third stage weight adjusted for the subsampling of students required at this stage.

4.2.1 First-Stage (School) Weight

The first stage weight represents the inverse of the first stage selection probability assigned to a sampled school. The TIMSS sample design required that school selection probabilities be proportional to school size, with school size being enrollment in the target grades. The basic first stage weight for the *i*th sampled school was thus defined as

$$BW_{sc}^{i} = \frac{M}{n * m_i}$$

where n is the number of sampled schools, m_i is the measure of size for the *i*th school and

$$M = \sum_{i=1}^{N} m_1$$

where N is the total number of schools in the stratum.

The basic first stage weight also incorporates a weighting factor or factors resulting from additional front-end sampling stages that were required by some TIMSS participants. This occurred when geographical regions were sampled before schools were se-

[3] For example, the United States sampled school districts as primary sampling units (PSUs), and then schools within the sampled PSUs.

lected. The calculation of such weighting factors is similar to the first stage weight since sampling geographical regions was also done with probability proportional to size (PPS). The resulting first stage weight is simply the product of the "region" weight and the first stage weight as described earlier.

In some countries, schools were selected with equal probabilities. This generally occurred when no reliable measure of school size is available. In this case, the basic first stage weight for the *i*th sampled school was defined as

$$BW_{sc}^{i} = \frac{N}{n}$$

where n is the number of sampled schools and N is the total number of schools in the stratum. It should be noted that in this case the basic weight for all sampled schools is identical.

4.2.1.1 School-Level Response Rate (Participation Rate)

A school-level response rate, weighted and unweighted, was calculated to measure the proportion of originally selected schools that ultimately participated in the assessment. Since replacement schools were used to maintain the sample size, school-level response rates have been reported both with and without the use of replacement schools. The calculation of the response rate used the following terms, derived from the data collection:

n_{ex} = number of sampled schools that should have been excluded

n_{op} = number of originally sampled schools that participated

n_{rp} = number of replacement schools that participated

n_{nr} = number of non-responding schools (neither the originally selected schools nor their replacements participating.)

Note that the following equation holds:

$$n_{ex} + n_{op} + n_{rp} + n_{nr} = n$$

The unweighted school-level response rate is defined as the ratio of originally sampled schools that participated to the total number of sampled schools minus any excluded schools. It was calculated by the following equation:

$$R_{unw}^{sc} = \frac{n_{op}}{n_{op} + n_{rp} + n_{nr}}$$

The weighted school-level response rate is defined in a similar manner. The weight assigned to the *i*th sampled school for this purpose is the sampling interval used to select it, SI_{sc}^{i} . The weighted school-level response rate, based solely on originally selected schools, is therefore the ratio of the weighted sum of originally sampled schools that participated to the weighted sum of all sampled schools less any excluded schools. It was calculated by the following equation:

$$R_{w}^{sc} = \frac{\sum_{i=1}^{n_{op}} SI_{sc}^{i}}{\sum_{i=1}^{n_{op}} SI_{sc}^{i} + \sum_{i=1}^{n_{rp}} SI_{sc}^{i} + \sum_{i=1}^{n_{nr}} SI_{sc}^{i}}$$

The weighted school-level response rate, including replacement schools, was calculated by the following equation:

$$R_{w,rp}^{sc} = \frac{\sum_{i=1}^{n_{op}} SI_{sc}^{i} + \sum_{i=1}^{n_{rp}} SI_{sc}^{i}}{\sum_{i=1}^{n_{op}} SI_{sc}^{i} + \sum_{i=1}^{n_{rp}} SI_{sc}^{i} + \sum_{i=1}^{n_{nr}} SI_{sc}^{i}}$$

4.2.1.2 School-Level Nonresponse Adjustment

First stage weights were calculated for originally sampled schools and replacement schools that participated. Any sampled schools that were no longer eligible were removed from the calculation of this nonresponse adjustment. Examples are secondary schools included in the sampling frame by mistake and schools that no longer existed. The school-level nonresponse adjustment was calculated separately for each design domain and explicit stratum.

The school-level nonresponse adjustment was calculated as follows:

$$A_{sc} = \frac{n - n_{ex}}{n_{op} + n_{rp}}$$

and the final first stage weight for the *i*th school thus becomes

$$FW_{sc}^{i} = A_{sc} * BW_{sc}^{i}$$

In the event that a sampled school had participating classrooms in only one grade when both grades were in fact present, the school-level nonresponse adjustment becomes grade-specific. Such a school was considered a participant for the grade in which students were tested but as a non-participant for the grade in which no students were tested. This led also to the calculation of separate school-level response rates by grade.

4.2.2 Second-Stage (Classroom) Weight

The second stage weight represents the inverse of the second stage selection probability assigned to a sampled classroom. Classrooms were sampled in one of two ways in Population 1 and Population 2:

- Equal probability if there was no subsampling of students within a classroom
- Probability proportional to classroom size if subsampling of students within a classroom was required

The second stage weight was calculated independently for each grade within a sampled school in Population 1 and Population 2.

A nonresponse adjustment was not required for the second stage weight. Where the classroom selected in one target grade did not participate but the sampled classroom in the other target grade did, the separate first stage nonresponse adjustments were applied by target grade.

4.2.2.1 Equal Probability Weighting

For grade g within the ith school, let $C^{g,i}$ be the total number of classrooms and c^g be the number of sampled classrooms. Using equal probability sampling, the final second stage weight assigned to all sampled classrooms from grade g in the ith school was

$$FW_{cl1}^{g,i} = \frac{C^{g,i}}{c^g}$$

As a rule, c^g takes the value 1 or 2 and remains fixed for all sampled schools. In cases where c^g has the value 2 and only one of the sampled classrooms participated, a classroom-level nonresponse adjustment was applied to the second stage weight by multiplying it by the factor 2.

4.2.2.2 Probability Proportional to Size (PPS) Weighting

For grade g within the ith school, let $k^{g,i,j}$ be the size of the jth classroom. Using PPS sampling, the final second stage weight assigned to the jth sampled classroom from grade g in the ith school was

$$FW_{cl2}^{g,i,j} = \frac{K^{g,i}}{c^g * k^{g,i,j}}$$

where c^g is the number of sampled classrooms as defined earlier and

$$K^{g,i} = \sum_{j=1}^{c^g} k^{g,i,j}$$

Again, as a rule, c^g takes the value 1 or 2 and will remain fixed for all sampled schools. In cases where c^g has the value 2, and only one of the sampled classrooms participated, a classroom-level nonresponse adjustment was applied to the second stage weight by multiplying it by the factor 2.

4.2.3 Third-Stage (Student) Weight

The third stage weight represents the inverse of the third stage selection probability attached to a sampled student. If intact classrooms were sampled as specified in Foy, Rust, and Schleicher (1996), then the basic third stage weight for the *j*th grade *g* classroom in the *i*th school was

$$BW_{st}^{g,i,j} = 1.0$$

If, on the other hand, subsampling of students was required within sampled classrooms, then the basic third stage weight for the *j*th grade *g* classroom in the *i*th school was

$$BW_{st}^{g,i,j} = \frac{k^{g,i,j}}{s^g}$$

where $k^{g,i,j}$ is the size of the *j*th grade *g* classroom in the *i*th school, as defined earlier, and s^g is the number of sampled students per sampled classroom. The latter number usually remains constant for all sampled classrooms in a grade.

4.2.3.1 Student-Level Response Rate (Participation Rate) and Adjustment

The basic third stage weight requires an adjustment to reflect the outcome of the data collection efforts. The following terms were derived from the data collection for each sampled classroom:

$s_{ex}^{g,i,j}$ = number of sampled students that should have been excluded

$s_{rs}^{g,i,j}$ = number of sampled students that participated

$s_{nr}^{g,i,j}$ = number of sampled students that did not participate.

Note that the following equation holds:

$$s_{ex}^{g,i,j} + s_{rs}^{g,i,j} + s_{nr}^{g,i,j} = s^{g,i,j}$$

where $s^{g,i,j}$ is the number of sampled students per sampled classroom. This number should be constant if subsampling of students is done within each sampled classroom and represents the classroom size, $k^{g,i,j}$, when intact classrooms are tested.

The student-level response rate, for a given classroom, was calculated as follows:

$$R^{st} = \frac{s_{rs}^{g,i,j}}{s_{rs}^{g,i,j} + s_{nr}^{g,i,j}}$$

Excluded students (i.e., those meeting the guidelines for student-level exclusions specified in Foy, Rust, and Schleicher, 1996) were not included in the calculation of the response rate.

The student-level nonresponse adjustment was calculated as follows:

$$A_{st}^{g,i,j} = \frac{s_{rs}^{g,i,j} + s_{nr}^{g,i,j}}{s_{rs}^{g,i,j}}$$

Note that the student-level nonresponse adjustment is simply the inverse of the student-level response rate. The final third stage weight for the jth grade g classroom in the ith school thus becomes

$$FW_{st}^{g,i,j} = A_{st}^{g,i,j} * BW_{st}^{g,i,j}$$

The weighted overall student-level response rate was computed as follows:

$$R_w^{st} = \frac{\sum_{i=1}^{rs} BW_{sc}^{i} * BW_{cl1}^{g,i,j} * BW_{st}^{g,i,j}}{\sum_{i=1}^{rs+nr} BW_{sc}^{i} * BW_{cl1}^{g,i,j} * BW_{st}^{g,i,j}}$$

where the numerator is the summation of the basic weights over all responding students, and the denominator is the summation of the basic weights over all responding and nonresponding students. Weighted student response rates were reported separately by grade in the TIMSS international reports.

4.2.4 Overall Sampling Weights

The overall sampling weight is simply the product of the final first stage weight, the appropriate final second stage weight, and the appropriate final third stage weight. If intact classrooms were tested, then the overall sampling weight was

$$W^{g,i,j} = FW_{sc}^{i} * FW_{sc}^{g,i,j} * FW_{st}^{g,i,j}$$

If subsampling within classrooms was done, then the overall sampling weight was

$$W^{g,i,j} = FW_{sc}^{i} * FW_{cl2}^{g,i,j} * FW_{st}^{g,i,j}$$

It is important to note that sampling weights varied by school, grade, and classroom. However, students within the same classroom have the same sampling weights.

The use of sampling weights is critical to obtaining proper survey estimates when sampling techniques other than simple random sampling are used. TIMSS has produced a sampling weight for each student sampled for the TIMSS main (written) assessment and subsampled for the performance assessment. Secondary analysts using the TIMSS data will need to be aware of this and use the proper weights when conducting analyses and reporting results.

REFERENCES

Foy, P., Rust, K., and Schleicher, A. (1996). Sample design. In M.O. Martin and D.L. Kelly (Eds.), *TIMSS technical report, volume I: Design and development.* Chestnut Hill, MA: Boston College.

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

Schleicher, A. and Siniscalco, M.T. (1996). Field operations. In M.O. Martin and D.L. Kelly (Eds.), *TIMSS technical report, volume I: Design and development.* Chestnut Hill, MA: Boston College.

5 Estimation of Sampling Variability, Design Effects, and Effective Sample Sizes

Eugenio J. Gonzalez
Boston College

Pierre Foy
Statistics Canada

5.1 OVERVIEW

In order to derive parameter estimates of the distribution of student achievement in each country that were both accurate and cost-effective, TIMSS made use of probability sampling techniques to sample students from national student populations.[1] The statistics computed from these national probability samples were used to estimate population parameters. Because there is some uncertainty involved in generalizing from samples to populations, the important statistics in the TIMSS international reports (Beaton, A.E. et al., 1996; Beaton, A.E. et al., 1996; Martin, M.O. et al., 1997; Mullis, I.V.S. et al., 1997) are presented together with their standard errors, which are a measure of this uncertainty.

The TIMSS sampling design applies stratified multistage cluster-sampling techniques to the problem of selecting efficient and accurate samples of students while working with schools and classes. Such complex designs capitalize on the structure of the student population (i.e., students grouped in classes within schools) to derive student samples that permit efficient and economical data collection. However, complex sampling designs make the task of computing standard errors to quantify sampling variability more difficult.

When, as in TIMSS, the sampling design involves multistage cluster sampling, there are several options for the estimation of sampling error that avoid the assumption of simple random sampling (see Wolter, 1985). The jackknife repeated replication technique (JRR) was chosen for estimating sampling errors in TIMSS because it is computationally straightforward, and provides approximately unbiased estimates of the sampling errors of means, totals, and percentages in complex sample designs.

The particular variation on the JRR technique used in TIMSS is described in Johnson and Rust (1992). This method assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sample design, and each pair regarded as members of a pseudo-stratum for variance estimation purposes. Note that when using the JRR technique for the estimation of sampling variability, the approach will appropriately reflect the combined effect of the between- and within-PSU contributions to the sampling variance. The general use of the JRR entails systematically assigning pairs of schools to sampling zones, and the random selection of one of these schools to have its contribution doubled, and the other zeroed, so as to construct a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for all of

[1] See Foy, Rust, and Schleicher (1996) for details of the TIMSS sampling design.

the original sample, and once more for each of the pseudo-replicate samples. The variation between the estimates from each of the replicate samples and the original sample estimate is the jackknife estimate of the sampling error of the statistic. Specific applications of the jackknife method are also discussed in the chapters describing the reporting of student achievement in subject-matter content areas (Chapter 9) and the Test-Curriculum Matching Analysis (Chapter 10).

Although the jackknife was the standard method of computing sampling errors in TIMSS, where standard errors were required for medians the balanced repeated replication (BRR) method was used instead. BRR was chosen over the JRR method in this instance because it produces asymptotically more consistent estimates for order statistics such as medians and percentiles.

5.2 CONSTRUCTION OF SAMPLING ZONES FOR SAMPLING VARIANCE ESTIMATION

An important step in applying the JRR and the BRR techniques to the estimation of sampling variability consists of assigning the schools to implicit strata, also known as sampling zones. Since the sample design called for 150 schools, a maximum of 75 zones was expected within each country, with two schools per zone. These zones were constructed by sequentially pairing the sampled schools. Because schools were generally sorted by a set of implicit stratification variables, the resulting assignment to sampling zones takes advantage of any benefit due to this implicit stratification. In countries where more than 150 schools were sampled, it was sometimes necessary to combine two schools for variance estimation purposes before assigning them to a sampling zone.

Zones were constructed within design domains, or explicit strata. In cases where there was an odd number of schools in an explicit stratum, either by design or because of school-level nonresponse, the students in the remaining school were randomly divided to make up two "quasi" schools for the purposes of calculating the jackknife standard error. Each zone then consisted of a pair of schools or "quasi" schools. Table 5.1 shows the number of sampling zones by grade in each country.

5.3 COMPUTING SAMPLING VARIANCE USING THE JRR METHOD

The JRR algorithm used in TIMSS assumes that there are *H* sampling zones within each country, each one containing two sampled schools selected independently. When computing a statistic "*t*" from the sample for a country, the formula for the JRR variance estimate of the statistic *t* is then given by the following equation:

$$Var_{jrr}(t) = \sum_{h=1}^{H} [t(J_h) - t(S)]^2$$

where *H* is the number of pairs in the sample for the country. The term $t(S)$ corresponds to the statistic computed for the whole sample (computed with any specific weights that may have been used to compensate for the unequal probability of selection of the different elements in the sample or any other post-stratification weight). The element $t(J_h)$ denotes the same statistic using the *h*th jackknife replicate, computed for all cases

Table 5.1 Sampling Zones by Grade Level*

Country	Third Grade	Fourth Grade	Seventh Grade	Eighth Grade
Australia	74	74	74	74
Austria	68	68	65	66
Belgium (Fl)	-	-	71	71
Belgium (Fr)	-	-	60	60
Bulgaria	-	-	52	58
Canada	75	75	75	75
Colombia	-	-	71	71
Cyprus	74	74	55	55
Czech Republic	73	73	75	75
Denmark	-	-	75	75
England	67	67	64	64
France	-	-	67	68
Germany	-	-	69	69
Greece	75	75	75	75
Hong Kong	62	62	43	43
Hungary	75	75	75	75
Iceland	75	75	75	75
Iran, Islamic Rep.	75	75	75	75
Ireland	73	73	66	66
Israel	-	44	-	23
Japan	74	74	75	75
Korea	75	75	75	75
Kuwait	-	75	-	36
Latvia (LSS)	59	59	64	64
Lithuania	-	-	73	73
Netherlands	52	52	48	48
New Zealand	75	75	75	75
Norway	70	70	72	74
Portugal	72	72	71	71
Romania	-	-	72	72
Russian Federation	-	-	41	41
Scotland	65	65	64	64
Singapore	75	75	69	69
Slovak Republic	-	-	73	73
Slovenia	61	61	61	61
South Africa	-	-	66	66
Spain	-	-	75	75
Sweden	-	-	75	60
Switzerland	-	-	75	75
Thailand	75	75	74	74
United States	59	59	55	55

A dash (-) means the country did not participate at this grade level

* Third, fourth, seventh, and eighth grades in most countries.

except those in the *h*th stratum of the sample, removing all cases associated with one of the randomly selected units of the pair within the *h*th stratum, and including, twice, the elements associated with the other unit in the *h*th stratum. In practice, this is effectively accomplished by recoding to zero the weights for the cases of the element of the pair to be excluded from the replication, and multiplying by two the weights of the remaining element within the *h*th pair.

The computation of the JRR variance estimate for any statistic from the TIMSS database requires the computation of any statistic up to 76 times for any given country: once to obtain the statistic for the full sample, and up to 75 times to obtain the statistics for each of the jackknife replicates (J_h). The number of times a statistic needs to be computed for a given country depends on the number of implicit strata or sampling zones defined for that country.

Doubling and zeroing the weights of the selected units within the sampling zones is accomplished effectively with the creation of replicate weights which are then used in the calculations. Gonzalez and Smith (1997) provide examples of how this approach allows standard statistical software such as SAS or SPSS to be used to compute JRR estimates of sampling variability in TIMSS. The replicate weight approach requires the user to temporarily create a new set of weights for each pseudo-replicate sample. Each replicate weight is equal to *k* times the overall sampling weight, where *k* can take values of zero, one or two depending on whether or not the case is to be removed from the computation, left as it is, or have its weight doubled. The value of *k* for an individual student record for a given replicate depends on the assignment of the record to the specific PSU and zone.

Within each zone the members of the pair of schools are assigned an indicator (u_i), coded randomly to 1 or 0 so that one of the members of each pair had values of 1 on the variable u_i, and the remaining member a value of 0. This indicator determines whether the weights for the elements in the school in this zone are to be doubled or zeroed. The replicate weight ($W_h^{g,i,j}$) for the elements in a school assigned to zone *h* is computed as the product of k_h times their overall sampling weight, where k_h can take values of zero, one, or two depending on whether the school is to be omitted, be included with its usual weight, or have its weight doubled for the computation of the statistic of interest. In TIMSS, the replicate weights are not permanent variables, but are created temporarily by the sampling variance estimation program as a useful computing device.

When creating the replicate weights the following procedure was followed:

Each sampled student was assigned a vector of 75 weights or $W_h^{g,i,j}$, where *h* takes values from 1 to 75.

The value of $W_0^{g,i,j}$ is the overall sampling weight which is simply the product of the final school weight, the appropriate final classroom weight, and the appropriate final student weight as described in chapter 4.

The replicate weights for a single case were then computed as:

$$W_h^{g,i,j} = W_0^{g,i,j} * k_{hi} ,$$

where the variable k_h for an individual *i* takes the value $k_{hi} = 2*u_i$ if the record belongs to zone *h*, and $k_{hi} = 1$ otherwise.

In TIMSS, a total of 75 replicate weights were computed for each country regardless of the number of actual zones within the country. If a country had fewer than 75 zones, then the replicate weights W_h, where *h* was greater than the number of zones within the country, were each the same as the overall sampling weight. Although this involved some redundant computation, having 75 replicate weights for each country has no effect on the size of the error variance computed using the jackknife formula, but facilitated the computation of standard errors for a number of countries at one time.

Figure 5.1 shows example SAS and SPSS computer code used to compute standard errors in TIMSS. Further examples are given in Gonzalez and Smith (1997). Although standard errors presented in the international reports were computed using SAS programs developed at the International Study Center, they were also verified against results produced by the WesVarPC software (Westat, 1997). Results were compared with each other for accuracy.[2]

Figure 5.1 Computer Code in SAS and SPSS to Generate JRR Replicate Weights

```
SAS Computer Code
data a;
  set datafile ;
array rwt  rwt1 - rwt75 ;         * Replicate Weights  ;
do i=1 to 75;
     if  jkzone  <>i                then rwt(i) = weight * 1;
     if (jkzone  = i & jkindic = 1) then rwt(i) = weight * 2;
     if (jkzone  = i & jkindic = 0) then rwt(i) = weight * 0;
end;

SPSS Computer Code
vector rwgt(75).
loop #i = 1 to 75.
if (jkzone = #i and jkindic = 0)  rwgt(#i) = weight * 0.
if (jkzone = #i and jkindic = 1)  rwgt(#i) = weight * 2.
if (jkzone <>#i               )  rwgt(#i) = weight * 1.
end loop.
```

5.4 COMPUTING SAMPLING VARIANCE USING THE BRR METHOD

Like the JRR method, balanced repeated replication (BRR) uses the variation between PSUs to estimate the sampling variation of a statistic. BRR forms a series of replicate half-samples by randomly selecting one of the pair of PSUs in each sampling zone. The weights of the selected PSUs are doubled to compensate for the omitted PSUs. When a statistic is computed independently from each of the replicate half-samples, the variation in the results may be used to estimate the sampling variance of that statistic. When computing a statistic t from the sample, the formula for the BRR variance estimate of the statistic t is given by the equation:

$$Var_{brr}(t) = \frac{\sum_{g=1}^{G} \text{t}[t(B_g) - t(S)]^2}{G}$$

[2] Minor differences were occasionally found between the results obtained with WesVar and those obtained with software developed in-house. However, these differences were in all cases due to the fact that the two programs did not always choose the same PSUs in forming jackknife replicates. When identical jackknife replicates were used for both programs, the results were identical.

where G is the number of replicate half-samples formed from the entire sample. The term $t(S)$ corresponds to the statistic computed for the whole sample weighted to compensate for unequal selection probabilities and post-stratification adjustments. The element $t(B_g)$ denotes the same statistic using the gth replicate half-sample, formed by including only half the units in the original sample.

Although each replicate half-sample contains only one unit from each of the H strata, there are $2H$ possible half-samples for a given sample. When the number of strata, H, is large, the number of possible half-samples becomes enormous (3.78×10^{22} in the case of TIMSS with 75 replicates), and the computation of estimates of sampling variability using all such half-samples is no longer feasible. However, by selecting a subsample of G orthogonally balanced half-replicates it is possible to obtain an unbiased estimate of the variance that would have been obtained if all possible replicate half-samples had been used (see Wolter, 1985). This is true whenever G is an integral multiple of 4 that is greater than H, where H is the number of strata in the sample. The selection of the G half-samples is facilitated by the use of Hadamard matrices. For the purpose of computing the standard errors of medians for selected age groups in TIMSS, a Hadamard matrix of order 76 was used. The WesVarPC (Westat, 1997) software was used to construct the replicate half-samples in TIMSS, although the BRR sampling errors themselves were computed using software developed at the TIMSS International Study Center.

5.5 DESIGN EFFECTS AND EFFECTIVE SAMPLE SIZES

Complex survey samples such as those in TIMSS typically have sampling errors much larger than a simple random sample of the same size. This is because the elements of the clusters that are the building blocks of complex samples (in TIMSS the elements are students grouped in classes within schools) usually resemble each other more than they do members of the population in general. Consequently, a sample of size n drawn using simple random sampling from a population will usually be more efficient (i.e., have smaller sampling errors) than a sample of the same size drawn by means of a sample of pre-existing clusters in the population. The degree to which members of a cluster resemble each other more than they do elements of the population in general on some criterion variable may be measured by the intra-class correlation coefficient (Kish, 1965). When the intra-class correlation for a variable in a population is large, it may be necessary to select a much larger sample using cluster-sample techniques than would be necessary using simple random sampling methods.

Although the design efficiency of a multistage cluster sample is generally less than that of a simple random sample of the same size, multistage samples have other advantages in terms of economy and operational efficiency that make them the method of choice for surveys of student populations such as TIMSS. One way to quantify the reduction in design efficiency is through the design effect (Kish, 1965). The design effect for a variable is the ratio of two estimates of the sampling variance for a particular sample statistic: one computed using a technique such as the jackknife that takes all components of variance in the sampling design into account, and the other computed using the simple random sampling formula. The design effect is specific to the statistic and the variable for which it is computed. Since in TIMSS the technique for estimating sampling variance for means and percentages was the JRR, the design effect for these statistics was computed as the ratio of the JRR variance estimate to the variance estimate computed under the assumptions of simple random sampling. The design effect was computed as follows:

$$DEff(t) = \frac{Var_{jrr}(t)}{Var_{srs}(t)}$$

where $Var_{jrr}(t)$ is the sampling variance computed using the JRR method, and $Var_{srs}(t)$ is the variance computed under the assumptions of simple random sampling. When computing the design effect for the proportion of students (p) responding correctly to an item,[3] the sampling variance of the statistic ($Var_{srs}(P)$) based on a sample with n cases, was computed as:

$$Var_{srs}(P) = \frac{P * (1 - P)}{n}$$

When computing the design effect of a mean ($\bar{x}$), the sampling variance of the statistic ($Var_{srs}(\bar{x})$) based on a simple random sample with n cases was computed as:

$$Var_{srs}(\bar{x}) = \frac{Var_x}{n}$$

Another, related, measure of the design efficiency is the effective sample size. The effective sample size is the ratio of the actual sample size to the design effect. It is the number of sampling elements that would be required in a simple random sample to provide the same precision obtained with the actual complex sampling design. The effective sample size is computed as:

$$EffN(t) = \frac{N}{DEff(t)}$$

The TIMSS standard for sampling precision required that all student samples have an effective sample size of at least 400 for the main criterion variables (Foy, Rust, and Schleicher, 1996). Note that these requirements were for the entire populations (i.e., grades three and four combined for Population 1, and grades seven and eight for Population 2). Design effects and effective sample sizes for the mean mathematics and science achievement scores by population are presented in Tables 5.2 through 5.13. Design effects and effective sample sizes by grade and by grade and gender are included in Appendix C.

[3] Proportion correct is defined here as the proportion of students obtaining the maximum score on the item.

Table 5.2 Design Effects and Effective Sample Sizes for Third and Fourth Grades* (Combined) - Mathematics Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	11248	516	9247.0	3.4	0.9	14.33	785
Austria	5171	524	7837.9	3.6	1.2	8.74	591
Canada	16002	502	7548.0	2.5	0.7	12.99	1232
Cyprus	6684	467	8028.1	2.5	1.1	5.20	1285
Czech Republic	6524	533	8376.5	2.8	1.1	6.10	1069
England	6182	485	8766.2	2.5	1.2	4.28	1445
Greece	6008	461	8703.9	3.4	1.2	8.02	749
Hong Kong	8807	556	6743.9	3.3	0.9	14.29	616
Hungary	6044	512	9176.7	3.4	1.2	7.63	792
Iceland	3507	442	5888.7	2.6	1.3	4.11	854
Iran, Islamic Rep.	6746	404	5179.4	3.4	0.9	15.44	437
Ireland	5762	513	8301.7	3.2	1.2	7.31	789
Israel	2351	531	7151.4	3.5	1.7	4.13	569
Japan	8612	568	7006.7	1.6	0.9	3.08	2795
Korea	5589	586	5812.0	1.9	1.0	3.32	1682
Kuwait	4318	400	4458.9	2.8	1.0	7.42	582
Latvia (LSS)	4270	498	7860.5	3.9	1.4	8.19	521
Netherlands	5314	535	6348.6	2.9	1.1	7.12	746
New Zealand	4925	470	8295.9	4.0	1.3	9.29	530
Norway	4476	462	6931.8	2.6	1.2	4.44	1009
Portugal	5503	452	7466.2	3.1	1.2	7.13	772
Scotland	6433	489	8128.2	3.2	1.1	8.20	784
Singapore	14169	588	11743.3	4.1	0.9	20.47	692
Slovenia	5087	520	7439.5	2.8	1.2	5.41	941
Thailand	5862	467	5482.5	4.4	1.0	20.46	287
United States	11115	512	8022.6	2.8	0.8	11.00	1010

*Third and fourth grades in most countries.

Table 5.3 Design Effects and Effective Sample Sizes for Third Grade* Mathematics Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	4741	484	8114.9	4.0	1.3	9.55	497
Austria	2526	487	6877.0	5.3	1.6	10.50	241
Canada	7594	469	6111.8	2.7	0.9	8.75	868
Cyprus	3308	430	5984.4	2.8	1.3	4.23	782
Czech Republic	3256	497	6853.4	3.3	1.5	5.23	622
England	3056	456	7634.3	3.0	1.6	3.67	833
Greece	2955	428	7254.6	4.0	1.6	6.36	464
Hong Kong	4396	524	5250.2	3.0	1.1	7.74	568
Hungary	3038	476	7980.5	4.2	1.6	6.78	448
Iceland	1698	410	4519.7	2.8	1.6	2.93	579
Iran, Islamic Rep.	3361	378	4302.7	3.5	1.1	9.77	344
Ireland	2889	476	6558.0	3.6	1.5	5.71	506
Japan	4306	538	5671.4	1.5	1.1	1.76	2452
Korea	2777	561	4922.8	2.3	1.3	2.95	940
Latvia (LSS)	2054	463	6544.7	4.3	1.8	5.72	359
Netherlands	2790	493	4209.3	2.7	1.2	4.90	569
New Zealand	2504	440	6771.7	4.0	1.6	6.01	417
Norway	2219	421	5116.7	3.1	1.5	4.11	540
Portugal	2650	425	7293.0	3.8	1.7	5.24	506
Scotland	3132	458	6321.9	3.4	1.4	5.60	559
Singapore	7030	552	9984.8	4.8	1.2	16.22	433
Slovenia	2521	488	5980.9	2.9	1.5	3.59	701
Thailand	2870	444	5075.9	5.1	1.3	14.61	196
United States	3819	480	6709.8	3.4	1.3	6.56	582

*Third grade in most countries.

Table 5.4 Design Effects and Effective Sample Sizes for Fourth Grade* Mathematics Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	6507	547	8399.9	3.2	1.1	7.93	820
Austria	2645	559	6212.5	3.1	1.5	4.05	653
Canada	8408	532	7000.5	3.3	0.9	13.11	641
Cyprus	3376	502	7461.4	3.1	1.5	4.43	761
Czech Republic	3268	567	7446.4	3.3	1.5	4.68	698
England	3126	513	8316.7	3.2	1.6	3.91	800
Greece	3053	492	8088.6	4.4	1.6	7.18	425
Hong Kong	4411	587	6240.4	4.3	1.2	13.11	336
Hungary	3006	548	7762.9	3.7	1.6	5.38	559
Iceland	1809	474	5232.1	2.7	1.7	2.50	725
Iran, Islamic Rep.	3385	429	4773.5	4.0	1.2	11.15	304
Ireland	2873	550	7283.4	3.4	1.6	4.68	614
Israel	2351	531	7151.4	3.5	1.7	4.13	569
Japan	4306	597	6590.6	2.1	1.2	2.80	1540
Korea	2812	611	5457.7	2.1	1.4	2.31	1219
Kuwait	4318	400	4458.9	2.8	1.0	7.42	582
Latvia (LSS)	2216	525	7199.9	4.8	1.8	7.15	310
Netherlands	2524	577	4974.4	3.4	1.4	5.74	440
New Zealand	2421	499	8022.9	4.3	1.8	5.60	432
Norway	2257	502	5497.9	3.0	1.6	3.61	624
Portugal	2853	475	6450.9	3.5	1.5	5.49	520
Scotland	3301	520	7994.1	3.9	1.6	6.25	528
Singapore	7139	625	10854.0	5.3	1.2	18.54	385
Slovenia	2566	552	6797.1	3.2	1.6	3.84	669
Thailand	2992	490	4834.7	4.7	1.3	13.59	220
United States	7296	545	7243.8	3.0	1.0	9.23	790

*Fourth grade in most countries.

Table 5.5 Design Effects and Effective Sample Sizes for Third and Fourth Grades* (Combined) Science Mean Scale Score - Population 1

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	11248	537	9809.8	3.3	0.9	12.33	913
Austria	5171	536	7904.7	3.4	1.2	7.35	704
Canada	16002	521	8434.2	2.2	0.7	9.41	1700
Cyprus	6684	445	6461.3	2.4	1.0	6.07	1101
Czech Republic	6524	526	7859.0	2.8	1.1	6.36	1025
England	6182	525	10343.8	2.5	1.3	3.75	1647
Greece	6008	472	7503.3	3.3	1.1	8.75	687
Hong Kong	8807	508	6399.1	3.0	0.9	12.06	730
Hungary	6044	498	8322.2	3.3	1.2	7.94	761
Iceland	3507	470	8176.1	3.0	1.5	3.86	908
Iran, Islamic Rep.	6746	387	6567.5	3.6	1.0	13.42	503
Ireland	5762	510	8360.8	3.3	1.2	7.53	765
Israel	2351	505	7450.2	3.6	1.8	4.19	561
Japan	8612	548	5956.0	1.4	0.8	2.64	3263
Korea	5589	575	5353.3	1.7	1.0	3.16	1767
Kuwait	4318	401	7250.5	3.1	1.3	5.86	737
Latvia (LSS)	4270	491	7474.7	4.1	1.3	9.47	451
Netherlands	5314	528	5008.0	2.8	1.0	8.12	654
New Zealand	4925	503	10495.7	4.8	1.5	10.65	463
Norway	4476	491	9347.5	2.8	1.4	3.82	1171
Portugal	5503	453	8861.4	3.5	1.3	7.43	740
Scotland	6433	510	9546.3	3.8	1.2	9.59	671
Singapore	14169	517	10473.8	4.1	0.9	23.01	616
Slovenia	5087	516	6797.7	2.8	1.2	5.71	891
Thailand	5862	452	5923.1	5.2	1.0	27.15	216
United States	11115	538	9646.5	2.8	0.9	9.34	1190

*Third and fourth grades in most countries.

Table 5.6 Design Effects and Effective Sample Sizes for Third Grade* Science Mean Scale Score - Population 1

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	4741	510	9561.3	4.4	1.4	9.54	497
Austria	2526	505	7667.5	4.6	1.7	7.06	358
Canada	7594	490	7766.0	2.5	1.0	6.31	1203
Cyprus	3308	415	5344.5	2.5	1.3	3.91	846
Czech Republic	3256	494	7156.4	3.4	1.5	5.35	609
England	3056	499	10118.3	3.5	1.8	3.63	842
Greece	2955	446	6800.1	3.9	1.5	6.70	441
Hong Kong	4396	482	5408.7	3.3	1.1	8.72	504
Hungary	3038	464	7886.0	4.1	1.6	6.35	478
Iceland	1698	435	6738.7	3.3	2.0	2.70	630
Iran, Islamic Rep.	3361	356	5772.2	4.2	1.3	10.14	331
Ireland	2889	479	7703.0	3.7	1.6	5.03	574
Japan	4306	522	5272.6	1.6	1.1	2.00	2156
Korea	2777	553	5103.3	2.4	1.4	3.14	885
Latvia (LSS)	2054	465	6817.4	4.5	1.8	6.20	331
Netherlands	2790	499	4022.8	3.2	1.2	7.01	398
New Zealand	2504	473	9913.8	5.2	2.0	6.87	365
Norway	2219	450	8069.1	3.9	1.9	4.12	538
Portugal	2650	423	9146.9	4.3	1.9	5.35	496
Scotland	3132	484	9021.1	4.2	1.7	6.19	506
Singapore	7030	488	9762.8	5.0	1.2	18.34	383
Slovenia	2521	487	6091.0	2.8	1.6	3.23	780
Thailand	2870	433	6010.7	6.6	1.4	20.63	139
United States	3819	511	8796.1	3.2	1.5	4.42	863

*Third grade in most countries.

Table 5.7 Design Effects and Effective Sample Sizes for Fourth Grade* Science Mean Scale Score - Population 1

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	6507	563	8699.4	3.0	1.2	6.78	960
Austria	2645	565	6370.7	3.3	1.6	4.43	597
Canada	8408	549	7381.8	3.0	0.9	10.14	829
Cyprus	3376	475	5730.1	3.3	1.3	6.44	524
Czech Republic	3268	557	6598.4	3.1	1.4	4.77	685
England	3126	551	9207.8	3.3	1.7	3.65	857
Greece	3053	497	6888.4	4.1	1.5	7.30	418
Hong Kong	4411	533	6046.9	3.7	1.2	10.03	440
Hungary	3006	532	6505.4	3.4	1.5	5.47	550
Iceland	1809	505	7207.9	3.3	2.0	2.74	660
Iran, Islamic Rep.	3385	416	5546.6	3.9	1.3	9.40	360
Ireland	2873	539	7205.7	3.3	1.6	4.41	651
Israel	2351	505	7450.2	3.6	1.8	4.19	561
Japan	4306	574	5296.3	1.8	1.1	2.53	1703
Korea	2812	597	4639.3	1.9	1.3	2.10	1342
Kuwait	4318	401	7250.5	3.1	1.3	5.86	737
Latvia (LSS)	2216	512	7022.1	4.9	1.8	7.65	290
Netherlands	2524	557	4319.8	3.1	1.3	5.45	463
New Zealand	2421	531	9418.7	4.9	2.0	6.14	394
Norway	2257	530	7432.4	3.6	1.8	3.85	586
Portugal	2853	480	7122.1	4.0	1.6	6.46	441
Scotland	3301	536	8731.0	4.2	1.6	6.58	501
Singapore	7139	547	9445.0	5.0	1.2	19.12	373
Slovenia	2566	546	5780.5	3.3	1.5	4.96	517
Thailand	2992	473	5012.2	4.9	1.3	14.26	210
United States	7296	565	9028.6	3.1	1.1	7.65	954

*Fourth grade in most countries.

Table 5.8 Design Effects and Effective Sample Sizes for Seventh and Eighth Grades* (Combined) Mathematics Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	12,852	514	9,287.0	3.5	0.9	17.27	744
Austria	5,786	524	8,080.8	2.5	1.2	4.50	1,285
Belgium (Fl)	5,662	562	7,270.7	4.0	1.1	12.16	465
Belgium (Fr)	4,883	518	6,907.2	3.0	1.2	6.31	774
Bulgaria	3,771	527	11,612.4	4.6	1.8	6.97	541
Canada	16,581	511	7,196.6	1.9	0.7	8.42	1,970
Colombia	5,304	376	4,103.4	2.8	0.9	10.25	518
Cyprus	5,852	459	7,394.3	1.4	1.1	1.55	3,770
Czech Republic	6,672	544	8,778.7	3.8	1.1	11.00	606
Denmark	4,370	485	6,911.4	1.9	1.3	2.32	1,885
England	3,579	491	8,587.4	2.4	1.5	2.40	1,493
France	6,014	514	6,136.6	2.4	1.0	5.51	1,091
Germany	5,763	497	7,780.5	4.1	1.2	12.41	464
Greece	7,921	461	8,019.5	2.6	1.0	6.91	1,146
Hong Kong	6,752	576	10,163.8	6.8	1.2	30.29	223
Hungary	5,978	519	8,745.0	3.0	1.2	6.34	943
Iceland	3,730	473	5,376.0	2.6	1.2	4.60	811
Iran, Islamic Rep.	7,429	414	3,551.4	1.8	0.7	6.59	1,127
Ireland	6,203	513	8,239.7	3.4	1.2	8.59	722
Israel	1,415	522	8,463.5	6.2	2.4	6.36	222
Japan	10,271	588	10,102.3	1.7	1.0	2.88	3,567
Korea	5,827	592	11,622.5	2.0	1.4	2.06	2,827
Kuwait	1,655	392	3,325.4	2.5	1.4	3.15	526
Latvia (LSS)	4,976	477	6,531.0	2.4	1.1	4.55	1,095
Lithuania	5,056	454	6,656.9	2.8	1.1	5.82	869
Netherlands	4,084	529	7,257.6	4.6	1.3	12.14	336
New Zealand	6,867	490	8,180.3	2.9	1.1	7.28	943
Norway	5,736	482	6,855.2	1.9	1.1	3.16	1,818
Portugal	6,753	438	4,058.8	2.0	0.8	6.71	1,007
Romania	7,471	468	7,709.6	3.3	1.0	10.49	712
Russian Federation	8,160	518	8,399.0	3.9	1.0	14.71	555
Scotland	5,776	481	7,481.5	4.1	1.1	13.19	438
Singapore	8,285	622	8,682.6	4.8	1.0	22.21	373
Slovak Republic	7,101	527	8,230.6	2.7	1.1	6.37	1,115
Slovenia	5,606	519	7,642.8	2.4	1.2	4.40	1,274
South Africa	9,792	351	4,167.8	3.1	0.7	23.21	422
Spain	7,596	468	5,504.4	1.9	0.9	4.83	1,574
Sweden	6,906	498	7,024.7	2.0	1.0	3.82	1,808
Switzerland	8,940	526	7,097.2	2.1	0.9	5.39	1,658
Thailand	11,643	508	6,952.1	4.9	0.8	40.70	286
United States	10,973	488	8,261.9	4.3	0.9	24.83	442

*Seventh and eighth grades in most countries.

Table 5.9 Design Effects and Effective Sample Sizes for Seventh Grade* Mathematics Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	5,599	498	8,437.6	3.8	1.2	9.59	584
Austria	3,013	509	7,260.4	3.0	1.6	3.70	815
Belgium (Fl)	2,768	558	5,877.2	3.5	1.5	5.91	469
Belgium (Fr)	2,292	507	6,085.4	3.5	1.6	4.73	484
Bulgaria	1,798	514	10,670.8	7.5	2.4	9.39	191
Canada	8,219	494	6,396.9	2.2	0.9	6.30	1,305
Colombia	2,655	369	3,967.1	2.7	1.2	4.89	543
Cyprus	2,929	446	6,747.6	1.9	1.5	1.61	1,823
Czech Republic	3,345	523	7,972.0	4.9	1.5	10.15	329
Denmark	2,073	465	6,030.0	2.1	1.7	1.56	1,330
England	1,803	476	8,084.6	3.7	2.1	2.98	606
France	3,016	492	5,460.0	3.1	1.3	5.46	552
Germany	2,893	484	7,237.0	4.1	1.6	6.77	428
Greece	3,931	440	7,289.8	2.8	1.4	4.34	905
Hong Kong	3,413	564	9,841.0	7.8	1.7	21.34	160
Hungary	3,066	502	8,232.0	3.7	1.6	5.01	613
Iceland	1,957	459	4,594.9	2.6	1.5	2.84	689
Iran, Islamic Rep.	3,735	401	3,232.4	2.0	0.9	4.59	815
Ireland	3,127	500	7,537.8	4.1	1.6	7.03	445
Japan	5,130	571	9,220.1	1.9	1.3	2.05	2,507
Korea	2,907	577	10,930.5	2.5	1.9	1.72	1,689
Latvia (LSS)	2,567	462	5,859.6	2.8	1.5	3.45	743
Lithuania	2,531	428	5,657.0	3.2	1.5	4.45	568
Netherlands	2,097	516	6,231.6	4.1	1.7	5.66	370
New Zealand	3,184	472	7,540.2	3.8	1.5	6.08	523
Norway	2,469	461	5,779.8	2.8	1.5	3.42	721
Portugal	3,362	423	3,569.6	2.2	1.0	4.62	727
Romania	3,746	454	7,091.3	3.4	1.4	5.99	625
Russian Federation	4,138	501	7,781.8	4.0	1.4	8.30	499
Scotland	2,913	463	6,670.6	3.7	1.5	6.06	480
Singapore	3,641	601	8,694.2	6.3	1.5	16.88	216
Slovak Republic	3,600	508	7,240.7	3.4	1.4	5.66	636
Slovenia	2,898	498	6,715.2	3.0	1.5	3.77	769
South Africa	5,301	348	4,023.3	3.8	0.9	19.06	278
Spain	3,741	448	4,836.5	2.2	1.1	3.87	968
Sweden	2,831	477	5,911.6	2.5	1.4	2.93	965
Switzerland	4,085	506	5,684.3	2.3	1.2	3.79	1,078
Thailand	5,810	495	6,178.2	4.9	1.0	22.14	262
United States	3,886	476	7,966.0	5.5	1.4	14.73	264

*Seventh grade in most countries.

Table 5.10 Design Effects and Effective Sample Sizes for Eighth Grade* Mathematics Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	7,253	530	9,651.1	4.0	1.2	12.18	596
Austria	2,773	539	8,462.9	3.0	1.7	3.05	910
Belgium (Fl)	2,894	565	8,435.6	5.7	1.7	11.00	263
Belgium (Fr)	2,591	526	7,431.9	3.4	1.7	4.03	644
Bulgaria	1,973	540	12,187.6	6.3	2.5	6.42	308
Canada	8,362	527	7,444.2	2.4	0.9	6.51	1,285
Colombia	2,649	385	4,120.9	3.4	1.2	7.64	347
Cyprus	2,923	474	7,684.9	1.9	1.6	1.36	2,155
Czech Republic	3,327	564	8,771.2	4.9	1.6	9.21	361
Denmark	2,297	502	7,007.4	2.8	1.7	2.61	879
England	1,776	506	8,641.6	2.6	2.2	1.44	1,234
France	2,998	538	5,781.2	2.9	1.4	4.33	693
Germany	2,870	509	8,025.5	4.5	1.7	7.22	398
Greece	3,990	484	7,798.5	3.1	1.4	4.81	829
Hong Kong	3,339	588	10,188.4	6.5	1.7	13.94	239
Hungary	2,912	537	8,641.1	3.2	1.7	3.52	826
Iceland	1,773	487	5,780.1	4.5	1.8	6.31	281
Iran, Islamic Rep.	3,694	428	3,513.5	2.2	1.0	4.88	758
Ireland	3,076	527	8,564.1	5.1	1.7	9.47	325
Israel	1,415	522	8,463.5	6.2	2.4	6.36	222
Japan	5,141	605	10,388.5	1.9	1.4	1.74	2,951
Korea	2,920	607	11,848.0	2.4	2.0	1.40	2,091
Kuwait	1,655	392	3,325.4	2.5	1.4	3.15	526
Latvia (LSS)	2,409	493	6,743.4	3.1	1.7	3.50	688
Lithuania	2,525	477	6,424.9	3.5	1.6	4.91	515
Netherlands	1,987	541	7,897.7	6.7	2.0	11.15	178
New Zealand	3,683	508	8,153.3	4.5	1.5	9.08	406
Norway	3,267	503	7,033.6	2.2	1.5	2.20	1,487
Portugal	3,391	454	4,075.6	2.5	1.1	5.15	659
Romania	3,725	482	7,958.2	4.0	1.5	7.63	488
Russian Federation	4,022	535	8,446.6	5.3	1.4	13.48	298
Scotland	2,863	498	7,639.1	5.5	1.6	11.25	254
Singapore	4,644	643	7,782.4	4.9	1.3	14.39	323
Slovak Republic	3,501	547	8,474.6	3.3	1.6	4.51	776
Slovenia	2,708	541	7,700.1	3.1	1.7	3.36	806
South Africa	4,491	354	4,270.1	4.4	1.0	20.79	216
Spain	3,855	487	5,397.9	2.0	1.2	2.87	1,341
Sweden	4,075	519	7,278.7	3.0	1.3	4.90	832
Switzerland	4,855	545	7,670.4	2.8	1.3	4.88	996
Thailand	5,833	522	7,365.0	5.7	1.1	25.79	226
United States	7,087	500	8,266.4	4.6	1.1	18.45	384

*Eighth grade in most countries.

Table 5.11 Design Effects and Effective Sample Sizes for Seventh and Eighth Grades* (Combined) - Science Mean Scale Score - Population 2

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	12,852	524	11,329.0	3.3	0.9	12.28	1,046
Austria	5,786	538	9,606.7	2.9	1.3	5.03	1,150
Belgium (Fl)	5,662	540	6,125.6	2.6	1.0	6.16	920
Belgium (Fr)	4,883	458	7,000.1	2.5	1.2	4.48	1,091
Bulgaria	3,771	548	11,746.9	4.0	1.8	5.22	722
Canada	16,581	515	8,596.0	2.0	0.7	7.40	2,239
Colombia	5,304	398	5,580.2	3.4	1.0	11.05	480
Cyprus	5,852	440	8,152.7	1.3	1.2	1.18	4,956
Czech Republic	6,672	553	7,549.6	2.7	1.1	6.68	999
Denmark	4,370	460	7,993.3	2.1	1.4	2.39	1,832
England	3,579	532	11,125.7	2.6	1.8	2.18	1,641
France	6,014	474	6,229.8	2.1	1.0	4.16	1,446
Germany	5,763	515	9,962.9	4.1	1.3	9.63	599
Greece	7,921	472	8,025.1	2.1	1.0	4.45	1,781
Hong Kong	6,752	509	7,870.6	4.6	1.1	18.14	372
Hungary	5,978	535	8,551.7	2.6	1.2	4.68	1,277
Iceland	3,730	478	6,195.1	2.5	1.3	3.89	959
Iran, Islamic Rep.	7,429	452	5,474.7	2.1	0.9	6.26	1,187
Ireland	6,203	516	9,161.1	3.0	1.2	6.03	1,028
Israel	1,415	524	10,758.9	5.7	2.8	4.33	327
Japan	10,271	552	8,175.0	1.6	0.9	3.13	3,285
Korea	5,827	550	8,821.1	1.7	1.2	1.97	2,958
Kuwait	1,655	430	5,459.9	3.7	1.8	4.18	396
Latvia (LSS)	4,976	459	6,945.4	2.1	1.2	3.13	1,591
Lithuania	5,056	441	7,788.4	2.8	1.2	5.14	983
Netherlands	4,084	540	7,216.3	3.6	1.3	7.43	550
New Zealand	6,867	504	10,140.0	3.0	1.2	5.97	1,150
Norway	5,736	505	7,894.2	1.8	1.2	2.26	2,539
Portugal	6,753	453	5,940.1	2.0	0.9	4.63	1,459
Romania	7,471	469	10,470.0	4.1	1.2	12.20	612
Russian Federation	8,160	510	9,710.2	3.6	1.1	10.92	747
Scotland	5,776	493	9,984.8	4.1	1.3	9.80	589
Singapore	8,285	576	10,542.6	5.3	1.1	21.76	381
Slovak Republic	7,101	527	8,127.0	2.7	1.1	6.14	1,157
Slovenia	5,606	544	7,762.2	2.0	1.2	2.78	2,019
South Africa	9,792	322	9,192.8	4.6	1.0	22.80	429
Spain	7,596	497	6,627.9	1.7	0.9	3.23	2,353
Sweden	6,906	512	8,184.2	2.0	1.1	3.45	2,000
Switzerland	8,940	503	7,867.9	1.9	0.9	4.30	2,078
Thailand	11,643	509	5,266.7	3.1	0.7	21.79	534
United States	10,973	521	11,268.9	4.6	1.0	20.22	543

*Seventh and eighth grades in most countries.

Table 5.12 Design Effects and Effective Sample Sizes for Seventh Grade* Science Mean Scale Score - Population 2

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	5,599	504	10,522.1	3.6	1.4	6.78	826
Austria	3,013	519	8,833.5	3.1	1.7	3.36	897
Belgium (Fl)	2,768	529	5,343.3	2.6	1.4	3.37	821
Belgium (Fr)	2,292	442	6,183.9	3.0	1.6	3.45	665
Bulgaria	1,798	531	10,607.9	5.4	2.4	5.02	358
Canada	8,219	499	8,045.0	2.3	1.0	5.46	1,505
Colombia	2,655	387	5,218.9	3.2	1.4	5.34	497
Cyprus	2,929	420	7,567.9	1.8	1.6	1.31	2,238
Czech Republic	3,345	533	6,684.3	3.3	1.4	5.56	602
Denmark	2,073	439	7,453.4	2.1	1.9	1.28	1,625
England	1,803	512	10,226.4	3.5	2.4	2.16	834
France	3,016	451	5,510.5	2.6	1.4	3.62	833
Germany	2,893	499	9,147.1	4.1	1.8	5.19	557
Greece	3,931	449	7,631.1	2.6	1.4	3.38	1,163
Hong Kong	3,413	495	7,471.9	5.5	1.5	13.77	248
Hungary	3,066	518	8,351.8	3.2	1.7	3.69	830
Iceland	1,957	462	5,643.0	2.8	1.7	2.68	730
Iran, Islamic Rep.	3,735	436	5,124.9	2.6	1.2	4.77	784
Ireland	3,127	495	8,288.2	3.5	1.6	4.50	695
Japan	5,130	531	7,427.5	1.9	1.2	2.41	2,129
Korea	2,907	535	8,419.3	2.1	1.7	1.57	1,848
Latvia (LSS)	2,567	435	6,087.5	2.7	1.5	3.07	835
Lithuania	2,531	403	6,313.6	3.4	1.6	4.59	551
Netherlands	2,097	517	6,248.5	3.6	1.7	4.33	484
New Zealand	3,184	481	9,316.0	3.4	1.7	4.00	797
Norway	2,469	483	7,195.8	2.9	1.7	2.88	857
Portugal	3,362	428	5,109.1	2.1	1.2	2.91	1,155
Romania	3,746	452	9,999.2	4.4	1.6	7.30	513
Russian Federation	4,138	484	8,890.2	4.2	1.5	8.06	514
Scotland	2,913	468	8,773.3	3.8	1.7	4.85	601
Singapore	3,641	545	10,030.6	6.6	1.7	15.94	228
Slovak Republic	3,600	510	7,218.0	3.0	1.4	4.59	784
Slovenia	2,898	530	7,387.2	2.4	1.6	2.19	1,322
South Africa	5,301	317	8,470.9	5.3	1.3	17.46	304
Spain	3,741	477	6,387.0	2.1	1.3	2.65	1,410
Sweden	2,831	488	7,110.8	2.6	1.6	2.62	1,082
Switzerland	4,085	484	6,709.2	2.5	1.3	3.67	1,113
Thailand	5,810	493	4,779.5	3.0	0.9	10.85	536
United States	3,886	508	11,014.6	5.5	1.7	10.51	370

*Seventh grade in most countries.

Table 5.13 Design Effects and Effective Sample Sizes for Eighth Grade* Science Mean Scale Score - Population 2

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	7,253	545	11,338.8	3.9	1.3	9.50	763
Austria	2,773	558	9,636.0	3.7	1.9	3.87	717
Belgium (Fl)	2,894	550	6,579.3	4.2	1.5	7.62	380
Belgium (Fr)	2,591	471	7,315.2	2.8	1.7	2.87	904
Bulgaria	1,973	565	12,273.1	5.3	2.5	4.49	439
Canada	8,362	531	8,644.9	2.6	1.0	6.46	1,295
Colombia	2,649	411	5,703.8	4.1	1.5	7.68	345
Cyprus	2,923	463	7,838.6	1.9	1.6	1.38	2,112
Czech Republic	3,327	574	7,574.0	4.3	1.5	8.11	410
Denmark	2,297	478	7,741.4	3.1	1.8	2.91	790
England	1,776	552	11,202.9	3.3	2.5	1.78	999
France	2,998	498	5,893.4	2.5	1.4	3.15	952
Germany	2,870	531	10,284.8	4.8	1.9	6.45	445
Greece	3,990	497	7,220.9	2.2	1.3	2.75	1,448
Hong Kong	3,339	522	7,908.8	4.7	1.5	9.26	361
Hungary	2,912	554	8,105.2	2.8	1.7	2.81	1,036
Iceland	1,773	494	6,246.6	4.0	1.9	4.64	382
Iran, Islamic Rep.	3,694	470	5,277.5	2.4	1.2	4.02	919
Ireland	3,076	538	9,132.9	4.5	1.7	6.89	447
Israel	1,415	524	10,758.9	5.7	2.8	4.33	327
Japan	5,141	571	8,108.4	1.6	1.3	1.72	2,992
Korea	2,920	565	8,774.9	1.9	1.7	1.22	2,395
Kuwait	1,655	430	5,459.9	3.7	1.8	4.18	396
Latvia (LSS)	2,409	485	6,589.1	2.7	1.7	2.69	897
Lithuania	2,525	476	6,564.2	3.4	1.6	4.51	560
Netherlands	1,987	560	7,225.6	5.0	1.9	6.80	292
New Zealand	3,683	525	9,958.0	4.4	1.6	7.04	523
Norway	3,267	527	7,628.7	1.9	1.5	1.63	2,010
Portugal	3,391	480	5,447.4	2.3	1.3	3.41	993
Romania	3,725	486	10,345.6	4.7	1.7	8.10	460
Russian Federation	4,022	538	9,075.2	4.0	1.5	7.02	573
Scotland	2,863	517	9,968.9	5.1	1.9	7.48	383
Singapore	4,644	607	9,097.9	5.5	1.4	15.65	297
Slovak Republic	3,501	544	8,458.0	3.2	1.6	4.36	804
Slovenia	2,708	560	7,695.7	2.5	1.7	2.16	1,252
South Africa	4,491	326	9,769.0	6.6	1.5	20.29	221
Spain	3,855	517	6,072.4	1.7	1.3	1.84	2,096
Sweden	4,075	535	8,145.7	3.0	1.4	4.41	923
Switzerland	4,855	522	8,266.9	2.5	1.3	3.67	1,324
Thailand	5,833	525	5,232.6	3.7	0.9	15.67	372
United States	7,087	534	11,178.9	4.7	1.3	14.29	496

*Eighth grades in most countries.

REFERENCES

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Smith, T.A., and Kelly, D.L. (1996a). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Beaton, A.E., Mullis, I.V.S., Martin, M.O. Gonzalez, E.J., Kelly, D.L., and Smith, T.A. (1996b). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

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 technical report, volume I: Design and development*. Chestnut Hill, MA: Boston College.

Gonzalez, E. J. and Smith, T. A., Eds. (1997). *User guide for the TIMSS international database: Primary and middle school years – 1995 assessment.* Chestnut Hill, MA: Boston College.

Johnson, E. G. and Rust, K.F. (1992). Population references and variance estimation for NAEP data. *Journal of Educational Statistics*, 17, 175-190.

Kish, L. (1965). *Survey sampling.* New York: John Wiley & Sons.

Martin, M.O., Mullis, I.V.S., Beaton, A.E., Gonzalez, E.J., Smith, T.A., and Kelly, D.L. (1997). *Science achievement in the primary school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L., and Smith, T.A. (1997). *Mathematics achievement in the primary school years IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Westat, Inc. (1997). *A user's guide to WesVarPC.* Rockville, MD: Westat, Inc.

Wolter, K. M. (1985). *Introduction to variance estimation.* New York: Springer-Verlag.

6 Item Analysis and Review

Ina V.S. Mullis
Michael O. Martin
Boston College

6.1 CROSS-COUNTRY ITEM STATISTICS

In order to assess the statistical properties of the items before proceeding with item response theory (IRT) scaling (see Chapter 7), TIMSS computed a series of statistics for every item in every country. These basic item statistics (see Figure 6.1 for an example item) were produced by the IEA Data Processing Center. For each item, the basic display presents the number of students that responded in each country, the difficulty level (the percentage of students that answered the item correctly), and the discrimination index (the point-biserial correlation between success on the item and a total score).[1] For multiple-choice items the display presents the percentage of students that chose each option, including the percentage that omitted or did not reach the item, and the point-biserial correlation between each option and the total score. For free-response items (which could have more than one score level), the display presents the difficulty and discrimination of each score level.

As a prelude to the main IRT scaling, the display presents some statistics from a preliminary Rasch analysis, including the Rasch item difficulty for each item, the standard error of this difficulty estimate, and an index of the goodness-of-fit of the item to the Rasch model (Wu, 1997).

The item-analysis display presents the difficulty level of each item separately for male and female students, and, because the TIMSS IRT scaling spans two grades at Population 1 and Population 2, separately for lower- and upper-grade students. As a guide to the overall statistical properties of the item, it also presents the international item difficulty (the mean of the item difficulties across countries) and the international item discrimination (the mean of the item discriminations).

As an aid to reviewers, the item-analysis display includes a series of "flags" signaling the presence of one or more conditions that might indicate a problem with an item. The following conditions are flagged:

- Item difficulty exceeds 95 percent in the sample as a whole
- Item difficulty is less than 25 percent for 4-option multiple-choice items in the sample as a whole (20 percent for 5-option items)

[1] For the purpose of computing the discrimination index, the total score was the percentage of items a student answered correctly in mathematics or science.

Figure 6.1 Examples of Cross-Country Item Analysis

Population: 2 (B) Subject: 'Mathematics' Cluster: H Item: BSMMH07

Country	Correct Answer			Flags	Percentages for each alternative								Point biserials for each alternative								Rasch			Group Difficulties				International Mean	
	N	DIFF	DISCR		A	B	C	D	E	W	OMIT	NR	A	B	C	D	E	W	OMIT	NR	RDIFF	SE	FIT	MAL	FEM	LOW	UPP	IDIFF	IDISCR
AUS	5112	78.5	0.36	qG	3.3	9.8	78.5*	2.4	5.2		0.6	0.1	-0.21	-0.14	.36*	-0.10	-0.22		-0.07	-0.04	-1.14	0.04	1.07	77.9	79.1	70.3	78.7	68.1	0.38
AUT	2241	70.3	0.41	q	4.6	12.4	70.3*	3.4	6.6		2.2	0.1	-0.19	-0.16	.41*	-0.11	-0.19		-0.14	-0.06	-0.34	0.05	1.06	71.5	69.2	70.3	70.3	68.1	0.38
BFL	2123	82.9	0.37	qG	2.7	8.4	82.9*	2.5	2.9		0.5	0.0	-0.21	-0.19	.37*	-0.13	-0.17		-0.07	-0.01	-0.79	0.06	1.00	83.3	82.6	82.1	83.8	68.1	0.38
BFR	1841	72.5	0.39	q	8.0	12.2	72.5*	2.2	3.3		1.3	0.1	-0.24	-0.14	.39*	-0.09	-0.17		-0.11	-0.02	-0.60	0.06	1.05	73.6	71.7	70.1	74.6	68.1	0.38
BGR	1285	59.2	0.46	qG	8.2	16.2	59.2*	4.4	6.8		3.1	0.0	-0.17	-0.15	.46*	-0.12	-0.21		-0.19	0.00	0.06	0.06	1.04	59.2	59.2	55.5	62.9	68.1	0.38
CAN	6180	71.5	0.41	qG	5.5	10.9	71.5*	4.1	7.3		0.4	0.1	-0.21	-0.18	.41*	-0.11	-0.20		-0.08	-0.03	-0.83	0.03	1.00	72.0	71.2	69.3	73.7	68.1	0.38
CHE	3718	77.4	0.42	qG	4.2	9.7	77.4*	3.1	3.6		0.7	0.2	-0.25	-0.19	.42*	-0.12	-0.15		-0.11	-0.04	-0.67	0.04	1.02	77.8	77.1	78.4	82.6	68.1	0.38
CHE	588	79.8	0.38	qG	3.9	8.3	79.8*	2.2	4.3		0.5	0.0	-0.21	-0.17	.38*	-0.11	-0.19		-0.08	0.00	-0.73	0.11	0.98	80.9	78.5	78.9	80.6	68.1	0.38
CHE	570	75.8	0.36	qS	3.3	11.8	75.8*	3.2	3.9		0.4	0.4	-0.16	-0.15	.36*	-0.10	-0.20		-0.10	-0.10	-0.50	0.11	1.03	74.2	77.5	73.0	78.6	68.1	0.38
COL	1978	35.9	0.25		14.0	19.2	35.9*	7.7	14.8		5.8	1.6	0.00	-0.02	.25*	-0.05	-0.15		-0.15	-0.09	-0.58	0.05	1.07	36.3	35.7	35.4	36.5	68.1	0.38
CSK	2465	64.8	0.38	qsSG	5.5	19.7	64.8*	3.3	3.8		1.1	0.2	-0.21	-0.18	.38*	-0.10	-0.14		-0.08	-0.05	-0.02	0.05	1.11	67.3	62.3	62.7	66.9	68.1	0.38
CYP	2129	53.9	0.35		9.1	16.6	53.9*	6.0	10.1		2.5	0.0	-0.12	-0.10	.35*	-0.08	-0.19		-0.12	-0.03	-0.43	0.05	1.10	54.7	53.0	52.1	55.6	68.1	0.38
DEU	2258	70.2	0.41	q	5.3	9.8	70.2*	3.7	6.2		2.3	0.4	-0.22	-0.11	.41*	-0.09	-0.16		-0.19	-0.09	-0.86	0.05	1.03	70.4	70.7	66.8	73.6	68.1	0.38
DNK	2372	77.4	0.43	q	3.7	9.6	77.4*	3.1	3.7		1.4	0.4	-0.20	-0.20	.43*	-0.13	-0.20		-0.14	-0.04	-1.28	0.05	0.97	78.4	76.3	78.0	86.4	68.1	0.38
ESP	2813	54.7	0.40	qsS	8.7	19.0	54.7*	4.5	8.7		4.4	0.1	-0.19	-0.10	.40*	-0.09	-0.19		-0.15	0.00	-0.26	0.04	1.03	56.8	52.6	50.5	58.8	68.1	0.38
FRA	2245	75.6	0.35	qsS	6.1	11.4	75.6*	1.7	3.4		0.4	0.3	-0.20	-0.14	.35*	-0.06	-0.18		-0.05	-0.07	-0.90	0.05	1.04	78.1	73.2	70.7	80.7	68.1	0.38
GBR	1322	77.5	0.42	qG	2.6	11.6	77.5*	2.7	5.2		0.4	0.0	-0.20	-0.20	.42*	-0.12	-0.23		-0.10	0.00	-1.48	0.07	0.99	78.6	76.2	77.5	77.4	68.1	0.38
GRC	3017	45.4	0.42	G	17.8	14.5	45.4*	6.5	11.2		2.8	0.1	-0.17	-0.07	.42*	-0.07	-0.22		-0.12	0.00	0.02	0.04	1.06	46.2	44.5	38.1	52.6	68.1	0.38
HKG	2537	85.6	0.42	qs	1.7	5.5	85.6*	2.8	3.9		0.4	0.0	-0.23	-0.24	.42*	-0.09	-0.22		-0.05	0.00	-1.09	0.06	1.00	86.8	84.0	84.3	86.8	68.1	0.38
HUN	2224	61.5	0.46	qSG	9.2	15.5	61.5*	5.0	1.9		6.6	0.4	-0.24	-0.16	.46*	-0.18	-0.10		-0.17	-0.06	-0.06	0.05	1.02	62.9	60.4	55.0	68.3	68.1	0.38
IRL	2332	77.7	0.41	qS	3.7	11.2	77.7*	3.1	3.6		0.1	0.0	-0.23	-0.20	.41*	-0.11	-0.19		-0.05	-0.03	-1.07	0.05	1.01	77.7	77.8	75.3	80.3	68.1	0.38
IRN	2755	42.1	0.30	SG	17.5	20.3	42.1*	6.0	11.7		2.1	0.0	-0.13	-0.05	.30*	-0.04	-0.20		-0.01	0.00	-0.37	0.04	1.05	41.3	43.1	35.5	48.8	68.1	0.38
ISL	1388	73.1	0.40	q	3.6	13.4	73.1*	4.0	5.0		0.4	0.1	-0.18	-0.20	.40*	-0.15	-0.15		-0.06	-0.04	-1.37	0.07	0.98	73.5	72.7	69.3	77.5	68.1	0.38
ISR	518	63.3	0.38		7.9	16.6	63.3*	5.0	5.6		1.2	0.4	-0.16	-0.15	.38*	-0.18	-0.10		-0.10	-0.16	-0.18	0.10	1.10	65.0	64.0	63.3	63.3	68.1	0.38
JPN	3913	83.3	0.33	qSG	1.9	8.6	83.3*	2.4	3.7		0.0	0.0	-0.17	-0.14	.33*	-0.13	-0.22		0.00	0.00	-0.73	0.05	1.09	82.4	84.2	82.1	84.4	68.1	0.38
KOR	2160	91.3	0.45	qsG	1.7	3.1	91.3*	2.0	1.8		0.1	0.0	-0.19	-0.20	.45*	-0.21	-0.28		-0.04	0.00	-1.84	0.08	0.91	92.9	89.3	90.7	91.9	68.1	0.38
KWT	635	42.8	0.27		14.8	16.4	42.8*	7.1	14.5		4.1	0.3	-0.10	-0.03	.27*	-0.09	-0.14		-0.06	-0.10	-0.66	0.09	1.06	41.4	43.5	.	.	68.1	0.38
LTU	1882	45.2	0.39	sS	10.2	24.0	45.2*	5.6	6.9		6.5	0.4	-0.15	-0.09	.39*	-0.04	-0.18		-0.22	-0.06	0.02	0.05	1.06	42.3	47.8	39.7	50.7	68.1	0.38
LVA	1867	47.3	0.35	q	8.2	26.0	47.3*	4.6	9.4		3.9	0.6	-0.16	-0.06	.35*	-0.07	-0.21		-0.12	-0.05	0.06	0.05	1.10	46.7	47.9	43.5	51.2	68.1	0.38
MEX	4371	48.2	0.33	q	16.4	18.6	48.2*	4.0	10.8		1.2	0.4	-0.17	-0.06	.33*	-0.05	-0.19		-0.07	-0.04	-0.88	0.03	1.02	48.3	48.0	45.8	50.7	68.1	0.38
NLD	1546	78.4	0.36	q	3.2	13.8	78.4*	1.6	2.7		0.3	0.2	-0.27	-0.17	.36*	-0.08	-0.14		-0.11	-0.08	-0.98	0.07	1.06	78.4	78.2	74.9	82.0	68.1	0.38
NOR	2144	72.8	0.37	qG	4.2	12.3	72.8*	3.3	6.5		0.9	0.1	-0.21	-0.15	.37*	-0.08	-0.18		-0.12	-0.03	-1.12	0.05	1.02	72.0	73.3	70.2	74.8	68.1	0.38
NZL	2543	77.1	0.41	qsS	4.3	10.5	77.1*	3.1	4.7		0.1	0.0	-0.20	-0.20	.41*	-0.14	-0.19		-0.02	-0.04	-1.41	0.05	1.00	79.0	75.2	73.9	79.8	68.1	0.38
PHL	4478	49.3	0.15	qbBFSG	6.5	25.8	49.3*	4.7	13.2		0.4	0.1	-0.06	0.03	.15*	-0.07	-0.16		-0.03	-0.02	-0.92	0.03	1.20	50.2	48.7	49.7	49.0	68.1	0.38
PRT	2496	64.0	0.35	qsS	6.9	17.7	64.0*	3.0	6.1		2.2	0.2	-0.17	-0.13	.35*	-0.11	-0.15		-0.12	0.01	-1.14	0.05	1.01	67.5	60.4	60.9	67.0	68.1	0.38
ROM	2789	43.5	0.35	sS	16.6	16.9	43.5*	6.9	9.3		4.0	0.5	-0.12	-0.08	.35*	-0.07	-0.18		-0.12	-0.04	0.19	0.04	1.09	45.9	41.1	39.9	47.1	68.1	0.38
RUS	3036	54.4	0.31	qF	8.5	24.4	54.4*	2.6	5.6		3.3	0.4	-0.17	-0.04	.31*	-0.09	-0.20		-0.15	-0.03	0.20	0.04	1.19	54.9	53.9	51.5	57.3	68.1	0.38
SCO	2127	76.4	0.39	q	2.8	12.4	76.4*	2.9	4.8		0.3	0.0	-0.22	-0.17	.39*	-0.11	-0.23		-0.07	-0.02	-1.48	0.06	1.01	76.1	76.6	73.9	78.9	68.1	0.38
SGP	3096	85.5	0.36	qs	1.8	7.2	85.5*	2.2	3.2		0.1	0.0	-0.12	-0.24	.36*	-0.14	-0.17		-0.04	0.01	-0.56	0.05	1.02	84.5	86.4	82.0	88.2	68.1	0.38
SLV	2650	61.1	0.37	qFG	5.5	20.8	61.1*	4.0	5.8		1.4	0.2	-0.17	-0.16	.37*	-0.08	-0.18		-0.09	-0.05	0.03	0.05	1.13	60.3	61.8	58.6	63.6	68.1	0.38
SVN	2086	68.2	0.38	qs	5.5	17.3	68.2*	3.5	3.2		0.8	0.0	-0.21	-0.19	.38*	-0.09	-0.13		-0.06	0.00	-0.46	0.05	1.06	69.9	66.7	62.8	73.7	68.1	0.38
SWE	3277	73.9	0.38	q	3.3	13.5	73.9*	4.0	4.7		0.3	0.1	-0.20	-0.16	.38*	-0.13	-0.20		-0.08	-0.04	-0.96	0.04	1.05	73.4	74.4	74.6	80.0	68.1	0.38
USA	4108	69.7	0.37	qsS	5.4	14.5	69.7*	3.8	5.8		0.7	0.0	-0.21	-0.19	.37*	-0.11	-0.13		-0.07	0.01	-1.04	0.04	1.06	67.6	72.0	66.4	71.6	68.1	0.38

- Item difficulty exceeds 95 percent or is less than 25 percent (20 percent for 5-option items) for students in the lower grade
- Item difficulty exceeds 95 percent or is less than 25 percent (20 percent for 5-option items) for students in the upper grade
- One or more of the distracter percentages is less than 5 percent
- One or more of the distracter percentages is greater than the percentage for the correct answer
- Point-biserial correlation for one or more of the distracters exceeds zero
- Item discrimination (i.e., the point-biserial for the correct answer) is less than 0.2
- Item discrimination does not increase with each score level (for an item with more than one score level)
- Rasch goodness-of-fit index is less than 0.88 or greater than 1.12
- Difficulty levels on the item are significantly different for males and females
- Difference in item difficulty levels between males and females diverge significantly from the average difference between males and females across all the items making up the total score
- Difference in item difficulty levels between lower and upper grades diverge significantly from the average difference between lower and upper grades on all the items making up the total score.

Although not all of these conditions necessarily indicate a problem, the flags are a useful way to draw a reviewer's attention to potential sources of concern. The IEA Data Processing Center also produced information about the inter-rater agreement for the free-response items.

6.2 GRAPHICAL DISPLAYS

As a further aid to reviewing the psychometric characteristics of the items, the Australian Council for Educational Research (ACER) produced graphical representations of selected item statistics for each participating country (see Figure 6.2). This display presents, for each item, the difficulty level and discrimination for every country, together with the Rasch goodness-of-fit statistic and an indication of the item-by-country interaction. The item-by-country interaction chart plots a confidence interval for the probability of success on the item in each country against the average probability of success across all countries. The graphical representations allow comparisons across countries on these statistics at a glance.

Figure 6.2 Example of Graphical Displays of Cross-Country Item Statistics - Mathematics - Population 2

6.3 SUMMARY INFORMATION FOR POTENTIALLY PROBLEMATIC ITEMS

Although the system of flagging potentially problematic conditions and the graphical summaries were both very helpful in identifying items with possible problems, the task of reviewing the characteristics of each item in each country was still considerable. To ensure that no serious item problem would go unnoticed, ACER also provided, for each item, a list of countries that exhibited one or more potentially serious characteristics (see Figure 6.3). Countries were listed in this display if the item had a significant item-by-country interaction (i.e., students in the country found the item easier or more difficult than items in general), or if they exhibited problematic discrimination (i.e., the point-biserial for a distracter was greater than .05, the point-biserial for the correct answer was negative, or, for items with more than one score point, the point-biserial did not increase with each score level). Countries were also listed if their data showed poor fit to the Rasch model for that item.

6.4 ITEM CHECKING PROCEDURES

Prior to the international scaling of the Population 1 and 2 achievement data by ACER, the International Study Center conducted a thorough review of the item statistics for all participating countries to ensure that items were performing comparably across countries. Although only a small number of items were found to be inappropriate for international comparisons, throughout the series of item-checking steps a number of reasons were discovered for differences in items across countries. Most of these were inadvertent changes in the items during the printing process, including omitting an item option or misprinting the graphics associated with an item. However, differences attributable to translation problems were found for an item or two in several countries.

In particular, items with the following problems were considered for possible deletion from the international database:

- Errors were detected during translation verification but were not corrected before test administration
- Data cleaning revealed more or fewer options than in the original version of the item
- The item analysis information showed the item to have a negative biserial
- The item-by-country interaction results showed a very large negative interaction for a given country
- The item-fit statistic indicated the item was not fitting the model
- For free-response items, the within-country scoring reliability data showed an agreement of less than 70% for the score level. Also, performance in items with more than one score level was not ordered by score, or correct levels were associated with negative point-biserials.

Figure 6.3 Example Summary Information for Items with Poor Statistics for Some Countries

	Item-by-Country Interactions		Discrimination			Fit
Country	Easier than Expected	Harder than Expected	Non-key PB is Positive	Key PB is Negative	Ability not Ordered	Fit Large
	Table=#Name					
Item 119	*BSMSQ15*		*BSMS/WHICH IS NOT A CHEMICAL CHANGE (A)*			
DEU			X			
HKG			X			
ISL			X			
ISR			X			
NOR			X			
PHL						
Item 120	*BSMSQ16*		*BSMS/HOW LONG TAKE LIGHT FROM STAR (D)*			
COL			X			
CYP			X			
DEU			X			
GRC			X			
HKG			X			
ISR			X			
KOR			X			
MEX			X			
ROM			X			
THA			X	X		

The statistics and translation verification documentation were used as pointers towards checking actual booklets and contacting National Research Coordinators. If a problem could be detected by the International Study Center (such as a negative point-biserial for a correct answer or too few options for the multiple-choice questions), the item was deleted from the international scaling. However, if there was a question about potential translation or cultural issues, then the NRC was queried, and the International Study Center abided by the decision made by the NRC. In several cases, NRCs consulted mathematics or science experts before making a decision.

Considering that the checking involved approximately 500 items for each of more than 40 countries, very few deviations from the international format were revealed. Table 6.1 contains a list of the changes made in the international database for Populations 1 and 2.

Table 6.1 Recodes Made to Free-Response Item Codes in the Written Assessment and Performance Assessment Items

	Item	Variable	Recode	Comment
General	All Items		37, 38 → 39	Country-specific diagnostic codes recoded to 'other' categories within the score level.
			27, 28 → 29	
			17, 18 → 19	
			77, 78 → 79	
Population 2 - Written Assessment Items	K10	BSMMK08	71 → 70	Training team found it difficult to distinguish between the 70 and 71 codes; both codes combined in 70.
	L04	BSESL04	20 → 10	Only 20s have positive point-biserial correlation; change to 1-point item codes.
			21 → 11	
			29 → 19	
			10 → 74	
			11 → 75	
			12 → 76	
			19 → 79	
	M11	BSESM11	10, 11, 12, 13 → 71	Only 30s have positive point-biserial correlation; change to 1-point item codes.
			20, 21, 22, 23, 24, 25 → 72	
			30 → 10	
			31 → 11	
	Y01	BSESY01	20 → 10	Only 20s have positive point-biserial correlation; change to 1-point item codes.
			21 → 11	
			22 → 12	
			29 → 19	
			10 → 73	
			11 → 74	
			19 → 75	
	Y02	BSESY02	21 → 19	Typographical error in category 21 in coding guide.
	J03	BSSSJ03	19 → 10	Typographical error in coding guide.
	M12	BSSSM12	19 → 10	Typographical error in coding guide.
	O14	BSES014	20 → 10	Only 20s have positive point-biserial correlation.
			29 → 19	
			10 → 72	
			11 → 73	
			19 → 74	
	Q18	BSSSQ18	19 → 10	Typographical error in coding guide.
			29 → 20	
	L16	BSSML16	19 → 10	Typographical error in coding guide.
	M06	BSSMM06	19 → 10	Typographical error in coding guide.
	M08	BSSMM08	19 → 10	Typographical error in coding guide.
	Q10	BSSMQ10	19 → 10	Typographical error in coding guide.
	R13	BSSMR13	74 → 79	Typographical error in code 74 (28 instead of 280); leaves gap in 7* diagnostic codes.
	S01A	BSEMS01A	19 → 10	Typographical error in coding guide.
	S02A	BSEMS02A	19 → 10	Typographical error in coding guide.
	T01A	BSEMT01A	29 → 20	Typographical error in coding guide.
	T02A	BSEMT02A	19 → 10	Typographical error in coding guide.
	U01A	BSEMU01A	19 → 10	Typographical error in coding guide.
	U02A	BSEMU02A	19 → 10	Typographical error in coding guide.
			29 → 20	
	U02B	BSEMU02B	19 → 10	Typographical error in coding guide.
			29 → 20	

Table 6.1 Recodes Made to Free-Response Item Codes in the Written Assessment and Performance Assessment Items (Continued)

	Item	Variable	Recode	Comment
Population 1 - Written Assessment Items	T04A	ASEMT04A	20 → 10 29 → 19 10 → 72 11 → 73	Only 20s have positive point-biserial correlation.
	T04B	ASEMT04B	20 → 10 29 → 19 10 → 72 11 → 73	Only 20s have positive point-biserial correlation.
	V04A	ASEMV04A	30 → 20 20 → 12 21 → 13	Differentiation between 30s, 20s, and 10s not clear.
	Y01	ASESY01	20 → 10 29 → 19 10 → 72 11 → 73 19 → 74	Only 20s have positive point-biserial correlation.
	Z02	ASESZ02	30 → 10 31 → 11 20 → 71 29 → 72 10, 11, 12, 13 → 73	Only 30s have positive point-biserial correlation.
Performance Assessment Items	Task M2 (Calculator) Item 5 (Population 2)	BSPM25	11 → 12	Error in coding guide: valid codes listed as 10,12, 19 (no code 11). Recoded 11 codes used in some countries.
	Task M5 (Packaging) Item 1 (Populations 2 & 1)	BSPM51 ASPM51	30 → 22 31 → 23	Two versions of task used across countries: original asked for 2 OR 3 boxes; revised asked for 3. Item changed to 2-point value for report tables; changed codes for 3 correct boxes (30,31) to 2-point codes (22,23).
	Task S5 (Solutions) Item 2A (Population 2)	BSPS52A	99 → 98	Administrator notes not coded consistently across countries; invalid 99 codes (blank) used in several countries recoded to not administered. Item omitted from report table but kept in data file.
	Task S5 (Solutions) Item 4 (Population 2)	BSPS54	10 → 21	Coding guide revised based on reports of problematic scoring during training development.
	Task S6 (Containers) Item 1A (Population 1)	ASPS61A	99 → 98	Administrator notes not coded consistently across countries; invalid 99 codes (blank) used in several countries recoded to not administered.

REFERENCES

Wu, M.L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalised item response models*. Unpublished master's dissertation, University of Melbourne.

7 Scaling Methodology and Procedures for the Mathematics and Science Scales

Raymond J. Adams
Margaret L. Wu
Greg Macaskill
Australian Council for Educational Research

The principal method by which student achievement is reported in TIMSS is through scale scores derived using Item Response Theory (IRT) scaling. With this approach, the performance of a sample of students in a subject area can be summarized on a common scale or series of scales even when different students have been administered different items. The common scale makes it possible to report on relationships between students' characteristics (based on their responses to the background questionnaires) and their overall performance in mathematics and science.

Because of the need to achieve broad coverage of both mathematics and science within a limited amount of student testing time, each student was administered relatively few items within each content area of each subject. In order to achieve reliable indices of student proficiency in this situation, it was necessary to make use of multiple imputation or "plausible values" methodology. Further information on plausible value methods may be found in Mislevy (1991), and in Mislevy, Johnson, and Muraki (1992). The proficiency scale scores or plausible values assigned to each student are actually random draws from the estimated ability distribution of students with similar item response patterns and background characteristics. The plausible values are intermediate values that may be used in statistical analyses to provide good estimates of parameters of student populations. Although intended for use in place of student scores in analyses, plausible values are designed primarily to estimate population parameters, and are not optimal estimates of individual student proficiency.

This chapter provides details of the IRT model used in TIMSS to scale the achievement data. For those interested in the technical background of the scaling, the chapter describes the model itself and the method of estimating the parameters of the model.

7.1 THE TIMSS SCALING MODEL

The scaling model used in TIMSS was the multidimensional random coefficients logit model as described by Adams, Wilson, and Wang (1997), with the addition of a multivariate linear model imposed on the population distribution. The scaling was done with the *ConQuest* software (Wu, Adams, and Wilson, 1997) that was developed in part to meet the needs of the TIMSS study.

The multidimensional random coefficients model is a generalization of the more basic unidimensional model.

7.1.1 The Unidimensional Random Coefficients Model

Assume that **I** items are indexed i=1,...,I with each item admitting $K_i + 1$ response alternatives $k = 0,1,...,K_i$. Use the vector valued random variable, $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iK_i})$,

$$\text{where } X_{ij} = \begin{cases} 1 & \text{if response to item } i \text{ is in category } j \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

to indicate the $K_i + 1$ possible responses to item i.

A response in category zero is denoted by a vector of zeroes. This effectively makes the zero category a reference category and is necessary for model identification. The choice of this as the reference category is arbitrary and does not affect the generality of the model. We can also collect the $\mathbf{X}_i$ together into the single vector $\mathbf{X}' = (\mathbf{X}_1', \mathbf{X}_2',...,\mathbf{X}_\mathbf{i}')$, which we call the response vector (or pattern). Particular instances of each of these random variables are indicated by their lower-case equivalents: $\mathbf{x}$, $\mathbf{x}_i$ and x_{ik}.

The items are described through a vector $\xi^T = (\xi_1, \xi_2, \ldots, \xi_p)$ of p parameters. Linear combinations of these are used in the response probability model to describe the empirical characteristics of the response categories of each item. These linear combinations are defined by design vectors $\mathbf{a}_{jk}$, ($j = 1,\ldots,I$; $k = 1,\ldots K_i$) each of length p, which can be collected to form a design matrix $\mathbf{A}' = (\mathbf{a}_{11}, \mathbf{a}_{12}, \ldots, \mathbf{a}_{1K_1}, \mathbf{a}_{21}, \ldots, \mathbf{a}_{2K_2}, \ldots, \mathbf{a}_{iK_i})$. Adopting a very general approach to the definition of items, in conjunction with the imposition of a linear model on the item parameters, allows us to write a general model that includes the wide class of existing Rasch models, for example, the item bundles models of Wilson and Adams (1995).

An additional feature of the model is the introduction of a scoring function, which allows the specification of the score or "performance level" that is assigned to each possible response to each item. To do this we introduce the notion of a response score b_{ij}, which gives the performance level of an observed response in category j of item i. The b_{ij} can be collected in a vector as $\mathbf{b}^T = (b_{11}, b_{12}, \ldots, b_{1K_1}, b_{21}, b_{22}, \ldots, b_{2K_2}, \ldots, b_{iK_i})$. (By definition, the score for a response in the zero category is zero, but other responses may also be scored zero.)

In the majority of Rasch model formulations there has been a one-to-one match between the category to which a response belongs and the score that is allocated to the response. In the simple logistic model, for example, it has been standard practice to use the labels 0 and 1 to indicate both the categories of performance and the scores. A similar practice has been followed with the rating scale and partial credit models, where each different possible response is seen as indicating a different level of performance, so that the category indicators 0, 1, 2, etc. that are used serve as both scores and labels. The use of **b** as a scoring function allows a more flexible relationship between the qualitative aspects of a response and the level of performance that it reflects. Examples of where this is applicable are given in Kelderman and Rijkes (1994) and Wilson (1992). A primary reason for implementing this feature in the model was to facilitate the analysis of the two-digit coding scheme that was used in the TIMSS short-answer and extended-response items. In the final analyses, however, only the first digit of the coding was used in the scaling, so this facility in the model and scaling software was not used in TIMSS.

Letting θ be the latent variable, the item response probability model is written as:

$$\Pr(\mathbf{X}_{ij} = 1;\mathbf{A},\mathbf{b},\xi|\theta) = \frac{\exp(b_{ij}\theta + \mathbf{a}_{ij}^T\ \xi)}{\sum_{k=1}^{K_i} \exp(b_{ik}\theta + \mathbf{a}_{ij}^T\ \xi)} \tag{2}$$

and a response vector probability model as

$$f(\mathbf{x};\xi|\theta) = \Psi(\theta,\xi)\exp[\mathbf{x}^T(\mathbf{b}\theta + \mathbf{A}\xi)] \tag{3}$$

with

$$\Psi(\theta,\xi) = \left\{\sum_{\mathbf{z}\in\Omega} \exp[\mathbf{z}^T(\mathbf{b}\theta + \mathbf{A}\xi)]\right\}^{-1} \tag{4}$$

where Ω is the set of all possible response vectors.

7.1.2 The Multidimensional Random Coefficients Multinomial Logit Model

The multidimensional form of the model is a straightforward extension of the model that assumes that a set of D traits underlie the individuals' responses. The D latent traits define a D-dimensional latent space, and the individuals' positions in the D-dimensional latent space are represented by the vector $\theta = (\theta_1,\theta_2,\ldots,\theta_D)$. The scoring function of response category k in item i now corresponds to a D by 1 column vector rather than a scalar as in the unidimensional model. A response in category k in dimension d of item i is scored b_{ikd}. The scores across D dimensions can be collected into a column vector $\mathbf{b}_{ik} = (b_{ik1},b_{ik2},\ldots,b_{ikD})^T$, again be collected into the scoring sub-matrix for item i, $\mathbf{B}_i = (\mathbf{b}_{i1},\mathbf{b}_{i2},\ldots,\mathbf{b}_{ik_i})^T$, and then be collected into a scoring matrix $\mathbf{B} = (\mathbf{B}_1^T,\mathbf{B}_2^T,\ldots\mathbf{B}_I^T)^T$ for the whole test. If the item parameter vector, ξ, and the design matrix, **A**, are defined as they were in the unidimensional model, the probability of a response in category k of item i is modeled as

$$\Pr(\mathbf{X}_{ij} = 1;\mathbf{A},\mathbf{B},\xi|\theta) = \frac{\exp(\mathbf{b}_{ij}\theta + \mathbf{a}^T_{ij}\ \xi)}{\sum_{k=1}^{K_i} \exp(\mathbf{b}_{ik}\theta + \mathbf{a}^T{}_{ik}\xi)} \tag{5}$$

And for a response vector we have:

$$f(\mathbf{x};\xi|\theta) = \Psi(\theta,\xi)\exp[\mathbf{x}'(\mathbf{B}\theta + \mathbf{A}\xi)] \tag{6}$$

with

$$\Psi(\theta,\xi) = \left\{\sum_{\mathbf{z}\in\Omega} \exp[\mathbf{z}^T(\mathbf{B}\theta + \mathbf{A}\xi)]\right\}^{-1} \tag{7}$$

The difference between the unidimensional model and the multidimensional model is that the ability parameter is a scalar, θ, in the former, and a D by 1 column vector, θ, in the latter. Likewise, the scoring function of response k to item i is a scalar, b_{ik}, in the former, whereas it is a D by 1 column vector, $\mathbf{b}_{ik}$, in the latter.

7.2 THE POPULATION MODEL

The item response model is a conditional model in the sense that it describes the process of generating item responses conditional on the latent variable, θ. The complete definition of the TIMSS model, therefore, requires the specification of a density, $f_\theta(\theta;\alpha)$, for the latent variable, θ. We use α to symbolize a set of parameters that characterize the distribution of θ. The most common practice when specifying unidimensional marginal item response models is to assume that the students have been sampled from a normal population with mean μ and variance σ^2. That is:

$$f_\theta(\theta;\alpha) \equiv f_\theta(\theta;\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] \tag{8}$$

or equivalently

$$\theta = \mu + E \tag{9}$$

where $E \sim N(0,\sigma 2)$.

Adams, Wilson, and Wu (1997) discuss how a natural extension of (8) is to replace the mean, μ, with the regression model, $\mathbf{Y}_n^T\beta$, where $\mathbf{Y}_n$ is a vector of u fixed and known values for student n, and β is the corresponding vector of regression coefficients. For example, $\mathbf{Y}_n$ could be constituted of student variables such as gender, socio-economic status, or major. Then the population model for student n becomes

$$\theta_n = \mathbf{Y}_n^T\beta + E_n \tag{10}$$

where we assume that the E_n are independently and identically normally distributed with mean zero and variance σ^2 so that (10) is equivalent to

$$f_\theta(\theta_n;\mathbf{Y}_n, \mathrm{b}, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}(\theta_n - \mathbf{Y}_n^T\beta)^T(\theta_n - \mathbf{Y}_n^T\beta)\right] \tag{11}$$

a normal distribution with mean $\mathbf{Y}_n^T\beta$ and variance σ^2. If (11) is used as the population model then the parameters to be estimated are β, σ^2, and ξ.

The TIMSS scaling model takes the generalization one step further by applying it to the vector valued θ rather than the scalar valued θ, resulting in the multivariate population model

$$f_{\theta}(\theta_n;\mathbf{W}_n,\gamma,\Sigma) = (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\theta_n-\gamma\mathbf{W}_n)^T\Sigma^{-1}(\theta_n-\gamma\mathbf{W}_n)\right] \tag{12}$$

where γ is a $u \times D$ matrix of regression coefficients, Σ is a $D \times D$ variance-covariance matrix and $\mathbf{W}_n$ is a $u \times 1$ vector of fixed variables. If (12) is used as the population model then the parameters to be estimated are γ, Σ, and ξ. In TIMSS we refer to the W_n variables as conditioning variables.

7.3 ESTIMATION

The ConQuest software uses maximum likelihood methods to provide estimates of γ, Σ, and ξ. Combining the conditional item response model (6) and the population model (12) we obtain the unconditional or marginal response model

$$f_{\mathrm{x}}(\mathbf{x};\xi,\gamma,\Sigma) = \int_{\theta} f_{\mathrm{x}}(\mathbf{x};\xi|\theta)f_{\theta}(\theta;\gamma,\Sigma)d\theta \tag{13}$$

and it follows that the likelihood is

$$\Lambda = \prod_{n=1}^{N} f_{\mathrm{x}}(\mathbf{x}_n;\xi,\gamma,\Sigma) \tag{14}$$

where N is the total number of sampled students.

Differentiating with respect to each of the parameters and defining the marginal posterior as

$$h_{\theta}(\theta_n;\mathbf{W}_n\xi,\gamma,\Sigma|\mathbf{x}_n) = \frac{f_{\mathrm{x}}(\mathbf{x}_n;\xi|\theta_n)f_{\theta}(\theta_n;\mathbf{W}_n,\gamma,\Sigma)}{f_{\mathrm{x}}(\mathbf{x}_n;\mathbf{W}_n\xi,\gamma,\Sigma)} \tag{15}$$

provides the following system of likelihood equations:

$$\mathbf{A}'\sum_{n=1}^{N}\left[\mathbf{x}_n - \int_{\theta_n} E_{\mathbf{z}}(\mathbf{z}|\theta_n)h_{\theta}(\theta_n;\mathbf{Y}_n,\xi,\gamma,\Sigma|\mathbf{x}_n)d\theta_n\right] = \mathbf{0} \tag{16}$$

$$\hat{\gamma} = \left(\sum_{n=1}^{N}\bar{\theta}_n\mathbf{W}_n^T\right)\left(\sum_{n=1}^{N}\mathbf{W}_n\mathbf{W}_n^T\right)^{-1} \tag{17}$$

and

$$\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\int_{\theta_n}(\theta_n-\gamma\mathbf{W}_n)(\theta_n-\gamma\mathbf{W}_n)^T h_{\theta}(\theta_n;\mathbf{Y}_n,\xi,\gamma,\Sigma|\mathbf{x}_n)d\theta_n \tag{18}$$

where

$$E_z(\mathbf{z}|\theta_n) = \Psi(\theta_n, \xi) \sum_{\mathbf{z} \in \Omega} \mathbf{z} \exp[\mathbf{z}'(\mathbf{b}\theta_n + \mathbf{A}\xi)] \tag{19}$$

and

$$\bar{\theta}_n = \int_{\theta_n} \theta_n h_\theta(\theta_n; \mathbf{Y}_n, \xi, \gamma, \Sigma | \mathbf{x}_n) d\theta_n \tag{20}$$

The system of equations defined by (16), (17), and (18) is solved using an EM algorithm (Dempster, Laird, and Rubin, 1977) following the approach of Bock and Aitken (1981).

7.3.1 Quadrature and Monte Carlo Approximations

The integrals in equations (16), (17) and (18) are approximated numerically using either quadrature or Monte Carlo methods. In each case we define, Θ_p, p=1,...,P a set of P D-dimensional vectors (which we call nodes), and for each node we define a corresponding weight $W_p(\gamma,\Sigma)$. The marginal item response probability (13) is then approximated using

$$f_{\mathbf{x}}(\mathbf{x};\xi, \gamma, \Sigma) = \sum_{p=1}^{P} f_{\mathbf{x}}(\mathbf{x};\xi|\Theta_p) W_p(\gamma, \Sigma) \tag{21}$$

and the marginal posterior (15) is approximated using

$$h_\Theta(\Theta_q; \mathbf{W}_n, \xi, \gamma, \Sigma|\mathbf{x}_n) = \frac{f_{\mathbf{x}}(\mathbf{x}_n;\xi|\Theta_q) W_q(\gamma, \Sigma)}{\sum_{p=1}^{P} f_{\mathbf{x}}(\mathbf{x}_n;\xi|\Theta_p) W_p(\gamma, \Sigma)} \tag{22}$$

for q=1,...,P.

The EM algorithm then proceeds as follows:

Step 1. Prepare a set of nodes and weights depending upon $\gamma^{(t)}$ and $\Sigma^{(t)}$ the estimates of γ and Σ at iteration t.

Step 2. Calculate the discrete approximation of the marginal posterior density of θ_n given $\mathbf{x}_n$ at iteration t using

$$h_\Theta(\Theta_p; \mathbf{W}_n, \xi^{(t)}\gamma^{(t)}, \Sigma^{(t)}|\mathbf{x}_n) = \frac{f_{\mathbf{x}}(\mathbf{x}_n;\xi^{(t)}|\Theta_p) W_p(\gamma^{(t)}, \Sigma^{(t)})}{\sum_{p=1}^{P} f_{\mathbf{x}}(\mathbf{x}_n;\xi^{(t)}|\Theta_p) W_p(\gamma^{(t)}, \Sigma^{(t)})} \tag{23}$$

where $\xi^{(t)}$, $\gamma^{(t)}$, $\Sigma^{(t)}$ and are estimates of $\xi^{(t)}$, $\gamma^{(t)}$, and $\Sigma^{(t)}$ at iteration t.

Step 3. Use a Newton-Raphson method to solve the following to produce estimates of $\hat{\xi}^{(t+1)}$.

$$A' \sum_{n=1}^{N} \left[\mathbf{x}_n - \sum_{r=1}^{P} E_{\mathbf{z}}(\mathbf{z}|\Theta_r) h_{\Theta}(\Theta_r; \mathbf{W}_n, \xi^{(t)}, \gamma^{(t)}, \Sigma^{(t)} | \mathbf{x}_n) \right] = \mathbf{0} \tag{24}$$

Step 4. Estimate $\gamma^{(t+1)}$ and $\Sigma^{(t+1)}$ using

$$\hat{\gamma}^{(t+1)} = \left(\sum_{n=1}^{N} \bar{\Theta}^n \mathbf{W}_n^T \right) \left(\sum_{n=1}^{N} \mathbf{W}_n \mathbf{W}_n^T \right)^{-1} \tag{25}$$

and

$$\hat{\Sigma}^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} \sum_{r=1}^{P} (\Theta_r - \gamma^{(t+1)} \mathbf{W}_n)(\Theta_r - \gamma^{(t+1)} \mathbf{W}_n)^T h_{\Theta}(\Theta_r; \mathbf{Y}_n, \xi^{(t)}, \gamma^{(t)}, \Sigma^{(t)} | \mathbf{x}_n) \tag{26}$$

where

$$\bar{\Theta}^n = \sum_{r=1}^{P} \Theta_r h_{\Theta}(\Theta_r; \mathbf{W}_n, \xi^{(t)}, \gamma^{(t)}, \Sigma^{(t)} | \mathbf{x}_n) \tag{27}$$

Step 5. Return to Step 1.

The difference between the quadrature and Monte Carlo methods lies in the way the nodes and weights are prepared. For the quadrature case we begin by choosing a fixed set of Q points, $(\Theta_{d1}, \Theta_{d2}, \ldots, \Theta_{dQ})$ for each latent dimension and then define a set of Q^D nodes that are indexed r = $1, \ldots, Q^D$, and are given by the Cartesian coordinates

$$\Theta_r = (\Theta_{1j_1}, \Theta_{2j_2}, \ldots, \Theta_{dj_d}) \text{ with } j_1 = 1, \ldots Q; j_2 = 1, \ldots, Q; \ldots; j_d = 1, \ldots, Q \quad .$$

The weights are then chosen to approximate the continuous latent population density (12). That is,

$$W_p = K(2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\Theta_p - \gamma \mathbf{W}_n)^T \Sigma^{-1} (\Theta_p - \gamma \mathbf{W}_n) \right] \tag{28}$$

where K is a scaling factor to ensure that the sum of the weights is 1.

In the Monte Carlo case the nodes are drawn at random from the standard multivariate normal distribution and at each iteration the nodes are rotated using standard methods so that they become random draws from a multivariate normal distribution with mean $\gamma \mathbf{W}_n$ and variance Σ. In the Monte Carlo case the weight for all nodes is $1/P$.

For further information on the quadrature approach to estimating the model see Adams, Wilson, and Wang (1997), and for further information on the Monte Carlo method see Volodin and Adams (1997). In the TIMSS scaling the Bock-Aitken quadrature approach was used for unidimensional models and the Volodin Monte Carlo methods was used when scaling in high dimensions.

7.3.2 Latent Estimation and Prediction

The marginal item response (13) does not include parameters for the latent values θ_n and hence the estimation algorithm does not result in estimates of the latent values. For TIMSS, expected a posteriori estimates (EAP) of each student's latent achievement was produced. The EAP prediction of the latent achievement for case n is

$$\theta_n^{EAP} = \sum_{r=1}^{P} \Theta_r h_\Theta(\Theta_r; \mathbf{W}_n, \hat{\xi}, \hat{\gamma}, \hat{\Sigma} | \mathbf{x}_n) \tag{29}$$

Variance estimates for these predictions were estimated using

$$\mathrm{var}(\theta_n^{EAP}) = \sum_{r=1}^{P} (\Theta_r - \theta_n^{EAP})(\Theta_r - \theta_n^{EAP})^T h_\Theta(\Theta_r; \mathbf{W}_n \hat{\xi}, \hat{\gamma}, \hat{\Sigma} | \mathbf{x}_n) \tag{30}$$

7.3.3 Drawing Plausible Values

Plausible values are random draws from the marginal posterior of the latent distribution, (15), for each student. For details on the use of plausible values the reader is referred to Mislevy (1991) and Mislevy et al. (1992).

Unlike previously described methods for drawing plausible values (Beaton, 1987; Mislevy et al., 1992) ConQuest does not assume normality of the marginal posterior distributions. Recall from (15) that the marginal posterior is given by

$$h_\theta(\theta_n; \mathbf{W}_n, \xi, \gamma, \Sigma | \mathbf{x}_n) = \frac{f_{\mathbf{x}}(\mathbf{x}_n; \xi | \theta_n) f_\theta(\theta_n; \mathbf{W}_n, \gamma, \Sigma)}{\int_\theta f_{\mathbf{x}}(\mathbf{x}; \xi | \theta) f_\theta(\theta, \gamma, \Sigma) d\theta} \tag{31}$$

The ConQuest procedure begins drawing M vector valued random deviates, $\{\varphi_{nm}\}_{m=1}^{M}$ from the multivariate normal distribution $f\theta(\theta_n, \mathbf{W}_n\gamma, \Sigma)$ for each case n. These vectors are used to approximate the integral in the denominator of (31) using the Monte Carlo integration

$$\int_\theta f_{\mathbf{x}}(\mathbf{x}; \xi | \theta) f_\theta(\theta, \gamma, \Sigma) d\theta \approx \frac{1}{M} \sum_{m=1}^{M} f_{\mathbf{x}}(\mathbf{x}; \xi | \varphi_{mn}) \equiv \Im \tag{32}$$

At the same time the values

$$p_{mn} = f_{\mathbf{x}}(\mathbf{x}_n;\xi|\varphi_{mn})f_{\theta}(\varphi_{mn};\mathbf{W}_n,\gamma,\Sigma) \tag{33}$$

are calculated, so that we obtain the set of pairs $\left(\varphi_{nm}, \frac{p_{mn}}{\mathfrak{I}}\right)_{m=1}^{M}$ which can be used as an approximation to the posterior density (31). The probability that φ_{nj} could be drawn from this density is given by

$$q_{nj} = \frac{p_{mn}}{\sum_{m=1}^{M} p_{mn}} \tag{34}$$

At this point, L uniformly distributed random numbers, $\{\eta_i\}_{i=1}^{L}$, are generated and for each random draw the vector φ_{ni_0} that satisfies the condition

$$\sum_{s=1}^{i_0-1} q_{sn} < \eta_i \leq \sum_{s=1}^{i_0} q_{sn} \tag{35}$$

is selected as a plausible vector.

7.4 SCALING STEPS

The item response model described above was fit to the data in two steps. In the first step the items were calibrated using a subsample of students drawn from the samples of the participating countries. These samples were called the *international calibration samples.* In a second step the model was fit separately for each country with the item parameters fixed at values estimated in the first step.

There were three principal reasons for using an international calibration sample for estimating international item parameters. First, it seemed unnecessary to estimate parameters using the complete data set; second, drawing equal-sized subsamples from each country for inclusion in the international calibration sample ensured that each country was given equal weight in the estimation of the international parameters; and third, the drawing of appropriately weighted samples meant that weighting would not be necessary in the international scaling runs.

7.4.1 Drawing the International Calibration Sample

At the time when the international scaling of the data commenced the TIMSS database of item response data contained information from 25 Population 1 countries and 39 Population 2 countries. Those countries are listed in Table 7.1.

For each target population, samples of 600 tested students were selected from the database for each participating country. This generally lead to roughly equal samples from each target grade. For Israel, where only the upper grade was tested, the sample size was reduced to 300 tested students. The sampled students were selected using a probability-proportional-to-size systematic selection method. The overall sampling weights were used as measures of size for this purpose. This resulted in equal selection probabilities, within national samples, for the students in the calibration samples. The Population 1 and 2 international calibration samples contained 14,700 and 23,100 students, respectively.

Table 7.1 Countries Included in the International Item Calibration

Population 1[1]	Population 2[2]	
Australia	Australia	Netherlands
Austria	Austria	New Zealand
Canada	Belgium (Flemish)	Norway
Cyprus	Belgium (French)	Portugal
Czech Republic	Bulgaria	Romania
England	Canada	Russian Federation
Greece	Colombia	Scotland
Hong Kong	Cyprus	Singapore
Hungary	Czech Republic	Slovak Republic
Iceland	Denmark	Slovenia
Iran	England	Spain
Ireland	France	Sweden
Israel*	Germany	Switzerland
Japan	Greece	United States
Korea	Hong Kong	
Latvia	Hungary	
Mexico	Iceland	
Netherlands	Iran	
New Zealand	Ireland	
Norway	Israel*	
Portugal	Japan	
Scotland	Korea	
Singapore	Latvia	
Slovenia	Lithuania	
United States	Mexico	

*A sample of 600 students was drawn from each country, excepting for Israel where only 300 students where drawn because Israel sampled students from only the higher of the two grade levels.

Note: Mexico's data was used to estimate the international item parameters, although Mexico subsequently withdrew its results from the international reports. Although results for Kuwait, the Philippines, and South Africa were reported in the international reports, their data were not used to estimate the international parameters.

[1] Third and fourth grades in most countries.

[2] Seventh and eighth grades in most countries.

7.4.2 International Scaling Results

Tables 7.2, 7.3, 7.4, and 7.5 show the basic statistics that resulted from international scaling for mathematics and science at Populations 1 and 2. The number of respondents shown for each item is the number of cases that were considered valid for calibration purposes. There are two reasons why this value is not equal to the total number of students in the calibration samples. First, the test rotation design was such that only items in cluster A were administered to all students (see Adams and Gonzalez, 1996), and second, items to which students did not respond because there were deemed to be "not reached" were treated as missing data in the calibration phase of the analysis. The percent correct figures that are reported were computed by summing the total of the scores achieved by all students who provided valid responses and dividing that by the number of students multiplied by the maximum score that could be achieved for that item; for most but not all items, the maximum possible scores was one. The difficulty estimate and asymptotic errors are in the logit metric, which is the natural metric for the ConQuest scaling software. The mean square fit statistic is an index of the fit of the data to the assumed scaling model; the statistic used here was derived by Wu (1997). Under the null hypothesis that the data and model are consistent, the expected value of these statistics is one. Values that are less than one are usually associated with items that have greater than average discrimination, while values that are greater than one may result from lower than average discrimination, guessing, or some other deviation from the model.

7.4.3 Fit of the Scaling Model

Tables 7.1 and 7.2 show the international item statistics and parameter estimates for Population 1 mathematics and science, respectively. Table 7.3 and 7.4 show the corresponding information for Population 2. The mean square fit statistics reported in Tables 7.2, 7.3, 7.4, and 7.5 show that the vast majority of items fit the Rasch model very well. Items with mean squares greater than one in Population 1 mathematics were B06, H05, I08, K07, M07, U01, and V04a. The reasons for the misfit of these items vary. For item B06, misfit is caused by the fact that the item does not discriminate as well as the other items. This may be seen in Figure 7.1 showing the modeled and empirical item characteristic curves for this item. For item H05, the modeled and empirical item characteristics curves are shown in Figure 7.2. There appear to be two reasons for this misfit of this item: first, it is slightly less discriminating than was assumed by the model; but second, interestingly, some students in the middle of the latent ability distribution did not perform as well as was expected, and these students would receive considerable weight in the estimation of the weighted mean square. Item K07 (Figure 7.3) was amongst the most difficult items, and it was multiple choice, so it is not surprising that some students are likely to have attempted to guess the correct response. A closer review of the item shows that one of the distracters proved to be attractive to some higher-achieving students – in fact the point-biserial for the distracter is positive for quite a few countries. This item survived the review process because of a policy decision to retain as many items as possible. Items M07, U01, and V04a all misfit because they had a slightly lower than modeled discrimination.

The items that had mean square statistics less than one were all found to be more discriminating than was modeled. Misfit of this form is not usually deemed to be of concern. Interestingly, however, the majority of the most discriminating items are short-answer or extended-response type. This may well be due to the fact that it is unlikely that students would have guessed the answers to these questions.

Table 7.2 Population 1 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample

Item Label	Number of Respondents in International Calibration Sample	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASMMA01	14637	79.78	-1.345	0.022	1.03
ASMMA02	14627	51.47	0.218	0.018	1.08
ASMMA03	14618	58.61	-0.131	0.018	1.01
ASMMA04	14603	78.33	-1.242	0.022	0.95
ASMMA05	14571	79.29	-1.307	0.022	1.06
ASMMB05	7283	60.66	-0.233	0.026	0.99
ASMMB06	7268	48.98	0.343	0.026	1.13
ASMMB07	7259	40.52	0.761	0.026	1.06
ASMMB08	7247	82.03	-1.511	0.033	0.94
ASMMB09	7240	56.66	-0.029	0.026	0.99
ASMMC01	5460	80.44	-1.401	0.037	0.89
ASMMC02	5447	64.79	-0.448	0.031	0.96
ASMMC03	5441	48.87	0.350	0.030	0.99
ASMMC04	5437	75.19	-1.040	0.034	0.90
ASMMD05	5463	74.81	-1.007	0.034	0.93
ASMMD06	5451	51.73	0.213	0.030	1.09
ASMMD07	5438	83.41	-1.609	0.039	0.94
ASMMD08	5428	55.56	0.031	0.030	0.99
ASMMD09	5411	47.92	0.405	0.030	1.01
ASMME01	5439	63.28	-0.373	0.031	0.98
ASMME02	5420	71.25	-0.807	0.033	0.90
ASMME03	5425	59.54	-0.177	0.031	0.96
ASMME04	5413	31.98	1.216	0.032	1.10
ASMMF05	5471	66.02	-0.515	0.031	0.97
ASMMF06	5459	59.11	-0.161	0.030	1.09
ASMMF07	5451	59.70	-0.189	0.030	0.99
ASMMF08	5445	60.39	-0.223	0.030	1.04
ASMMF09	5437	68.81	-0.662	0.032	1.03
ASMMG01	5181	36.33	0.987	0.032	1.07
ASMMG02	5170	69.83	-0.707	0.033	1.06
ASMMG03	5152	87.36	-1.990	0.044	0.93
ASMMG04	5365	47.83	0.414	0.030	1.01
ASMMH05	5434	45.99	0.464	0.030	1.15
ASMMH06	5422	48.51	0.345	0.030	1.06
ASMMH07	5407	66.08	-0.518	0.031	1.01
ASMMH08	5394	64.29	-0.422	0.031	0.99
ASMMH09	5383	65.11	-0.464	0.031	1.00
ASMMI01	1864	49.79	0.306	0.051	1.01
ASMMI02	1859	31.47	1.239	0.055	1.07
ASMMI03	1859	53.68	0.118	0.051	1.03
ASMMI04	1859	81.33	-1.461	0.064	0.90
ASMMI05	1859	46.26	0.480	0.051	0.98
ASMMI06	1858	69.48	-0.697	0.055	0.97
ASMMI07	1858	54.36	0.086	0.051	0.89
ASMMI08	1854	49.30	0.334	0.051	1.17
ASMMI09	1853	61.09	-0.247	0.052	1.00
ASMMJ01	1814	84.45	-1.768	0.069	0.94

Table 7.2 Population 1 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 1)

Item Label	Number of Respondents in International Calibration Sample	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASMMJ02	1811	60.41	-0.237	0.053	1.07
ASMMJ03	1728	68.34	-0.698	0.057	0.85
ASMMJ04	1804	39.36	0.831	0.054	0.97
ASMMJ05	1724	36.31	0.966	0.056	1.05
ASMMJ06	1799	66.54	-0.555	0.055	1.08
ASMMJ07	1796	53.73	0.112	0.053	0.97
ASMMJ08	1793	44.28	0.588	0.053	0.99
ASMMJ09	1783	71.34	-0.817	0.058	0.99
ASMMK01	1803	62.17	-0.287	0.054	1.07
ASMMK02	1797	77.13	-1.147	0.061	0.98
ASMMK03	1796	45.38	0.560	0.053	0.98
ASMMK04	1718	44.88	0.628	0.054	0.93
ASMMK05	1787	73.81	-0.928	0.059	1.07
ASMMK06	1784	58.18	-0.069	0.053	0.99
ASMMK07	1779	22.60	1.848	0.062	1.18
ASMMK08	1771	68.83	-0.629	0.056	0.91
ASMMK09	1764	35.43	1.088	0.055	1.04
ASSML01	1791	42.16	0.699	0.053	0.85
ASMML02	1789	45.95	0.515	0.052	1.01
ASMML03	1714	83.84	-1.638	0.070	0.97
ASMML04	1784	66.70	-0.512	0.055	0.94
ASMML05	1782	38.61	0.886	0.053	1.07
ASMML06	1705	47.39	0.423	0.053	1.06
ASMML07	1772	41.25	0.755	0.053	0.89
ASMML08	1767	28.13	1.459	0.058	0.96
ASMML09	1761	59.68	-0.144	0.053	1.02
ASMMM01	1829	73.32	-0.905	0.057	1.04
ASSMM02	1826	47.21	0.415	0.051	0.91
ASMMM03	1819	58.71	-0.133	0.052	1.03
ASSMM04	1811	36.33	0.953	0.053	1.01
ASMMM05	1805	36.95	0.923	0.053	1.13
ASMMM06	1798	65.46	-0.463	0.054	0.95
ASMMM07	1788	36.86	0.934	0.053	1.15
ASMMM08	1779	84.54	-1.664	0.069	0.92
ASMMM09	1778	65.75	-0.472	0.054	0.93
ASEMS01	3501	43.32	0.600	0.024	1.01
ASSMS02	3356	59.06	-0.068	0.039	0.86
ASEMS03	3297	28.45	1.280	0.026	0.95
ASSMS04	3096	48.87	0.496	0.040	0.91
ASSMS05	3016	52.06	0.360	0.040	1.08
ASEMT01	3336	70.92	-0.734	0.042	0.85
ASEMT01	3266	40.55	0.760	0.025	0.94
ASSMT02	3328	41.17	0.827	0.039	0.94
ASSMT03	3257	45.96	0.601	0.039	0.92
ASEMT04a	3082	18.23	2.231	0.050	0.91
ASEMT04b	2984	12.40	2.771	0.059	0.89
ASSMT05	3033	62.45	-0.179	0.041	0.99

Table 7.2 Population 1 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 2)

Item Label	Number of Respondents in International Calibration Sample	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASEMU01	3483	53.37	0.209	0.023	1.19
ASSMU02	3418	51.20	0.300	0.038	1.10
ASEMU03a	3323	58.47	-0.035	0.039	0.84
ASEMU03b	3274	40.16	0.877	0.039	0.83
ASEMU03c	3250	75.75	-0.975	0.044	0.98
ASSMU04	3237	54.71	0.167	0.039	0.94
ASSMU05	3152	80.43	-1.277	0.048	0.95
ASEMV01	3486	37.74	0.872	0.026	1.04
ASSMV02	3438	41.97	0.711	0.038	0.87
ASSMV03	3347	60.86	-0.188	0.039	0.88
ASEMV04a	3305	49.26	0.390	0.030	1.16
ASEMV04b	3232	45.39	0.578	0.039	0.90
ASSMV05	3104	47.29	0.515	0.039	0.99

Figure 7.1 Empirical and Modelled Item Characteristic Curves for Mathematics Population 1 Item: ASMMB06. Fit MNSQ=1.13

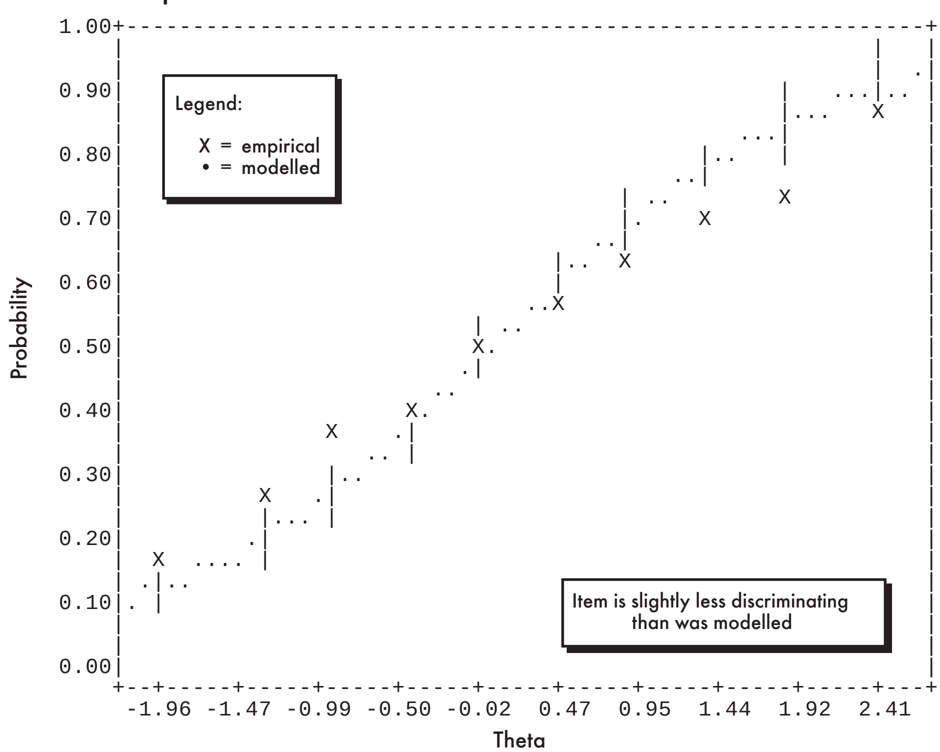


Figure 7.2 Empirical and Modelled Item Characteristic Curves for Mathematics Population 1 Item: ASEMH05. Fit MNSQ=1.15

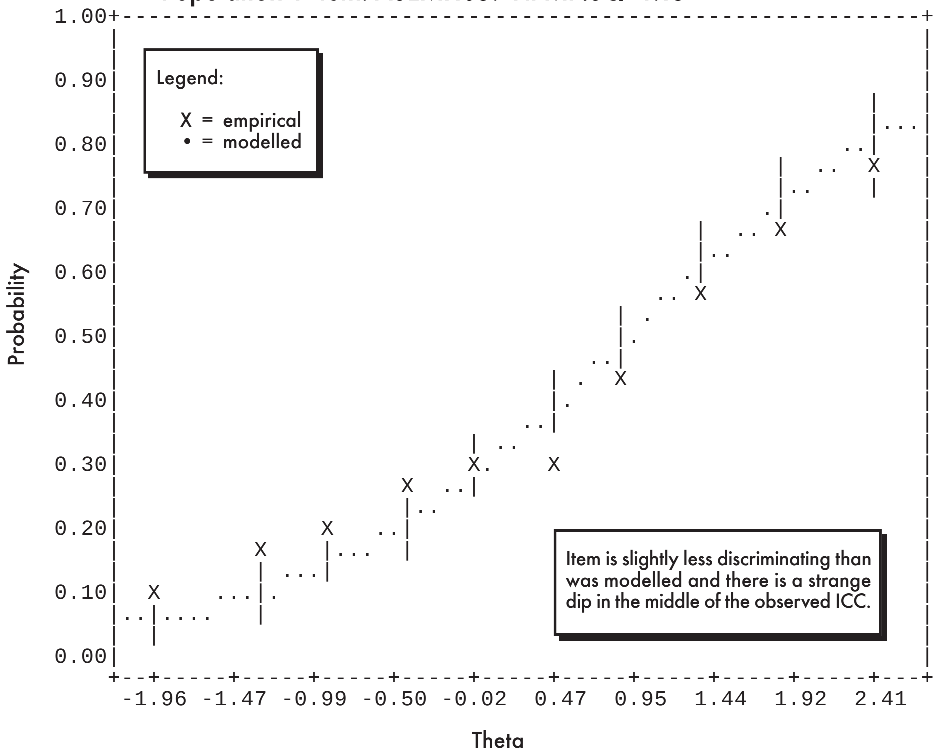
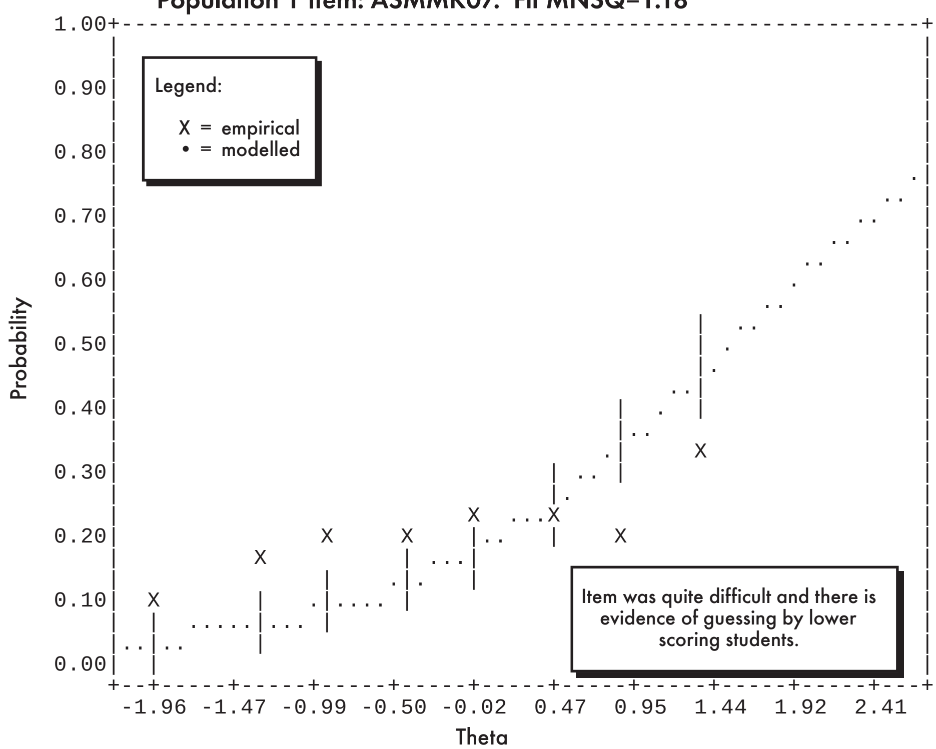


Figure 7.3 Empirical and Modelled Item Characteristic Curves for Mathematics Population 1 Item: ASMMK07. Fit MNSQ=1.18



For Population 1 science there are fewer misfitting items than for mathematics. The worst-fitting items are G08, H04, O05, and R03. Item G08 (Figure 7.4) was a difficult item and its misfit may be caused by guessing, while the misfit of H04 is caused by lower than modeled discrimination. An examination of the country-level data for these items shows that both have distracters that have positive point-biserials in a number of countries. The misfit of items O05 and R03, which is illustrated in Figures 7.5 and 7.6, is more difficult to explain – both of these items performed quite well in each of the participating countries. Item O05 does show some evidence of lower than expected discrimination, but there is also a large “blip” in the observed percentage correct for student in the second-lowest ability grouping. Item R03 has fewer than expected students at the upper achievement levels. An examination of the item-by-country interactions shows that students in the countries that had high average scores found this item more difficult than expected. Notably, this was the case in Japan, Korea, Singapore and Hong Kong, while the item was easier than expected in Slovenia, Hungary, and the Czech Republic.

Table 7.3 Population 1 Science: Item Statistics and Parameter Estimates for the International Calibration Sample

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASMSA06	14554	84.15	-1.546	0.024	0.97
ASMSA07	14516	82.86	-1.439	0.023	0.96
ASMSA08	14501	62.02	-0.197	0.018	1.02
ASMSA09	14478	66.13	-0.404	0.019	1.00
ASMSA10	14454	74.32	-0.856	0.020	0.96
ASMSB01	7312	70.69	-0.658	0.028	0.99
ASMSB02	7298	72.06	-0.734	0.028	0.96
ASMSB03	6987	68.33	-0.524	0.028	0.98
ASMSB04	6975	57.49	0.029	0.026	1.08
ASMSC05	5429	68.06	-0.494	0.031	1.01
ASMSC06	5420	91.09	-2.248	0.049	0.97
ASMSC07	5410	43.51	0.694	0.030	1.05
ASMSC08	5387	81.60	-1.322	0.037	0.97
ASMSC09	5380	48.83	0.448	0.029	1.06
ASMSD01	5282	58.54	-0.026	0.030	0.96
ASMSD02	5483	68.19	-0.512	0.031	1.01
ASMSD03	5479	88.37	-1.944	0.044	0.95
ASMSD04	5474	72.84	-0.770	0.033	1.02
ASMSE05	5411	61.50	-0.193	0.030	0.99
ASMSE06	5401	92.33	-2.467	0.053	0.98
ASMSE07	5399	57.47	0.004	0.030	0.96
ASSSE08	5364	73.60	-0.832	0.033	0.98
ASMSE09	5349	43.30	0.678	0.030	1.07
ASMSF01	5507	84.60	-1.610	0.039	0.99
ASMSF02	5495	65.13	-0.375	0.031	1.06
ASMSF03	5491	61.28	-0.181	0.030	0.97
ASMSF04	5481	68.87	-0.569	0.031	0.93
ASMSG05	5356	79.41	-1.191	0.036	0.98
ASMSG06	5354	86.33	-1.743	0.042	1.03
ASMSG07	5346	70.61	-0.649	0.032	0.99
ASMSG08	5338	26.60	1.548	0.033	1.14
ASMSG09	5323	86.02	-1.709	0.042	0.94
ASMSH01	5460	77.34	-1.059	0.034	1.03
ASMSH02	5449	83.56	-1.506	0.039	0.93
ASSSH03	5438	39.48	0.870	0.030	0.96
ASMSH04	5434	42.69	0.717	0.030	1.21
ASMSN01	1862	38.56	0.933	0.051	1.03
ASMSN02	1860	69.68	-0.581	0.054	0.94
ASMSN03	1858	40.90	0.820	0.051	1.10
ASMSN04	1855	32.40	1.246	0.053	1.10
ASMSN05	1850	70.81	-0.642	0.055	1.05
ASMSN06	1849	40.02	0.866	0.051	1.01
ASMSN07	1842	51.09	0.346	0.050	1.03
ASMSN08	1838	58.60	-0.008	0.051	1.05
ASMSN09	1832	59.55	-0.052	0.051	0.99
ASMSO01	1829	42.81	0.676	0.051	1.03
ASMSO02	1825	38.03	0.912	0.052	1.07

Table 7.3 Population 1 Science: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 1)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASMSO03	1823	62.97	-0.281	0.052	0.96
ASMSO04	1820	66.76	-0.473	0.054	1.04
ASMSO05	1815	54.16	0.154	0.051	1.15
ASSSO06	1808	54.31	0.153	0.051	0.94
ASMSO07	1808	71.35	-0.708	0.056	1.08
ASMSO08	1804	51.39	0.297	0.051	1.04
ASSSO09	1783	38.70	0.909	0.052	0.93
ASMSP01	1780	83.99	-1.472	0.068	0.92
ASMSP02	1780	84.16	-1.480	0.068	0.89
ASMSP03	1777	32.86	1.254	0.054	1.07
ASSSP04	1773	31.64	1.323	0.055	1.01
ASMSP05	1768	45.76	0.630	0.052	1.00
ASMSP06	1698	33.69	1.214	0.055	1.01
ASMSP07	1765	39.60	0.929	0.052	0.99
ASMSP08	1763	53.66	0.269	0.052	0.95
ASMSP09	1759	42.30	0.807	0.052	1.02
ASMSQ01	1850	60.65	-0.071	0.051	0.95
ASMSQ02	1845	63.58	-0.211	0.052	1.07
ASMSQ03	1841	49.48	0.456	0.050	1.00
ASSSQ04	1834	58.34	0.044	0.051	0.92
ASMSQ05	1824	69.19	-0.494	0.054	1.02
ASMSQ06	1822	41.93	0.812	0.051	1.06
ASMSQ07	1819	40.74	0.870	0.051	1.03
ASSSQ08	1808	46.13	0.617	0.051	0.97
ASMSQ09	1795	52.70	0.318	0.051	1.00
ASSSR01	1830	18.80	2.106	0.063	1.03
ASMSR02	1814	39.14	0.944	0.052	1.04
ASMSR03	1805	55.62	0.173	0.051	1.15
ASMSR04	1797	73.46	-0.735	0.057	0.89
ASMSR05	1790	56.09	0.149	0.051	0.99
ASMSR06	1776	39.92	0.908	0.052	1.07
ASMSR07	1765	55.30	0.190	0.051	0.96
ASMSR08	1670	53.77	0.242	0.053	1.09
ASMSR09	1734	44.87	0.678	0.052	1.07
ASESW01	3512	55.25	0.178	0.026	1.09
ASSSW02	3431	38.41	0.989	0.038	0.95
ASSSW03	3218	26.51	1.658	0.043	1.00
ASSSW04	3143	50.81	0.458	0.038	0.88
ASESW05	2900	52.14	0.440	0.040	0.95
ASESW05	2747	36.77	1.188	0.042	0.95
ASESX01	3581	66.23	-0.721	0.031	0.90
ASSSX02	3557	73.35	-0.787	0.041	0.92
ASESX03	3471	42.64	0.736	0.025	0.97
ASSSX04	3397	82.31	-1.344	0.048	0.93
ASMSX05	3323	59.43	-0.009	0.038	1.07
ASESY01	3399	27.83	1.519	0.041	0.91
ASESY02	3258	66.73	-0.353	0.040	0.96

Table 7.3 Population 1 Science: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 2)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASESY02	3126	39.38	0.985	0.039	1.01
ASESY03	3021	65.54	-0.231	0.041	0.94
ASESY03	2884	45.67	0.725	0.040	0.95
ASESZ01	3479	58.47	0.026	0.037	0.94
ASESZ01	3406	20.35	1.996	0.045	1.02
ASESZ02	3390	63.54	-0.203	0.038	0.94
ASESZ03	3361	49.02	0.485	0.024	0.99

Figure 7.4 Empirical and Modelled Item Characteristic Curves for Science Population 1 Item: ASMSG08. Fit MNSQ=1.14

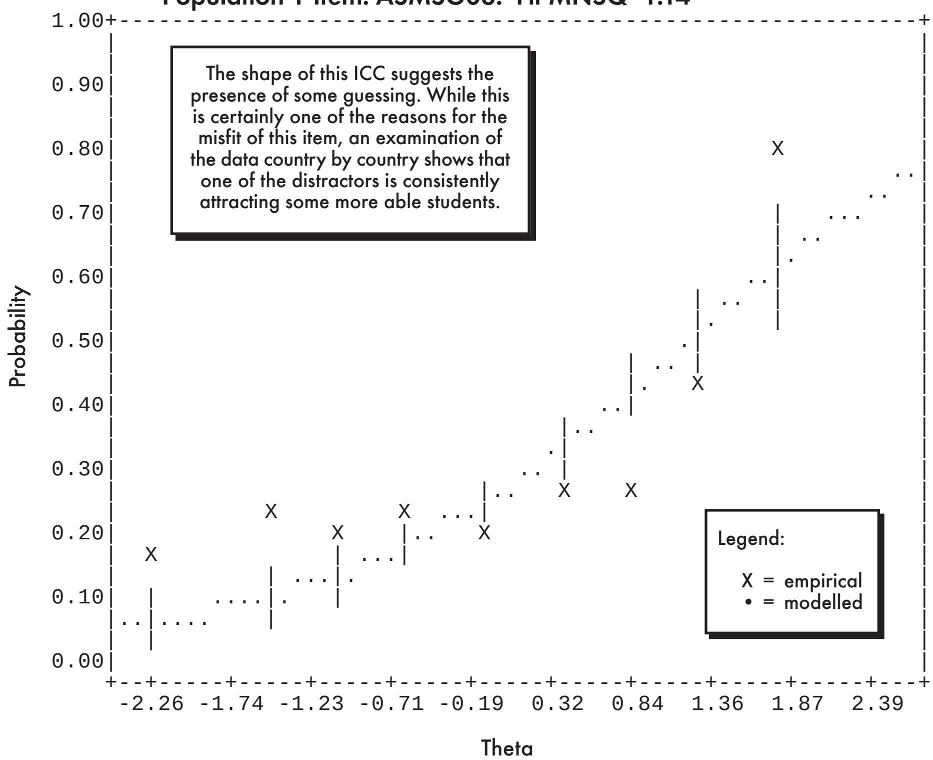


Figure 7.5 Empirical and Modelled Item Characteristic Curves for Science Population 1 Item: ASMSO05. Fit MNSQ=1.15

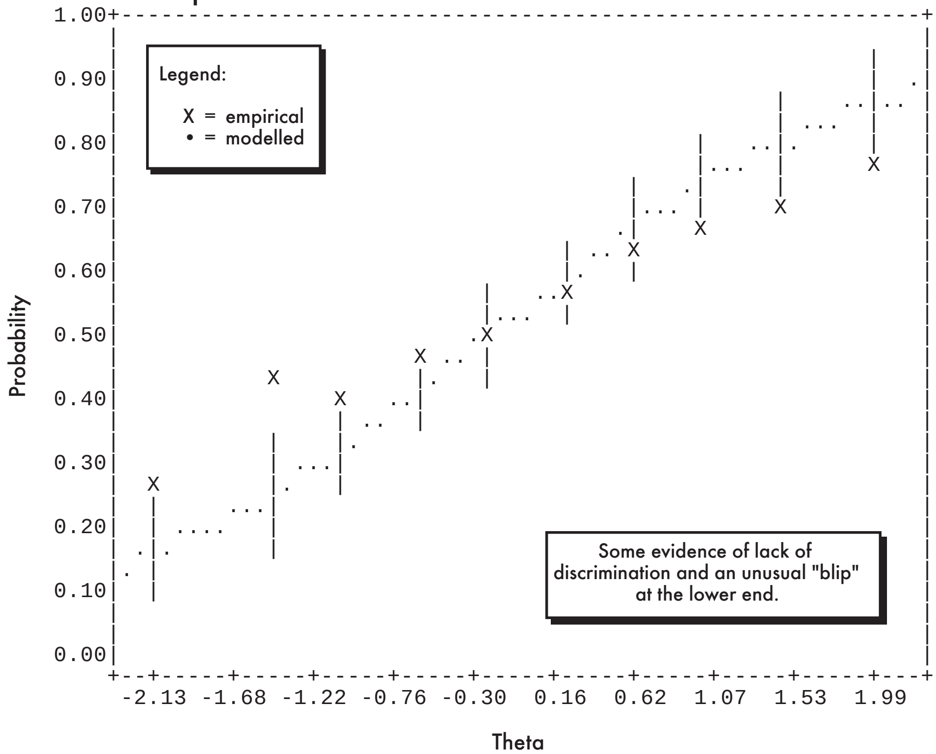
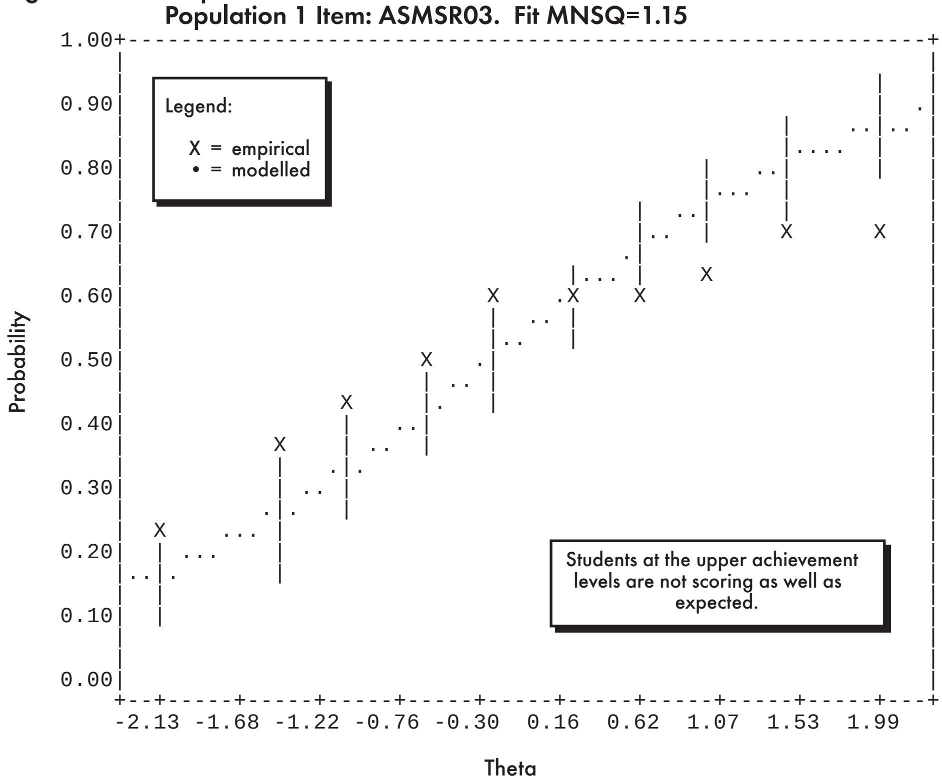


Figure 7.6 Empirical and Modelled Item Characteristic Curves for Mathematics Population 1 Item: ASMSR03. Fit MNSQ=1.15



The fit of the items for Population 2 mathematics is quite acceptable, although it is the least favorable of the four data sets. There are eight items with fit that is less than or equal to 0.85 – six of them short-answer or extended-response – and ten items with a fit greater than 1.15 – nine of them multiple-choice. For the items that have weighted fit mean squares greater than 1.15 the reason for that misfit is quite varied. Items I03, J18, L11, and N17 are all relatively difficult multiple-choice questions and exhibit evidence of guessing. As with the questions that showed elements of guessing characteristics in the Population 1 data sets, each of these items has a distracter that had a positive point-biserial in a large number of countries. Items L11 and N17 also showed bad fit in a number of countries. Item N16 is the only item with fit above 1.15, where it is reasonably clear that the misfit is due to the item having lower than modeled discrimination. The misfit for items B07, D10, and N15, which cannot be easily characterized, is illustrated in Figure 7.8, which shows the observed and expected item characteristic curves for item D10. Examining this item at the country level we note that in a number of countries it has a distracter with a positive point-biserial. This distracter has probably attracted some of the more able students, resulting in the empirical item characteristic curve being lower than the modeled curve for students toward the upper end of the achievement distribution. Plots for items B07, N15, and P09 show a similar pattern, but in examining the data we have not been able to find an explanation for the unusual shape of the observed item characteristic curve.

Table 7.4 Population 2 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMMA01	23039	56.91	-0.127	0.015	0.87
BSMMA02	23036	74.91	-1.120	0.017	1.02
BSMMA03	22437	61.57	-0.359	0.015	0.97
BSMMA04	23033	53.16	0.062	0.015	1.03
BSMMA05	23032	58.68	-0.217	0.015	1.07
BSMMA06	23026	76.54	-1.225	0.017	1.05
BSMMB07	11400	62.71	-0.421	0.022	1.01
BSMMB08	11389	68.82	-0.754	0.022	1.16
BSMMB09	11383	57.46	-0.148	0.021	1.08
BSMMB10	11377	49.42	0.258	0.021	0.85
BSMMB11	11372	63.32	-0.451	0.022	0.95
BSMMB12	11366	64.17	-0.496	0.022	0.91
BSMMC01	8671	57.57	-0.158	0.024	0.96
BSMMC02	8670	76.21	-1.196	0.027	0.95
BSMMC03	8667	60.61	-0.313	0.024	0.94
BSMMC04	8661	54.28	0.009	0.024	1.14
BSMMC05	8660	52.81	0.082	0.024	1.02
BSMMC06	8654	71.13	-0.883	0.026	1.03
BSMMD07	8773	60.97	-0.345	0.024	1.02
BSMMD08	8761	66.00	-0.610	0.025	1.02
BSMMD09	8756	66.71	-0.649	0.025	0.86
BSMMD10	8753	44.27	0.499	0.024	1.16
BSMMD11	8743	85.38	-1.906	0.032	1.02
BSMMD12	8740	72.15	-0.957	0.026	1.04
BSMME01	8715	69.15	-0.791	0.026	1.00
BSMME02	8710	44.32	0.492	0.024	0.99
BSMME03	8705	56.01	-0.096	0.024	1.00
BSMME04	8698	65.56	-0.590	0.025	0.95
BSMME05	8687	53.02	0.056	0.024	0.97
BSMME06	8679	40.45	0.693	0.024	0.95
BSMMF07	8615	30.47	1.243	0.026	1.03
BSMMF08	8606	59.57	-0.253	0.024	1.15
BSMMF09	8602	65.21	-0.546	0.025	0.99
BSMMF10	8596	53.52	0.052	0.024	1.01
BSMMF11	8398	45.83	0.444	0.024	0.89
BSMMF12	8583	54.44	0.007	0.024	1.07
BSMMG01	8638	52.92	0.072	0.024	1.15
BSMMG02	8633	75.98	-1.192	0.027	0.98
BSMMG03	8631	49.70	0.234	0.024	1.02
BSMMG04	8629	67.44	-0.682	0.025	0.95
BSMMG05	8622	58.37	-0.202	0.024	0.94
BSMMG06	8621	40.82	0.686	0.025	1.01
BSMMH07	8581	66.37	-0.613	0.025	1.04
BSMMH08	8575	73.84	-1.046	0.027	0.92
BSMMH09	8570	84.75	-1.837	0.032	0.96
BSMMH10	8564	43.57	0.559	0.024	0.97
BSMMH11	8549	60.51	-0.297	0.025	0.97

Table 7.4 Population 2 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 1)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMMH12	8536	73.11	-0.996	0.027	0.94
BSMMI01	2884	34.71	1.008	0.044	1.09
BSMMI02	2883	55.64	-0.070	0.042	0.94
BSMMI03	2882	39.73	0.738	0.043	1.16
BSSMI04	2880	42.43	0.597	0.042	1.03
BSMMI05	2877	72.54	-0.981	0.046	0.95
BSSMI06	2876	76.29	-1.217	0.048	1.09
BSMMI07	2877	63.30	-0.464	0.043	1.13
BSMMI08	2871	41.14	0.666	0.043	1.12
BSMMI09	2872	64.28	-0.516	0.043	0.94
BSMMJ10	2937	40.86	0.661	0.042	0.87
BSMMJ11	2865	45.62	0.405	0.042	1.08
BSSMJ12	2929	41.62	0.624	0.042	1.03
BSSMJ13	2928	81.69	-1.576	0.051	0.99
BSMMJ14	2930	43.38	0.536	0.041	1.02
BSMMJ15	2926	63.91	-0.486	0.042	1.07
BSMMJ16	2924	51.74	0.126	0.041	0.98
BSMMJ17	2916	65.84	-0.585	0.043	1.01
BSMMJ18	2913	41.57	0.631	0.042	1.18
BSMMK01	2958	68.80	-0.804	0.044	1.05
BSSMK02	2958	64.16	-0.549	0.042	0.98
BSMMK03	2956	65.93	-0.644	0.043	1.08
BSMMK04	2957	38.35	0.772	0.042	1.12
BSSMK05	2956	36.87	0.851	0.043	0.79
BSMMK06	2956	38.94	0.741	0.042	1.08
BSMMK07	2955	50.59	0.147	0.041	1.03
BSMMK08	2955	31.78	1.134	0.044	1.05
BSMMK09	2952	47.63	0.297	0.041	0.90
BSMML08	2857	57.75	-0.168	0.042	1.10
BSMML09	2857	84.35	-1.805	0.055	0.98
BSMML10	2855	86.76	-2.024	0.059	1.00
BSMML11	2854	32.20	1.146	0.044	1.24
BSMML12	2855	72.71	-0.976	0.046	0.93
BSMML13	2854	89.66	-2.332	0.065	0.98
BSMML14	2852	22.72	1.735	0.049	1.08
BSMML15	2845	35.75	0.959	0.044	1.02
BSSML16	2844	37.48	0.868	0.043	0.93
BSMML17	2745	47.14	0.379	0.043	0.92
BSMMM01	2832	85.56	-1.887	0.057	0.95
BSMMM02	2831	62.91	-0.410	0.043	1.13
BSMMM03	2830	75.69	-1.142	0.048	0.97
BSMMM04	2830	37.77	0.868	0.043	0.87
BSMMM05	2768	48.48	0.316	0.042	1.07
BSSMM06	2828	33.80	1.084	0.044	0.90
BSMMM07	2827	71.45	-0.878	0.046	1.01
BSSMM08	2827	46.44	0.427	0.042	1.10
BSMMN11	2831	82.20	-1.600	0.053	0.94

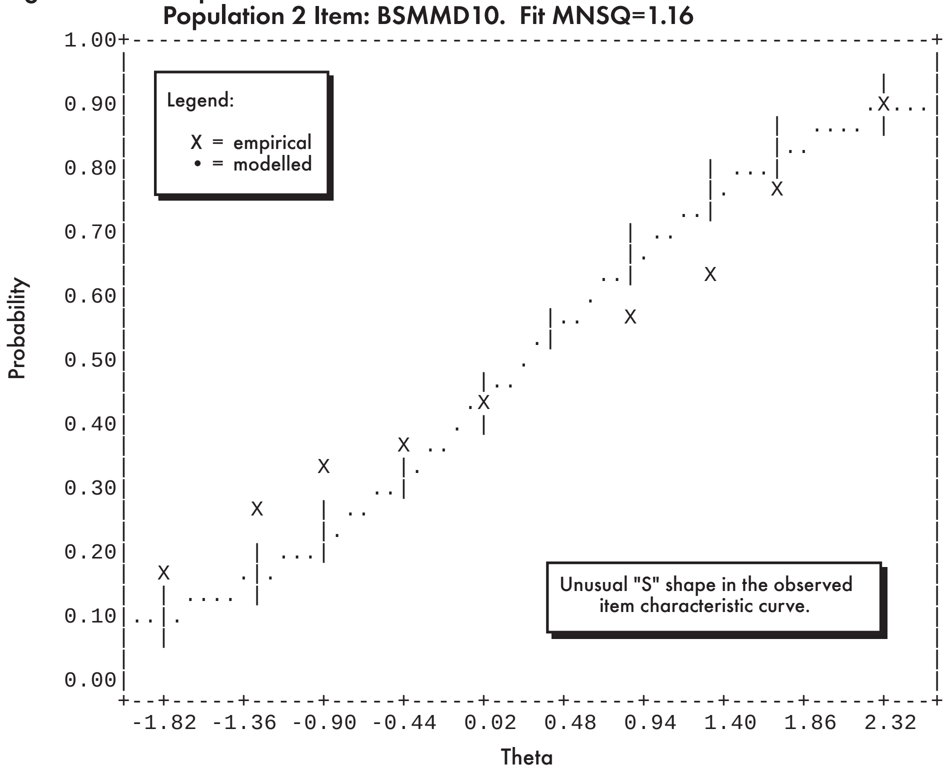
Table 7.4 Population 2 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 2)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMMN12	2829	65.04	-0.522	0.044	1.10
BSSMN13	2825	46.27	0.439	0.042	0.90
BSMMN14	2821	66.43	-0.593	0.044	0.96
BSMMN15	2822	64.56	-0.492	0.044	1.19
BSMMN16	2746	45.59	0.482	0.043	1.19
BSMMN17	2728	38.56	0.819	0.044	1.22
BSMMN18	2799	54.31	0.048	0.042	0.99
BSSMN19	2782	50.32	0.253	0.042	0.79
BSMMO01	2881	55.22	-0.015	0.042	0.98
BSMMO02	2879	25.81	1.589	0.047	0.92
BSMMO03	2881	45.33	0.489	0.042	0.99
BSMMO04	2808	44.16	0.555	0.043	1.08
BSMMO05	2881	43.91	0.562	0.042	0.85
BSSMO06	2879	69.78	-0.793	0.045	0.95
BSMMO07	2879	68.04	-0.694	0.044	1.01
BSMMO08	2877	66.28	-0.595	0.044	1.02
BSSMO09	2876	47.46	0.383	0.042	0.84
BSMMP08	2765	54.94	-0.005	0.043	0.98
BSMMP09	2764	37.34	0.895	0.044	1.16
BSMMP10	2757	54.12	0.037	0.043	0.96
BSMMP11	2754	53.96	0.044	0.043	1.15
BSMMP12	2752	70.17	-0.809	0.046	1.00
BSMMP13	2740	67.12	-0.636	0.045	0.95
BSMMP14	2736	77.60	-1.262	0.050	0.99
BSMMP15	2730	62.78	-0.401	0.044	0.98
BSSMP16	2720	33.57	1.110	0.045	0.90
BSMMP17	2674	84.89	-1.798	0.057	1.06
BSMMQ01	2784	41.81	0.651	0.043	1.07
BSMMQ02	2778	47.70	0.353	0.043	1.09
BSMMQ03	2776	33.43	1.104	0.045	1.05
BSMMQ04	2774	84.35	-1.777	0.056	1.01
BSMMQ05	2770	65.42	-0.549	0.044	1.03
BSMMQ06	2762	38.88	0.810	0.044	1.01
BSMMQ07	2752	58.76	-0.199	0.043	0.96
BSMMQ08	2746	43.81	0.556	0.043	0.86
BSMMQ09	2683	50.09	0.238	0.043	0.99
BSSMQ10	2728	43.80	0.560	0.043	1.00
BSMMR06	2786	74.80	-1.098	0.047	1.01
BSMMR07	2785	44.34	0.516	0.043	0.99
BSMMR08	2784	48.38	0.312	0.042	1.10
BSMMR09	2783	40.14	0.734	0.043	0.94
BSMMR10	2783	51.49	0.155	0.042	1.02
BSMMR11	2783	44.05	0.529	0.043	1.04
BSMMR12	2781	86.19	-1.954	0.058	0.95
BSSMR13	2779	32.13	1.167	0.045	0.91
BSSMR14	2778	37.08	0.892	0.044	0.82
BSEMS01	2829	77.48	-1.273	0.049	1.05

Table 7.4 Population 2 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 3)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSEMS01	2739	23.80	1.722	0.050	0.97
BSEMS02	2534	65.04	-0.421	0.046	0.92
BSEMS02	2382	30.31	1.420	0.050	0.82
BSEMS02	2171	28.33	1.590	0.053	0.90
BSEMT01	5661	31.90	0.915	0.019	1.01
BSEMT01	5053	35.84	1.047	0.033	0.79
BSEMT02	4998	23.41	1.799	0.037	0.88
BSEMT02	4221	9.90	3.076	0.055	1.00
BSEMU01	5585	34.45	1.045	0.031	0.96
BSEMU01	5330	33.66	1.132	0.032	0.99
BSEMU02	5009	37.10	0.778	0.020	1.13
BSEMU02	4671	20.65	1.745	0.026	0.95
BSSMV01	5477	52.67	0.110	0.030	0.91
BSEMV02	5582	27.11	1.113	0.017	1.17
BSMMV03	5538	40.75	0.732	0.031	0.95
BSSMV04	5512	39.26	0.813	0.031	0.90

Figure 7.7 Empirical and Modelled Item Characteristic Curves for Mathematics Population 2 Item: BSMMD10. Fit MNSQ=1.16



The compatibility of the model and data for Population 2 science is better than for any of the other three data sets. There is just one item, item Q18, with a weighted mean square greater than 1.15, and there are no items with weighted mean squares as low as 0.85. Figure 7.8 is a plot of the observed and expected item characteristic curves for Q18. The plot shows evidence of guessing. Examining the behavior at the country level again reveals that there is a distracter that is positive in many countries.

As a set, the data appear to be quite compatible with the assumed Rasch scaling model. Certainly the extent of deviation from the model will have had no influence on the substantive outcomes of the study. A few isolated items that were retained in the scaling did not fit the model. The source of this misfit can generally be traced to multiple-choice item distracters that were attractive to some more able students.

Table 7.5 Population 2 Science: Item Statistics and Parameter Estimates for the International Calibration Sample

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMSA07	23016	67.11	-0.562	0.015	1.02
BSMSA08	23024	65.54	-0.484	0.015	1.04
BSMSA09	23015	76.66	-1.089	0.016	0.93
BSMSA10	22998	68.38	-0.626	0.015	1.03
BSMSA11	22982	57.92	-0.119	0.014	0.99
BSMSA12	22968	58.51	-0.146	0.014	0.98
BSMSB01	11410	86.92	-1.845	0.029	0.98
BSMSB02	11406	53.40	0.095	0.020	1.08
BSMSB03	11407	26.47	1.394	0.022	1.00
BSMSB04	11404	88.48	-1.998	0.030	0.95
BSMSB05	11362	48.88	0.299	0.020	1.10
BSMSB06	11402	83.44	-1.547	0.026	1.01
BSMSC07	8642	37.75	0.816	0.024	1.01
BSMSC08	8637	71.63	-0.786	0.025	0.98
BSMSC09	8633	72.95	-0.859	0.025	1.02
BSMSC10	8636	77.08	-1.102	0.027	1.03
BSMSC11	8624	45.26	0.468	0.023	0.98
BSMSC12	8618	52.70	0.131	0.023	1.09
BSMSD01	8794	40.22	0.674	0.023	1.02
BSMSD02	8787	73.39	-0.914	0.025	0.94
BSMSD03	8787	36.95	0.830	0.023	0.98
BSMSD04	8779	54.99	-0.001	0.023	1.02
BSMSD05	8769	66.27	-0.535	0.024	0.97
BSMSD06	8770	72.75	-0.877	0.025	0.97
BSMSE07	8669	41.38	0.634	0.023	1.09
BSMSE08	8666	79.17	-1.247	0.028	1.01
BSMSE09	8662	77.80	-1.158	0.027	0.96
BSMSE10	8651	53.59	0.080	0.023	0.99
BSMSE11	8642	57.30	-0.089	0.023	1.04
BSMSE12	8396	53.93	0.074	0.023	1.06
BSMSF01	8637	66.61	-0.549	0.024	0.98
BSMSF02	8634	63.19	-0.380	0.024	0.91
BSMSF03	8392	66.17	-0.515	0.024	1.08
BSMSF04	8399	68.33	-0.632	0.025	0.96
BSMSF05	8631	80.44	-1.349	0.028	0.97
BSMSF06	8630	68.81	-0.661	0.025	0.95
BSMSG07	8619	86.99	-1.856	0.033	0.98
BSMSG08	8619	59.38	-0.189	0.023	1.00
BSMSG09	8614	74.32	-0.948	0.026	1.00
BSMSG10	8612	51.64	0.166	0.023	0.98
BSMSG11	8605	49.56	0.260	0.023	1.05
BSMSG12	8596	51.14	0.189	0.023	1.06
BSMSH01	8264	69.26	-0.648	0.025	0.99
BSMSH02	8496	79.26	-1.240	0.028	1.01
BSMSH03	8494	79.15	-1.233	0.028	0.96
BSMSH04	8587	50.73	0.216	0.023	1.04
BSMSH05	8586	22.99	1.603	0.027	1.02

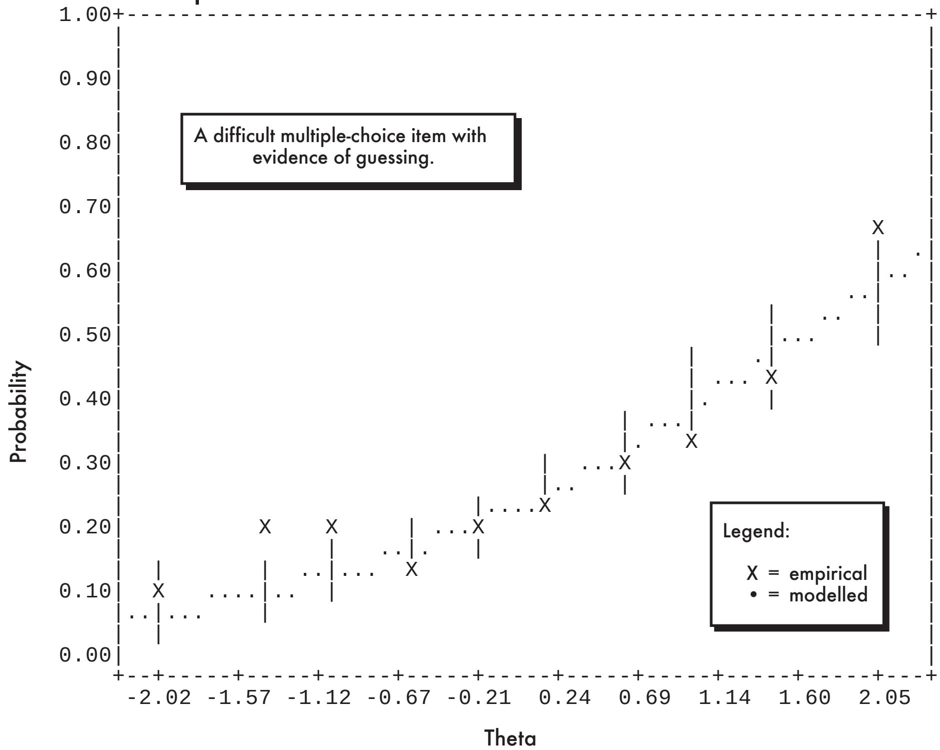
Table 7.5 Population 2 Science: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 1)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMSH06	8583	51.40	0.186	0.023	0.96
BSMSI10	2871	74.54	-0.921	0.045	1.01
BSMSI11	2869	45.59	0.477	0.040	0.98
BSMSI12	2866	35.35	0.957	0.041	0.96
BSMSI13	2795	60.86	-0.217	0.041	0.94
BSMSI14	2862	54.72	0.066	0.040	1.05
BSMSI15	2859	49.11	0.320	0.040	0.97
BSMSI16	2854	85.00	-1.632	0.054	0.99
BSMSI17	2851	40.72	0.704	0.040	1.06
BSSSI18	2847	35.30	0.964	0.041	0.96
BSMSI19	2759	51.40	0.223	0.040	0.91
BSMSJ01	2784	39.22	0.752	0.041	1.04
BSMSJ02	2942	62.71	-0.369	0.040	0.97
BSSSJ03	2941	27.13	1.337	0.044	0.97
BSMSJ04	2941	41.11	0.629	0.040	1.00
BSMSJ05	2940	64.35	-0.447	0.041	0.98
BSMSJ06	2940	22.69	1.603	0.046	1.00
BSMSJ07	2873	47.16	0.362	0.040	1.00
BSMSJ08	2936	45.16	0.442	0.040	0.98
BSSSJ09	2935	76.12	-1.079	0.045	0.94
BSSSK10	2876	34.14	0.947	0.042	1.08
BSMSK11	2946	55.02	-0.012	0.039	0.96
BSMSK12	2945	51.85	0.132	0.039	0.99
BSMSK13	2944	74.01	-0.953	0.044	0.93
BSMSK14	2942	79.74	-1.306	0.048	0.94
BSMSK15	2940	59.35	-0.210	0.040	0.99
BSMSK16	2938	35.94	0.867	0.041	0.99
BSMSK17	2936	51.63	0.143	0.039	1.01
BSMSK18	2925	54.22	0.029	0.039	1.02
BSSSK19	2907	72.38	-0.852	0.044	0.97
BSMSL01	2859	47.22	0.365	0.040	1.00
BSMSL02	2858	51.68	0.163	0.040	0.99
BSMSL03	2859	68.42	-0.631	0.043	0.95
BSESL04	2858	32.96	1.041	0.042	0.94
BSMSL05	2857	64.30	-0.425	0.041	1.04
BSMSL06	2856	52.21	0.139	0.040	1.00
BSMSL07	2857	68.15	-0.618	0.042	0.96
BSMSM10	2822	46.03	0.425	0.040	1.01
BSESM11	2821	65.65	-0.483	0.042	0.92
BSSSM12	2817	52.36	0.141	0.040	0.92
BSMSM13	2813	46.82	0.391	0.040	0.96
BSSSM14	2810	71.14	-0.764	0.044	1.02
BSMSN01	2839	43.68	0.550	0.040	1.08
BSMSN02	2837	39.76	0.732	0.041	1.04
BSMSN03	2837	60.35	-0.207	0.041	0.98
BSMSN04	2834	50.53	0.241	0.040	1.07
BSMSN05	2688	31.29	1.161	0.044	1.08

Table 7.5 Population 2 Science: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 2)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMSN06	2833	64.03	-0.382	0.041	0.98
BSSSN07	2833	88.92	-2.017	0.061	0.91
BSMSN08	2834	70.82	-0.727	0.043	1.00
BSMSN09	2774	47.08	0.399	0.040	0.95
BSSSN10	2830	53.11	0.123	0.040	0.96
BSSSO10	2874	53.58	0.087	0.040	0.96
BSMSO11	2797	36.11	0.905	0.042	1.02
BSMSO12	2812	24.64	1.516	0.046	0.98
BSMSO13	2872	58.67	-0.147	0.040	1.01
BSESO14	2870	54.36	0.052	0.040	0.91
BSMSO15	2862	38.29	0.791	0.041	1.01
BSSSO16	2862	57.97	-0.112	0.040	0.95
BSSSO17	2803	55.26	0.025	0.040	1.04
BSMSP01	2780	82.77	-1.501	0.052	0.92
BSSSP02	2776	21.65	1.669	0.048	0.95
BSSSP03	2777	79.40	-1.264	0.049	0.91
BSMSP04	2774	54.25	0.047	0.040	0.98
BSSSP05	2770	55.96	-0.031	0.041	0.99
BSSSP06	2764	48.20	0.304	0.025	0.98
BSMSP07	2766	52.39	0.132	0.040	1.06
BSMSQ11	2713	42.87	0.569	0.041	1.03
BSSSQ12	2709	45.99	0.427	0.041	0.99
BSMSQ13	2703	59.67	-0.193	0.042	0.98
BSMSQ14	2678	45.29	0.463	0.041	1.07
BSMSQ15	2612	32.50	1.080	0.044	1.03
BSMSQ16	2597	25.53	1.444	0.047	1.05
BSSSQ17	2567	65.41	-0.460	0.044	0.99
BSSSQ18	2433	36.33	0.728	0.026	1.17
BSMSR01	2786	68.88	-0.654	0.043	1.03
BSMSR02	2786	38.30	0.771	0.041	1.05
BSESR03	2784	29.76	1.160	0.030	0.94
BSSSR04	2784	50.11	0.231	0.040	0.91
BSSSR05	2783	49.08	0.277	0.040	0.88
BSESW01	5652	80.08	-1.306	0.035	1.00
BSESW01	5408	42.77	0.607	0.029	1.05
BSESW02	5176	43.51	0.517	0.018	1.05
BSESX01	5690	22.12	1.398	0.022	0.98
BSESX02	5123	68.79	-0.609	0.032	1.03
BSESX02	4923	34.51	1.019	0.032	0.99
BSESY01	5562	6.53	3.162	0.055	0.94
BSESY02	5291	28.34	1.208	0.021	1.13
BSESZ01	2725	62.35	-0.287	0.042	0.99
BSESZ01	2449	42.63	0.642	0.043	0.90
BSESZ01	2335	28.09	1.379	0.048	0.91
BSESZ02	2341	75.22	-0.895	0.050	1.02
BSESZ02	2260	50.18	0.344	0.045	1.00

Figure 7.8 Empirical and Modelled Item Characteristic Curves for Science Population 2 Item: BSSQ18. Fit MNSQ=1.17



7.4.4 Reliability

Table 7.6 reports a variety of reliability indices for the four tests. The median Cronbach Alpha coefficients were computed by calculating the Cronbach Alpha coefficient for each test booklet within each country. The median of these values was then used as a reliability index for each country. The median of those country medians is reported in Table 7.6.

The separation reliability for the international calibration sample was computed by fitting the scaling model without the use of any conditioning variables, drawing five plausible values for each student, and then computing the median of the ten correlations between pairs of plausible values. In general, these statistics show that the science tests are slightly less reliable than the mathematics tests and that the Population 1 tests are slightly less reliable than the Population 2 tests.

Table 7.6 Unconditional Reliabilities

TIMSS Test	Median of Lower Grade National Cronbach Alpha Coefficients	Median of Upper Grade National Cronbach Alpha Coefficients	Separation Reliability in the International Calibration Sample
Mathematics Population 1	0.82	0.84	0.83
Science Population 1	0.78	0.77	0.77
Mathematics Population 2	0.86	0.89	0.89
Science Population 2	0.77	0.78	0.80

7.4.5 The Population Model for Population 2

For Population 2 it was considered expedient to proceed with a scaling that did not make extensive use of conditioning. There were two reasons for this. First, the reliability of the Population 2 data was relatively high so that the possible effect of conditioning would be ignorable. Second, the background data were not fully cleaned and checked at the time of processing, and extensive conditioning would have delayed publication of the international reports.

For each participating country the scaling was undertaken with all item parameters set at the values obtained from fitting the model to the international calibration sample. In the population model sampling weights were used, and student grade was used as a conditioning variable. Five plausible values were drawn and an EAP estimate of achievement was obtained for each student. As illustrated in Table 7.7, conditioning on grade led to little improvement in the person separation reliability. This conditioning was, however, necessary to ensure that consistent results were obtained when plausible values were used to estimate characteristics of the achievement distributions for the upper and lower grades separately.

Table 7.7 Population 2 Reliabilities For Three Countries With and Without Conditioning on Grade

Country	Mathematics		Science	
	Conditioning on Grade	No Conditioning	Conditioning on Grade	No Conditioning
Australia	0.88	0.88	0.80	0.80
Cyprus	0.86	0.86	0.78	0.77
Hong Kong	0.87	0.87	0.76	0.76

7.4.6 The Population Model for Population 1

In Population 1 conditioning was used much more extensively. For both mathematics and science the variables sex, grade, and the interaction between sex and grade were used as conditioning variables.[1] Additionally, for mathematics the mean of the mathematics score variable ASMRAWST was computed for each class, assigned to each student in that class, and then used as a conditioning variable. This variable was called ASMRAWAV. Similarly, for science the mean of the mathematics score variable ASSRAWST was computed for each class, assigned to each student in that class, and then used as a conditioning variable. This variable was called ASSRAWAV. This conditioning was undertaken so as to improve the estimation of between-class and between-school variance components that would be obtained from secondary analyses using plausible values. Each individual student's science score ASSRAWST was also used as a conditioning variable for mathematics, and in the case of science, each individual student's mathematics score ASMRAWST was used.

1 The gender variable ASBGSEX is trichotomous (male, female, missing). When used in conditioning, this variable was replaced with two dummy coded variables.

Table 7.8 Number of Principal Components Retained In Conditioning - Population 1

Country	Number of Retained Principal Components
Australia	62
Austria	85
Canada	84
Cyprus	69
Czech Republic	84
England	51
Greece	78
Hong Kong	81
Hungary	86
Iceland	69
Iran	83
Ireland	83
Israel	62
Japan	59
Korea	78
Kuwait	88
Latvia	74
Mexico	103
Netherlands	83
New Zealand	87
Norway	75
Portugal	91
Scotland	46
Singapore	106
Slovenia	77
Thailand	73
United States	72

For both mathematics and science, the pool of over 100 student-level background variables was also represented in the conditioning. For each student-level variable a set of dummy variables was constructed from the original variables (see Appendix D). This new set of dummy variables retained all of the information in the original set of variables but made them appropriate for use in a principal components analysis. A principal components analysis of the set of dummy variables was then undertaken for each country and as many components retained as explained 90% of the variance. Scores on each of the retained components were then computed for each student. The number of retained components for each country is shown in Table 7.8.

These components, and the products of these components and ASMRAWAV (in the case of mathematics) and ASSRAWAV (in the case of science), were used as conditioning variables. Table 7.9 shows the conditioning variables that were used for mathematics and science. For some countries the total was in excess of 200. Table 7.10 illustrates for three selected countries the increase in reliability that was attained by conditioning, first by grade and then with the full set of conditioning variables.

Table 7.9 Variables Used in Conditioning - Population 1

Variables	Mathematics	Science
Grade	✓	✓
Gender	✓	✓
Gender by grade interaction	✓	✓
Mathematics score	✗	✓
Science score	✓	✗
Class mean mathematics score	✓	✗
Class mean science score	✗	✓
Principal components	✓	✓
Principal component by class mean mathematics score	✓	✗
Principal component by class mean science score	✗	✓

Table 7.10 Variables Used in Conditioning - Population 1

Country	Mathematics			Science		
	No Conditioning	Conditioning on Grade	Full Conditioning	No Conditioning	Conditioning on Grade	Full Conditioning
Australia	0.83	0.84	0.87	0.77	0.78	0.83
Cyprus	0.82	0.83	0.86	0.74	0.75	0.81
Hong Kong	0.78	0.79	0.84	0.73	0.74	0.81

REFERENCES

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

Adams, R.J., Wilson, M.R., and Wang, W.C. (1997). The multidimensional random coefficients multinomial logit. *Applied Psychological Measurement*, 21, 1-24.

Adams, R.J., Wilson, M.R., and Wu, M.L. (1997). Multilevel item responses models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22 (1), 46-75.

Beaton, A.E. (1987). *Implementing the new design: The NAEP 1983-84 technical report.* Report No. 15-TR-20. Princeton, NJ: Educational Testing Service.

Bock, R.D. and Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, 46, 443-459.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

Kelderman, H. and Rijkes, C.P.M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, 59, 149-176.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177-196.

Mislevy, R.J., Beaton, A.E., Kaplan, B., and Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29 (2), 133-161.

Mislevy, R.J., Johnson, E.G., and Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17, 131-154.

Mislevy, R.J. and Sheehan, K.M. (1987). Marginal estimation procedures. In A.E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report.* Report No. 15-TR-20. Princeton, NJ: Educational Testing Service.

Mislevy, R.J. and Sheehan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, 54 (4), 661-679.

Rubin, D.B. (1987). *Multiple imputation for non-response in surveys.* New York: John Wiley & Sons.

Volodin, N. and Adams, R.J. (1997). *The estimation of polytomous item response models with many dimensions.* Paper presented at the Annual Meeting of the Psychometric Society, Gatlinburg, TN.

Wilson, M.R. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement*, 16 (3).

Wu, M.L., Adams, R.J., and Wilson, M. (1997). *Conquest: Generalized item response modelling software – Manual.* Melbourne: Australian Council for Educational Research.

Wu, M.L. (1997). *The development and application of a fit test for use with marginal maximum estimation and generalized item response models.* Unpublished Master's dissertation, University of Melbourne.

8 Reporting Student Achievement in Mathematics and Science

Eugenio J. Gonzalez
Boston College

8.1 STANDARDIZING THE TIMSS INTERNATIONAL SCALE SCORES

The item response theory (IRT) scaling procedures described in the Chapter 7 yielded imputed scores or plausible values in a logit metric, with the majority of scores falling in the range from -3 to +3. These scores were transformed onto an international achievement scale with mean 500 and standard deviation 100, which was more suited to reporting international results. This scale avoids negative values for student scale scores and eliminates the need for decimal points in reporting student achievement.

Since a plausible value is an imputed score that includes a random component, it is customary when using this methodology to draw a number of plausible values for each respondent (usually five). Each analysis is then carried out five times, once with each plausible value, and the results averaged to get the best overall result. The variability among the five results is a measure of the error due to imputation and, where this is large, it may be combined with jackknife estimates of sampling error to give a more realistic indication of the total variability of a statistic. In TIMSS at Population 1 and 2 there was little variability between results from the five plausible values, and so it was decided to simplify the analytic procedures by ignoring this variability and using the first plausible value as the international student score in mathematics and science.

In order to ensure that the mean of the TIMSS international achievement scale was close to the average student achievement level across countries, it was necessary to estimate the mean and standard deviation of the logit scores for all participating students. To accomplish this, the logit scores from all students from all countries at both grade levels were combined into a standardization sample. This sample consisted of student data from 40 countries, each country equally weighted. South Africa and the Philippines were not included in the sample. The means and standard deviations derived from this procedure are shown in Tables 8.1 through 8.4. These tables show the average logit for each of the five plausible values, and for the international student score (which is simply a copy of the first plausible value).

Table 8.1 Standardization Parameters of International Mathematics Scores Population 1

Variable	Mean Logit	Standard Deviation
International Mathematics Score	0.228345	1.070685
Mathematics Plausible Value #1	0.228345	1.070685
Mathematics Plausible Value #2	0.227183	1.069980
Mathematics Plausible Value #3	0.228378	1.069806
Mathematics Plausible Value #4	0.229702	1.070308
Mathematics Plausible Value #5	0.228632	1.072624
Average	0.228448	1.070681

Table 8.2 Standardization Parameters of International Science Scores Population 1

Variable	Mean Logit	Standard Deviation
International Science Score	0.288556	0.958956
Science Plausible Value #1	0.288556	0.958956
Science Plausible Value #2	0.283356	0.959373
Science Plausible Value #3	0.283130	0.959993
Science Plausible Value #4	0.286728	0.959670
Science Plausible Value #5	0.283406	0.960045
Average	0.285035	0.959607

Table 8.3 Standardization Parameters of International Mathematics Scores Population 2

Variable	Mean Logit	Standard Deviation
International Mathematics Score	0.214809	1.105079
Mathematics Plausible Value #1	0.214809	1.105079
Mathematics Plausible Value #2	0.215036	1.106252
Mathematics Plausible Value #3	0.215540	1.108284
Mathematics Plausible Value #4	0.215463	1.106881
Mathematics Plausible Value #5	0.213658	1.104365
Average	0.214901	1.106172

Table 8.4 Standardization Parameters of International Science Scores Population 2

Variable	Mean Logit	Standard Deviation
International Science Score	0.211454	0.770235
Science Plausible Value #1	0.211454	0.770235
Science Plausible Value #2	0.211574	0.770093
Science Plausible Value #3	0.211886	0.771142
Science Plausible Value #4	0.213772	0.769263
Science Plausible Value #5	0.210969	0.771090
Average	0.211931	0.770365

Each country was weighted to contribute equally to the calculation of the international mean and standard deviation, except for Kuwait and Israel, which tested only one grade at each population. These two countries were weighted to make only half the contribution of the countries with both grades. The contribution of the students from each grade within each country was proportional to the number of students at each grade level within the country. The transformation applied to the plausible value logit scores was

$$S_{ijk} = 500 + 100 * \left(\frac{\theta_{ijk} - \bar{\theta}_j}{SD_{\theta_j}}\right)$$

where S_{ijk} is the standardized scale score with mean 500 and standard deviation 100 for student i, in plausible value j, in country k; θ_{ijk} is the logit score for the same student, $\bar{\theta}_j$ is the weighted average across all countries on plausible value j, and SD_{θ_j} is the standard deviation across all countries on plausible value j. Since five plausible values (logit scores) were drawn for each student, each of these was transformed so that the international mean of the result scores was 500, with standard deviation 100.

Because plausible values are actually random draws from the estimated distribution of student achievement and not actual student scores, student proficiency estimates were occasionally obtained that were unusually high or low. Where a transformed plausible value fell below 50, the value was recoded to 50, therefore making 50 the lowest score on the transformed scale. This happened in very few cases across the countries. The highest transformed scores did not exceed 1000 points, so the transformed values were left untouched at the upper end of the distribution.

8.2 STANDARDIZING THE INTERNATIONAL ITEM DIFFICULTIES

To help readers of the TIMSS international reports understand the international achievement scales, TIMSS produced item difficulty maps that showed the location on the scales of several items from the subject matter content areas covered by the mathematics and science tests. In order to locate the example items on the achievement scales, the item difficulty parameter for each item had to be transformed from its original logit metric to the metric of the international achievement scales (a mean of 500 and standard deviation of 100).

The procedure for deriving the international item difficulties is described in Chapter 7. The international item difficulties obtained from the scaling procedure represent the proficiency level of a person who has a 50 percent chance of responding to the item correctly. For the item difficulty maps it was preferred that the difficulty correspond to the proficiency level of a person showing greater mastery of the item. For this reason it was decided to calibrate these item difficulties in terms of the proficiency of a person with a 65 percent chance of responding correctly.

In order to derive item difficulties for the item difficulty maps, the original item difficulties from the scaling were transformed in two ways. First, they were moved along the logit scale from the point where a student would have a 50 percent chance to the point where the student would have a 65 percent chance of responding correctly. This was achieved by adding the natural log of the odds of a 65 percent response rate to the original log odds since the logit metric allows this addition to take place in a straightforward manner. Second, the new logit item difficulty was transformed onto the international achievement scale. The means and standard deviations for this transformation were the average of the plausible value means, and the average of the plausible value standard deviations from Table 8.1 through Table 8.4 above. This resulted in the following transformations for the mathematics and science items.

For the Populations 1 and 2 mathematics item difficulties, dm_j, the transformed item difficulty dm'_j was computed as follows:

$$\text{Population 1} \quad dm'_j = 500 + 100 \times \left(\frac{dm_j + \ln\left(\frac{0.65}{0.35}\right) - 0.228448}{1.07068} \right)$$

$$\text{Population 2} \quad dm'_j = 500 + 100 \times \left(\frac{dm_j + \ln\left(\frac{0.65}{0.35}\right) - 0.214901}{1.106172} \right)$$

For the Populations 1 and 2 science item difficulties, ds_j, the transformed item difficulty ds'_j was computed as follows:

Population 1 $$ds'_j = 500 + 100 \times \left(\frac{ds_j + \ln\left(\frac{0.65}{0.35}\right) - 0.285035}{0.959607} \right)$$

Population 2 $$ds'_j = 500 + 100 \times \left(\frac{ds_j + \ln\left(\frac{0.65}{0.35}\right) - 0.211931}{0.770365} \right)$$

The resulting values are the item difficulties presented in the item difficulty maps in the international reports.

8.3 MULTIPLE COMPARISONS OF ACHIEVEMENT

An essential purpose of the TIMSS international reports is to provide fair and accurate comparisons of student achievement across the participating countries. Most of the tables in the reports summarize student achievement by means of a statistic such as a mean or percentage, and each summary statistic is accompanied by its standard error, which is a measure of the variability in the statistic resulting from the sampling process. In comparisons of student performance from two countries, the standard errors can be used to assess the statistical significance of the difference between the summary statistics.

The multiple comparison charts presented in the TIMSS international reports are designed to help the reader compare the average performance of a country with that of other participating countries of interest.The significance tests reported in these charts are based on a Bonferroni procedure for multiple comparisons that holds to 5 percent the probability of erroneously declaring the mean of one country to be different from another country.

If we were to take repeated samples from two populations with the same mean and test the hypothesis that the means from these two samples are significantly different at the $\alpha = .05$ level, i.e. with 95 percent confidence, then in about 5 percent of the comparisons we would expect to find significant differences between the sample means even though we know that there is no difference between the population means. In this example with one test of the difference between two means, the probability of finding significant differences in the samples when none exist in the populations (the so-called type I error) is given by $\alpha = .05$. Conversely, the probability of not making a type I error is $1 - \alpha$, which in the case of a single test is .95. However, if we wish to compare the means of three countries, this involves three tests (country A versus country B, country B versus country C, and country A versus country C). Since these are independent tests, the probability of **not** making a type I error in any of these tests is the product of the individual probabilities, which is $(1 - \alpha)(1 - \alpha)(1 - \alpha)$. With $\alpha = .05$, the overall probability of not making a type I error is only .873, which is considerably less than the probability for a single test. As the number of tests increases, the probability of not making a type I error decreases, and conversely, the probability of making a type I error increases.

Several methods can be used to correct for the increased probability of a type I error while making many simultaneous comparisons. Dunn (1961) developed a procedure that is appropriate for testing a set of a priori hypotheses while controlling the probability that the type I error will occur. When using this procedure, the researcher adjusts the value α when making multiple simultaneous comparisons to compensate for the increase in the probability of making a type I error. This is known as the Dunn-Bonferroni procedure for multiple a priori comparisons (Winer, Brown, and Michels, 1991).

In this procedure the significance level of the test of the difference between means is adjusted by dividing the significance level (α) by the number of comparisons that are planned and then looking up the appropriate quantile from the normal distribution. In deciding the number of comparisons, and hence the appropriate adjustment to the significance level for TIMSS, it was necessary to decide how the multiple comparison tables would most likely be used. One approach would have been to adjust the significance level to compensate for all possible comparisons between the countries presented in the table. This would have meant adjusting the significance level for 820 comparisons at the eighth-grade, 741 at the seventh-grade, 325 at the fourth-grade, and 276 at the third-grade. In decision-making terms this would be a very conservative procedure, however, and would run the risk of making an error of a different kind, i.e., of concluding that a difference between sample means is not significant when in fact there is a difference between the population means.

Since most users probably are interested in comparing a single country with all other countries and would not be making all possible between-country comparisons at any one time, a more realistic approach, which was adopted in TIMSS, seemed to be to adjust the significance level for a number of comparisons equal to the number of countries (minus one). From this perspective the number of simultaneous comparisons to be adjusted for at eighth grade, for example, is 40 rather than 820, and at seventh grade is 38 rather than 741. The number of comparisons is 25 for the fourth-grade table, and 23 for the third-grade table. As a consequence, we used the critical values shown in Table 8.5, given by the appropriate quantiles from the normal (Gaussian) distribution.

Table 8.5 Critical Values Used for the Multiple Comparison Figures in TIMSS International Reports

Grade Level	Alpha Level	Number of Comparisons	Critical Value
3rd Grade	0.05	23	3.0654
4th Grade	0.05	25	3.0902
7th Grade	0.05	38	3.2125
8th Grade	0.05	40	3.2273

Two means were considered significantly different from each other if the absolute differences between them was greater than the critical value multiplied by the standard error of the difference. The standard error of the difference between the two means was computed as the square root of the sum of the squared standard errors of the mean:

$$se_{diff} = \sqrt{se_1^2 + se_2^2}$$

where se_1 and se_2 are the standard errors for each of the means being compared, respectively, computed using the jackknife method of variance estimation. Tables 8.6a and 8.6b show the means and standard errors used in the calculation of statistical significance between means for mathematics and science, at Population 2 and Population 1, respectively. By applying the Bonferroni correction, we were able to state that, for any given row or column of the multiple comparison chart, the differences between countries shown in the chart are statistically significant at the 95 percent level of codence.

Table 8.6a Means and Standard Errors for Multiple Comparison Figures Mathematics and Science - Population 2

Country	Mathematics 7th Grade Mean	S.E.	Mathematics 8th Grade Mean	S.E.	Science 7th Grade Mean	S.E.	Science 8th Grade Mean	S.E.
Australia	497.9	3.8	529.6	4.0	504.4	3.6	544.6	3.9
Austria	509.2	3.0	539.4	3.0	518.8	3.1	557.7	3.7
Belgium (Fl)	557.6	3.5	565.2	5.7	528.7	2.6	550.3	4.2
Belgium (Fr)	507.1	3.5	526.3	3.4	442.0	3.0	470.6	2.8
Bulgaria	513.8	7.5	539.7	6.3	530.8	5.4	564.8	5.3
Canada	494.0	2.2	527.2	2.4	499.2	2.3	530.9	2.6
Colombia	368.5	2.7	384.8	3.4	387.5	3.2	411.1	4.1
Cyprus	445.7	1.9	473.6	1.9	419.9	1.8	462.6	1.9
Czech Republic	523.4	4.9	563.7	4.9	532.9	3.3	573.9	4.3
Denmark	464.8	2.1	502.3	2.8	439.0	2.1	478.3	3.1
England	476.2	3.7	505.7	2.6	512.0	3.5	552.1	3.3
France	492.2	3.1	537.8	2.9	451.5	2.6	497.7	2.5
Germany	484.4	4.1	509.2	4.5	499.5	4.1	531.3	4.8
Greece	439.9	2.8	483.9	3.1	448.6	2.6	497.3	2.2
Hong Kong	563.6	7.8	588.0	6.5	495.3	5.5	522.1	4.7
Hungary	501.8	3.7	537.3	3.2	517.9	3.2	553.7	2.8
Iceland	459.4	2.6	486.8	4.5	462.0	2.8	493.6	4.0
Iran, Islamic Rep.	400.9	2.0	428.3	2.2	436.3	2.6	469.7	2.4
Ireland	499.7	4.1	527.4	5.1	495.2	3.5	537.8	4.5
Israel	.	.	521.6	6.2	.	.	524.5	5.7
Japan	571.1	1.9	604.8	1.9	531.0	1.9	571.0	1.6
Korea	577.1	2.5	607.4	2.4	535.0	2.1	564.9	1.9
Kuwait	.	.	392.2	2.5	.	.	429.6	3.7
Latvia (LSS)	461.6	2.8	493.4	3.1	434.9	2.7	484.8	2.7
Lithuania	428.2	3.2	477.2	3.5	403.1	3.4	476.4	3.4
Netherlands	516.0	4.1	541.0	6.7	517.2	3.6	560.1	5.0
New Zealand	471.7	3.8	507.8	4.5	481.0	3.4	525.5	4.4
Norway	460.7	2.8	503.3	2.2	483.2	2.9	527.2	1.9
Portugal	423.1	2.2	454.4	2.5	427.9	2.1	479.6	2.3
Romania	454.4	3.4	481.6	4.0	451.6	4.4	486.1	4.7
Russian Federation	500.9	4.0	535.5	5.3	484.0	4.2	538.1	4.0
Scotland	462.9	3.7	498.5	5.5	468.1	3.8	517.2	5.1
Singapore	601.0	6.3	643.3	4.9	544.7	6.6	607.3	5.5
Slovak Republic	507.8	3.4	547.1	3.3	509.7	3.0	544.4	3.2
Slovenia	498.2	3.0	540.8	3.1	529.9	2.4	560.1	2.5
South Africa	347.5	3.8	354.1	4.4	317.1	5.3	325.9	6.6
Spain	448.0	2.2	487.3	2.0	477.2	2.1	517.0	1.7
Sweden	477.5	2.5	518.6	3.0	488.4	2.6	535.4	3.0
Switzerland	505.5	2.3	545.4	2.8	483.7	2.5	521.7	2.5
Thailand	494.7	4.8	522.5	5.7	492.8	3.0	525.5	3.7
United States	475.7	5.5	499.8	4.6	508.2	5.5	534.4	4.7

Table 8.6b Means and Standard Errors for Multiple Comparison Figures Mathematics and Science - Population 1

Country	Mathematics				Science			
	3th Grade Mean	S.E.	4th Grade Mean	S.E.	3rd Grade mean	S.E.	4th Grade Mean	S.E.
Australia	483.4	4.0	546.3	3.1	509.7	4.3	562.5	2.9
Austria	487.0	5.3	559.3	3.1	504.6	4.6	564.8	3.3
Canada	469.5	2.7	532.1	3.3	490.4	2.5	549.3	3.0
Cyprus	430.4	2.8	502.4	3.1	414.7	2.5	475.4	3.3
Czech Republic	497.2	3.3	567.1	3.3	493.7	3.4	556.5	3.1
Greece	428.1	4.0	491.9	4.4	445.9	3.9	497.2	4.1
Hong Kong	524.0	3.0	586.6	4.3	481.6	3.3	533.0	3.7
Hungary	476.1	4.2	548.4	3.7	464.4	4.1	531.6	3.4
Iceland	410.1	2.8	473.8	2.7	435.4	3.3	504.7	3.3
Iran, Islamic Rep.	378.0	3.5	428.5	4.0	356.2	4.2	416.5	3.9
Ireland	475.8	3.6	549.9	3.4	479.1	3.7	539.5	3.3
Israel	.	.	531.4	3.5	.	.	504.8	3.6
Japan	537.9	1.5	596.8	2.1	521.8	1.6	573.6	1.8
Korea	560.9	2.3	610.7	2.1	552.9	2.4	596.9	1.9
Kuwait	.	.	400.2	2.8	.	.	401.3	3.1
Latvia (LSS)	463.3	4.3	525.4	4.8	465.3	4.5	512.2	4.9
Netherlands	492.9	2.7	576.7	3.4	498.8	3.2	556.7	3.1
New Zealand	439.5	4.0	498.7	4.3	473.1	5.2	531.0	4.9
Norway	421.3	3.1	501.9	3.0	450.3	3.9	530.3	3.6
Portugal	425.3	3.8	475.4	3.5	423.0	4.3	479.8	4.0
Singapore	552.1	4.8	624.9	5.3	487.7	5.0	546.7	5.0
Thailand	444.3	5.1	490.2	4.7	432.6	6.6	472.9	4.9
England	456.5	3.0	512.7	3.2	499.2	3.5	551.5	3.3
Scotland	458.0	3.4	520.4	3.9	483.9	4.2	535.6	4.2
United States	479.8	3.4	544.6	3.0	511.2	3.2	565.5	3.1
Slovenia	487.6	2.9	552.4	3.2	486.9	2.8	545.7	3.3

8.4 INTERNATIONAL MARKER LEVELS OF ACHIEVEMENT

For both populations, international marker levels of achievement were computed at each grade level for mathematics and science. In order to compute the marker levels, all of the student data from all participating countries for a subject at a grade level were pooled, and then the pooled data were used to estimate the 90th, the 75th, and the 50th international percentiles. These percentiles were chosen as international markers because they have a ready interpretation. The 90th percentile in this instance corresponds to the "Top 10% Level," since it is the scale score above which the highest-scoring 10 percent of the students across all countries combined are to be found. Similarly, the 75th percentile corresponds to the "Top Quarter Level," since this is the score above which the top 25 percent of students are to be found, and the 50th percentile corresponds to the "Top Half Level," since this is the score above which the top 50 percent of students are to be found. If student proficiencies were distributed in the same way across countries we would expect about 10 percent of students in each country to score at or above the Top 10% Level, about 25 percent of students to score at or above the Top Quarter marker, and about 50 percent of students to score at or above the Top Half marker. In pooling the data, countries were weighted in accordance with their estimated enrollment size, as shown in Table 8.7.

Table 8.7 Estimated Enrollment by Grade Level Within Country

Country	Third Grade	Fourth Grade	Seventh Grade	Eighth Grade
Australia	237828	245635	238294	231349
Austria	86044	91391	89593	86739
Belgium (Fl)	-	-	64177	75069
Belgium (Fr)	-	-	49898	59270
Bulgaria	-	-	140979	147094
Canada	371166	389160	377732	377426
Colombia	-	-	619462	527145
Cyprus	9740	9995	10033	9347
Czech Republic	116052	120406	152492	152494
Denmark	-	-	44980	54172
England	531682	534922	465457	485280
France	-	-	860657	815510
Germany	-	-	742346	726088
Greece	99000	106181	130222	121911
Hong Kong	83847	89901	88591	88574
Hungary	116779	117228	118727	112436
Iceland	3735	3739	4212	4234
Iran, Islamic Rep.	1391859	1433314	1052795	935093
Ireland	58503	60497	68477	67644
Israel	-	66967	-	60584
Japan	1388749	1438465	1562418	1641941
Korea	607007	615004	798409	810404
Kuwait	-	24071	-	13093
Latvia (LSS)	15121	18883	17041	15414
Lithuania	-	-	36551	39700
Netherlands	171561	173407	175419	191663
New Zealand	48386	52254	48508	51133
Norway	49036	49896	51165	50224
Portugal	114775	133186	146882	137459
Romania	-	-	295348	296534
Russian Federation	-	-	2168163	2004792
Scotland	59393	59054	61938	64638
Singapore	41904	41244	36181	36539
Slovak Republic	-	-	83074	79766
Slovenia	27453	27685	28049	26011
South Africa	-	-	649180	766334
Spain	-	-	549032	547114
Sweden	-	-	96494	98193
Switzerland	-	-	66681	69733
Thailand	883765	864525	680225	657748
United States	3643393	3563795	3156847	3188297

Having established the international marker levels, the next step was to compute the percentage of students in each country scoring at or above the marker levels. Countries with proportionately large numbers of high-achieving students had higher percentages of students scoring above the marker levels. For example, it was not unusual for high-achieving countries to have more than 30 percent of their students scoring at or above the Top 10% marker. Conversely, countries with lower achievement levels sometimes had very few students reaching that marker level.

Using these three marker levels, then, the students were classified into one of four groups: those below the international median or 50th percentile; those at or above the international median but below the third quartile or 75th percentile; those at or above the third quartile, but below the 90th percentile; and those at or above the 90th percentile. Standard errors for the percentage of students in each country were also computed using the jackknife method for sampling variance estimation. The international marker levels are presented in Table 8.8 below.

Table 8.8 International Marker Levels (Percentiles) of Achievement

Population 1

Mathematics			
Grade	P50	P75	P90
3	474	538	592
4	535	601	658
Science			
Grade	P50	P75	P90
3	488	554	610
4	541	607	660

Population 2

Mathematics			
Grade	P50	P75	P90
7	476	551	619
8	509	587	656
Science			
Grade	P50	P75	P90
7	483	553	615
8	521	592	655

8.5 REPORTING MEDIAN ACHIEVEMENT BY AGE

The target populations in TIMSS are defined in terms of adjacent grade levels (the two grades with the most 13-year-olds for Population 2 and the two grades with the most 9-year-olds for Population 1), and student achievement in the international reports is reported for the most part by grade. Since grades are primarily measures of years of schooling, they provide an appropriate basis on which to compare student achievement across countries. However, because of differences internationally in age of entry to formal schooling, and in promotion and retention practices through the grades, there is considerable variation across countries in the ages of students within comparable grade levels. Although TIMSS addressed this issue by using age as the primary basis for choosing the grades to be compared, there was still considerable variation between countries in the average age of their students within any given grade level.

Since TIMSS tested two adjacent grades at each of Populations 1 and 2, in many participating countries most or all 9-year-olds and 13-year-olds were included in the tested grades. Therefore, it was possible to extract just the students in these age groups from the total sample and make reasonable comparisons on the basis of age group. Although some countries had 100 percent of the age group in the grades tested, most countries had some, usually small, percentage of students in the age group outside of the tested grades. For example, in Population 2, some countries had a percentage of 13-year-olds below seventh grade, and a percentage above eighth grade. There was no way to estimate reliably the scores of the students missing from the age group, but it was possible to estimate how many students were involved by extrapolating from the distribution of ages within each of the tested grades.

Since the computation of the mean requires that all elements of the target group be present, it was not possible to compute the mean for 13-year-olds or for 9-year-olds without making assumptions about the scores of the students who were outside the tested grades. However, the median is a measure of the central tendency of a distribution which is less dependent on the values of the elements making up the distribution. In order to compute a median one need only be able to order the elements on the attribute of interest; it is not necessary to know their actual values. By capitalizing on this property of the median it was possible to estimate a median score for 9- and 13-year-olds while assuming only that those students who were in grades below the lower grade tested would score below the median, and those in grades above the upper grade tested would score above the median.

The first step was to estimate, from the age distribution within the tested grades, the percentages of students in the age group in grades below the lower grade tested and in grades above the upper grade tested. To do this it was assumed that the age distribution in the grades below the grades tested was similar to the age distribution in the lower grade lagged by one year for each grade below, and that the age distribution in the grades above the grades tested was similar to that of the upper grade increased by one year for each grade above. The next step was to adjust the median to compensate for the missing out-of-grade students. If there were no such missing students, that is, if the tested grades included all students in the age group, then the median would as usual be set to the 50th percentile, the score below which 50 percent of the student scores are found. However, when some percentage of the age group is outside the grades tested, the 50 percent refers to the entire age group, and not just to the tested students. In this case, the estimate for the number of out-of-grade students in the age group must be added to the number in the age group within the tested grades to estimate the size of the age group, and the percentage in the grades below the lower grade must be subtracted from 50 percent to find the percentile within the tested group that corresponds to 50 percent of the total age group.

8.5.1 Computational Example

Let us take for example a country in which the grades tested for Population 2 were the seventh and eighth. Table 8.9 shows the distribution of students by age in these two grades.[1] We can see that although the modal age of students in the grades tested is 13 at the time of testing, these are not the majority of the students. In fact, there are more students that are older or younger than the target age (53 percent).

Table 8.9 Observed Distribution of Age Groups Within Target Grades

Grade	Age					
	11	12	13	14	15	16
7	0	6506	28601	647	340	0
8	0	0	5121	25292	3702	2226
Total	0	6506	33722	25939	4042	2226

[1] The age of a student for the purpose of this analysis was considered to be the number of whole years between the date of birth of the student and the time of testing. For example, a student 13 years and 11 months old and a student 13 years and 1 month old were both considered to be 13 years old.

In Table 8.10 the age distribution of the seventh-grade students has been projected into the previous three grades with appropriate lags, and the figures from the eighth grade have been projected into the following school year with appropriate increases, until there are expected to be no more 13-year-old students. We notice from this table that the selection of grades to be tested in this country was right on target insofar as no other pair of adjacent grades would have more 13-year-olds. The two grades selected in this country included approximately 97 percent of the 13-year-olds in the country. Selecting the sixth and seventh grades would have yielded a coverage of only 84 percent of the 13-year-olds in the country, and selecting the eighth and ninth grades would have yielded a coverage of only 15 percent of the 13-year-olds.

Table 8.10 Observed and Estimated Distribution of Age Groups by Grade

Grade	Age 10	11	12	13	14	15	16	17	% of 13-Year-Olds
4	28601	647	340	0	.	.	.	.	0.00%
5	6506	28601	647	340	0	.	.	.	0.98%
6	0	6506	28601	647	340	0	.	.	1.86%
7	.	0	6506	28601	647	340	0	.	82.40%
8	.	.	0	5121	25292	3702	2226	0	14.75%
9	.	.	.	0	5121	25292	3702	2226	0.00%
Total of 13-Year-Olds:	.	.	.	34709	.	.	.	.	.

After the corresponding lags and increases are projected to the grades adjacent to the grades tested, we estimate that there are approximately 34,709 13-year-olds in the country (340 + 647 + 28601 + 5121). Of those 34,709 13-year-olds, about 3 percent are in grades below the lower grade, there are none in grades above the upper grade, and about 97 percent are in the two grades tested. With this information we can estimate the median achievement of the 13-year-olds, but we need to make one further assumption. We know that, in general, as the students move along the educational system their performance on the test improves. So it is reasonable to assume that those 13-year-olds who are in grades below the lower grade will perform below the median of all 13-year-olds, and those above the target grades will perform above the median of all 13-year-olds. Based on this assumption we can then compute the median of the 13-year-olds by looking at the percentile (*Px*) from the 13-year-olds in the target grades given by the following formula:

$$Px = \left(\frac{(50 - PBTG) * 100}{PITG}\right)$$

where *PBTG* is the estimated percent of 13-year-old students below the target grades, and *PITG* is the percent of students in the target grades. To complete our example, we would then look up the *Px* percentile in the distribution of 13-year-olds within the country. This works out to be

$$48.54 = \left(\frac{(50 - 2.84) * 100}{97.15}\right)$$

The median for the 13-year-olds in this particular country corresponds to the 48.54th percentile in the distribution of 13-year-olds in the tested grades. For the purpose of the tables presented in the international reports, the median for the students in the age group was computed only if both grades were tested within the country, the appropriate target grades were selected for the testing, and at least an estimated 75 percent of the 13-year-olds were in the target grades. The distribution of students by age across the grades tested is presented in Tables 8.11 and 8.12.

Table 8.11 Coverage of 9-Year-Olds in the Population 1 Sample

Country	Coverage of 9-Year-Olds					
	% Below Lower Grade*	% in Lower Grade	% in Upper Grade	% Above Upper Grade*	Percent of 9-Year-Olds Tested	Percentile in 9-Year-Olds Sample Representing Median for 9-Year-Olds in Country
Australia	5.8%	64.9%	28.9%	0.4%	93.8%	47.1
Austria	13.2%	71.5%	15.2%	0.0%	86.8%	42.4
Canada	4.8%	46.3%	47.5%	1.3%	93.8%	48.1
Cyprus	1.4%	35.1%	62.5%	0.9%	97.7%	49.8
Czech Republic	9.2%	75.5%	15.4%	0.0%	90.8%	45.0
England	0.9%	57.8%	41.2%	0.1%	99.0%	49.6
Greece	0.8%	10.9%	87.6%	0.7%	98.6%	50.0
Hong Kong	6.2%	43.2%	50.0%	0.7%	93.1%	47.0
Hungary	10.5%	70.2%	19.0%	0.3%	89.2%	44.3
Iceland	0.4%	14.8%	84.4%	0.4%	99.2%	50.0
Iran, Islamic Rep.	16.9%	50.7%	32.0%	0.4%	82.7%	40.0
Ireland	8.4%	68.4%	23.2%	0.0%	91.6%	45.4
Israel	.	.	.	.	.	.
Japan	0.5%	90.8%	8.7%	0.0%	99.5%	49.7
Korea	7.9%	67.2%	24.3%	0.7%	91.5%	46.1
Kuwait	.	.	.	.	.	.
Latvia	23.8%	54.7%	21.2%	0.3%	75.9%	34.5
Netherlands	6.9%	63.0%	30.1%	0.0%	93.1%	46.3
New Zealand	0.3%	50.2%	49.1%	0.3%	99.4%	50.0
Norway	0.1%	38.1%	61.7%	0.1%	99.9%	50.0
Portugal	6.7%	45.0%	47.9%	0.4%	92.9%	46.6
Scotland	0.3%	22.9%	75.7%	1.1%	98.6%	50.4
Singapore	2.1%	80.5%	17.4%	0.1%	97.8%	49.0
Slovenia	40.0%	59.6%	0.4%	0.0%	60.0%	.
Thailand	29.2%	60.1%	10.6%	0.2%	70.6%	.
United States	4.5%	61.1%	34.2%	0.2%	95.3%	47.8

*Data are estimated; students below the lower grade and above the upper grade were not included in the sample.

Table 8.12 Coverage of 13-Year-Olds in the Population 2 Sample

Country	Coverage of 13-Year-Olds					
	% Below Lower Grade*	% in Lower Grade	% in Upper Grade	% Above Upper Grade*	Percent or 13-Year-Olds Tested	Percentile in 13-Year-Olds Sample Representing Median for 13-Year-Olds in Country
Australia	7.5%	63.9%	28.2%	0.4%	92.1%	46.2
Austria	10.4%	62.5%	27.1%	0.0%	89.6%	44.2
Belgium (Fl)	5.4%	45.4%	48.8%	0.4%	94.2%	47.3
Belgium (Fr)	13.3%	40.5%	46.0%	0.2%	86.5%	42.4
Bulgaria	2.9%	58.6%	36.6%	1.9%	95.2%	49.4
Canada	8.0%	48.5%	42.9%	0.6%	91.4%	45.9
Colombia	51.5%	30.5%	15.8%	2.2%	46.3%	.
Cyprus	1.6%	27.3%	70.4%	0.8%	97.7%	49.6
Czech Republic	9.7%	72.9%	17.3%	0.0%	90.3%	44.6
Denmark	1.0%	33.9%	64.2%	0.9%	98.1%	49.9
England	0.6%	57.2%	41.7%	0.5%	98.9%	50.0
France	20.2%	43.6%	34.7%	1.5%	78.3%	38.1
Germany	26.1%	71.5%	2.2%	0.2%	73.7%	.
Greece	2.9%	10.3%	85.6%	1.2%	95.9%	49.1
Hong Kong	10.0%	44.2%	45.5%	0.3%	89.6%	44.6
Hungary	10.2%	65.2%	24.3%	0.3%	89.5%	44.5
Iceland	0.0%	16.6%	82.9%	0.5%	99.5%	50.3
Indonesia	10.2%	58.3%	27.5%	4.0%	85.8%	46.3
Iran	28.1%	47.0%	24.9%	0.1%	71.9%	
Ireland	13.9%	68.8%	17.3%	0.1%	86.1%	42.0
Israel	.	.	.	.	.	.
Japan	0.3%	90.9%	8.8%	0.0%	99.7%	49.8
Korea	1.5%	69.9%	28.2%	0.4%	98.1%	49.4
Kuwait	.	.	.	.	.	.
Latvia	10.7%	60.0%	29.1%	0.1%	89.1%	44.0
Lithuania	10.2%	64.2%	25.5%	0.2%	89.6%	44.4
Mexico	18.9%	40.4%	37.0%	3.7%	77.4%	40.2
Netherlands	9.8%	58.7%	31.2%	0.4%	89.8%	44.8
New Zealand	0.5%	51.7%	47.4%	0.4%	99.1%	50.0
Norway	0.2%	42.4%	57.2%	0.1%	99.7%	49.9
Philippines	.	.	.	.	.	.
Portugal	22.9%	43.7%	33.1%	0.3%	76.8%	35.3
Romania	24.4%	66.2%	9.4%	0.0%	75.6%	33.8
Russian Federation	4.5%	50.5%	44.3%	0.7%	94.8%	48.0
Scotland	0.2%	22.7%	76.8%	0.3%	99.5%	50.0
Singapore	3.1%	82.2%	14.7%	0.0%	96.9%	48.4
Slovak Republic	4.4%	73.2%	22.4%	0.0%	95.6%	47.7
Slovenia	33.1%	65.3%	1.6%	0.1%	66.9%	.
South Africa	40.6%	35.1%	21.1%	3.2%	56.2%	.
Spain	14.9%	45.8%	39.0%	0.3%	84.7%	41.4
Sweden	0.8%	45.0%	54.1%	0.1%	99.1%	49.6
Switzerland	8.3%	47.5%	44.0%	0.2%	91.5%	45.6
Thailand	18.0%	58.4%	19.6%	4.0%	78.0%	41.0
United States	8.7%	57.5%	33.5%	0.3%	91.0%	45.4

*Data are estimated; Students below the lower grade and above the upper grade were not included in the sample.

8.6 REPORTING GENDER DIFFERENCES WITHIN COUNTRIES

Gender differences were reported in overall student achievement in mathematics and science, as well as in several subject matter content areas. The computational procedures differed in several ways because of the different approaches to summarizing student performance: IRT scaling for the overall mathematics and science scores, and average percent correct for the subject matter content areas. This chapter describes the procedure for computing gender differences for the overall scores. The procedure for reporting gender differences in content areas is described in Chapter 9.

The analysis of overall gender differences focused on significant differences in mathematics and science achievement within each country using the international scale scores. These results are presented for each country in a table with an accompanying graph indicating where the difference between the boys' achievement and the girls' achievement was statistically significant. The significance of the difference was determined by comparing the absolute value of the standardized difference between the two means with a critical value of 1.96, corresponding to a 95 percent confidence level (two-tailed test; $\alpha = 0.05$, with infinite degrees of freedom). The same critical value was used for the third, fourth, seventh, and eighth grade results. The standardized difference between the mean for boys and girls (*t)* was computed as

$$t_k = \frac{\bar{x}_{kb} - \bar{x}_{kg}}{\sqrt{se_{kb}^2 + se_{kg}^2}}$$

where t_k is the standardized difference between two means for country k, $\bar{x}_{kb}$ and $\bar{x}_{kg}$ are the means for boys and girls within country k, and se_{kb} and se_{kg} are the standard errors for the boys' and girls' means in country k computed using the jackknife error estimation method described earlier. The above formula assumes independent samples of boys and girls, and was used in TIMSS due to time constraints. However, since in most countries boys and girls attended the same schools, in fact the samples of boys and girls are not completely independent. It would have been more correct to jackknife the difference between boys and girls. The appropriate test is then the difference between the mean for boys and the mean for girls divided by the jackknife standard error of the difference. Tables 8.13 through 8.20 show the standard errors of the differences computed under the assumption of independent sampling for boys and girls and computed using the jackknife technique for correlated samples. No corrections for multiple comparisons were made when comparing the achievement for boys and girls.

Table 8.13 Standard Error of the Gender Difference Mathematics - Third Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	487.0 (4.5)	479.8 (4.4)	4.0	6.3
Austria	493.6 (9.2)	481.3 (3.8)	9.6	10.0
Canada	476.7 (3.2)	462.9 (3.0)	3.4	4.4
Cyprus	433.3 (3.3)	428.0 (3.1)	3.2	4.5
Czech Republic	502.0 (3.7)	492.5 (3.8)	3.4	5.3
Greece	432.2 (4.4)	423.9 (4.2)	3.4	6.0
Hong Kong	528.5 (3.2)	518.4 (3.5)	2.9	4.8
Hungary	479.0 (4.9)	476.2 (4.4)	3.7	6.6
Iceland	417.9 (3.5)	402.5 (3.0)	3.4	4.7
Iran, Islamic Rep.	384.2 (4.4)	372.7 (4.9)	6.2	6.6
Ireland	473.2 (4.3)	478.7 (4.5)	5.2	6.3
Japan	539.5 (2.0)	536.3 (1.7)	2.2	2.7
Korea	566.9 (2.8)	554.3 (2.5)	2.7	3.8
Latvia (LSS)	462.4 (5.3)	464.1 (4.5)	4.9	7.0
Netherlands	496.7 (2.9)	488.9 (3.2)	2.8	4.3
New Zealand	435.8 (4.4)	443.0 (4.5)	3.9	6.3
Norway	429.9 (3.5)	411.4 (3.8)	4.0	5.2
Portugal	430.0 (3.5)	420.4 (5.0)	4.1	6.1
Singapore	550.8 (5.4)	553.5 (5.0)	4.1	7.4
Thailand	440.2 (5.0)	448.3 (5.6)	3.2	7.5
England	460.7 (3.5)	452.3 (3.4)	3.2	4.8
Scotland	461.9 (3.8)	453.7 (3.5)	3.0	5.2
United States	480.2 (3.1)	479.3 (4.4)	3.3	5.4
Slovenia	492.4 (3.1)	482.6 (3.5)	3.0	4.7

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.14 Standard Error of the Gender Difference Mathematics - Fourth Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	547.2 (3.5)	545.5 (3.7)	3.7	5.1
Austria	563.2 (3.6)	555.5 (3.6)	3.6	5.1
Canada	533.5 (3.4)	530.9 (3.9)	2.9	5.2
Cyprus	506.4 (3.5)	498.7 (3.3)	2.7	4.8
Czech Republic	568.5 (3.4)	565.8 (3.6)	2.7	5.0
Greece	491.0 (5.0)	492.7 (4.5)	3.9	6.8
Hong Kong	586.5 (4.7)	587.3 (4.2)	2.6	6.3
Hungary	551.6 (4.2)	546.4 (3.9)	3.6	5.8
Iceland	474.3 (3.3)	473.3 (3.0)	3.4	4.5
Iran, Islamic Rep.	432.9 (6.0)	423.8 (5.0)	7.8	7.8
Ireland	548.5 (3.9)	551.4 (4.3)	4.6	5.8
Israel	537.2 (4.4)	528.0 (4.1)	4.5	6.0
Japan	600.6 (2.5)	593.1 (2.2)	2.3	3.3
Korea	618.2 (2.5)	603.0 (2.6)	2.9	3.6
Kuwait	398.8 (4.6)	401.6 (2.5)	5.1	5.3
Latvia (LSS)	520.7 (5.5)	530.2 (5.2)	4.5	7.5
Netherlands	584.7 (3.8)	569.5 (3.4)	2.6	5.1
New Zealand	493.8 (5.7)	503.5 (4.3)	5.3	7.1
Norway	504.2 (3.5)	499.1 (3.6)	3.5	5.0
Portugal	477.6 (3.8)	473.1 (3.7)	2.6	5.3
Singapore	620.2 (5.5)	630.2 (6.4)	5.4	8.4
Thailand	484.8 (5.8)	495.6 (4.2)	3.9	7.1
England	515.1 (3.4)	510.3 (4.4)	4.4	5.5
Scotland	520.3 (4.3)	520.2 (3.8)	2.6	5.8
United States	545.4 (3.1)	543.8 (3.3)	1.9	4.5
Slovenia	551.1 (3.4)	553.9 (4.0)	3.6	5.2

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.15 Standard Error of the Gender Difference Science - Third Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	509.8 (5.6)	509.6 (4.3)	5.1	7.1
Austria	508.3 (6.9)	501.2 (4.0)	6.7	7.9
Canada	496.0 (3.2)	485.9 (2.9)	3.3	4.3
Cyprus	417.6 (2.7)	412.0 (3.0)	2.6	4.0
Czech Republic	503.0 (4.1)	484.7 (3.9)	3.9	5.6
Greece	452.7 (4.6)	438.8 (3.9)	3.6	6.0
Hong Kong	488.3 (3.4)	473.5 (3.8)	3.1	5.1
Hungary	472.0 (4.2)	459.6 (4.7)	3.4	6.3
Iceland	439.9 (4.0)	431.0 (3.9)	4.5	5.6
Iran, Islamic Rep.	358.7 (5.7)	354.0 (5.7)	7.8	8.1
Ireland	481.2 (4.6)	476.8 (4.4)	5.3	6.4
Japan	523.0 (2.1)	520.6 (2.0)	2.5	2.8
Korea	561.8 (2.8)	543.1 (2.7)	2.7	3.9
Latvia (LSS)	461.7 (5.2)	468.7 (4.8)	4.2	7.1
Netherlands	504.4 (3.8)	493.4 (3.1)	2.4	4.9
New Zealand	469.6 (5.9)	476.3 (5.7)	5.2	8.2
Norway	456.8 (4.6)	444.0 (4.5)	4.6	6.4
Portugal	430.8 (4.3)	415.0 (5.4)	4.7	6.9
Singapore	490.8 (5.8)	484.5 (5.2)	4.3	7.7
Thailand	428.4 (6.5)	436.6 (7.1)	3.8	9.6
England	503.3 (4.8)	495.3 (3.4)	4.7	5.9
Scotland	485.3 (4.4)	482.0 (4.7)	3.5	6.5
United States	514.2 (4.2)	508.1 (3.2)	3.8	5.2
Slovenia	495.7 (3.4)	477.7 (3.4)	3.7	4.8

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.16 Standard Error of the Gender Difference Science - Fourth Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S. E. of the Difference Using JRR	S. E. of the Difference Assuming SRS
Australia	568.9 (3.3)	555.8 (3.2)	3.0	4.6
Austria	571.8 (3.9)	556.4 (3.7)	3.7	5.3
Canada	552.7 (3.7)	545.0 (3.2)	3.0	4.9
Cyprus	480.3 (4.0)	470.6 (3.1)	2.9	5.1
Czech Republic	565.5 (3.4)	548.3 (3.6)	3.3	5.0
Greece	500.7 (4.5)	493.8 (4.3)	3.2	6.2
Hong Kong	539.7 (4.1)	525.7 (3.8)	2.9	5.6
Hungary	539.3 (3.8)	525.1 (3.9)	3.5	5.4
Iceland	513.8 (4.3)	496.2 (3.3)	3.8	5.4
Iran, Islamic Rep.	420.7 (5.9)	412.0 (4.7)	7.5	7.6
Ireland	542.8 (3.5)	536.2 (4.5)	4.5	5.7
Israel	512.2 (4.5)	501.1 (3.8)	4.0	5.9
Japan	580.4 (2.0)	566.8 (2.0)	2.0	2.9
Korea	603.8 (2.2)	589.9 (2.5)	2.9	3.3
Kuwait	389.1 (5.8)	414.3 (3.1)	7.0	6.6
Latvia (LSS)	511.7 (5.4)	512.7 (5.5)	4.7	7.7
Netherlands	569.8 (3.6)	544.3 (3.5)	3.3	5.0
New Zealand	527.0 (6.1)	535.0 (4.8)	4.9	7.7
Norway	533.6 (4.7)	525.7 (3.7)	4.4	5.9
Portugal	481.3 (4.5)	478.2 (4.2)	3.3	6.2
Singapore	548.5 (5.4)	544.5 (6.3)	5.8	8.3
Thailand	471.2 (5.9)	474.5 (4.3)	3.3	7.3
England	555.0 (4.0)	548.1 (3.4)	3.6	5.3
Scotland	537.6 (4.5)	533.4 (4.3)	2.9	6.2
United States	571.5 (3.3)	559.6 (3.3)	2.4	4.6
Slovenia	547.9 (3.3)	544.1 (4.0)	3.0	5.2

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.17 Standard Error of the Gender Difference Mathematics - Seventh Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	495.1 (5.2)	500.5 (4.3)	5.8	6.8
Austria	510.0 (4.6)	508.6 (3.3)	4.8	5.6
Belgium (Fl)	556.7 (4.5)	558.5 (4.7)	5.9	6.5
Belgium (Fr)	513.8 (4.1)	501.1 (4.2)	4.1	5.9
Bulgaria	508.0 (6.9)	518.3 (8.7)	5.1	11.1
Canada	495.1 (2.7)	493.4 (2.6)	3.1	3.8
Colombia	371.7 (3.8)	365.0 (3.9)	5.3	5.4
Cyprus	445.9 (2.5)	445.6 (2.6)	3.4	3.6
Czech Republic	526.6 (4.8)	520.3 (5.6)	3.6	7.4
Slovak Republic	510.9 (4.4)	504.9 (3.3)	3.9	5.5
Denmark	468.5 (2.8)	461.8 (2.9)	3.7	4.0
France	497.0 (3.6)	488.8 (3.3)	2.7	4.9
Germany	486.3 (4.8)	483.8 (4.5)	4.3	6.6
Greece	439.5 (3.2)	440.4 (3.0)	2.7	4.4
Hong Kong	569.7 (9.7)	555.8 (8.3)	9.6	12.8
Hungary	502.5 (3.8)	501.1 (4.4)	3.8	5.8
Iceland	460.5 (2.7)	458.3 (3.2)	2.9	4.2
Iran, Islamic Rep.	407.1 (2.7)	393.1 (2.3)	3.7	3.5
Ireland	506.7 (6.0)	493.7 (4.8)	6.9	7.7
Japan	576.4 (2.7)	565.4 (2.0)	3.0	3.4
Korea	584.4 (3.7)	567.1 (4.4)	6.2	5.7
Latvia (LSS)	463.3 (3.5)	459.6 (3.3)	3.8	4.8
Lithuania	423.3 (3.6)	433.1 (3.5)	3.2	5.0
Netherlands	517.5 (5.2)	514.6 (4.3)	4.8	6.7
New Zealand	473.1 (4.6)	470.1 (3.8)	3.7	5.9
Norway	462.4 (3.3)	458.8 (3.2)	3.2	4.6
Portugal	426.3 (2.7)	420.2 (2.2)	2.2	3.5
Romania	456.6 (3.7)	452.4 (3.7)	2.9	5.2
Russian Federation	502.4 (5.1)	499.5 (3.5)	3.5	6.1
Singapore	601.3 (7.1)	600.8 (8.0)	8.2	10.7
South Africa	351.8 (5.3)	344.2 (3.3)	4.1	6.2
Spain	450.7 (2.7)	445.2 (2.7)	3.1	3.8
Sweden	480.1 (2.8)	474.8 (3.2)	3.4	4.2
Switzerland	512.5 (2.9)	498.5 (2.6)	2.9	3.9
Thailand	494.3 (4.8)	495.4 (5.7)	4.4	7.5
England	483.9 (6.2)	467.0 (4.3)	8.3	7.5
Scotland	464.5 (4.6)	461.7 (3.8)	3.8	5.9
United States	478.1 (5.7)	473.3 (5.7)	3.2	8.1
Slovenia	500.6 (3.5)	495.8 (3.2)	3.2	4.7

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.18 Standard Error of the Gender Difference Mathematics - Eighth Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	527.4 (5.1)	532.0 (4.6)	5.4	6.9
Austria	543.6 (3.2)	535.6 (4.5)	4.9	5.6
Belgium (Fl)	563.1 (8.8)	567.2 (7.4)	11.7	11.5
Belgium (Fr)	530.0 (4.7)	523.5 (3.7)	5.1	6.0
Bulgaria	533.2 (7.0)	546.2 (6.7)	5.2	9.6
Canada	526.0 (3.2)	529.6 (2.7)	3.4	4.2
Colombia	385.7 (6.9)	384.0 (3.6)	8.2	7.7
Cyprus	472.2 (2.8)	475.3 (2.5)	3.7	3.7
Czech Republic	569.0 (4.5)	558.4 (6.3)	4.5	7.7
Slovak Republic	549.0 (3.7)	545.3 (3.6)	3.2	5.2
Denmark	511.5 (3.2)	494.3 (3.4)	3.4	4.7
France	541.9 (3.1)	535.7 (3.8)	3.2	4.9
Germany	511.6 (5.1)	509.1 (5.0)	4.7	7.1
Greece	489.7 (3.7)	477.8 (3.1)	2.9	4.8
Hong Kong	597.2 (7.7)	577.2 (7.7)	8.6	10.9
Hungary	537.3 (3.6)	537.2 (3.6)	3.3	5.1
Iceland	487.6 (5.5)	485.9 (5.6)	6.3	7.8
Iran, Islamic Rep.	434.1 (2.9)	420.8 (3.3)	4.5	4.4
Ireland	534.6 (7.2)	520.3 (6.0)	8.2	9.3
Israel	538.7 (6.6)	509.4 (6.9)	5.8	9.6
Japan	609.2 (2.6)	600.0 (2.1)	2.9	3.3
Korea	615.2 (3.2)	597.9 (3.4)	4.8	4.7
Kuwait	389.0 (4.3)	395.5 (2.6)	5.0	5.0
Latvia (LSS)	495.6 (3.8)	491.2 (3.5)	3.7	5.2
Lithuania	476.8 (4.0)	477.6 (4.1)	4.0	5.7
Netherlands	544.8 (7.8)	536.4 (6.4)	4.4	10.1
New Zealand	512.2 (5.9)	503.0 (5.3)	6.7	7.9
Norway	505.3 (2.8)	501.3 (2.7)	3.3	3.9
Portugal	459.8 (2.8)	448.9 (2.7)	2.4	3.9
Romania	482.9 (4.8)	480.2 (4.0)	3.4	6.2
Russian Federation	534.8 (6.3)	536.0 (5.0)	3.7	8.0
Singapore	642.2 (6.3)	644.6 (5.4)	6.5	8.3
South Africa	359.8 (6.3)	349.2 (4.1)	5.7	7.5
Spain	492.2 (2.5)	482.7 (2.6)	3.2	3.6
Sweden	519.5 (3.6)	517.7 (3.1)	3.1	4.7
Switzerland	547.8 (3.5)	543.0 (3.1)	3.7	4.7
Thailand	517.0 (5.6)	526.2 (7.0)	6.5	9.0
England	507.7 (5.1)	503.5 (3.5)	7.1	6.2
Scotland	506.2 (6.6)	490.3 (5.2)	4.9	8.4
United States	502.0 (5.2)	497.5 (4.5)	2.9	6.9
Slovenia	544.9 (3.8)	536.9 (3.3)	3.4	5.0

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.19 Standard Error of the Gender Difference Science - Eighth Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	506.6 (5.2)	502.3 (4.0)	5.9	6.6
Austria	522.3 (4.3)	515.5 (4.1)	5.2	6.0
Belgium (Fl)	535.8 (3.3)	521.4 (3.1)	3.9	4.5
Belgium (Fr)	452.7 (3.6)	432.1 (3.5)	3.5	5.0
Bulgaria	529.2 (5.5)	532.0 (6.7)	5.7	8.7
Canada	505.5 (2.9)	493.1 (2.5)	2.9	3.8
Colombia	396.4 (3.8)	378.5 (4.4)	4.8	5.8
Cyprus	420.1 (2.8)	420.1 (2.6)	4.1	3.9
Czech Republic	543.2 (3.2)	522.9 (4.1)	3.2	5.2
Slovak Republic	520.3 (4.0)	499.4 (3.1)	3.9	5.1
Denmark	452.0 (3.0)	427.4 (2.8)	3.9	4.1
France	460.8 (3.1)	442.7 (3.0)	3.1	4.3
Germany	504.9 (4.9)	495.4 (4.5)	4.5	6.6
Greece	451.7 (3.2)	445.5 (2.8)	3.1	4.2
Hong Kong	503.5 (6.6)	485.0 (5.8)	6.3	8.7
Hungary	525.3 (3.9)	510.5 (3.4)	3.4	5.1
Iceland	467.7 (4.4)	455.9 (2.4)	4.5	5.0
Iran, Islamic Rep.	443.0 (2.9)	427.8 (4.1)	4.9	5.0
Ireland	504.4 (4.6)	487.3 (4.5)	5.9	6.4
Japan	536.0 (2.6)	525.8 (1.9)	2.7	3.2
Korea	545.4 (2.8)	520.8 (3.2)	4.4	4.2
Latvia (LSS)	439.6 (3.6)	430.1 (3.0)	3.8	4.7
Lithuania	405.4 (3.5)	400.7 (4.2)	3.8	5.5
Netherlands	522.8 (4.0)	512.2 (4.4)	4.4	5.9
New Zealand	489.1 (4.3)	471.7 (3.7)	4.3	5.7
Norway	488.9 (3.6)	477.2 (3.6)	4.3	5.1
Portugal	436.3 (2.4)	420.1 (2.4)	2.4	3.4
Romania	455.8 (4.7)	447.7 (4.9)	3.5	6.7
Russian Federation	492.9 (5.3)	475.4 (3.8)	3.8	6.5
Singapore	548.1 (7.9)	541.3 (8.2)	9.2	11.4
South Africa	323.8 (6.4)	312.5 (5.2)	4.9	8.3
Spain	487.5 (2.9)	466.7 (2.3)	3.3	3.7
Sweden	493.3 (2.9)	483.6 (3.3)	3.5	4.4
Switzerland	492.4 (2.9)	474.8 (2.9)	3.0	4.1
Thailand	494.8 (3.3)	491.7 (3.5)	3.1	4.8
England	522.2 (5.6)	499.9 (4.6)	8.0	7.3
Scotland	477.0 (4.4)	459.2 (4.1)	3.8	6.0
United States	514.4 (6.3)	502.2 (5.8)	5.2	8.6
Slovenia	539.2 (3.0)	521.2 (2.8)	3.2	4.1

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.20 Standard Error of the Gender Difference Science - Eighth Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	549.6 (5.2)	539.5 (4.1)	5.3	6.6
Austria	566.4 (4.0)	548.7 (4.6)	4.3	6.1
Belgium (Fl)	557.6 (6.0)	542.9 (5.8)	8.7	8.4
Belgium (Fr)	478.9 (4.8)	463.0 (2.9)	5.5	5.6
Bulgaria	563.2 (5.7)	566.8 (6.6)	6.3	8.7
Canada	537.4 (3.1)	525.4 (3.7)	4.3	4.8
Colombia	417.6 (7.3)	404.9 (4.6)	8.4	8.6
Cyprus	461.0 (2.2)	464.8 (2.7)	3.0	3.4
Czech Republic	585.9 (4.2)	561.6 (5.8)	4.5	7.2
Slovak Republic	552.2 (3.5)	536.9 (3.9)	3.6	5.2
Denmark	494.2 (3.6)	463.3 (3.9)	4.5	5.3
France	505.9 (2.7)	490.1 (3.3)	3.1	4.3
Germany	541.7 (5.9)	523.9 (4.9)	4.8	7.6
Greece	504.8 (2.6)	489.3 (3.1)	3.3	4.0
Hong Kong	534.7 (5.5)	507.3 (5.1)	5.8	7.5
Hungary	563.0 (3.1)	544.6 (3.4)	3.6	4.7
Iceland	501.1 (5.1)	485.5 (4.6)	5.2	6.9
Iran, Islamic Rep.	477.3 (3.8)	460.5 (3.2)	5.2	4.9
Ireland	543.6 (6.6)	532.0 (5.2)	7.6	8.4
Israel	544.8 (6.4)	512.2 (6.1)	7.1	8.9
Japan	579.0 (2.4)	562.4 (2.0)	3.0	3.1
Korea	575.9 (2.7)	551.5 (2.3)	3.8	3.6
Kuwait	416.0 (6.6)	443.5 (3.3)	7.4	7.4
Latvia (LSS)	492.4 (3.3)	477.8 (3.2)	3.5	4.6
Lithuania	483.9 (3.8)	470.3 (4.0)	3.9	5.5
Netherlands	570.2 (6.4)	549.8 (4.9)	5.1	8.1
New Zealand	537.6 (5.4)	512.3 (5.2)	6.2	7.6
Norway	534.0 (3.2)	520.5 (2.0)	3.7	3.8
Portugal	490.5 (2.8)	468.4 (2.7)	2.8	3.9
Romania	492.0 (5.3)	480.1 (5.0)	3.8	7.3
Russian Federation	544.0 (4.9)	532.9 (3.7)	3.4	6.2
Singapore	611.9 (6.7)	602.7 (7.0)	8.1	9.7
South Africa	336.6 (9.5)	315.4 (6.0)	8.6	11.3
Spain	526.4 (2.1)	508.1 (2.3)	2.9	3.1
Sweden	542.5 (3.4)	528.0 (3.4)	3.4	4.8
Switzerland	529.0 (3.2)	514.0 (3.0)	3.7	4.4
Thailand	524.4 (3.9)	526.3 (4.3)	3.6	5.8
England	561.6 (5.6)	541.6 (4.2)	7.7	7.1
Scotland	527.3 (6.4)	506.9 (4.7)	5.1	7.9
United States	538.8 (4.9)	530.0 (5.2)	3.6	7.2
Slovenia	573.2 (3.2)	547.8 (3.2)	4.1	4.5

JRR = jackknife repeated replicate method

SRS = simple random sample

8.7 REPORTING POPULATION 1 ACHIEVEMENT ON THE POPULATION 2 SCALE

In order to establish a link between the reporting scales for Population 1 and Population 2, a number of items in the TIMSS tests were administered to students in both populations. A total of 15 mathematics and 18 science items were administered in both populations, at grades three and four and at grades seven and eight. The 15 mathematics items were exclusively multiple choice, while the 18 science items consisted of 10 multiple-choice items and 8 free-response items. All of these items were dichotomously scored, and were worth one score point each. Because of the existence of these "link items," it was possible to link the Population 1 results to those of Population 2.

8.7.1 Estimating the Shift in Item Difficulties

The scaling of the student achievement data for Population 2 and the reporting of results on that scale were completed before those for Population 1. Because of this, the scales from the two populations were linked by reporting the Population 1 results on the Population 2 scale. In order to achieve this, the item difficulties were first calibrated separately in each population. The link items were then identified and the average of the differences between the item difficulties from each of the calibrations was computed, separately for mathematics and science. This average is an estimate of the shift in item difficulty that would have to be made in order to report the results from the Population 1 scaling on a scale based on the calibration of the Population 2 items. Table 8.21 and 8.22 present the mathematics and science link items with their item difficulties calibrated separately for Population 1 and Population 2, the difference between them, and the average of the difference (the shift) calculated as

$$Shift(s) = \frac{\sum_{L}(d_l^{pop1} - d_l^{pop2})}{L}$$

where L is the number of link items, d_l^{pop1} is the item difficulty of item L at Population 1, d_l^{pop2} is the item difficulty of item L at Population 2. This shift is applied to the logit metric in which the Population 1 scores are first computed.

Table 8.21 Mathematics Link Items

Item Name in Population 1	Item Name in Population 2	Difficulty at Population 1	Difficulty at Population 2	Difference in Difficulty Between Populations
F08	D11	-0.227	-1.906	-1.679
K07	E06	1.852	0.693	-1.159
C03	H08	0.352	-1.047	-1.399
G04	H12	0.414	-0.996	-1.410
U02	I06	0.297	-1.216	-1.513
G01	J17	0.984	-0.585	-1.569
H05	K03	0.461	-0.644	-1.105
F05	L10	-0.516	-2.025	-1.509
L08	L12	1.461	-0.977	-2.438
L04	L13	-0.516	-2.332	-1.816
L02	M03	0.516	-1.143	-1.659
B07	P12	0.758	-0.809	-1.567
F06	P14	-0.164	-1.262	-1.098
B05	Q04	-0.234	-1.777	-1.543
I09	R12	-0.250	-1.953	-1.703
Average		0.346	-1.199	-1.544
Standard Error				0.085

Table 8.22 Science Link Items

Item Name in Population 1	Item Name in Population 2	Difficulty at Population 1	Difficulty at Population 2	Difference in Difficulty Between Populations
D04	B01	-0.773	-1.845	-1.072
E07	B04	0.000	-1.998	-1.998
N08	C10	-0.008	-1.102	-1.094
P05	D02	0.633	-0.914	-1.547
P09	D06	0.805	-0.877	-1.682
B04	F03	0.031	-0.515	-0.546
O04	H03	-0.477	-1.233	-0.756
Q02	I10	-0.211	-0.921	-0.710
O06	K19	0.156	-0.852	-1.008
Q08	M14	0.617	-0.764	-1.381
Q04	N07	0.047	-2.016	-2.063
O01	N08	0.680	-0.727	-1.407
R01	N10	2.109	0.123	-1.986
Y01	O14	1.516	0.053	-1.463
W03	O16	1.656	-0.112	-1.768
O05	R01	0.156	-0.654	-0.810
Z01A	W01A	0.023	-1.306	-1.329
Z01B	W01B	2.000	0.606	-1.394
Average		0.498	-0.836	-1.334
Standard Error				0.109

8.7.2 Estimating the Variance of the Shift

Because the student responses from which the item difficulty parameters are estimated are derived from random samples of students, the estimates of the relative item difficulty of the items in the two samples are subject to sampling variation. It is important to take this variation into account when reporting results on a scale that has been constructed by means of a shift from another scale. This variance component, known as the variance of the shift, is computed as the variance of the differences in item difficulty with respect to the mean difference in item difficulty. The formula for this calculation is as follows

$$Var_{Shift(s)} = \frac{\sum_{L}((d_l^{pop1} - d_l^{pop2}) - Shift(s))^2}{L^2}$$

where L is the number of link items, d_l^{pop1} is the item difficulty of item L at Population 1, d_l^{pop2} is the item difficulty of item L at Population 2, and *Shift (s)* is the average difference between two calibrations. The variance of the shift is used only when reporting the scores from one scale onto another scale. This variance component is added to the standard variance estimate.

REFERENCES

Dunn, O.J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*, 52-64.

Winer, B. J., Brown, D. R., and Michels, K. M. (1991). *Statistical principles in experimental design.* New York: McGraw Hill.

9 Reporting Achievement in Mathematics and Science Content Areas

Albert E. Beaton
Eugenio J. Gonzalez
Boston College

9.1 ADAPTING AVERAGE PROPORTION-CORRECT TECHNOLOGY FOR TIMSS

Although item response theory (IRT) methods were used to scale the student achievement data for purposes of international reporting, TIMSS also made use of an approach whereby the proportion of items answered correctly by the students in a country was averaged over the set of items in a subject matter content area. This "average-proportion-correct technology" was used for reporting performance in each of the 11 content areas of mathematics and science that were assessed at the seventh and eighth grades, and each of the 10 content areas that were assessed at the third and fourth grades. The content scales assessed in each subject, at each grade level, are presented in Table 9.1. Average proportion-correct technology was also used for the Test Curriculum Matching Analyses (TCMA) described in Chapter 10. This approach allows the averaging across items, even though the items are located in different assessment booklets and individual students do not respond to all of the items being averaged. Using this technology, it is also possible to obtain standard errors for the proportion correct with a slight modification of the jackknife repeated replicate (JRR) variance estimation procedures outlined in Chapter 5.

Table 9.1 Mathematics and Science Content Areas

Third and Fourth Grades (Population 1)

Mathematics	Science
• Whole Numbers	• Earth Science
• Fractions and Proportionality	• Life Science
• Measurement, Estimation, and Number Sense	• Physical Science
• Data Representation, Analysis, and Probability	• Environmental Issues and the Nature of Science
• Geometry	
• Patterns, Relations, and Functions	

Seventh and Eighth Grades (Population 2)

Mathematics	Science
• Fractions and Number Sense	• Earth Science
• Geometry	• Life Science
• Algebra	• Physics
• Data Representation, Analysis, and Probability	• Chemistry
• Measurement	• Environmental Issues and the Nature of Science
• Proportionality	

Unlike the TIMSS IRT scaling, the average proportion-correct approach does not provide scores or plausible values for individual students, and is also sensitive to ceiling effects on sets of items, in particular when a subpopulation of interest responds correctly to most or all of the items in a set. However, the average proportion-correct approach was used in TIMSS for reporting student performance in subject matter content areas and for the TCMA analyses in preference to IRT scaling because of cost considerations, and because of the extra time the more complex scaling approach would have required.

Adapting the average proportion-correct technology for TIMSS posed two particular problems. The first was that some of the TIMSS items had graded responses, that is, the students were assigned a score ranging from 0 to 3 points depending on the item and the degree of correctness of their responses to the item. When an item response can have only two values, 0 for incorrect and 1 for correct, the average score on the item for a sample of students is also the proportion correct. However, this does not hold for an item where responses can score more than 1. For such items, it was necessary to find a way to use the proportion correct to represent the responses.

The second problem was that occasionally an item was found to be unusable for some countries. The item review process (see Chapter 6) revealed that from time to time an item for a country was misprinted, mistranslated, or missing by mistake from the booklet, or had other problems that prevented them from being comparable with the items administered in other countries. Such items were deleted for the country concerned; however, they could affect the overall proportion correct for a specific country if, for example, a country happened to have mistranslated the most difficult item in a content area. While such missing items are handled readily by IRT methods, they cause difficulties for the average proportion-correct approach. The items deleted at each population are documented in Chapter 6.

9.1.1 Treating Graded Response Items

A simple way to handle graded responses would be to compute the average score on each of the items in a particular area and then add up these averages to obtain the average score on the scale. The statistic computed for each country would then be the sum of its averages for the items involved in the area. The average for a binary (right/wrong) item in this situation would be its proportion correct, and for a graded response the average score on the item. However, an average computed this way would have an upper bound equal to the total number of score points possible divided by the total number of items. If any of the items were graded-response items with maximum scores greater than 1, the upper bound for the average would be greater than 1, and the average would not be interpretable as a proportion-correct.

By transforming the graded responses into a series of binary items, TIMSS was able to use the proportion-correct technology without losing information, and, in fact ,some additional information was gained. Consider that an item may be assigned a score of 0, 1, 2, or 3. We can code a student's response as if it were three variables ($v_{j,1}$, $v_{j,2,}$ $v_{j,3}$) as follows:

$v_{j,1}$ equals 1 if the student receives a 1, 2, or 3, and 0 otherwise;

$v_{j,2}$ equals 1 if the student receives a 2 or 3, and 0 otherwise; and

$v_{j,3}$ equals 1 if the student receives a 3, and 0 otherwise.

We can then call $p_{vj,1}$, $p_{vj,2}$, *and* $p_{vj,3}$ the proportions of students who received a 1 on $v_{j,1}$, $v_{j,2}$, and $v_{j,3}$ respectively. Note in particular that $p_{vj,1} \geq p_{vj,2} \geq p_{vj,3}$. The average value of item v_j can then be computed from the proportions of students who receive a score of 1 on $v_{j,1}$, $v_{j,2}$, $v_{j,3}$, that is,

$$\bar{v}_j = \sum_i p_{vj,i}$$

We can also compute the average proportion correct on these three items (p_j) as

$$p_j = \frac{1}{I}\sum_i p_{vj,i}$$

where I is the maximum score points on the item.

As a numerical example, let us assume the frequency distribution shown in Table 9.2 for a graded-response item administered to 1,000 students within a country.

Table 9.2 Sample Frequency of Responses to Item v_j

Score	Frequency
0	200
1	300
2	400
3	100
Total	1000

The score on this item is 1.4, computed as follows:

$$\bar{v}_j = \frac{200 * 0 + 300 * 1 + 400 * 2 + 100 * 3}{1000} = \frac{1400}{1000} = 1.4$$

The three proportions for this item would then be

$p_{vj,1} = 0.80$

$p_{vj,2} = 0.50$

$p_{vj,3} = 0.10$

from which the same average can be computed by

$$\bar{v}_j = \sum_i p_{vj,\,i} = 0.80 + 0.50 + 0.10 = 1.4$$

Using this method of coding we can treat the graded-response items as binary items and still compute the average score for an item allowing the full range of values. If all graded-response items are coded in this way, then the average proportion correct over any set of items will be proportionally the same as if the averages of graded items were mixed with the percentages correct of binary items. Yet we have gained the advantage of working only with proportions.

We note that the three proportions in our example ($p_{vj,1}$, $p_{vj,2}$, and $p_{vj,3}$) contain some information that the average graded response does not. From these proportions, we can see what proportion of students in a country responded at each score level for that item. When this procedure is used, the number of items is then effectively increased from j to j', where j' is equal to the total number of possible scores points on the set of items.

9.1.2 Missing Proportions Correct

A second problem with the reporting of average proportion correct was what to do on those rare occasions when an item has to be deleted for a country. It is important that the deletion should neither penalize nor benefit the country. Where an item was found to be unusable for a country, that item could be omitted from the analysis for all countries without any threat to fairness, but since different items exhibited problems in different countries, this would reduce the total item pool unacceptably, and would necessitate discarding perfectly good data for the unaffected countries. On the other hand, if the item is deleted only for the affected country, there is the possibility of unduly influencing the country's overall score. To minimize the effect of deleted items on overall average proportion correct, TIMSS derived a method of estimating the proportion of students in the country that would have performed successfully on the items if they had been included. To achieve this, TIMSS used the information on how the country performed on the remaining items, and how the other countries performed on the item in question. Transforming all items into binary variables as described in the preceding section greatly facilitated the implementation of this procedure.

Note that this approach was used when average proportions correct were used for cross-national comparisons. The IRT scaling did not require this procedure since one of the advantages of IRT scaling is its capacity to deal with missing items.

9.1.3 Computational Method

The TIMSS approach was as follows: Let us assume that we want to estimate the average proportion correct over a set of items for a set of countries but that one country, country k, has mistranslated item j' and therefore the proportion correct for country k on item j' cannot be known from the available data.[1] Different countries may have dif-

[1] We will use the notation j' for an item to signify the dichotomized version of the item, as described in the section on *Treating Graded Response Items*.

ferent unusable items and thus different missing proportions. The TIMSS procedure may be used as long as there is at least one known proportion for each country and for each item, although of course it works best when there are just a few missing items.

The TIMSS approach begins by filling in the missing values using the model

$$p_{kj'} = p_{k0} + p_{0j'} - p_{00}$$

where $p_{kj'}$ is the estimated proportion correct of country k on its unusable item j', p_{k0} is the average proportion correct of country k on all of its usable items, $p_{0j'}$ is the average proportion correct of all other countries on item j', and p_{00} is the average proportion correct for all available items over all countries. Imputation under this model implies that there is no interaction between the proportion correct on the imputed item and the countries.

The above model was improved in two ways. First, filling in an estimated value of $p_{kj'}$ affects the values of p_{k0}, $p_{0j'}$, and p_{00}, so the method should be iterative, making successive estimates until all values stabilize. Second, proportion correct is not a good statistic for an additive model such as is specified above; in fact, unless the proportions are transformed to an additive metric, estimated proportions of greater that 1 or less than 0 are possible. The use of the logit transformation of the proportions avoids this problem.

The logit transformation used to transform the percents correct into an additive scale is

$$z_{kj'} = Logit(p_{kj'}) = \ln\left(\frac{p_{kj'}}{1 - p_{kj'}}\right)$$

Using this equation transforms a proportion correct for an item ($p_{kj'}$) to a logit value ($z_{kj'}$) that may range from minus to plus infinity. The logit for p = .50 is zero. The logit for 0 is minus infinity and for 1 is plus infinity, and so values of 0 and 1 are not usable. In the unusual case when there is a value of 1.0 or 0.0 for a proportion correct for an item, 0.9999 is substituted for 1.0, and 0.0001 is substituted for 0.0. This logit transformation permits simple and appropriate arithmetic calculations on proportions.

If we now define a matrix of proportions $P_{kj'}$ where k is the number of countries and j' is the number of items, and some of the elements of $P_{kj'}$ are missing, the method used to estimate the missing proportion correct works as described below.

Step 1: The matrix with logit scores $Z_{kj'}$ is produced from the usable elements of the matrix $P_{kj'}$ by the transformation of the elements in $P_{kj'}$ into logit scores as defined above. The elements $z_{kj'}$ when item j' is deemed unusable in country k are left blank in this $Z_{kj'}$ matrix. The matrix $Z_{kj'}$ also has a "zeroth" row and column. The elements in z_{k0} contain the average of the elements on the kth row of the $Z_{kj'}$ matrix. These are the country averages across the usable items. The elements in $z_{0j'}$ contain the average of the elements of the j'th column of the $Z_{kj'}$

matrix. These are the item averages across all countries. The element z_{00} contains the overall average for the elements in vector $z_{0j'}$ and z_{k0}. In the initial matrix $Z_{kj'}$, the averages are defined over the usable $z_{kj'}$ elements and the missing values are not used.

Step 2: The first estimation for the logits of the missing $z_{kj'}$ values is then given by the formula

$$z'_{kj'} = z_{k0} + z_{0j'} + z_{00}$$

Step 3: At this point a new matrix $Z'_{kj'}$ is created where each of the $Z'_{kj'}$ elements are the same as those in $Z_{kj'}$, but where the missing $z_{kj'}$ elements are replaced with the newly estimated $z'_{kj'}$.

Step 4: New averages are computed for the vectors z'_{k0}, $z'_{0j'}$, and z'_{00} with the elements of the newly created $Z'_{kj'}$, matrix. These averages can now be computed over all available values in $Z'_{kj'}$ which is now a complete matrix with no missing elements.

Step 5: New estimates for the missing elements in the $Z_{kj'}$ matrix are then computed as

$$z'_{kj'} = z'_{k0} + z'_{0j'} - z'_{00}$$

where $z'_{kj'}$, z'_{k0}, $z'_{0j'}$, and z'_{00} are the values obtained from the $Z'_{kj'}$ matrix on the succeeding iterations.

Steps 3 through 5 above are repeated until a stable solution has been reached. The criterion for convergence is that none of the elements in the $z'_{kj'}$ vectors changes more than .001 from one iteration to the next.

Once a stabilized $Z'_{kj'}$ matrix is obtained, the estimates for the missing elements in $P_{kj'}$ are obtained by creating the matrix $P'_{kj'}$ using the inverse logit transformation

$$p'_{kj'} = \frac{\exp(z'_{kj'})}{1 + \exp(z'_{kj'})}$$

and applying it to each of the elements of the $Z'_{kj'}$ matrix.

The average percent correct on a scale for each country is then obtained by averaging the rows of the $P'_{kj'}$ matrix.

In doing this, notice that the average proportions correct for countries that have all usable data, or for items that were usable for all countries, remain unchanged. In TIMSS the missing proportion-correct values for the unusable items were imputed using only the information for the content area to which the item was assigned. These imputed percent-correct values were then used in the computation of the average percent correct at the content area level and overall for each subject.

9.1.4 Computing Standard Errors

Once the estimates for the missing elements of the $P_{kj'}$ matrix are obtained, the average percent correct for the items of a scale in a country can be computed. These average percents correct are the elements of the vector P_{k0} from the matrix $P_{kj'}$. Each of the $p_{kj'}$ values was computed using the overall sampling weight.

In order to obtain variance estimates for the average percent corrects, it is possible to make use of the replicate weights approach used by the jackknife algorithm to estimate the sampling variability of the data used to fill in the blanks in the $P_{kj'}$ matrix. It is important to keep in mind that the estimated elements of the matrix $P_{kj'}$ are computed using the elements in the vectors $P_{0j'}$ *and* P_{k0} and therefore are subject to variability as repeated replicate samples are drawn from each country. To implement the jackknife repeated replication (JRR) procedure in this case, the sampling zones across the countries are randomly sorted, and information from different zones by country is used to obtain each of the 75 estimates from which the sampling errors are computed. When the sampling zones within a country are sorted they are renumbered and treated as an international zone or international replicate.

The JRR procedure was implemented as follows. TIMSS assigned the schools within each country in pairs to one of up to H jackknife zones, where H is equal to 75. The 75 sampling zones were used to create 75 “pseudo-replicates” of the original sample. Each of the pseudo-replicates consists of a copy of the original data, except that in one of the sampling zones (a different one each time) one school of the pair of schools, chosen at random, is omitted, and the weights for the other member of the pair are doubled. In computing a jackknife estimate of the sampling variability of a statistic such as a mean or a proportion, the statistic is computed once for the data in the original sample, and once again for each of the pseudo-replicate samples. The variation between the original sample estimate and the estimates from each of the replicate samples is the jackknife estimate of the sampling error of the statistic.

Doubling or omitting the weights of the selected school within each sampling zone is accomplished effectively in computational terms by the creation of replicate weights. The replicate-weight approach requires the temporary creation of a new set of weights for each replicate sample. To create the replicate weights for the first replicate sample, one of the pair of schools in the first sampling zone is chosen at random to have its weights doubled, while the other member of the pair has its weights set to zero to compensate. The weights of the schools in all other sampling zones are left unchanged. The replicate weights for the second replicate sample are created in a similar manner. Again, the weights for the schools in all other zones are unchanged from the original weights. This procedure is repeated for all 75 sampling zones, resulting in 75 sets of replicate weights (W_h) for each country.

Using these 75 replicate weights we then compute for each country k a matrix $T^{k}_{h'j'}$ where the row elements across the hth row are the proportion correct of each of the j' items in the scale computed using the hth replicate weight, and the elements down each j'th column are the proportion correct for the j'th item computed using each of the hth replicate weights. The row vectors of this matrix for the kth country are then randomly sorted so that the order of the replicate weights used to compute each row now varies by country. The rows of each of these matrices are now renumbered using the indexing variable h', and the newly sorted matrix is called $T^{k}_{h'j'}$.

At this point we then proceed to form each of the 75 $P^{h'}_{kj'}$ matrices by taking the h'^{th} row from the $T^{k}_{h'j'}$ matrices. After the estimation of the missing elements of each of the $P^{h'}_{kj'}$ matrices takes place, the resulting 75 $P'^{h'}_{k0}$ vectors will contain the H replicates for each of the k countries in the sample. At this point the standard method for estimating the sampling variance is used by applying the following equation for each country:

$$jse_{P'_{k0}} = \sqrt{\sum_{h'}(p'_{k0} - p'^{h'}_{k0})}$$

9.2 PROFILES OF RELATIVE PERFORMANCE BY CONTENT AREAS

In addition to performance on mathematics and science overall, it was of interest to see how countries performed on the content areas within each subject relative to their performance on the subject overall. If the results for all countries are summarized in a table of average percents correct organized by country and by content area, then differences in relative performance across content areas for a country may be thought of as a country-by-content area interaction. There were six content areas in mathematics at each population, and four science content areas at Population 1 and five at Population 2, that were used in this analysis. The relative performance for the countries on the content areas was examined separately for each subject.

Suppose now that we have computed the vector of average percent corrects (P'_{k0}) for each of the content areas on the test using the procedures described earlier, and that we join each of these column vectors to form a new matrix called R_{ks} where a row contains the average percent correct for country k on scale s for a specific subject. This R_{ks} matrix has also a "zeroth" row and column. The elements in r_{k0} contain the average of the elements on the kth row of the R_{ks} matrix. These are the country averages across the content areas. The elements in r_{0s} contain the average of the elements of the sth column of the R_{ks} matrix. These are the content area averages across all countries. The element r_{00} contains the overall average for the elements in vector r_{0j} or r_{k0}. Based on this information we can then construct the matrix R'_s in which the elements are computed as

$$r'_{ks} = r_{ks} + r_{00} - r_{0s} - r_{k0}$$

Each of these elements can be considered as the interaction between the performance of country k on content area s. A value of zero for an element r'_s indicates a level of performance for country k on content area s that would be expected given its performance on other content areas and its performance relative to other countries on that content area. A negative value for an element r'_s indicates a performance for country k on content area s lower than would be expected on the basis of the country's overall performance. A positive value for an element r'_s indicates a performance for country k on content area s better than expected.

Although we can compute the values for the country by content area interaction, this value is of little interest unless we can determine whether it is significantly different from zero. To do this we need to compute the corresponding standard error for each of the r'_s elements and perform a test of significance, taking into account the multiple comparisons by using the Dunn-Bonferroni procedure (see Chapter 8).

To compute the JRR standard error, suppose that we have computed the vector of average percents correct for each of the international replicates $P'^{h'}_{k0}$ for each of the content areas on the test using the procedures described in the previous chapter, and that we join each of these column vectors to form a new set of matrices each called R^h_{ks} where a row contains the average percent correct for country *k* on content area *s* for a specific subject, for the *h*th international set of replicates. Each of these R^h_{ks} matrices has also a "zeroth" row and column. The elements in r^h_{k0} contain the average of the elements on the *k*th row of the R^h_{ks} matrix. These are the country averages across the content areas. The elements in r^h_{0s} contain the average of the elements of the *s*th column of the matrix. These are the content area averages across all countries. The element r^h_{00} contains the overall average for the elements in vector r^h_{0j} or r^h_{k0}. Based on this information we can then construct the set of matrices R'^h_{ks} in which the elements are computed as

$$r'^h_{ks} = r^h_{ks} + r^h_{00} - r^h_{0s} - r^h_{k0}$$

The JRR standard error is given by the formula

$$jse_{r_{ks}} = \sqrt{\sum_h (r'_{ks} - r'^h_{ks})^2}$$

A relative performance was considered significantly different from the expected if the 95 percent confidence interval built around it did not include zero. The confidence interval for each of the r'_{ks} elements was computed by adding and subtracting to the r'_{ks} element its jackknifed standard error multiplied by the critical value for the number of comparisons.

The critical values were determined by adjusting the critical value for a two-tailed test, at the alpha 0.05 level of significance for multiple comparisons according the Dunn-Bonferroni procedure. Since the number of scales varied by subject, and the number of countries varied by grade, eight different critical values were computed. Table 9.3 summarizes the number of comparisons performed by subject at each grade level.

Table 9.3 Number of Comparisons and Critical Values Used for the Test of Significance of the Relative Performance Within Country

Subject	Grade	Countries	Scales	Comparisons	Critical Value
Science	8th	41	5	205	3.6683
Science	7th	39	5	195	3.6554
Mathematics	8th	41	6	246	3.7148
Mathematics	7th	39	6	234	3.7020
Science	3rd	24	4	96	3.4698
Science	4th	26	4	104	3.4913
Mathematics	3rd	24	6	144	3.5774
Mathematics	4th	26	6	156	3.5984

9.3 PERCENT CORRECT FOR INDIVIDUAL ITEMS

To portray student achievement as fully as possible, the TIMSS international reports present many examples of the items used in the TIMSS tests, together with the percentage of students in each country responding correctly to the item. For multiple-choice items this was the weighted percentage of students that answered the item correctly. This percentage was based on the total number of students that were administered the items. Omitted and not-reached items were treated as incorrect. For free-response items with more than one score level the percent correct for these example items was computed as the weighted percentage of students that achieved the highest score possible on the item.

When the percent correct for example items were computed, student responses were classified in the following way. For multiple-choice items, the responses to item *j* were classified as correct (C_j) when the correct option for an item was selected, incorrect (W_j) when the incorrect option for an item was selected, invalid (I_j) when two or more choices were made on the same question, not reached (R_j) when it was determined that the student stopped working on the test before reaching the question, and not administered (A_j) when the question was not included in the student's booklet or the question was mistranslated or misprinted. For free-response items student responses to item *j* were classified as correct (C_j) when the maximum number of points was obtained on the question, incorrect (W_j) when the wrong answer or an answer not worth all the points in the question was given, invalid (N_j) when, although something was written in the answer sheet, what was written was not legible or interpretable, not reached (R_j) when it was determined that the student stopped working on the test before reaching the question, and not administered (A_j) when the question was not included in the student's booklet or the question was mistranslated or misprinted. The percent correct for an item (P_j) was computed as

$$P_j = \frac{c_j}{c_j + w_j + i_j + r_j + n_j}$$

where c_j, w_j, i_j, r_j and n_j are the weighted counts of the correct, wrong, invalid, not reached, and not interpretable responses to item j, respectively.

Note that although the not-reached responses were treated as missing for the purpose of estimating the item parameters in the international IRT scaling, they were considered to be wrong answers for an individual when percents correct for an item were computed.

9.4 REPORTING GENDER DIFFERENCES BY CONTENT AREAS

Differences between the performance of boys and girls in the subject matter content areas were also examined using the average percent-correct approach. The performance difference was determined to be significant if the standardized difference between the average percent correct for boys and girls within a country exceeded the critical value, corrected using the Dunn-Bonferroni procedure for multiple comparisons.

The standardized difference between the average percent corrects (t_k) was computed as

$$t_k = \frac{\bar{p}_{kb} - \bar{p}_{kg}}{\sqrt{pse_{kb}^2 + pse_{kg}^2}}$$

where $\bar{p}_{kp}$ and $\bar{p}_{kg}$ are the average percents correct within the content area for boys and girls, respectively, within country k, and pse_{kb} and pse_{kg} are the standard errors of the average percents correct for boys and girls, respectively, within country k computed using the jackknife procedure for estimating sampling error. The critical value for the seventh grade was 3.22005, and for the eighth grade was 3.23431. These critical values are corrected using the Dunn-Bonferroni procedure for multiple comparisons. At the seventh grade, the critical value was corrected for 39 comparisons, and at the eighth grade for 41 comparisons. The critical value used for the third and fourth grade tests of significance was 1.960. This critical value was not adjusted for multiple comparisons.

10 TIMSS Test-Curriculum Matching Analysis

Albert E. Beaton
Eugenio J. Gonzalez
Boston College

10.1 INTRODUCTION

TIMSS developed international tests of mathematics and science that reflect as far as possible the various curricula of the participating countries. The subject matter coverage of these tests was reviewed by the TIMSS Subject Matter Advisory Committee (SMAC), which consists of mathematics and science educators and practitioners from around the world, and the test was approved for use by the National Research Coordinators (NRCs) of the participating countries. Although every effort was made in TIMSS to ensure the widest possible subject matter coverage, no test can measure all that is taught or learned in every participating country.

Given that no test can cover the curriculum in every country completely, the question arises as to how well the items on the tests match the curricula of each of the participating countries. To address this issue, TIMSS asked each country to indicate which items on the tests, if any, were inappropriate to its curriculum. For each country in turn, TIMSS took the list of remaining items, and computed the average percentage correct on these items for that country and all other countries. This allowed each country to select only those items on the tests that they would like included, and to compare the performance of their students on those items with the performance of the students in each of the other participating countries on that set of items. However, in addition to comparing the performance of all countries on the set of items chosen by each country, the Test-Curriculum Matching Analysis (TCMA) also shows each country's performance on the items chosen by each of the other countries. In these analyses, each country was able to see the performance of all countries on the items appropriate for its curriculum, but to see also the performance of its students on items judged appropriate for the curriculum in other countries.

Each NRC was given a questionnaire with all the items included in the TIMSS tests and was asked to indicate, for each item, whether it was considered an appropriate item for their curriculum. The questionnaire sought the information separately for each item at each grade level at which the items were administered. The results from these questionnaires were then used to assess the curricular coverage of the items in the tests, and the effect on the test results of all countries of omitting those items identified by each NRC or their representative. It must be stressed that this analysis was not intended to replace the carefully constructed and agreed-upon tests that TIMSS used for its international comparisons and research analyses. The IRT scaling and research analyses used all items that were included in the tests and that met psychometric standards. In the TCMA analysis, items identified by NRCs were omitted from test results only in the series of analyses designed to illuminate and explain the international comparisons based on the entire test.

10.2 THE ANALYTICAL METHOD OF THE TCMA

The TCMA makes use of the average proportion-correct technology described in Chapter 9. The basic item-level data for a participating country at a grade level were represented by the matrix D_{ikj}. This matrix contains elements d_{ikj} , which represent the scored response of student i in country k to item j. The possible values for item j are 0 or 1 for multiple-choice items, and between 0 and 3 for multiple-score items. Most of the elements of D are missing since each student took only one of eight possible booklets administered at a grade level. Depending on the booklet, each student took between one-fifth and two-fifths of the total item pool (Adams and Gonzalez, 1996).

The information provided by the NRC as to whether or not an item should be omitted from these analyses for the particular grade were summarized in a matrix T_{kj} where the elements t_{kj} represent the information that the NRC in country k submitted about item j (for a particular grade). The actual responses of the NRCs for an item were 0 (meaning omit this item for my country) or 1 (meaning include it). Given that multiple-score items were included in the TIMSS tests, both matrices D_{ikj} and T_{kj} were then converted to $D_{ikj'}$ and $T_{kj'}$ matrices as described in the previous chapter. In that conversion, the score points on each item in the matrix $D_{ikj'}$ were transformed into their binary representation, and the item selection by the NRC, contained in the matrix $T_{kj'}$, was transformed into a matrix that matched the $D_{ikj'}$.

Although the procedure described here will work generally for any item selection proportion from 0 to 1, the TCMA analysis in TIMSS was limited to a binary choice of either including or excluding the item at the specific grade level. The computational procedure used for the TCMA analysis was as follows. First form the $P'_{kj'}$ matrix. The elements in matrix $P'_{kj'}$ are computed from the D_{ikj} matrix after the transformations and estimation outlined in the Chapter 9 are applied to the data. The elements of $P'_{kj'}$ are the weighted averages of the student responses in country k to item j', that is, the average of the student responses $d_{ikj'}$, estimated for some elements. Under the TIMSS design, students not administered particular items may be considered missing at random and treated as not having taken the item. Item responses coded as not reached or omitted are treated as incorrect responses.

The next step is to compute a Test Coverage Index. A reasonable index is the percentage of the total possible test points that were deemed appropriate by each country. The total possible test points in a TIMSS test are equal to C_t, and the total possible score on the items deemed appropriate in country k is computed as

$$C_k = \sum_{j'} t_{kj'} \quad .$$

The Test Coverage Index can then be computed as the ratio of the total possible score on the items deemed appropriate in country *k* to the total possible test points in the TIMSS test:

$$\text{Test Coverage Index} = \frac{C_k}{C_t} \qquad .$$

The Test Coverage Index indicates the proportion of score points of the test that was considered appropriate to the curriculum in the country. The TCI for each country is presented in Tables 10.1 and 10.2.

Table 10.1 Test Coverage Index for the TIMSS Mathematics Tests

Country	3rd grade	4th Grade	7th Grade	8th Grade
Australia	0.56	0.98	0.70	0.95
Austria	-	-	0.81	0.91
Belgium (Fl)	-	-	0.64	0.86
Belgium (Fr)	-	-	0.64	0.85
Bulgaria	-	-	0.73	0.73
Canada	0.56	0.88	0.54	0.91
Colombia	-	-	0.55	0.82
Cyprus	0.68	0.88	0.62	0.77
Czech Republic	0.54	0.74	0.90	0.93
Denmark	-	-	0.36	0.83
England	0.48	0.76	0.57	0.80
France	-	-	0.79	0.86
Germany	-	-	0.81	0.96
Greece	0.45	0.81	0.67	0.47
Hong Kong	0.40	0.81	0.86	0.93
Hungary	0.65	0.85	0.98	1.00
Iceland	0.30	0.65	0.63	0.82
Iran, Islamic Rep.	0.58	0.86	0.79	0.91
Ireland	0.36	0.76	0.70	0.90
Israel	-	0.74	0.00	0.98
Japan	0.74	0.89	0.90	0.94
Korea	0.61	0.43	0.89	0.91
Kuwait	-	0.58	0.00	0.86
Latvia	0.45	0.93	0.93	0.99
Lithuania	-	-	0.90	0.96
Netherlands	0.23	0.52	0.50	0.72
New Zealand	0.63	0.87	0.71	0.90
Norway	0.72	0.88	0.73	0.93
Portugal	0.90	0.90	0.91	0.94
Romania	-	-	0.54	0.88
Russian Federation	-	-	0.75	0.78
Scotland	0.41	0.81	0.47	0.77
Singapore	0.51	0.74	0.78	0.89
Slovak Republic	-	-	0.94	0.94
Slovenia	0.71	0.79	0.89	0.93
South Africa	-	-	0.50	0.80
Spain	-	-	0.93	0.98
Sweden	-	-	0.62	0.78
Switzerland	-	-	0.56	0.82
Thailand	-	-	-	-
United States	1.00	1.00	1.00	1.00

Table 10.2 Test Coverage Index for the TIMSS Science Tests

Country	3rd Grade	4th Grade	7th Grade	8th Grade
Australia	0.47	0.76	0.64	0.91
Austria	-	-	0.36	0.90
Belgium (Fl)	-	-	0.32	0.67
Belgium (Fr)	-	-	0.16	0.40
Bulgaria	-	-	0.72	0.77
Canada	0.61	0.89	0.53	0.83
Colombia	-	-	0.74	0.77
Cyprus	0.42	0.58	0.20	0.53
Czech Republic	0.38	0.90	0.74	0.93
Denmark	-	-	0.14	0.48
England	0.30	0.50	0.71	0.85
France	-	-	0.18	0.50
Germany	-	-	0.60	0.88
Greece	0.29	0.68	0.49	0.76
Hong Kong	0.32	0.40	0.22	0.47
Hungary	0.39	0.47	0.67	0.88
Iceland	0.89	0.90	1.00	1.00
Iran, Islamic Rep.	0.37	0.81	0.33	0.60
Ireland	0.13	0.26	0.41	0.62
Israel	0.23	0.32	-	0.70
Italy	0.34	0.93	0.64	0.88
Japan	0.16	0.28	0.31	0.59
Korea	0.10	0.24	0.29	0.40
Kuwait	-	0.81	-	0.90
Latvia (LSS)	0.70	0.99	0.32	0.77
Lithuania	-	-	0.54	0.82
Mexico	-	-	0.36	0.39
Netherlands	0.32	0.65	0.23	0.70
New Zealand	0.65	0.86	0.75	0.86
Norway	0.47	0.59	0.63	0.76
Portugal	0.80	0.80	0.55	0.91
Romania	-	-	0.62	0.68
Russian Federation	-	-	0.34	0.66
Scotland	0.31	0.43	0.34	0.66
Singapore	0.30	0.50	0.57	0.75
Slovak Republic	-	-	0.76	0.88
Slovenia	0.88	0.93	0.90	0.96
South Africa	-	-	0.18	0.51
Spain	-	-	0.90	1.00
Sweden	-	-	0.59	0.86
Switzerland	-	-	0.25	0.72
Thailand	-	-	-	-
United States	1.00	1.00	1.00	1.00

After computing the TCI, the next step was to compute the normalized weight matrix. To facilitate cross-national comparisons, it is useful to anchor the various national proficiency estimates in a common manner. The national proficiency estimates described in the next section have the property that, if the students in a country correctly answer all of the items deemed appropriate for that country, then the country will receive a value of 100; if the students answer all of those items incorrectly, then the country will receive a value of zero. Items not deemed appropriate to the curriculum of a country are not used in computing these values. In situations where the information in T is either 1 (include) or 0 (omit), the country values may be considered percentages of possible points attained on included items. If T contains proportions other than 0 and 1, then the country values may be greater than 100, in which case the students answered more items correctly than was expected from the values in T.

To compute such country estimates, it is necessary to compute the matrix $W_{kj'}$, with the elements $w_{kj'}$, where the matrix elements are computed as follows:

$$w_{kj} = \frac{t_{kj'}}{\sum_{j'} t_{kj'}^2}$$

where the denominator of this equation is the sum of the squares of the NRCs' judgments to the items.

The Country Comparison Matrix can be computed from $P_{kj'}$ and $W_{kj'}$ by the matrix multiplication

$$C_{kk'} = 100 * (W_{kj'} * P'_{kj'})$$

where the elements of $C_{kk'}$ indicate how the students in country k' scored on the items deemed appropriate in country k.

Another way to directly estimate the $C_{kk'}$ matrix without going through the intermediate step of computing the w_{kj} matrix is as follows:

$$C_{kk'} = \frac{\sum_{j'} t_{kj'} * p_{kj'}}{\sum_{j'} t_{kj'}^2} * 100$$

The estimates in the resulting Country Comparison Matrix are unbiased estimators of average student performance based on the items selected by each country for inclusion in the TCMA. The precision of estimates varies as a result of the test booklet rotation as well as the different school and student sampling plans.

10.3 COMPUTING STANDARD ERRORS

The computation of the standard error for the TCMA is a continuation of the procedure described for computing the standard error for the average percent correct. Once the $P_{kj}'^{h'}$ matrices are obtained, we then continue to compute each of the $C_{kk'}^{h'}$ matrices, which can be computed with each of the different $P_{kj}'^{h'}$ replicate matrices. This is accomplished in a straightforward manner by use of the following multiplication:

$$C_{kk'}^{h'} = \frac{\sum_{j'} t_{kj'} * p_{kj'}^{h'}}{\sum_{j'} t_{kj'}^{2}} * 100$$

The jackknifed standard errors for each of the elements in the $C_{kk'}$ matrix are then computed by applying the following formula

$$jse_{C_{kk'}} = \sqrt{\sum_{h'} (c_{kk'} - c'^{h'}_{kk'})^{2}}$$

REFERENCES

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

11 Reporting Student and Teacher Questionnaire Data

Dana L. Kelly
Ina V.S. Mullis
Teresa A. Smith
Boston College

This chapter documents the development of the TIMSS international reports for the primary and middle school years (third, fourth, seventh, and eighth grades in most countries) and analysis and reporting of the background questionnaire data.[1] In particular, it provides an overview of the consensus process used to develop the report outlines and table prototypes; describes special considerations in reporting the student and teacher background data; and explains how TIMSS handled issues of non-response in reporting these data.

11.1 CONTEXT QUESTIONNAIRES

TIMSS administered questionnaires to students, their mathematics and science teachers, and the principals of their schools to gather contextual information related to the teaching and learning of mathematics and science. Table 11.1 lists the background questionnaires administered at each population.

Table 11.1 TIMSS Background Questionaires

Population 1	Population 2
Student Questionnaire	Student Questionnaire (nonspecialized)
School Questionnaire	Student Questionnaire (specialized)
Teacher Questionnaire	School Questionnaire
	Teacher Questionnaire (Mathematics)
	Teacher Questionnaire (Science)

Students in Populations 1 and 2 completed questions about their attitudes towards mathematics and science, home background, out-of-school activities, and classroom activities and experiences. At Population 2 there were two versions of the student questionnaires; one version was intended for systems where science is taught as an integrated subject and the other for systems where science is taught as separate subjects (biology, chemistry, earth science, and physics). These are referred to as the nonspecialized and specialized versions, respectively. Although these two versions of the questionnaire differed with respect to the science questions, the general background and mathematics-related questions were identical across the two forms. In the nonspecialized version, science-related questions pertaining to students' attitudes and classroom activities are based on single questions asking about "general or integrated

[1] Reporting of background questionnaire data for the assessment of students in their final year of secondary school will be described in the forthcoming *TIMSS Technical Report, Volume III.*

science," while in the specialized version a series of questions is asked about each of the separate science subject areas. This structure accommodated the diverse systems that participated in TIMSS but did pose challenges in reporting the data, as is further described later in this chapter.

Teachers of students in the lower and upper grades of Populations 1 and 2 answered questions about their education, instructional practices, classroom organization, and views on mathematics and science education. At Population 2, there were two versions of the teacher questionnaire, one for mathematics teachers and one for science teachers. Although the general background questions were the same for the two versions, questions pertaining to instructional practices, content coverage, classroom organization, and views of subject matter were geared towards mathematics or science. At Population 1, there was only one version of the questionnaire. It included general background questions and questions related to mathematics and science instruction. Section 11.5.1 of this chapter discusses the complications that arose from having one teacher questionnaire for Population 1 and how those complications were handled in the analysis and reporting.

The school questionnaire included questions regarding school characteristics and policies, resources, and course offerings.

The development of these questionnaires and the variables included in each instrument are described in Schmidt and Cogan (1996).

11.2 TIMSS REPORTING APPROACH

The TIMSS results were reported separately by grade. Because every country participated in Population 2, the core of TIMSS, the International Study Center published the results for the lower and upper grades of Population 2 (seventh and eighth grades) first, followed by the results for the lower and upper grades of Population 1 (third and fourth grades) and Population 3 (final year of secondary school). The mathematics results and science results were published in separate volumes.

Background data were reported for the students in the upper grade of the target populations only (fourth and eighth grades in most countries), but not for those in the lower grade for several reasons. First, reporting data for both grades in a population would have doubled the size of the report or limited the number of variables that could be reported. It was therefore decided that in order to present as wide a range of information as possible, data would be reported for only one grade of the target population, but would address as many issues as possible. In addition, more questions in the context questionnaires were geared towards the upper-grade students, particularly in the teacher questionnaire. Data for the lower grade of the target populations are available in the international database.

11.3 DEVELOPMENT OF THE INTERNATIONAL REPORTS

The International Study Center's initial plans for reporting the background data were based on the TIMSS conceptual model, and on research questions developed early in the study and used as the basis for instrument development. The documentation on the TIMSS conceptual model developed by the Survey of Mathematics and Science Opportunity (SMSO) project at Michigan State University, and the various documents presenting alternative reporting and analysis plans that had been written during the years of the study, were reviewed and used as the basis for the initial round of outlines for the international reports. These documents included:

- *TIMSS: Concepts, Measurements, and Analyses, Abbreviated Version* (Schmidt, 1993)
- *TIMSS Educational Opportunity Model: Detailed Instrumentation and Indices Development* (Schmidt, 1994)
- *TIMSS Monograph No. 1: Curriculum Frameworks for Mathematics and Science* (Robitaille et al., 1993)
- *TIMSS Monograph No. 2: TIMSS Research Design* (Robitaille and Garden, 1996)
- *TIMSS Analysis Plan IV: The First U.S. TIMSS Reports* (Williams, 1995).
- *Research Questions for TIMSS – Draft (*Robitaille and Nicol, 1993)
- *TIMSS ICC Publications Plan – Draft (*Robitaille, 1993)

In addition, reports of previous IEA studies and the research literature were used as a basis for the initial outlines.

As described in Schmidt and Cogan (1996), TIMSS was designed to investigate student learning of mathematics and science and the way in which education systems, schools, teachers, and the students themselves all influence the learning opportunities and experiences of individual students. This explanatory framework offers four major research questions used to undergird the development of the data collection instruments: What are students expected to learn? Who delivers instruction? How is instruction organized? What have students learned?

In attempting to address the influences on student learning put forth by the model as key determinants of achievement – the system, schools, teachers, and students – the TIMSS International Study Center included in the initial report outlines as much information as possible about these aspects of the education system. In particular, the major areas included were the following:

- The curricular context of students' learning
- Students' characteristics and attitudes towards mathematics and science

- System-level characteristics
- School characteristics
- Teacher qualifications and characteristics
- Instructional organization and activities.

Within each of these categories those aspects described in the model as key features of the educational process were included in the outlines as proposed subsections.

The goal of the international reports was to present as much descriptive data related to the TIMSS model as possible, without overburdening the reader, and taking into consideration the time and resources available to produce the reports. The intention was that these initial descriptive reports would provide the basis for more complex secondary analyses to be undertaken at a later date.

Towards this end, tables presenting descriptive data related to each feature (e.g., parents' education, instruction time) were planned and table prototypes prepared. This required a careful review of the questionnaires and detailed documentation of the variables and categories, recodes, and analyses to be undertaken. These plans were documented in analysis notes for each proposed table.

Drafts of the analysis plans, report outlines, and table prototypes reporting results for the upper and lower grades of Population 2 were developed by the International Study Center and underwent a lengthy review process involving the TIMSS Technical Advisory Committee, Subject Matter Advisory Committee, the International Steering Committee, and the NRCs. Through this review process, consensus was built among the constituents as to the reporting priorities for the first international reports, including which variables should be reported and how much information to include. The list of meetings during which the analysis plans, outlines, and tables prototypes were reviewed follows.

June 1995, Ottawa	Technical Advisory Committee
July 1995, Boston	Subject Matter Advisory Committee
August 1995, Vancouver	International Steering Committee
August 1995, Vancouver	National Research Coordinators
January 1996, Cyprus	National Research Coordinators

Following each review meeting, the report outlines and table prototypes were modified to reflect the perspectives of the various committee members and NRCs.

After the data became available for analysis in the spring of 1996, the International Study Center conducted the analyses documented in the analysis notes. The tables with results and accompanying text underwent a review process similar to that conducted for the outlines and table prototypes, and as a result, some tables and figures were modified and some were deleted from the report. For example, for some categorical variables, categories were modified to reflect the distribution of student responses. Also, it was not possible to report the data collected via the school questionnaire in the first international reports, mainly because many of the questions were asked in open-ended format and would have required more time to clean and prepare for analysis than was available. The school data are available in the TIMSS international database. NRCs had several opportunities to review the draft tables in the light of their national data and to provide feedback on the quality and consistency of the background data.

The draft reports (text and tables) were reviewed by the International Steering Committee and the NRCs at a meeting in Prague in August 1996. Further refinements were made to the tables following that meeting and final drafts were sent out for review in September 1996. This review resulted in several additional modifications to the interpretations and presentation of the data. The reports were published in November 1996 as *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study* (Beaton et al., 1996a) and *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study* (Beaton et al., 1996b).

The reports presenting results for the upper and lower grades of Population 1 were modeled for the most part on the Population 2 reports. Some modifications were made to reflect the issues relevant to the primary school years, and some tables that appeared in the middle school reports were not available for the primary school report because certain questions were not asked of the younger students or their teachers. As with the middle school reports, a series of meetings was held during which NRCs and TIMSS committee members had the opportunity to review the plans for the primary school reports. These reports were published in June 1997 as *Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study* (Mullis et al., 1997) and *Science Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study* (Martin et al., 1997).

11.4 REPORTING STUDENT BACKGROUND DATA

Reporting the data that students provided through the student questionnaire was fairly straightforward. Most of the tables in the international reports present percentages of students in each country responding to each category of each variable, together with the mean achievement (mathematics or science) of those students. Some tables present percentages or averages based on derived variables. *The User Guide for the TIMSS International Database, Supplement 4* (Gonzalez and Smith, 1997) documents all derived variables that were published in the TIMSS international reports and included in the database. In general, jackknife standard errors accompany the statistics reported. (See Chapter 5 of this volume for a description of the methodology and additional references.)

While reporting of the general background and mathematics-related variables was also straightforward, reporting of the student responses to questions about their attitudes and self-perceptions related to science was more complicated. As described earlier in this chapter, for the two grades at Population 2 countries could administer a student questionnaire that accommodated the manner in which science instruction is organized. One version of the questionnaire asked questions about science as an integrated subject (nonspecialized version); the other version asked questions about science taught as separate subject areas (specialized version). That countries administered different questionnaires posed a challenge for the international data processing and for the analysis. Moreover, the tables reporting those variables for which countries administered different versions had to present both types of data. As a result, those tables have a column where data are reported for the countries that administered the nonspecialized student questionnaire and a section where data are reported for the countries that administered the specialized student questionnaire.

In the tables and figures in the international report, countries that administered the nonspecialized version are included in the column reporting students' responses based on integrated science, while countries that administered the specialized version are included in the columns displaying students' responses based on separate science subject areas. Based on the form of the majority of science-related questions, 18 countries administered the specialized version and 22 countries the nonspecialized version of the student questionnaire (see Table 11.2). The classification of countries in Table 11.2 is based on whether the questions related to activities in science classes are based on integrated science classes or separate science subject areas.

Table 11.2 Countries Administering the Specialized and Nonspecialized Student Questionnaires - Population 2

Nonspecialized Version (Science as an Integrated Subject)	Specialized Version (Science as Separate Subjects)
Australia	Belgium (Flemish)
Austria	Belgium (French)
Canada	Czech Republic
Colombia	Denmark
Cyprus	France
England	Germany
Hong Kong	Greece
Iran	Hungary
Ireland	Iceland
Israel	Latvia
Japan	Lithuania
Korea	Netherlands
Kuwait	Portugal
New Zealand	Romania
Norway	Russian Federation
Scotland	Slovak Republic
Singapore	Slovenia
Spain	Sweden
Switzerland	
Thailand	
United States	

11.5 REPORTING TEACHER BACKGROUND DATA

Because the sampling for the teacher questionnaires was based on participating students, the responses to the teacher questionnaire do not necessarily represent all of the fourth- and eighth-grade teachers in each of the TIMSS countries. Rather, they represent teachers of the representative samples of students assessed. It is important to note that in the international reports, the student is always the unit of analysis, even when information from the teachers' questionnaires is being reported. Using the student as the unit of analysis makes it possible to describe the instruction received by representative samples of students. Although this approach may provide a different perspective from that obtained by simply collecting information from teachers, it is consistent with the TIMSS goals of providing information about the educational contexts and performance of students.

Another consequence of the TIMSS design, particularly at Population 2, was that since students were often taught mathematics or science by different teachers, and sometimes by more than one teacher (e.g., students were taking two or more mathematics classes or two or more science classes), they frequently needed to be linked to more than one teacher. When a student was taught one or the other subject by more than one teacher, the student's sampling weight was distributed among the teachers that taught the student. In this way, the student's contribution to student population estimates remained constant regardless of the number of teachers he or she had. This was consistent with the policy of reporting attributes of teachers and their classrooms in terms of the percentages of students taught by teachers possessing these attributes.

11.5.1 Population 1 Teacher Data

In the two grades tested for Population 1 (third and fourth grades in most countries), students generally are taught mathematics and science by a single classroom teacher who provides instruction in all subjects. Accordingly, the international version of the teacher questionnaire for the primary grades was prepared as a single document asking about demographic information and instruction in both mathematics and science. Reporting data for these situations was straightforward in the sense that for one teacher the variables pertaining to mathematics instruction were included in the international mathematics report and the variables pertaining to science instruction were included in the science report. General background data for that teacher were included in both reports.

In some countries, however, a portion or even all of third- and fouth-grade students are taught mathematics and science by different teachers, and it was difficult to make provision for both teachers to complete the questionnaire. In these cases, one of the teachers was usually given the questionnaire and completed it as fully as possible, in most cases omitting those questions pertaining to the subject not taught to the class (i.e., if the teacher was a mathematics teacher he or she would omit most questions pertaining to science instruction and vice versa). Although an examination of which questions a teacher completed could have indicated which subject the teacher taught to the target class, TIMSS instead used data provided by the schools to determine whether a teacher taught mathematics, science, or both to the target class. Accordingly, all tables in the Population 1 international mathematics report (Mullis et al., 1997) that contain teacher data are based only on those teachers identified by schools as either mathematics teachers or mathematics and science teachers. Likewise, tables in the Population 1 international science report (Martin et al., 1997) that contain teacher data are based only on those teachers identified by schools as either science teachers or mathematics and science teachers. By identifying teachers as teaching the sampled students in mathematics, science, or both, TIMSS was able to report teacher background, instructional, and classroom variables and, where relevant, the relationship with achievement in mathematics or science.

Because countries were required to sample two classes (from adjacent grades) in each school, it was possible for an individual to be the mathematics and/or science teacher of both the upper- and lower-grade classes. In order to keep the response burden for teachers to a minimum, no teacher was asked to respond to more than one questionnaire, even where that teacher taught mathematics and/or science to more than one of the sampled classes. This had implications for response rates, as described in section 11.6.

11.5.2 Population 2 Teacher Data

In the two grades tested for Population 2 (seventh and eighth grades in most countries), students are generally taught mathematics and science by different teachers. Accordingly, there was a questionnaire for mathematics teachers and another for science teachers, each with the same general questions but with different subject-matter-related questions. Data collected from mathematics teachers were presented in the international mathematics report and those collected from science teachers in the corresponding science report. Where possible and relevant, the mean achievement of students was reported for each category in a table to show the relationship with achievement.

For each sampled student, his or her mathematics and science teachers were assigned a questionnaire. However, if a teacher taught sampled classes in both mathematics and science, then that teacher was randomly assigned either a mathematics or a science questionnaire. If a teacher taught either mathematics or science at both the lower and upper grade then that teacher was assigned a questionnaire for the upper-grade target class. The assignment of questionnaires to teachers of sampled students had implications for response rates; this is further explained in section 11.6.

As explained earlier, for students with more than one mathematics or science teacher the student weight was distributed among the teachers that taught the student (in that subject) so that the student's contribution to the population estimates remained constant regardless of the number of teachers.

11.6 REPORTING RESPONSE RATES FOR BACKGROUND QUESTIONNAIRE DATA

While it is desirable that all questions included in a data collection instrument be answered by all intended respondents, a certain percentage of nonresponse is inevitable. In addition to the problem of unanswered questions, sometimes entire questionnaires are not completed or not returned. In TIMSS, the teachers, students, or principals sometimes did not complete the questionnaire assigned to them or some questions within it, resulting in certain variables having less than a 100% response rate. The tables in the TIMSS international reports contain special notation regarding response rates for the background variables. The following section describes the types of nonresponse and how the variables with varying response rates are labeled in the TIMSS reports.

11.6.1 Teacher Data

Because teachers were asked to complete no more than one questionnaire even if they taught mathematics or science to more than one sampled class, and because teachers sometimes did not complete the questionnaire assigned to them, each country had some percentage of students for whom no teacher questionnaire information was available. The following special notation was used to convey information about response rates in tables in the international reports.

- For a country where teacher responses were available for 70% to 84% of the students, an "r" appears next to the data for that country.
- When teacher responses were available for 50% to 69% of the students, an "s" appears next to the data for that country.
- When teacher responses were available for fewer than 50% of the students, an "x" replaces the data.
- When the percentages of students in a particular category fell below 2%, achievement data were not reported in that category. The data were replaced by a tilde (~).

11.6.2 Student Data

Although in general there were high response rates for the student background variables, some variables and some countries exhibited less than acceptable response rates. The notation in the tables of the reports is similar to that for the teacher data.

- For a country where responses were available for 70% to 84% of the students, an "r" appears next to the data for that country.
- When responses were available for 50% to 69% of the students, an "s" appears next to the data for that country.
- When responses were available for fewer than 50% of the students, an "x" replaces the data.
- When the percentages of students in a particular category fell below 2%, achievement data were not reported in that category. The data were replaced by a tilde (~).

REFERENCES

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Smith, T.A., and Kelly, D.L. (1996a). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College.

Beaton, A.E., Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Kelly, D.L., and Smith, T.A. (1996b). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College.

Gonzalez, E.J. and Smith, T.A., Eds. (1997). *User Guide for the TIMSS international database: Primary and middle school years – 1995 assessment.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Beaton, A.E., Gonzalez, E.J., Smith, T.A., and Kelly, D.L. (1997). *Science achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L., and Smith, T.A. (1997). *Mathematics achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS). Chestnut Hill, MA: Boston College.*

Robitaille, D.F. (1993). *TIMSS ICC publications plan – draft.* Unpublished document.

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 Nicol, C. (1993). *Research questions for TIMSS – Draft* (Doc. Ref.: ICC502/NRC161). Unpublished document.

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. (1994, June). *TIMSS educational opportunity model: Detailed instrumentation and indices development* (SMSO Research Report Series No. 58). East Lansing, MI: Michigan State University.

Schmidt, W.H. (1993, April). *TIMSS: Concepts, measurements, and analyses, abbreviated version* (SMSO Research Report Series No. 56). Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Schmidt, W.H. and Cogan, L.S. (1996). Development of the TIMSS context questionnaires. In M.O. Martin and D.L. Kelly (Eds.), *TIMSS technical report, volume I: Design and development.* Chestnut Hill, MA: Boston College.

Williams, T. (1995). TIMSS analysis plan IV: The first U.S. TIMSS reports. Unpublished document.

Appendix A: Table of Contents for Volume I of the Technical Report

9. FIELD OPERATIONS
Andreas Schleicher and Maria Teresa Siniscalco

9.1 OVERVIEW
9.2 DOCUMENTATION
9.3 SELECTING THE SCHOOL SAMPLE
9.4 IMPLICATIONS OF THE TIMSS DESIGN FOR WITHIN-SCHOOL FIELD OPERATIONS
9.5 WITHIN-SCHOOL SAMPLING PROCEDURES FOR POPULATIONS 1 AND 2
9.6 THE GENERAL PROCEDURE FOR WITHIN-SCHOOL SAMPLING
9.7 PROCEDURE A FOR WITHIN-SCHOOL SAMPLING
9.8 PROCEDURE B FOR WITHIN-SCHOOL SAMPLING
9.9 EXCLUDING STUDENTS FROM TESTING
9.10 CLASS, STUDENT, AND TEACHER ID AND TEACHER LINK NUMBER
9.11 WITHIN-SCHOOL SAMPLING PROCEDURES FOR POPULATION 3
9.12 RESPONSIBILITIES OF SCHOOL COORDINATORS AND TEST ADMINISTRATORS
9.13 PACKAGING AND SENDING MATERIALS
9.14 CODING, DATA ENTRY, DATA VERIFICATION, AND SUBMISSION OF DATA FILES AND MATERIALS
9.15 CODING THE FREE-RESPONSE ITEMS
9.16 DATA ENTRY
9.17 CONCLUSION

10. TRAINING SESSIONS FOR FREE-RESPONSE SCORING AND ADMINISTRATION OF PERFORMANCE ASSESSMENT
Ina V.S. Mullis, Chancey Jones, and Robert A. Garden

10.1 OVERVIEW
10.2 THE TIMSS FREE-RESPONSE CODING TRAINING TEAM
10.3 THE SCHEDULE OF THE REGIONAL TRAINING SESSIONS
10.4 DESCRIPTION OF EACH TRAINING SESSION
10.5 THE TRAINING MATERIALS
10.6 CONCLUDING REMARKS

11. QUALITY ASSURANCE PROCEDURES
Michael O. Martin, Ina V.S. Mullis, and Dana L. Kelly

11.1 OVERVIEW
11.2 STANDARDIZATION OF THE TIMSS PROCEDURES
11.3 PROCEDURES FOR TRANSLATION AND ASSEMBLY OF THE ASSESSMENT INSTRUMENTS
11.4 SCORING THE OPEN-ENDED RESPONSES
11.5 NATIONAL QUALITY CONTROL PROGRAM
11.6 TIMSS QUALITY CONTROL MONITORS
11.7 THE QUALITY CONTROL MONITOR’S VISIT TO THE SCHOOLS

APPENDIX A: **ACKNOWLEDGMENTS**

APPENDIX B: **TIMSS TEST BLUEPRINTS**

APPENDIX C: **TIMSS SURVEY OPERATIONS FORMS**

Appendix B: Characteristics of the National Samples

In Chapter 2, the TIMSS target populations were described and the participation rates and sample sizes were documented for Populations 1 and 2. This appendix describes, for each country and each population in which it participated, the target population definitions, coverage and exclusions, use of stratification variables, and any deviations from the general TIMSS design.

AUSTRALIA

Target Population

Table B.1 identifies the defined target grades by state for Population 1 and Population 2 in Australia. The target grades in the two populations varied by state. This variation is due to different age entrance rules applied in the Australian States and Territories. Allowing these state variations maximized coverage of the age-13 cohort.

Table B.1 Target Grades in Australia

State or Territory	Population 1	Population 2
New South Wales	3 and 4	7 and 8
Victoria	3 and 4	7 and 8
Queensland	4 and 5	8 and 9
South Australia	4 and 5	8 and 9
Western Australia	4 and 5	8 and 9
Tasmania	3 and 4	7 and 8
Northern Territory	4 and 5	8 and 9
Australian Capital Territory	3 and 4	7 and 8

Coverage and Exclusions

School-level exclusions in Population 1 consisted of extremely small schools, distance-education schools, and Victorian schools involved in another study. School-level exclusions in Population 2 consisted of extremely small schools and distance-education schools.

Sample Design - Population 1

- Explicit stratification by eight states and territories and three types of school (government, Catholic, and independent), for a total of 24 strata
- No implicit stratification

- Schools sorted on the sampling frame by geography
- Sample allocation of schools as presented in Table B.2
- Additional schools sampled after a first selection (these schools were included in the TIMSS sample for Population 1)
- School participation adjustments for weighting computed only at the state and territory level because the type-of-school level of stratification became too fine
- Sampled two upper-grade classrooms per school
- Sampled one lower-grade classroom per school except in Queensland, South Australia, Western Australia, and the Northern Territory, where two classrooms per school were sampled

Table B.2 Allocation of School Sample in Australia

State or Territory	Population 1 Schools	Population 2 Schools
New South Wales	40	40
Victoria	40	40
Queensland	40	40
Western Australia	40	35
South Australia	40	35
Tasmania	30	12
Northern Territory	20	8
Australian Capital Territory	18	4
All Australia	268	214

Sample Design - Population 2

- Explicit stratification by eight states and territories and three types of school (government, Catholic, and independent), for a total of 24 strata
- No implicit stratification
- Schools sorted on the sampling frame by geography
- Sample allocation of schools as presented in Table B.2
- Additional schools sampled after a first selection (these schools could not be included in the TIMSS sample for Population 2 because of time constraints; students from those schools were not assigned any sampling weights)

- School participation adjustments for weighting computed only at the state and territory level because the type-of-school level of stratification became too fine
- Sampled two upper-grade classrooms per school
- Sampled one lower grade classroom per school, except in Queensland, South Australia, Western Australia and the Northern Territory, where two classrooms per school were sampled

AUSTRIA

Coverage and Exclusions

School-level exclusions in both populations consisted of schools labeled "Sonders-chulen."

Sample Design - Population 1

- Explicit stratification by three levels of urbanization (Vienna, urban, and rural)
- Sampled 150 schools, 50 per explicit stratum
- Schools sorted on the sampling frame by geography
- Sampled one classroom per grade per school

Sample Design - Population 2

- Explicit stratification by two school types and three levels of urbanization, for a total of six strata (see Table B.3)
- Sampled 159 schools, based on the allocation presented in Table B.3
- Schools sorted on the sampling frame by geography
- Sampled one classroom per grade per school
- Sampled science classrooms in Population 2, rather than mathematics classrooms as in other countries, because streaming in mathematics class-es would have resulted in the inclusion of an inordinate number of science teachers in the data collection

Table B.3 Allocation of School Sample in Austria - Population 2

Explicit Stratum: School Type	Explicit Stratum: Urbanization (Number of Inhabitants)	Number of Schools
Hauptschulen (HS)	Up to 5,000	33
	From 5,001 to 1,000,000	33
	More than 1,000,000 (Vienna)	33
AHS-Unterstufe (Lower Step)	Up to 5,000	10
	From 5,001 to 1,000,000	25
	More than 1,000,000 (Vienna)	25
All Austria		159

BELGIUM (FLEMISH)

Coverage and Exclusions

School-level exclusions consisted mostly of lower-grade students in a track labeled 1B. These students had encountered failure in primary schooling and had been moved to the secondary system because of age. Since their curriculum was largely a review of primary education, the Flemish part of Belgium chose to exclude them. Small schools and schools with only vocational programs also were excluded.

Sample Design - Population 2

- No explicit stratification
- Implicit stratification by three types of school (state, local board, and Catholic) and two programs (schools with or without the technical program), for a total of six strata
- Sampled 150 schools to contribute a classroom from each grade in the general program
- Subsampled 15 schools among the 79 sampled schools with the technical program, to contribute a classroom from the technical program

BELGIUM (FRENCH)

Coverage and Exclusions

School-level exclusions consisted mostly of lower-grade students in a track labeled 1B. These students had failures in primary schooling and had been moved to the secondary system because of age. Since their curriculum was largely a review of primary education, the French part of Belgium chose to exclude them. Small schools and schools with only vocational programs also were excluded.

Sample Design - Population 2

- No explicit stratification
- Implicit stratification by three types of school (state, local board, and Catholic) and two programs (schools with or without the technical program), for a total of six strata
- Sampled 150 schools to contribute a classroom from each grade in the general program
- Subsampled 35 schools among the 70 sampled schools with the technical program, to contribute a classroom from the technical program

BULGARIA

Coverage and Exclusions

School-level exclusions consisted of schools for the disabled, sport schools, and art schools.

Sample Design - Population 2

- Explicit stratification by two types of schools (schools with both grades and schools with only the upper grade)
- Implicit stratification by three levels of urbanization (national capital, urban, and rural) and three levels of school size (since no valid measure of size was available)
- Sampled 150 schools with both grades and 17 schools with only the upper grade, for a total sample of 167 schools
- Sampled one classroom per grade per school

CANADA

Coverage and Exclusions

School-level exclusions consisted of offshore schools, schools where students are taught in their aboriginal language, very small schools, schools in Prince Edward Island, and French schools in New Brunswick.

Sample Design - Population 1 and Population 2

- Explicit stratification by province or territory, language (in Ontario), and three types of school (Population 1 only, Population 2 only, Population 1 and Population 2), for a total of 39 strata over both populations (see Table B.4)
- Type-of-school stratification allowing maximum overlap of sampled schools between Population 1 and Population 2
- No implicit stratification

- Sample allocation of schools as presented in Table B.4
- A total of 428 schools sampled for Population 1 and 429 sampled for Population 2
- The 40 Population 1 and Population 2 schools sampled in Alberta divided equally between populations since that province wanted to reduce the school participation burden
- The 14 Population 1 and Population 2 schools in British Columbia more finely stratified because of odd combinations of target grades present in those schools
- Sampled one classroom per grade per school
- Sampled two upper-grade classrooms per school in Ontario

Table B.4 Allocation of School Sample in Canada

Province or Territory	Population 1 Only Schools	Populations 1 and 2 Schools	Population 2 Only Schools
Newfoundland	25	15	25
Nova Scotia	3	2	3
New Brunswick	12	10	12
Québec	35	2	40
Ontario (French)	20	75	6
Ontario (English)	40	80	40
Manitoba	2	4	2
Saskatchewan	2	4	2
Alberta	35	40	35
British Columbia	4	10	14
Yukon Territory	2	2	2
Northwest Territories	2	2	2
All Canada	182	246	183

COLOMBIA

Coverage and Exclusions

School-level exclusions consisted of schools located in remote areas.

Sample Design - Population 2

- No explicit stratification
- Implicit stratification by five regions, two types of school (public and private), and four types of schedule (morning, afternoon, evening, and all day), for a total of 48 strata

- The fifth region further stratified by calendar since it is split between a Northern Hemisphere calendar and a Southern Hemisphere calendar (hence, 48 implicit strata)
- Sampled 150 schools
- Sampled one classroom per grade per school
- Subsampled 20 students per sampled classroom; classrooms sampled with PPS

CYPRUS

Coverage and Exclusions

School-level exclusions in Population 1 consisted of single-classroom schools. There were no school-level exclusions in Population 2.

Sample Design - Population 1

- No explicit stratification
- Implicit stratification by four regions and two levels of urbanization (urban and rural), for a total of eight strata
- Sampled 150 schools
- 74 schools were sampled with certainty because of their large size
- Sampled one classroom per grade per school

Sample Design - Population 2

- All 55 Population 2 schools included in TIMSS
- Sampled two classrooms per grade per school

CZECH REPUBLIC

Coverage and Exclusions

School-level exclusions consisted of schools for the disabled.

Sample Design - Population 1

- No explicit stratification
- Implicit stratification by four levels of urbanization and two types of school
- Sampled 150 schools
- Pseudo-schools constructed in Population 1
- Sampled one classroom per grade per school

Sample Design - Population 2

- No explicit stratification
- Implicit stratification by four levels of urbanization, two types of school, and two levels of school stream
- Sampled 150 schools
- Sampled one classroom per grade per school

DENMARK

Coverage and Exclusions

There were no school-level exclusions in Denmark.

Sample Design - Population 2

- Explicit stratification by two geographical levels (Copenhagen and the rest)
- No implicit stratification
- Schools sampled using a stratified simple random sample design
- Sampled 24 schools from Copenhagen and 134 from the rest of the country
- Sampled one classroom per grade per school
- Classrooms sampled by the school headmasters
- Grade 8 classrooms also sampled for national purposes
- A national test booklet added to the booklet rotation; students assigned the TIMSS booklets were considered a random subsample within classrooms

ENGLAND

Coverage and Exclusions

School-level exclusions consisted of special-needs schools, very small schools, and schools that were selected for their national evaluation samples. The last category accounts for the relatively high exclusion rates in both populations.

Sample Design - Population 1

- No explicit stratification
- Implicit stratification by three regions, two types of school, and two levels of urbanization
- Sampled 150 schools

- Sampled one classroom per grade per school
- Two classrooms sampled in single-grade schools

Sample Design - Population 2

- No explicit stratification
- Implicit stratification by three regions, two types of school, and two levels of urbanization
- Sampled 150 schools
- Students sampled across classrooms within grades in sampled schools, resulting in 16 students randomly sampled per grade per school
- 32 students randomly sampled in single-grade schools

FRANCE

Coverage and Exclusions

School-level exclusions consisted of schools in a track labeled CPPN, as well as schools in their offshore territories (*térritoires outre-mer).*

The target grades are *5iéme générale (5g)*, *4iéme générale (4g)*, and *4iéme technologique (4t).* Not all schools offer the 4t program, and this was accounted for in explicit stratification for sampling purposes.

Sample Design - Population 2

- Sampled three independent samples: *collèges*, *collèges* with 4t, *lycées professionnels*
- Overlap in the sampling frames for the first two samples, the second sampling frame being a subset of the first
- Explicit stratification by two levels of urbanization (rural and urban) and two types of school (public and private), for a total of four strata
- No implicit stratification
- Sample allocation of schools as presented in Table B.5
- Schools sampled using a Lahiri method of PPS selection
- All schools in the first sample contributing one 5g classroom; only 136 of them contributing a 4g classroom via a random drop method
- All seven schools in the second sample contributing one 5g classroom and one 4t classroom
- All eight schools in the third sample contributing a single 4t classroom, since these schools do not have the *général* track

- Overlap in the first two sampling frames, causing all *collèges* with 4t classrooms to have two chances of being sampled and contributing a 5g classroom; their school selection probabilities computed accordingly

Table B.5 Allocation of School Sample in France - Population 2

Sampling Frame	Sampled Schools	Sampled Classrooms		
		5g	4g	4t
All collèges	144	144	136	0
Collèges with 4t	7	7	0	7
Lycées Professionnels	8	0	0	8
All France	159	151	136	15

GERMANY

Coverage and Exclusions

One region, Baden-Württemberg, did not participate in TIMSS, thereby reducing national coverage of the target population.

School-level exclusions in Germany consisted of:

- Non-graded private schools
- Special schools for the disabled
- Schools in small strata where no schools were actually sampled
  - Realschulen in Brandenburg
  - Integrierte Gesamtschules and Integrierte Klassen in Hauptund Realschulen in Mecklenburg-Vorpommern and Niedersachsen
  - Integrierte Gesamtschulen in Rheinland-Pfalz and Saarland
- Schools in strata where none of the sampled schools participated
  - Realschulen in Berlin
  - Hauptschulen and Integrierte Gesamtschulen in Schleswig-Holstein

Sample Design - Population 2

- Explicit stratification by 14 regions and 5 types of school, for a total of 45 strata (Table B.6)
- No schools sampled in some of the explicit strata because they were small (see exclusions above)

Table B.6 Allocation of School Sample in Germany - Population 2

Region	Type of School: Hauptschulen	Realschulen	Gymnasien	Integrierte Gesamtschulen	Integrierte Klasse Haupt- und Realschulen	Total
Bayern	11	8	8	1	---	28
Berlin	1	1	2	2	---	6
Brandenburg	---	0	2	4	---	6
Bremen-Hamburg	2	2	1	1	---	6
Hessen	2	3	4	3	---	12
Mecklenburg-Vorpommern	2	4	4	0	0	10
Niedersachsen	5	5	3	0	0	13
Nordrhein-Westfalen	12	7	9	3	---	31
Rheinland-Pfalz	4	2	2	0	---	8
Saarland	1	1	1	0	---	3
Sachsen	---	---	4	---	7	11
Sachsen-Anhalt	---	---	1	---	5	6
Schleswig-Holstein	2	2	2	1	---	7
Thuringen	2	---	2	2	---	6
All Germany	44	35	45	17	12	153

- No implicit stratification
- Sample allocation of schools as presented in Table B.6
- Sampled one classroom per grade per school
- Upper-grade classrooms sampled with PPS and lower grade classrooms sampled with equal probabilities within schools
- Explicit strata considered as implicit in the construction of replicate strata for the jackknife estimation method, since there were an inordinate number of strata

GREECE

Coverage and Exclusions

School-level exclusions in Population 1 and Population 2 consisted of special schools where a different curriculum is used. Evening schools were also excluded in Population 2.

Sample Design - Population 1

- Explicit stratification by 11 regions
- No implicit stratification
- Proportional allocation of 187 schools to the 11 explicit strata

- Sampled one classroom per grade per school
- Computed an overall school participation adjustment for weighting, thereby ignoring the relatively fine explicit stratification

Sample Design - Population 2

- Explicit stratification by 11 regions
- No implicit stratification
- Proportional allocation of 180 schools to the 11 explicit strata
- Sampled one classroom per grade per school
- Always sampled the first classroom listed in the school administrative records from each grade
- Computed an overall school participation adjustment for weighting, thereby ignoring the relatively fine explicit stratification

HONG KONG

Coverage and Exclusions

School-level exclusions consisted of "international" schools that follow overseas curricula.

Sample Design - Population 1

- Explicit stratification by two levels of gender (co-educational and single-sex) and three levels of school administration (aided, government, and private), for a total of five strata (single-sex government schools do not exist)
- No implicit stratification
- A proportional allocation of 156 schools to the five explicit strata
- Eight of the sampled schools no longer in operation
- Sampled one classroom per grade per school
- Computed an overall school participation adjustment for weighting, thereby ignoring the relatively fine explicit stratification

Sample Design - Population 2

- Explicit stratification by two levels of gender (co-educational and single-sex), two levels of language (Chinese and English), and three levels of school administration (aided, government, and private) for a total of 10 strata (single-sex/Chinese/ government and single-sex/Chinese/private schools do not exist)
- No implicit stratification

- A proportional allocation of 105 schools to the 10 explicit strata
- One sampled school no longer in operation
- Sampled one classroom per grade per school
- Computed an overall school participation adjustment for weighting, thereby ignoring the relatively fine explicit stratification

HUNGARY

Coverage and Exclusions

School-level exclusions consisted of very small schools.

Sample Design - Population 1 and Population 2

- No explicit stratification
- Implicit stratification by three levels of urbanization (national capital, urban, and rural)
- Sampled 150 schools, to be used for both populations
- Sampled one classroom per grade per school
- Grade 8 classrooms sampled with PPS, using class size as the measure of size; grades 3, 4, and 7 classrooms sampled using the grade 8 selection probabilities
- Whenever the grade 8 selection probabilities were inappropriate for the other grades, assumed selection with equal probabilities for those grades; this was not a significant issue for grade 7, but did become an issue for grades 3 and 4

ICELAND

Coverage and Exclusions

School-level exclusions consisted of very small schools.

Sample Design - Population 1 and Population 2

- All eligible schools are included in TIMSS
- Sampled one classroom per grade per school

IRAN, ISLAMIC REPUBLIC OF

Coverage and Exclusions

School-level exclusions consisted of schools for the physically and mentally disabled.

Sample Design - Population 1

- Six regions as explicit strata
- Three implicit strata: rural schools, urban girls' schools, and urban boys' schools
- Sampled 180 schools, 30 per region
- Sampled one classroom per grade per school
- Subsampled 20 students per sampled classroom; classrooms sampled with PPS

Sample Design - Population 2

- Six regions as explicit strata
- Four implicit strata: rural girls' schools, rural boys' schools, urban girls' schools, and urban boys' schools
- Sampled 192 schools in Population 2, 32 per region
- Sampled one classroom per grade per school
- Subsampled 20 students per sampled classroom; classrooms were sampled with PPS

IRELAND

Coverage and Exclusions

School-level exclusions in Population 1 consisted of private schools, schools for the physically and mentally disabled, and very small schools. There are no school-level exclusions in Population 2.

Sample Design - Population 1

- Two explicit strata based on school size – small/medium schools and large schools
- Three implicit strata based on gender: boys' schools, girls' schools, and co-educational schools
- Sampled 91 small/medium schools and 59 large schools
- Pseudo-schools constructed
- Sampled one classroom per grade per school

Sample Design - Population 2

- No explicit stratification
- Five implicit strata based on gender and type of school: secondary boys' schools, secondary girls' schools, secondary coeducational schools, vocational schools, and community/comprehensive schools
- Sampled 150 schools
- Sampled one classroom per grade per school

ISRAEL

Coverage and Exclusions

Coverage in Israel is restricted to the Hebrew public education system. This means that the non-Jewish education system and the Jewish Orthodox Independent Education system are not covered. School-level exclusions consisted of special education schools for the physically and mentally disabled. Israel included only the upper grade (eighth grade) in Population 2 and the upper grade (fourth grade) in Population 1.

Sample Design - Population 1

- No explicit stratification
- No implicit stratification
- Sampled 100 schools
- Some sampled schools replacing schools participating in a longitudinal study; these alternate schools are recognized as non-procedural replacement schools
- Sampled one classroom per school
- Alternate classrooms sampled by the local school authorities in 27 of 87 participating schools

Sample Design - Population 2

- No explicit stratification
- Two implicit strata: junior high schools and elementary schools
- Sampled 100 schools
- Sampled one classroom per school
- Alternate classrooms sampled by the local school authorities in 35 of 46 participating schools

JAPAN

Coverage and Exclusions

School-level exclusions consisted of very small schools and schools for the physically and mentally disabled. Private schools also were excluded in Population 1.

Sample Design - Population 1

- Explicit stratification by three school sizes (small, medium, and large) and three levels of urbanization (rural, urban, and large urban), for a total of nine strata
- No implicit stratification
- Schools sampled using a stratified simple random sample design
- Sampled 150 schools
- Sampled one classroom per grade per school

Sample Design - Population 2

- Explicit stratification by three school sizes (small, medium, and large) and three levels of urbanization (rural, urban, and large urban), for a total of nine strata
- No small/large urban schools, but private schools added as a ninth stratum
- No implicit stratification
- Schools sampled using a stratified simple random sample design
- Sampled 158 schools
- Sampled one classroom per grade per school

KOREA

Coverage and Exclusions

School-level exclusions consisted of schools in remote places, islands, and border areas. Additional Population 2 school-level exclusions consisted of evening schools and physical education schools.

Sample Design - Population 1

- No explicit stratification
- Implicit stratification by region and urbanization, for a total of 24 strata
- Sampled 150 schools
- Sampled one classroom per grade per school
- Subsampled 20 students per sampled classroom; classrooms sampled with PPS

Sample Design - Population 2

- No explicit stratification
- Implicit stratification by region, urbanization, and type of school (national and private), for a total of 48 strata
- Sampled 150 schools
- Sampled one classroom per grade per school
- Subsampled 20 students per sampled classroom; classrooms sampled with PPS

KUWAIT

Coverage and Exclusions

There were no exclusions of any kind in Kuwait. Kuwait included only the upper grade (ninth grade) in Population 2 and the upper grade (fifth grade) in Population 1.

Sample Design - Population 1 and Population 2

- All eligible schools included in TIMSS
- Girls' schools and boys' schools
- Sampled one classroom per school
- Classrooms sampled based on the weekly school schedule; i.e., the Monday morning mathematics class was generally sampled

LATVIA

Coverage and Exclusions

Coverage in Latvia was restricted to students whose language of instruction is Latvian. School-level exclusions consisted of schools for the physically and mentally disabled and very small schools.

Sample Design - Population 1 and Population 2

- No explicit stratification
- Implicit stratification by five regions, two levels of urbanization (rural and urban), and three types of school (beginner, basic, and secondary)
- Sampled 150 schools
- Some schools sampled with certainty
- Pseudo-schools constructed
- Sampled one classroom per grade per school

LITHUANIA

Coverage and Exclusions

Coverage in Lithuania was restricted to students whose language of instruction is Lithuanian. School-level exclusions consisted of schools with more than one language of instruction, schools for the physically and mentally disabled, and very small schools.

Sample Design - Population 2

- Explicit stratification by three levels of urbanization (big urban, urban, and rural)
- No implicit stratification
- Proportional allocation of 151 schools to the three explicit strata
- Sampled one classroom per grade per school
- Computed an overall school participation adjustment for weighting

NETHERLANDS

Coverage and Exclusions

School-level exclusions consisted of special education schools for the physically and mentally disabled and very small schools.

Sample Design - Population 1

- No explicit stratification
- Implicit stratification by four levels of denomination, three levels of urbanization, and two levels of socio-economic composition
- Sampled 150 schools
- Pseudo-schools constructed
- Sampled all eligible students in sampled schools
- A national test booklet added to the booklet rotation in the upper grade; students assigned the TIMSS booklets considered a random subsample within classrooms

Sample Design - Population 2

- No explicit stratification
- Implicit stratification by three types of school and two levels of urbanization

- Sampled 150 schools
- Sampled one classroom per grade per school
- A national test booklet added to the booklet rotation in the upper grade; students assigned the TIMSS booklets considered a random subsample within classrooms

NEW ZEALAND

Coverage and Exclusions

School-level exclusions consisted of correspondence schools and very small schools. One geographically remote school was also excluded in Population 1.

Sample Design - Population 1

- No explicit stratification
- Implicit stratification by two levels of community size and three levels of school size
- Sampled 150 schools
- Sampled one classroom per grade per school

Sample Design - Population 2

- Explicit stratification by three types of school (both grades present, only upper grade present, only lower grade present)
- Implicit stratification varying by explicit stratum as described in Table B.7
- The sample allocation of schools as presented in Table B.7
- Sampled one classroom per grade per school

Table B.7 Allocation of School Sample in New Zealand - Population 2

Explicit Stratum	Sampled Schools	Implicit Stratification
Both Grades Present	23	Authority (state & private) Community size (2 levels) School gender (co-ed, boys, girls)
Upper Grade Only	127	–
Lower Grade Only	127	Authority (state & private) Community size (5 levels) School type (full primary & intermediate)

NORWAY

Coverage and Exclusions

School-level exclusions consisted of special schools for the disabled and schools with Sami (Lapp) as the language of instruction. Special schools with an alternative pedagogy were also excluded in Population 1.

Sample Design - Population 1

- Explicit stratification by three school sizes (see Table B.8)
- Implicit stratification by six regions and two levels of urbanization
- Sample allocation of schools as presented in Table B.8
- Sampled one classroom per grade per school

Table B.8 Allocation of School Sample in Norway - Population 1

Explicit Stratum	Sampled Schools
Schools with Small Classrooms	40
Schools with Mid-Sized Classrooms	83
Schools with Large Classrooms	27
All Norway	150

Sample Design - Population 2

- Explicit stratification by five types of school (see Table B.9)
- Implicit stratification by six regions and two levels of urbanization
- Sample allocation of schools as presented in Table B.9
- Sampled one classroom per grade per school

Table B.9 Allocation of School Sample in Norway - Population 2

Explicit Stratum		Sampled Schools
Dual-Grade Schools	Small Classrooms	13
	Large Classrooms	27
Upper-Grade Schools		110
Lower-Grade Schools	Small Classrooms	91
	Large Classrooms	19
All Norway		260

PHILIPPINES

Coverage and Exclusions

Regions 8 and 12 and the Autonomous Region of Muslim Mindanao were removed from their national coverage. School-level exclusions consisted of schools under the responsibility of the Agriculture, Fisheries, and Industrial Arts/Trade ministries. These exclusions affected only the upper grade, which is found in the secondary school system.

Sample Design - Population 2

- Preliminary sampling of 57 school divisions from a frame of 114 school divisions; some school divisions sampled randomly, others based on the advice of the Department of Education, Culture and Sports
- Explicit stratification by school system: elementary schools for the lower grade and secondary schools for the upper grade
- No implicit stratification
- Sampled 200 secondary schools and 200 elementary schools
- Generally, three to five secondary schools sampled per school division
- Elementary schools sampled based on the notion that they are feeder schools for the sampled secondary schools
- Sampled one classroom per grade per school
- Subsampled 32 students per sampled classroom, but classrooms sampled with equal probabilities within schools

Special note: Sampling weights could not be computed for the Philippines. The selection of elementary schools could not be considered random, nor was it possible to derive their selection probabilities.

PORTUGAL

Coverage and Exclusions

School-level exclusions in Population 1 consisted of very small schools. There were no school-level exclusions in Population 2.

Sample Design - Population 1

- Explicit stratification by seven regions
- Implicit stratification by two levels of urbanization (rural and urban) and three levels of socio-economic status
- Sampled 150 schools

- Pseudo-schools constructed
- Sampled one classroom per grade per school

Sample Design - Population 2

- No explicit stratification
- Implicit stratification by five regions, two levels of urbanization (rural and urban), and two levels of type of school (basic and secondary)
- Sampled 150 schools
- Pseudo-schools constructed
- Sampled one classroom per grade per school

ROMANIA

Coverage and Exclusions

School-level exclusions consisted of schools for the disabled, orphanages, schools with only one of the target grades, schools with multigrade classrooms, and very small schools.

Sample Design - Population 2

- No explicit stratification
- No implicit stratification
- Sampled 150 schools
- Pseudo-schools constructed
- Sampled one classroom per grade per school

RUSSIAN FEDERATION

Coverage and Exclusions

School-level exclusions consisted of schools where the language of instruction is other than Russian and schools in regions Nord Osetia and Chechnia.

Sample Design - Population 2

- Preliminary sampling of 40 regions from a frame of 79 regions; ten regions large enough to be sampled with certainty
- No explicit stratification
- Implicit stratification by two levels of urbanization (urban and rural)
- Sampled 175 schools

- Generally, four schools sampled per region; more schools sampled in most certainty regions
- Pseudo-schools constructed
- Sampled one classroom per grade per school

SCOTLAND

Coverage and Exclusions

School-level exclusions consisted of very small schools.

Sample Design - Population 1 and Population 2

- Explicit stratification by two types of school (state and independent)
- No implicit stratification
- Sampled 150 schools
- Pseudo-schools constructed
- Sampled one classroom per grade per school

SINGAPORE

Coverage and Exclusions

There are no school-level exclusions in Population 1. School-level exclusions in Population 2 consisted of newly-opened schools without the upper grade.

Sample Design - Population 1 and Population 2

- All eligible schools included in TIMSS
- Sampled one classroom per grade per school

SLOVAK REPUBLIC

Coverage and Exclusions

School-level exclusions consisted of schools where the language of instruction is other than Slovakian.

Sample Design - Population 2

- No explicit stratification
- Implicit stratification by 4 regions
- Sampled 150 schools
- Sampled one classroom per grade per school

SLOVENIA

Coverage and Exclusions

School-level exclusions consisted of schools for the disabled and schools where the language of instruction is Italian or Hungarian.

Sample Design - Population 1 and Population 2

- No explicit stratification
- Implicit stratification by four levels of urbanization and two types of school (dislocated or not)
- Sampled 150 schools, to be used for both populations
- Sampled one classroom per grade per school

SOUTH AFRICA

Coverage and Exclusions

School-level exclusions consisted of very small schools.

Sample Design - Population 2

- Explicit stratification by school system-elementary schools for the lower grade and secondary schools for the upper grade
- Implicit stratification by nine provinces
- Sampled 150 elementary schools and 150 secondary schools
- Some elementary schools with upper-grade classrooms; some secondary schools with lower-grade classrooms
- Sampled one classroom per grade per school
- Not all absent students recorded in the TIMSS database, so student participation rates are overestimated

SPAIN

Coverage and Exclusions

School-level exclusions consisted of schools where the language of instruction is Euskera, very small schools, and schools in 15 very small explicit strata (see notes below).

Sample Design - Population 2

- Explicit stratification by eight regions, two types of school (public and private), and three levels of school size, for a total of 43 strata
- No schools sampled from 15 of these strata because they were so small (see exclusions above)

- No implicit stratification
- Proportional allocation of 150 schools to the remaining 28 explicit strata
- Pseudo-schools constructed
- Sampled one classroom per grade per school
- Computed an overall school participation adjustment for weighting, thereby ignoring the relatively fine explicit stratification

SWEDEN

Coverage and Exclusions

School-level exclusions consisted of schools for the disabled.

Sample Design - Population 2

- Explicit stratification by school system: elementary schools for the lower grade and secondary schools for the upper grade
- No implicit stratification
- Sampled 160 elementary schools and 120 secondary schools
- Schools sampled using a PPS Lahiri method
- Sampled one classroom per elementary school and two classrooms per secondary school
- Eighth-grade classrooms also sampled for national purposes
- A national test booklet added to the booklet rotation; students assigned the TIMSS booklets considered a random subsample within classrooms

SWITZERLAND

Target Population

The target grades vary in Switzerland. In the German parts, they are 6 and 7. In all other parts of Switzerland, the target grades are 7 and 8.

Coverage and Exclusions

Four cantons – Jura, Waadt, Neuchatel and Freiburg – did not participate, thereby reducing national coverage of the target population. School-level exclusions consisted of schools for the disabled, schools where the language of instruction is not one of the official languages, and very small schools.

Sample Design - Population 2

- Explicit stratification by region, type of school, and track, for a total of 15 strata (see Table B.10)

- No implicit stratification
- Sample allocation of schools as presented in Table B.10
- In each stratum from the canton of Basle, all 16 sampled schools contributing a grade 7 classroom, 8 of them contributing a grade 8 classroom (see note below), and 2 of them contributing a grade 6 classroom
- Additional schools sampled for national purposes; students from such schools were not assigned sampling weights
- Sampled one classroom per grade per school
- Grade 8 classrooms also sampled in the German cantons for national purposes

Table B.10 Allocation of School Sample in Switzerland - Population 1

Explicit Stratum	Sampled Schools
Private schools, with lower grade	2
Private schools, with upper grade	2
Private schools, with both grades	2
Canton of Bern, German part	30
Canton of Basle, lower track	16
Canton of Basle, medium track	16
Canton of Basle, higher track	16
Other German cantons, with lower grade	80
Other German cantons, with upper grade	80
Other German cantons, with both grades	18
Canton of Bern, French part	12
Canton of Valais, French part	10
Geneva	18
Canton of Grison, Italian part	2
Canton of Ticino	37
All Switzerland	341

THAILAND

Coverage and Exclusions

School-level exclusions consisted of special education schools, demonstration schools run by the Department of Teacher Education and the Ministry of University Affairs, and private schools.

Sample Design - Population 1

- Explicit stratification by 13 regions and two levels of urbanization (rural and urban), for a total of 25 strata (Bangkok region is all urban)
- No implicit stratification
- Schools sampled using a stratified simple random sample design
- Proportional allocation of 150 schools to the first 24 explicit strata; five schools sampled from Bangkok
- Sampled one classroom per grade per school
- Always sampled the first classroom listed in the school administrative records from each grade
- Computed an overall school participation adjustment for weighting for the first 24 explicit strata, thereby ignoring the relatively fine explicit stratification

Sample Design - Population 2

- No explicit stratification
- No implicit stratification
- Schools sampled using a simple random sample design
- Sampled 150 schools
- Sampled one classroom per grade per school
- Always sampled the first classroom listed in the school administrative records from each grade

UNITED STATES

Coverage and Exclusions

School-level exclusions consisted of ungraded schools.

Sample Design - Population 1 and Population 2

- Preliminary sampling of 59 primary sampling units (PSU), from a frame of 1026 PSUs
- Explicit stratification of PSUs, prior to sampling, by four regions: northeast, southeast, midwest, and west
- Eleven PSUs sampled with certainty – essentially large urban centers
- Explicit stratification of schools by type – public and private

- Implicit stratification by two levels of minority status (high and low) and three levels of split grades (lower, upper, and both)
- Increased (i.e., doubled) school selection probabilities in the high minority strata
- Sampled 220 schools
- Sampled one lower-grade classroom and two upper-grade classrooms per school

Appendix C: Design Effects and Effective Sample Size Tables

C

Table C.1 Design Effects and Effective Sample Sizes by Grade and Gender
Third Grade - Girls - Mathematics Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	2392	480	7920.6	4.5	1.8	6.12	391
Austria	1261	481	5616.8	3.8	2.1	3.29	384
Canada	3691	463	5815.5	3.0	1.3	5.79	637
Cyprus	1640	428	5364.4	3.1	1.8	2.99	548
Czech Republic	1652	493	6587.2	3.8	2.0	3.55	465
England	1544	452	7073.2	3.4	2.1	2.50	619
Greece	1444	424	7234.4	4.2	2.2	3.45	419
Hong Kong	1969	518	4778.2	3.5	1.6	5.16	381
Hungary	1492	476	7508.2	4.4	2.2	3.84	388
Iceland	854	403	3818.9	3.0	2.1	2.06	415
Iran, Islamic Rep.	1744	373	4073.2	4.9	1.5	10.39	168
Ireland	1367	479	6047.2	4.5	2.1	4.60	297
Japan	2109	536	5373.6	1.7	1.6	1.17	1804
Korea	1325	554	4678.3	2.5	1.9	1.79	741
Latvia (LSS)	1043	464	6438.0	4.5	2.5	3.22	324
Netherlands	1379	489	4158.4	3.2	1.7	3.45	399
New Zealand	1289	443	6621.1	4.5	2.3	4.00	322
Norway	1069	411	5018.2	3.8	2.2	3.09	346
Portugal	1288	420	7233.3	5.0	2.4	4.47	288
Scotland	1576	454	6008.1	3.5	2.0	3.29	479
Singapore	3378	553	9151.0	5.0	1.6	9.28	364
Slovenia	1233	483	5623.2	3.5	2.1	2.65	466
Thailand	1439	448	5077.4	5.6	1.9	8.77	164
United States	1857	479	6724.8	4.4	1.9	5.33	349

*Third grade in most countries.

Table C.2 Design Effects and Effective Sample Sizes by Grade and Gender
Third Grade - Boys - Mathematics Mean Scale Score- Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	2348	488	8289.4	4.6	1.9	6.00	391
Austria	1243	494	8020.2	9.2	2.5	13.08	95
Canada	3754	477	6446.7	3.2	1.3	5.81	647
Cyprus	1636	433	6582.9	3.3	2.0	2.67	613
Czech Republic	1604	502	7085.4	3.7	2.1	3.12	515
England	1512	461	8168.3	3.5	2.3	2.21	685
Greece	1508	432	7236.7	4.4	2.2	4.00	377
Hong Kong	2412	528	5554.8	3.2	1.5	4.48	538
Hungary	1456	479	8359.1	4.9	2.4	4.18	348
Iceland	844	418	5117.9	3.5	2.5	2.07	408
Iran, Islamic Rep.	1616	384	4500.3	4.4	1.7	7.04	229
Ireland	1522	473	6997.4	4.3	2.1	4.10	371
Japan	2197	539	5953.4	2.0	1.6	1.50	1469
Korea	1452	567	5068.9	2.8	1.9	2.22	653
Latvia (LSS)	1010	462	6656.3	5.3	2.6	4.33	233
Netherlands	1391	497	4261.7	2.9	1.8	2.75	505
New Zealand	1213	436	6903.5	4.4	2.4	3.39	358
Norway	1102	430	5027.0	3.5	2.1	2.71	407
Portugal	1362	430	7306.1	3.5	2.3	2.27	600
Scotland	1537	462	6546.3	3.8	2.1	3.38	455
Singapore	3645	551	10745.7	5.4	1.7	9.88	369
Slovenia	1288	492	6275.2	3.1	2.2	2.00	644
Thailand	1430	440	5042.5	5.0	1.9	7.14	200
United States	1962	480	6695.5	3.1	1.8	2.86	686

*Third grade in most countries.

Table C.3 Design Effects and Effective Sample Sizes by Grade and Gender
Fourth Grade - Girls - Mathematics Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	3252	546	8241.4	3.9	1.6	5.88	553
Austria	1262	555	6209.2	3.6	2.2	2.58	490
Canada	4063	531	6741.8	3.9	1.3	9.18	442
Cyprus	1657	499	6940.7	3.3	2.0	2.63	630
Czech Republic	1707	566	7469.9	3.6	2.1	3.02	565
England	1582	510	8059.0	4.4	2.3	3.73	424
Greece	1575	493	7828.8	4.5	2.2	4.11	383
Hong Kong	2013	587	5795.3	4.2	1.7	6.21	324
Hungary	1462	546	7278.3	3.9	2.2	3.07	476
Iceland	929	473	5219.4	3.0	2.4	1.64	567
Iran, Islamic Rep.	1655	424	4346.1	5.0	1.6	9.54	173
Ireland	1421	551	6884.7	4.3	2.2	3.89	365
Israel	1097	528	7387.1	4.1	2.6	2.48	442
Japan	2153	593	5879.8	2.2	1.7	1.74	1238
Korea	1388	603	5244.1	2.6	1.9	1.75	795
Kuwait	2252	402	3730.9	2.5	1.3	3.87	581
Latvia (LSS)	1088	530	6745.3	5.2	2.5	4.35	250
Netherlands	1238	569	4790.8	3.4	2.0	3.00	413
New Zealand	1238	504	6946.6	4.3	2.4	3.27	379
Norway	1025	499	5065.8	3.6	2.2	2.56	401
Portugal	1393	473	6272.1	3.7	2.1	3.12	447
Scotland	1639	520	7442.4	3.8	2.1	3.20	512
Singapore	3383	630	10149.8	6.4	1.7	13.47	251
Slovenia	1282	554	6688.4	4.0	2.3	3.06	420
Thailand	1480	496	4731.1	4.2	1.8	5.40	274
United States	3749	544	7014.0	3.3	1.4	5.69	659

*Fourth grade in most countries.

Table C.4 Design Effects and Effective Sample Sizes by Grade and Gender
Fourth Grade - Boys - Mathematics Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	3240	548	8560.7	3.6	1.6	4.89	663
Austria	1341	563	6238.2	3.6	2.2	2.86	469
Canada	4172	534	7311.5	3.4	1.3	6.64	628
Cyprus	1705	506	7904.9	3.5	2.2	2.64	645
Czech Republic	1561	568	7416.8	3.4	2.2	2.50	624
England	1544	515	8569.1	3.4	2.4	2.08	743
Greece	1478	491	8357.3	5.0	2.4	4.47	330
Hong Kong	2375	586	6578.2	4.7	1.7	7.99	297
Hungary	1474	552	8161.0	4.2	2.4	3.23	456
Iceland	880	474	5245.0	3.3	2.4	1.82	482
Iran, Islamic Rep.	1730	433	5133.8	6.0	1.7	11.96	145
Ireland	1452	548	7685.2	3.9	2.3	2.86	508
Israel	1085	537	6743.6	4.4	2.5	3.18	342
Japan	2153	601	7271.4	2.5	1.8	1.90	1131
Korea	1424	618	5553.3	2.5	2.0	1.64	871
Kuwait	2066	399	5138.2	4.6	1.6	8.59	240
Latvia (LSS)	1128	521	7591.3	5.5	2.6	4.45	254
Netherlands	1258	585	5052.5	3.8	2.0	3.67	342
New Zealand	1183	494	9077.0	5.7	2.8	4.25	278
Norway	1167	504	5830.9	3.5	2.2	2.39	488
Portugal	1459	478	6616.2	3.8	2.1	3.16	461
Scotland	1651	520	8524.4	4.3	2.3	3.62	456
Singapore	3750	620	11439.1	5.5	1.7	9.96	376
Slovenia	1258	551	6910.2	3.4	2.3	2.08	605
Thailand	1510	485	4881.2	5.8	1.8	10.47	144
United States	3547	545	7478.8	3.1	1.5	4.49	789

*Fourth grade in most countries.

Table C.5 Design Effects and Effective Sample Sizes for Third Grade
Third Grade - Girls - Science Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	2392	510	8480.4	4.4	1.9	5.42	441
Austria	1261	501	6815.5	4.0	2.3	2.96	426
Canada	3691	486	7081.3	2.9	1.4	4.27	865
Cyprus	1640	412	5023.8	3.0	1.8	2.99	549
Czech Republic	1652	485	6719.7	3.9	2.0	3.70	447
England	1544	495	9085.1	3.4	2.4	1.99	776
Greece	1444	439	6244.4	3.9	2.1	3.59	403
Hong Kong	1969	473	5037.1	3.8	1.6	5.57	354
Hungary	1492	460	7694.0	4.7	2.3	4.33	344
Iceland	854	431	6215.0	3.9	2.7	2.07	412
Iran, Islamic Rep.	1744	354	5325.5	5.7	1.7	10.71	163
Ireland	1367	477	7012.8	4.4	2.3	3.81	359
Japan	2109	521	5021.6	2.0	1.5	1.60	1316
Korea	1325	543	4745.0	2.7	1.9	2.08	637
Latvia (LSS)	1043	469	6715.3	4.8	2.5	3.56	293
Netherlands	1379	493	4005.3	3.1	1.7	3.26	423
New Zealand	1289	476	9191.5	5.7	2.7	4.58	281
Norway	1069	444	7822.6	4.5	2.7	2.83	378
Portugal	1288	415	8854.6	5.4	2.6	4.17	309
Scotland	1576	482	9221.2	4.7	2.4	3.77	419
Singapore	3378	484	8626.1	5.2	1.6	10.43	324
Slovenia	1233	478	5630.6	3.4	2.1	2.55	483
Thailand	1439	437	5796.3	7.1	2.0	12.45	116
United States	1857	508	8156.9	3.2	2.1	2.34	795

*Third grade in most countries.

Table C.6 Design Effects and Effective Sample Sizes by Grade and Gender
Third Grade - Boys - Science Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	2348	511	10681.9	5.7	2.1	7.24	324
Austria	1243	508	8383.9	6.9	2.6	6.98	178
Canada	3754	496	8245.4	3.2	1.5	4.77	786
Cyprus	1636	418	5641.8	2.7	1.9	2.09	783
Czech Republic	1604	503	7440.8	4.1	2.2	3.62	444
England	1512	503	11134.2	4.8	2.7	3.17	478
Greece	1508	453	7238.1	4.6	2.2	4.34	347
Hong Kong	2412	488	5557.3	3.4	1.5	5.13	470
Hungary	1456	472	7907.7	4.2	2.3	3.21	454
Iceland	844	440	7234.9	4.0	2.9	1.91	443
Iran, Islamic Rep.	1616	359	6287.3	5.7	2.0	8.41	192
Ireland	1522	481	8306.6	4.6	2.3	3.91	389
Japan	2197	523	5511.5	2.1	1.6	1.68	1306
Korea	1452	562	5261.1	2.8	1.9	2.17	671
Latvia (LSS)	1010	462	6902.6	5.2	2.6	3.95	256
Netherlands	1391	504	4006.0	3.8	1.7	4.93	282
New Zealand	1213	470	10635.2	5.9	3.0	3.95	307
Norway	1102	457	8321.2	4.6	2.7	2.75	401
Portugal	1362	431	9308.7	4.3	2.6	2.75	495
Scotland	1537	485	8756.5	4.4	2.4	3.47	442
Singapore	3645	491	10774.5	5.8	1.7	11.25	324
Slovenia	1288	496	6372.6	3.4	2.2	2.27	568
Thailand	1430	428	6201.3	6.5	2.1	9.85	145
United States	1962	514	9369.8	4.2	2.2	3.62	542

*Third grade in most countries.

Table C.7 Design Effects and Effective Sample Sizes by Grade and Gender
Fourth Grade - Girls - Science Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	3252	556	7786.5	3.3	1.5	4.58	710
Austria	1262	556	6235.8	3.7	2.2	2.72	463
Canada	4063	545	6794.4	3.2	1.3	5.98	679
Cyprus	1657	471	5174.6	3.1	1.8	3.05	544
Czech Republic	1707	548	6520.7	3.6	2.0	3.43	498
England	1582	548	8066.4	3.4	2.3	2.30	689
Greece	1575	494	6724.6	4.3	2.1	4.27	369
Hong Kong	2013	526	5329.0	3.8	1.6	5.35	376
Hungary	1462	525	6269.7	3.9	2.1	3.47	421
Iceland	929	496	6552.0	3.3	2.7	1.53	609
Iran, Islamic Rep.	1655	412	5212.4	4.7	1.8	7.09	233
Ireland	1421	536	6743.7	4.5	2.2	4.22	337
Israel	1097	501	7313.7	3.8	2.6	2.19	501
Japan	2153	567	4638.2	2.0	1.5	1.92	1120
Korea	1388	590	4331.6	2.5	1.8	1.94	717
Kuwait	2252	414	5642.2	3.1	1.6	3.88	581
Latvia (LSS)	1088	513	6470.9	5.5	2.4	5.11	213
Netherlands	1238	544	4074.8	3.5	1.8	3.72	333
New Zealand	1238	535	7932.0	4.8	2.5	3.58	346
Norway	1025	526	6646.3	3.7	2.5	2.07	495
Portugal	1393	478	6630.5	4.2	2.2	3.64	383
Scotland	1639	533	7938.8	4.3	2.2	3.87	423
Singapore	3383	545	8672.1	6.3	1.6	15.28	221
Slovenia	1282	544	5550.8	4.0	2.1	3.63	353
Thailand	1480	474	4761.9	4.3	1.8	5.87	252
United States	3749	560	8555.8	3.3	1.5	4.77	786

*Fourth grade in most countries.

Table C.8 Design Effects and Effective Sample Sizes by Grade and Gender
Fourth Grade - Boys - Science Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	3240	569	9512.0	3.4	1.7	3.92	826
Austria	1341	572	6436.0	3.9	2.2	3.10	432
Canada	4172	553	7962.9	3.7	1.4	7.10	588
Cyprus	1705	480	6193.5	4.0	1.9	4.43	385
Czech Republic	1561	565	6530.1	3.4	2.0	2.83	552
England	1544	555	10354.3	4.0	2.6	2.42	638
Greece	1478	501	7034.7	4.5	2.2	4.19	352
Hong Kong	2375	540	6471.7	4.1	1.7	6.31	377
Hungary	1474	539	6562.3	3.8	2.1	3.21	459
Iceland	880	514	7745.3	4.3	3.0	2.11	417
Iran, Islamic Rep.	1730	421	5823.6	5.9	1.8	10.33	167
Ireland	1452	543	7653.8	3.5	2.3	2.37	612
Israel	1085	512	7498.8	4.5	2.6	2.90	375
Japan	2153	580	5860.0	2.0	1.6	1.47	1469
Korea	1424	604	4845.5	2.2	1.8	1.48	960
Kuwait	2066	389	8452.5	5.8	2.0	8.19	252
Latvia (LSS)	1128	512	7549.6	5.4	2.6	4.35	260
Netherlands	1258	570	4267.7	3.6	1.8	3.77	334
New Zealand	1183	527	10907.7	6.1	3.0	3.99	296
Norway	1167	534	8014.0	4.7	2.6	3.19	366
Portugal	1459	481	7591.0	4.5	2.3	3.97	367
Scotland	1651	538	9535.3	4.5	2.4	3.49	473
Singapore	3750	549	10125.2	5.4	1.6	10.78	348
Slovenia	1258	548	6033.5	3.3	2.2	2.30	546
Thailand	1510	471	5256.3	5.9	1.9	9.87	153
United States	3547	571	9443.4	3.3	1.6	4.02	883

*Fourth grade in most countries.

Table C.9 Design Effects and Effective Sample Sizes by Grade and Gender Seventh Grade - Girls - Mathematics Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	3039	500	8028.7	4.3	1.6	7.07	430
Austria	1545	509	6629.4	3.3	2.1	2.50	618
Belgium (Fl)	1344	559	6029.3	4.7	2.1	4.95	272
Belgium (Fr)	1196	501	5806.2	4.2	2.2	3.60	332
Bulgaria	960	518	10583.9	8.7	3.3	6.82	141
Canada	3957	493	6416.9	2.6	1.3	4.19	944
Colombia	1359	365	4029.5	3.9	1.7	5.05	269
Cyprus	1428	446	6137.9	2.6	2.1	1.62	883
Czech Republic	1682	520	7757.4	5.6	2.1	6.91	243
Denmark	1039	462	5807.6	2.9	2.4	1.53	681
England	825	467	7713.5	4.3	3.1	2.00	413
France	1439	489	5193.6	3.3	1.9	3.06	471
Germany	1427	484	6937.2	4.5	2.2	4.12	346
Greece	1902	440	6822.5	3.0	1.9	2.57	739
Hong Kong	1499	556	8894.4	8.3	2.4	11.54	130
Hungary	1533	501	7727.3	4.4	2.2	3.91	392
Iceland	947	458	4576.4	3.2	2.2	2.11	449
Iran, Islamic Rep.	1646	393	3048.4	2.3	1.4	2.94	560
Ireland	1678	494	7375.4	4.8	2.1	5.34	314
Japan	2500	565	8335.0	2.0	1.8	1.17	2133
Korea	1254	567	10791.0	4.4	2.9	2.23	563
Latvia (LSS)	1317	460	5728.4	3.3	2.1	2.53	521
Lithuania	1277	433	5355.0	3.5	2.0	2.90	440
Netherlands	1037	515	5978.8	4.3	2.4	3.17	327
New Zealand	1498	470	7104.9	3.8	2.2	3.03	494
Norway	1212	459	5696.5	3.2	2.2	2.17	559
Portugal	1732	420	3457.3	2.2	1.4	2.50	692
Romania	1931	452	7069.2	3.7	1.9	3.68	525
Russian Federation	2137	499	7254.5	3.5	1.8	3.52	607
Scotland	1440	462	6213.2	3.8	2.1	3.30	437
Singapore	1873	601	8525.2	8.0	2.1	13.97	134
Slovak Republic	1823	505	6849.4	3.3	1.9	2.90	629
Slovenia	1486	496	6649.1	3.2	2.1	2.32	641
South Africa	2818	344	3633.6	3.3	1.1	8.31	339
Spain	1892	445	4511.7	2.7	1.5	3.06	618
Sweden	1374	475	5806.3	3.2	2.1	2.47	557
Switzerland	2019	498	5433.0	2.6	1.6	2.46	822
Thailand	3301	495	6186.0	5.7	1.4	17.34	190
United States	1976	473	7400.7	5.7	1.9	8.80	224

*Seventh grade in most countries.

Table C.10 Design Effects and Effective Sample Sizes by Grade and Gender
Seventh Grade - Boys - Mathematics Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	2560	495	8863.9	5.2	1.9	7.82	327
Austria	1358	510	7984.1	4.6	2.4	3.57	380
Belgium (Fl)	1424	557	5727.0	4.5	2.0	4.97	286
Belgium (Fr)	1052	514	6254.9	4.1	2.4	2.88	365
Bulgaria	820	508	10781.7	6.9	3.6	3.58	229
Canada	4144	495	6354.5	2.7	1.2	4.79	865
Colombia	1265	372	3903.3	3.8	1.8	4.73	268
Cyprus	1496	446	7319.7	2.5	2.2	1.30	1153
Czech Republic	1663	527	8172.0	4.8	2.2	4.64	358
Denmark	998	468	6299.4	2.8	2.5	1.21	825
England	978	484	8266.8	6.2	2.9	4.52	217
France	1484	497	5565.7	3.6	1.9	3.48	426
Germany	1426	486	7385.4	4.8	2.3	4.50	317
Greece	2022	440	7728.9	3.2	2.0	2.76	732
Hong Kong	1910	570	10521.1	9.7	2.3	17.25	111
Hungary	1533	503	8736.1	3.8	2.4	2.52	609
Iceland	1010	460	4610.4	2.7	2.1	1.62	622
Iran, Islamic Rep.	2074	407	3292.0	2.7	1.3	4.47	464
Ireland	1449	507	7636.7	6.0	2.3	6.76	214
Japan	2630	576	9990.9	2.7	1.9	1.95	1349
Korea	1653	584	10905.9	3.7	2.6	2.08	796
Latvia (LSS)	1244	463	5971.9	3.5	2.2	2.55	488
Lithuania	1254	423	5909.5	3.6	2.2	2.72	461
Netherlands	1053	517	6466.6	5.2	2.5	4.35	242
New Zealand	1686	473	7918.9	4.6	2.2	4.44	380
Norway	1257	462	5852.6	3.3	2.2	2.30	547
Portugal	1630	426	3669.4	2.7	1.5	3.28	496
Romania	1812	457	7094.4	3.7	2.0	3.44	526
Russian Federation	2001	502	8325.3	5.1	2.0	6.18	324
Scotland	1462	465	7097.7	4.6	2.2	4.30	340
Singapore	1768	601	8862.3	7.1	2.2	10.15	174
Slovak Republic	1777	511	7629.3	4.4	2.1	4.58	388
Slovenia	1411	501	6776.2	3.5	2.2	2.53	557
South Africa	2432	352	4482.7	5.3	1.4	15.10	161
Spain	1849	451	5141.5	2.7	1.7	2.68	689
Sweden	1444	480	5883.7	2.8	2.0	1.87	773
Switzerland	2059	513	5840.9	2.9	1.7	2.95	698
Thailand	2440	494	6133.0	4.8	1.6	9.21	265
United States	1910	478	8526.8	5.7	2.1	7.41	258

*Seventh grade in most countries.

**Table C.11 Design Effects and Effective Sample Sizes by Grade and Gender
Eighth Grade - Girls - Mathematics Mean Scale Score - Population 2**

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	3722	532	9302.1	4.6	1.6	8.40	443
Austria	1321	536	8115.5	4.5	2.5	3.37	392
Belgium (Fl)	1437	567	7708.7	7.4	2.3	10.29	140
Belgium (Fr)	1291	524	6949.1	3.7	2.3	2.53	510
Bulgaria	1015	546	12872.6	6.7	3.6	3.52	288
Canada	4088	530	7071.2	2.7	1.3	4.08	1001
Colombia	1383	384	3965.7	3.6	1.7	4.45	311
Cyprus	1424	475	7414.2	2.5	2.3	1.22	1171
Czech Republic	1637	558	8624.3	6.3	2.3	7.51	218
Denmark	1120	494	6476.3	3.4	2.4	2.01	558
England	853	504	8193.6	3.5	3.1	1.24	688
France	1430	536	6011.3	3.8	2.1	3.50	408
Germany	1423	509	7826.6	5.0	2.3	4.47	318
Greece	1952	478	7267.8	3.1	1.9	2.62	745
Hong Kong	1508	577	9471.3	7.7	2.5	9.50	159
Hungary	1489	537	8771.5	3.6	2.4	2.26	659
Iceland	868	486	5183.7	5.6	2.4	5.17	168
Iran, Islamic Rep.	1637	421	3453.7	3.3	1.5	5.05	324
Ireland	1535	520	7872.5	6.0	2.3	6.99	220
Israel	668	509	8153.0	6.9	3.5	3.87	173
Japan	2495	600	9371.2	2.1	1.9	1.22	2052
Korea	1335	598	11732.9	3.4	3.0	1.32	1008
Kuwait	897	395	3035.4	2.6	1.8	2.01	447
Latvia (LSS)	1259	491	6749.7	3.5	2.3	2.32	543
Lithuania	1385	478	6512.4	4.1	2.2	3.57	388
Netherlands	977	536	7782.7	6.4	2.8	5.21	188
New Zealand	1775	503	7697.4	5.3	2.1	6.42	276
Norway	1634	501	6436.7	2.7	2.0	1.81	902
Portugal	1663	449	4045.5	2.7	1.6	3.03	550
Romania	1914	480	7590.0	4.0	2.0	3.99	480
Russian Federation	2151	536	7548.9	5.0	1.9	7.09	304
Scotland	1380	490	7301.7	5.2	2.3	5.20	265
Singapore	2307	645	7716.2	5.4	1.8	8.87	260
Slovak Republic	1785	545	8027.6	3.6	2.1	2.90	616
Slovenia	1381	537	7587.4	3.3	2.3	1.97	701
South Africa	2319	349	3899.5	4.1	1.3	9.97	233
Spain	2007	483	5174.3	2.6	1.6	2.58	778
Sweden	1979	518	7408.4	3.1	1.9	2.61	758
Switzerland	2411	543	7205.7	3.1	1.7	3.27	738
Thailand	3390	526	7565.4	7.0	1.5	22.19	153
United States	3561	497	7835.0	4.5	1.5	9.09	392

*Eighth grade in most countries.

Table C.12 Design Effects and Effective Sample Sizes by Grade and Gender Eighth Grade - Boys - Mathematics Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	3529	527	9985.3	5.1	1.7	9.21	383
Austria	1385	544	8761.6	3.2	2.5	1.65	838
Belgium (Fl)	1457	563	9152.1	8.8	2.5	12.30	118
Belgium (Fr)	1269	530	7792.1	4.7	2.5	3.62	351
Bulgaria	942	533	11266.3	7.0	3.5	4.05	233
Canada	4137	526	7791.3	3.2	1.4	5.60	739
Colombia	1240	386	4301.5	6.9	1.9	13.62	91
Cyprus	1494	472	7922.9	2.8	2.3	1.43	1041
Czech Republic	1690	569	8857.7	4.5	2.3	3.91	432
Denmark	1147	511	7370.5	3.2	2.5	1.57	731
England	923	508	9040.6	5.1	3.1	2.66	347
France	1449	542	5523.3	3.1	2.0	2.50	581
Germany	1410	512	7917.4	5.1	2.4	4.67	302
Greece	2037	490	8222.2	3.7	2.0	3.40	599
Hong Kong	1829	597	10604.4	7.7	2.4	10.20	179
Hungary	1423	537	8507.3	3.6	2.4	2.20	646
Iceland	905	488	6336.3	5.5	2.6	4.37	207
Iran, Islamic Rep.	2043	434	3480.5	2.9	1.3	4.97	411
Ireland	1541	535	9160.1	7.2	2.4	8.65	178
Israel	672	539	8009.0	6.6	3.5	3.70	182
Japan	2646	609	11296.9	2.6	2.1	1.53	1731
Korea	1585	615	11807.6	3.2	2.7	1.39	1142
Kuwait	758	389	3587.4	4.3	2.2	3.87	196
Latvia (LSS)	1148	496	6731.8	3.8	2.4	2.42	474
Lithuania	1140	477	6318.6	4.0	2.4	2.91	392
Netherlands	980	545	8010.3	7.8	2.9	7.43	132
New Zealand	1908	512	8530.1	5.9	2.1	7.70	248
Norway	1633	505	7630.9	2.8	2.2	1.66	983
Portugal	1728	460	4046.0	2.8	1.5	3.44	502
Romania	1809	483	8337.4	4.8	2.1	4.97	364
Russian Federation	1871	535	9470.6	6.3	2.2	7.81	240
Scotland	1477	506	7843.3	6.6	2.3	8.09	182
Singapore	2334	642	7831.0	6.3	1.8	11.72	199
Slovak Republic	1716	549	8928.0	3.7	2.3	2.68	640
Slovenia	1324	545	7799.4	3.8	2.4	2.41	550
South Africa	2089	360	4607.3	6.3	1.5	18.18	115
Spain	1848	492	5584.6	2.5	1.7	2.15	860
Sweden	2084	520	7174.4	3.6	1.9	3.67	568
Switzerland	2443	548	8096.7	3.5	1.8	3.69	662
Thailand	2407	517	6963.9	5.6	1.7	10.96	220
United States	3526	502	8677.3	5.2	1.6	11.04	319

*Eighth grade in most countries.

Table C.13 Design Effects and Effective Sample Sizes by Grade and Gender
Seventh Grade - Girls - Science Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	3039	502	9598.9	4.0	1.8	5.02	606
Austria	1545	516	8144.0	4.1	2.3	3.23	479
Belgium (Fl)	1344	521	4989.4	3.1	1.9	2.58	521
Belgium (Fr)	1196	432	6013.7	3.5	2.2	2.45	489
Bulgaria	960	532	11059.2	6.7	3.4	3.90	246
Canada	3957	493	7081.5	2.5	1.3	3.54	1118
Colombia	1359	378	4801.4	4.4	1.9	5.38	252
Cyprus	1428	420	6702.3	2.6	2.2	1.47	974
Czech Republic	1682	523	6470.0	4.1	2.0	4.42	381
Denmark	1039	427	6882.8	2.8	2.6	1.17	885
England	825	500	9404.8	4.6	3.4	1.86	444
France	1439	443	5146.2	3.0	1.9	2.56	563
Germany	1427	495	8645.7	4.5	2.5	3.36	425
Greece	1902	446	7212.3	2.8	1.9	2.01	945
Hong Kong	1499	485	6902.6	5.8	2.1	7.27	206
Hungary	1533	510	7850.7	3.4	2.3	2.21	695
Iceland	947	456	5275.5	2.4	2.4	1.04	914
Iran, Islamic Rep.	1646	428	4407.0	4.1	1.6	6.21	265
Ireland	1678	487	8188.9	4.5	2.2	4.20	400
Japan	2500	526	6834.2	1.9	1.7	1.28	1957
Korea	1254	521	8123.3	3.2	2.5	1.57	798
Latvia (LSS)	1317	430	5541.3	3.0	2.1	2.13	619
Lithuania	1277	401	5986.9	4.2	2.2	3.79	337
Netherlands	1037	512	6017.9	4.4	2.4	3.26	318
New Zealand	1498	472	8435.2	3.7	2.4	2.47	606
Norway	1212	477	6495.1	3.6	2.3	2.47	491
Portugal	1732	420	4681.3	2.4	1.6	2.08	832
Romania	1931	448	9803.8	4.9	2.3	4.65	415
Russian Federation	2137	475	7896.0	3.8	1.9	3.86	553
Scotland	1440	459	8033.4	4.1	2.4	2.97	484
Singapore	1873	541	9661.7	8.2	2.3	13.18	142
Slovak Republic	1823	499	6791.5	3.1	1.9	2.66	685
Slovenia	1486	521	7294.2	2.8	2.2	1.54	963
South Africa	2818	312	8343.5	5.2	1.7	9.21	306
Spain	1892	467	5840.6	2.3	1.8	1.77	1066
Sweden	1374	484	6542.8	3.3	2.2	2.31	596
Switzerland	2019	475	6404.6	2.9	1.8	2.62	769
Thailand	3301	492	4578.6	3.5	1.2	8.71	379
United States	1976	502	10022.5	5.8	2.3	6.73	294

*Seventh grade in most countries.

Table C.14 Design Effects and Effective Sample Sizes by Grade and Gender Seventh Grade - Boys - Science Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	2560	507	11508.3	5.2	2.1	6.12	419
Austria	1358	522	9589.6	4.3	2.7	2.61	520
Belgium (Fl)	1424	536	5587.0	3.3	2.0	2.79	510
Belgium (Fr)	1052	453	6106.0	3.6	2.4	2.22	473
Bulgaria	820	529	10112.7	5.5	3.5	2.44	336
Canada	4144	505	8850.7	2.9	1.5	3.91	1059
Colombia	1265	396	5438.0	3.8	2.1	3.31	383
Cyprus	1496	420	8350.1	2.8	2.4	1.44	1039
Czech Republic	1663	543	6695.9	3.2	2.0	2.54	655
Denmark	998	452	7845.4	3.0	2.8	1.17	850
England	978	522	10692.2	5.6	3.3	2.88	339
France	1484	461	5770.1	3.1	2.0	2.39	620
Germany	1426	505	9470.3	4.9	2.6	3.59	398
Greece	2022	452	8012.7	3.2	2.0	2.53	799
Hong Kong	1910	503	7787.9	6.6	2.0	10.56	181
Hungary	1533	525	8743.1	3.9	2.4	2.63	583
Iceland	1010	468	5927.2	4.4	2.4	3.29	307
Iran, Islamic Rep.	2074	443	5567.5	2.9	1.6	3.13	662
Ireland	1449	504	8247.1	4.6	2.4	3.69	393
Japan	2630	536	7934.0	2.6	1.7	2.27	1157
Korea	1653	545	8379.9	2.8	2.3	1.52	1087
Latvia (LSS)	1244	440	6567.0	3.6	2.3	2.44	509
Lithuania	1254	405	6627.3	3.5	2.3	2.34	536
Netherlands	1053	523	6411.8	4.0	2.5	2.68	392
New Zealand	1686	489	9947.8	4.3	2.4	3.12	540
Norway	1257	489	7792.2	3.6	2.5	2.10	597
Portugal	1630	436	5428.7	2.4	1.8	1.75	934
Romania	1812	456	10204.2	4.7	2.4	3.85	471
Russian Federation	2001	493	9767.5	5.3	2.2	5.72	350
Scotland	1462	477	9373.9	4.4	2.5	3.00	487
Singapore	1768	548	10374.7	7.9	2.4	10.69	165
Slovak Republic	1777	520	7438.7	4.0	2.0	3.88	458
Slovenia	1411	539	7314.7	3.0	2.3	1.72	822
South Africa	2432	324	8581.3	6.4	1.9	11.64	209
Spain	1849	487	6710.8	2.9	1.9	2.36	783
Sweden	1444	493	7554.1	2.9	2.3	1.60	901
Switzerland	2059	492	6857.1	2.9	1.8	2.55	806
Thailand	2440	495	5067.2	3.3	1.4	5.14	475
United States	1910	514	11944.2	6.3	2.5	6.30	303

*Seventh grade in most countries.

Table C.15 Design Effects and Effective Sample Sizes by Grade and Gender
Eighth Grade - Girls - Science Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	3722	540	10513.8	4.1	1.7	5.89	632
Austria	1321	549	9605.5	4.6	2.7	2.90	456
Belgium (Fl)	1437	543	6257.4	5.8	2.1	7.82	184
Belgium (Fr)	1291	463	6553.6	2.9	2.3	1.69	762
Bulgaria	1015	567	12463.5	6.6	3.5	3.52	288
Canada	4088	525	7980.0	3.7	1.4	7.00	584
Colombia	1383	405	5085.8	4.6	1.9	5.68	243
Cyprus	1424	465	6817.8	2.7	2.2	1.48	962
Czech Republic	1637	562	7271.7	5.8	2.1	7.54	217
Denmark	1120	463	6918.3	3.9	2.5	2.49	450
England	853	542	10490.9	4.2	3.5	1.46	584
France	1430	490	5864.9	3.3	2.0	2.66	538
Germany	1423	524	9847.1	4.9	2.6	3.43	415
Greece	1952	489	7083.1	3.1	1.9	2.59	754
Hong Kong	1508	507	7348.2	5.1	2.2	5.40	279
Hungary	1489	545	8179.2	3.4	2.3	2.15	691
Iceland	868	486	5479.2	4.6	2.5	3.39	256
Iran, Islamic Rep.	1637	461	4540.2	3.2	1.7	3.66	448
Ireland	1535	532	8392.9	5.2	2.3	4.97	309
Israel	668	512	9559.9	6.1	3.8	2.62	255
Japan	2495	562	7380.0	2.0	1.7	1.34	1865
Korea	1335	551	8213.4	2.3	2.5	0.90	1490
Kuwait	897	444	4820.0	3.3	2.3	1.97	455
Latvia (LSS)	1259	478	6267.9	3.2	2.2	1.99	631
Lithuania	1385	470	6502.9	4.0	2.2	3.39	409
Netherlands	977	550	6933.5	4.9	2.7	3.36	291
New Zealand	1775	512	8964.8	5.2	2.2	5.42	328
Norway	1634	520	6875.8	2.0	2.1	0.96	1703
Portugal	1663	468	5394.9	2.7	1.8	2.31	721
Romania	1914	480	9889.9	5.0	2.3	4.76	403
Russian Federation	2151	533	8690.2	3.7	2.0	3.45	623
Scotland	1380	507	9287.9	4.7	2.6	3.23	427
Singapore	2307	603	9058.1	7.0	2.0	12.54	184
Slovak Republic	1785	537	8404.9	3.9	2.2	3.26	547
Slovenia	1381	548	7147.1	3.2	2.3	2.00	689
South Africa	2319	315	8785.8	6.0	1.9	9.66	240
Spain	2007	508	5997.1	2.3	1.7	1.84	1093
Sweden	1979	528	7871.6	3.4	2.0	2.88	688
Switzerland	2411	514	7600.5	3.0	1.8	2.81	857
Thailand	3390	526	5233.5	4.3	1.2	11.83	287
United States	3561	530	10269.7	5.2	1.7	9.56	373

*Eighth grade in most countries.

**Table C.16 Design Effects and Effective Sample Sizes by Grade and Gender
Eighth Grade - Boys - Science Mean Scale Score - Population 2**

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	3529	550	12105.8	5.2	1.9	7.97	443
Austria	1385	566	9472.1	4.0	2.6	2.29	604
Belgium (Fl)	1457	558	6792.1	6.0	2.2	7.77	187
Belgium (Fr)	1269	479	7945.0	4.8	2.5	3.72	341
Bulgaria	942	563	12051.1	5.7	3.6	2.50	377
Canada	4137	537	9095.2	3.1	1.5	4.35	952
Colombia	1240	418	6294.6	7.3	2.3	10.42	119
Cyprus	1494	461	8717.2	2.2	2.4	0.82	1819
Czech Republic	1690	586	7575.8	4.2	2.1	3.99	424
Denmark	1147	494	8108.4	3.6	2.7	1.85	619
England	923	562	11659.4	5.6	3.6	2.52	367
France	1449	506	5815.9	2.7	2.0	1.88	770
Germany	1410	542	10144.9	5.9	2.7	4.78	295
Greece	2037	505	7233.9	2.6	1.9	1.83	1112
Hong Kong	1829	535	8014.9	5.5	2.1	6.78	270
Hungary	1423	563	7859.3	3.1	2.4	1.79	793
Iceland	905	501	6846.9	5.1	2.8	3.48	260
Iran, Islamic Rep.	2043	477	5716.0	3.8	1.7	5.08	402
Ireland	1541	544	9812.7	6.6	2.5	6.90	223
Israel	672	545	10654.2	6.4	4.0	2.59	260
Japan	2646	579	8655.3	2.4	1.8	1.78	1488
Korea	1585	576	8967.1	2.7	2.4	1.27	1250
Kuwait	758	416	5709.8	6.6	2.7	5.82	130
Latvia (LSS)	1148	492	6804.9	3.3	2.4	1.88	611
Lithuania	1140	484	6538.1	3.8	2.4	2.56	445
Netherlands	980	570	7295.0	6.4	2.7	5.54	177
New Zealand	1908	538	10562.9	5.4	2.4	5.35	356
Norway	1633	534	8300.1	3.2	2.3	2.05	798
Portugal	1728	490	5259.4	2.8	1.7	2.53	684
Romania	1809	492	10726.4	5.3	2.4	4.79	378
Russian Federation	1871	544	9449.0	4.9	2.2	4.75	394
Scotland	1477	527	10320.9	6.4	2.6	5.87	251
Singapore	2334	612	9069.5	6.7	2.0	11.68	200
Slovak Republic	1716	552	8393.3	3.5	2.2	2.49	688
Slovenia	1324	573	7952.9	3.2	2.5	1.69	781
South Africa	2089	337	10448.0	9.5	2.2	18.08	116
Spain	1848	526	5980.2	2.1	1.8	1.31	1408
Sweden	2084	542	8332.6	3.4	2.0	2.94	709
Switzerland	2443	529	8782.2	3.2	1.9	2.81	868
Thailand	2407	524	5186.1	3.9	1.5	7.20	335
United States	3526	539	12027.6	4.9	1.8	7.09	497

*Eighth grade in most countries.

Appendix D: Dummy Variables Constructed for Conditioning

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1

Variable Name	Variable Label	Original Coding	New Coding
ASBGBRN1	GEN\BORN IN COUNTRY	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGBRN2	GEN\BORN IN COUNTRY\AGE	age when moved to country: 1-15; missing:99; not admin.:98;	1-15 0 0 1 0 1
ASBGLANG	GEN\SPEAK LANGUAGE OF TEST AT HOME	always or almost always:1; sometimes:2; never:3; missing:9; not admin.:8;	1 0 0 0 1 0 0 0 1 0 0 0 0 0 0
ASBMEXTR	MAT\OUTSIDE SCHL\EXTRA LESSONS	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBSEXTR	SCI\OUTSIDE SCHL\EXTRA LESSONS	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGCLUB	GEN\OUTSIDE SCHL\CLUBS PARTICIPATION	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGDAY1	GEN\OUTSIDE SCHL\WATCH TY OR VIDEOS	no time:1; less than 1 hour:2; 1-2 hours:3; 3-4 hours:4; more than 4 hours:5; missing:9; not admin.:8;	0 0 0.5 0 1.5 0 4 0 6 0 0 1 0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBGDAY2	GEN\OUTSIDE SCHL\PLAY COMPUTER GAMES	no time:1;	0 0
		less than 1 hour:2;	0.5 0
		1-2 hours:3;	1.5 0
		3-4 hours:4;	4 0
		more than 4 hours:5;	6 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGDAY3	GEN\OUTSIDE SCHL\PLAY WITH FRIENDS	no time:1;	0 0
		less than 1 hour:2;	0.5 0
		1-2 hours:3;	1.5 0
		3-4 hours:4;	4 0
		more than 4 hours:5;	6 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGDAY4	GEN\OUTSIDE SCHL\DOING JOBS AT HOME	no time:1;	0 0
		less than 1 hour:2;	0.5 0
		1-2 hours:3;	1.5 0
		3-4 hours:4;	4 0
		more than 4 hours:5;	6 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGDAY5	GEN\OUTSIDE SCHL\PLAYING SPORTS	no time:1;	0 0
		less than 1 hour:2;	0.5 0
		1-2 hours:3;	1.5 0
		3-4 hours:4;	4 0
		more than 4 hours:5;	6 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGDAY6	GEN\OUTSIDE SCHL\READING A BOOK	no time:1;	0 0
		less than 1 hour:2;	0.5 0
		1-2 hours:3;	1.5 0
		3-4 hours:4;	4 0
		more than 4 hours:5;	6 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBMDAY7	MAT\OUTSIDE SCHL\STUDYING MATH	no time:1;	0 0
		less than 1 hour:2;	0.5 0
		1-2 hours:3;	1.5 0
		3-4 hours:4;	4 0
		more than 4 hours:5;	6 0
		missing:9;	0 1
		not admin.:8;	0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBSDAY8	SCI\OUTSIDE SCHL\STUDYING SCIENCE	no time:1; less than 1 hour:2; 1-2 hours:3; 3-4 hours:4; more than 4 hours:5; missing:9; not admin.:8;	0 0 0.5 0 1.5 0 4 0 6 0 0 1 0 1
ASBGDAY9	GEN\OUTSIDE SCHL\STUDYING OTHER SUBJ	no time:1; less than 1 hour:2; 1-2 hours:3; 3-4 hours:4; more than 4 hours:5; missing:9; not admin.:8;	0 0 0.5 0 1.5 0 4 0 6 0 0 1 0 1
ASBGADU1	GEN\STUDENT LIVES WITH\MOTHER	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGADU2	GEN\STUDENT LIVES WITH\FATHER	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGADU3	GEN\STUDENT LIVES WITH\BROTHER(S)	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGADU4	GEN\STUDENT LIVES WITH\SISTER(S)	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGADU5	GEN\STUDENT LIVES WITH\STEPMOTHER	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGADU6	GEN\STUDENT LIVES WITH\STEPFATHER	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGADU7	GEN\STUDENT LIVES WITH\GRANDPRNT(S)	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGADU8	GEN\STUDENT LIVES WITH\RELATIVE(S)	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBGADU9	GEN\STUDENT LIVES WITH\OTHER(S)	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGHOME	GEN\# OF PEOPLE LIVING IN HOME	number of people:1-60; missing:99; not admin.:98;	1-60 0 0 1 0 1
ASBGBRNM	GEN\BORN IN COUNTRY\MOTHER	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGBRNF	GEN\BORN IN COUNTRY\FATHER	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGBOOK	GEN\# OF BOOKS IN STUDENT'S HOME	0-10 books:1; 11-25 books:2; 26-100 books:3; 101-200 books:4; more than 200 books:5; missing:9; not admin.:8;	1 1 0 2 4 0 3 9 0 4 16 0 5 25 0 0 0 1 0 0 1
ASBGPS01	GEN\HOME POSSESS\CALCULA-TOR	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGPS02	GEN\HOME POSSESS\COMPUTER	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGPS03	GEN\HOME POSSESS\STUDY DESK	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBGPS04	GEN\HOME POSSESS\DICTIONARY	yes:1; no:2; missing:9; not admin.:8;	1 0 0 1 0 0 0 0
ASBSMIP1	SCI\MOTHER IMPT\DO WELL IN SCI-ENCE	yes:1; no:2; missing:9; not admin.:8;	3 0 2 0 0 1 0 1
ASBMMIP2	MAT\MOTHER IMPT\DO WELL IN MATH	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBGMIP3	GEN\MOTHER IMPT\GOOD IN SPORTS	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1
ASBGMIP4	GEN\MOTHER IMPT\HAVE TIME FOR FUN	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1
ASBSFIP1	SCI\FRIENDS IMPT\DO WELL IN SCI-ENCE	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1
ASBMFIP2	MAT\FRIENDS IMPT\DO WELL IN MATH	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1
ASBGFIP3	GEN\FRIENDS IMPT\GOOD IN SPORTS	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1
ASBGFIP4	GEN\FRIENDS IMPT\HAVE TIME FOR FUN	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1
ASBSSIP1	SCI\SELF IMPT\DO WELL IN SCI-ENCE	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1
ASBMSIP2	MAT\SELF IMPT\DO WELL IN MATH	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1
ASBGSIP3	GEN\SELF IMPT\GOOD IN SPORTS	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1
ASBGSIP4	GEN\SELF IMPT\HAVE TIME FOR FUN	yes:1; no:2; missing:9; not admin.:8;	1 0 0 0 0 1 0 1
ASBM-GOOD	MAT\USUALLY DO WELL IN MATH	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBSGOOD	SCI\USUALLY DO WELL IN SCIENCE	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1
ASBGSSTL	GEN\STUDENT HAD SOMETHING STOLEN	yes:1; no:2; missing:9; not admin.:8;	0 0 1 0 0 1 0 1
ASBGSHRT	GEN\STUDENT THOUGHT MIGHT GET HURT	yes:1; no:2; missing:9; not admin.:8;	0 0 1 0 0 1 0 1
ASBGFSTL	GEN\FRIEND HAD SOMETHING STOLEN	yes:1; no:2; missing:9; not admin.:8;	0 0 1 0 0 1 0 1
ASBGFHRT	GEN\FRIEND THOUGHT MIGHT GET HURT	yes:1; no:2; missing:9; not admin.:8;	0 0 1 0 0 1 0 1
ASBMDOW 1	MAT\DO WELL\NATURAL TALENT	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1
ASBMDOW 2	MAT\DO WELL\GOOD LUCK	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1
ASBMDOW 3	MAT\DO WELL\HARD WORK STUDYING	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1
ASBMDOW 4	MAT\DO WELL\MEMORIZE NOTES	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBSDOW1	SCI\DO WELL\NATURAL TALENT	strongly agree:1;	3 0
		agree:2;	2 0
		disagree:3;	1 0
		strongly disagree:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBSDOW2	SCI\DO WELL\GOOD LUCK	strongly agree:1;	3 0
		agree:2;	2 0
		disagree:3;	1 0
		strongly disagree:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBSDOW3	SCI\DO WELL\HARD WORK STUDYING	strongly agree:1;	3 0
		agree:2;	2 0
		disagree:3;	1 0
		strongly disagree:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBSDOW4	SCI\DO WELL\MEMORIZE NOTES	strongly agree:1;	3 0
		agree:2;	2 0
		disagree:3;	1 0
		strongly disagree:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBMLIKE	MAT\LIKE MATHEMATICS	like a lot:1;	0 0
		like:2;	1 0
		dislike:3;	2 0
		dislike a lot:4;	3 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBSLIKE	SCI\LIKE SCIENCE	like a lot:1;	0 0
		like:2;	1 0
		dislike:3;	2 0
		dislike a lot:4;	3 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBMCMLK	MAT\LIKE COMPUTERS\MATH CLASS	don't use computers:1;	1 0 0
		like a lot:2;	0 1 0
		like:3;	0 2 0
		dislike:4;	0 3 0
		dislike a lot:5;	0 4 0
		missing:9;	0 0 1
		not admin.:8;	0 0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBSCMLK	SCI\LIKE COMPUTERS\SCIENCE CLASS	don't use computers:1; like a lot: 2;like: 3;dislike:4; dislike a lot:5; missing:9; not admin.:8;	1 0 0 0 1 0 0 2 0 0 3 0 0 4 0 0 0 1 0 0 1
ASBMENJY	MAT\THINK\ENJOY LEARNING MATH	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1
ASBMBORE	MAT\THINK\MATH IS BORING	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1
ASBMEASY	MAT\THINK\MATH IS AN EASY SUBJECT	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1
ASBSENJY	SCI\THINK\ENJOY LEARNING SCIENCE	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1
ASBSBORE	SCI\THINK\SCIENCE IS BORING	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1
ASBSEASY	SCI\THINK\SCIENCE IS AN EASY SUBJECT	strongly agree:1; agree:2; disagree:3; strongly disagree:4; missing:9; not admin.:8;	3 0 2 0 1 0 0 0 0 1 0 1
ASBMPROB	MAT\TEACHER SHOW HOW TO DO PROBLEMS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBMNOTE	MAT\COPY NOTES FROM THE BOARD	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMTEST	MAT\HAVE A QUIZ OR TEST	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMWSHT	MAT\WORK FROM WORKSHEETS ON OWR OWN	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMPROJ	MAT\WORK ON PROJECTS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMCALC	MAT\USE CALCULATORS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMCOMP	MAT\USE COMPUTERS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMSGRP	MAT\WORK IN PAIRS OR SMALL GROUPS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMEVLF	MAT\SOLVE WITH EVERYDAY LIFE THINGS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMH-WGV	MAT\TEACHER GIVES HOMEWORK	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBMHWCL	MAT\BEGIN HOMEWORK IN CLASS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMHWTC	MAT\TEACHER CHECKS HOME-WORK	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMHWFC	MAT\CHECK EACH OTHER'S HOME-WORK	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBMHWDS	MAT\DISCUSS COMPLETED HOME-WORK	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSPROB	SCI\TEACHER SHOW HOW TO DO PROBLEMS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSNOTE	SCI\COPY NOTES FROM THE BOARD	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSTEST	SCI\HAVE A QUIZ OR TEST	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSPROJ	SCI\WORK ON PROJECTS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSWSHT	SCI\WORK FROM WORKSHEETS ON OWR OWN	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBSCALC	SCI\USE CALCULATORS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSCOMP	SCI\USE COMPUTERS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSEVLF	SCI\SOLVE WITH EVERYDAY LIFE THINGS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSSGRP	SCI\WORK IN PAIRS OR SMALL GROUPS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSHWGV	SCI\TEACHER GIVES HOMEWORK	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSHWCL	SCI\BEGIN HOMEWORK IN CLASS	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSHWTC	SCI\TEACHER CHECKS HOME-WORK	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSHWFC	SCI\CHECK EACH OTHER'S HOME-WORK	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1
ASBSHWDS	SCI\DISCUSS COMPLETED HOME-WORK	most lessons:1; some lessons:2; never:3; missing:9; not admin.:8;	2 0 1 0 0 0 0 1 0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBSDEMO	SCI\TEACHER GIVES DEMONSTRA-TION	most lessons:1;	2 0
		some lessons:2;	1 0
		never:3;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBSEXPR	SCI\DO EXPERIMENT IN CLASS	most lessons:1;	2 0
		some lessons:2;	1 0
		never:3;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGACT1	GEN\READ A BOOK	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGACT2	GEN\VISIT A MUSEUM	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGACT3	GEN\ATTEMD A CONCERT	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGACT4	GEN\GO TO THE THEATRE	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGACT5	GEN\GO TO THE MOVIES	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGNEWS	GEN\WATCH NEWS OR DOCU-MENTARIES	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1

Table D.1 Dummy Variable Construction for Input into Principal Components Population 1(Continued)

Variable Name	Variable Label	Original Coding	New Coding
ASBGOPER	GEN\WATCH OPERA, BALLET OR CLASSICS	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGNATR	GEN\WATCH NATURE, WILDLIFE OR HISTORY	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGPOPU	GEN\WATCH POPULAR MUSIC	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGSPRT	GEN\WATCH SPORTS	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGVIDE	GEN\WATCH VIDEO GAMES	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGCRTN	GEN\WATCH CARTOONS	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASBGCMDY	GEN\WATCH COMEDY, ADVENTURE OR SUSPENSE	about every day:1;	3 0
		about once a week:2;	2 0
		about once a month:3;	1 0
		rarely:4;	0 0
		missing:9;	0 1
		not admin.:8;	0 1
ASDAGE	GEN\STUDENTS AGE	number 1-97;	1-97 0
		missing 99;	0 1
		not admin 98;	0 1

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.

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, Co-Deputy International Study Director
Eugenio J. Gonzalez, Director of Operations and Data Analysis
Dana L. Kelly, Research Associate
Teresa A. Smith, Research Associate
Cheryl L. Flaherty, Research Associate
Maryellen Harmon, Performance Assessment Coordinator
Robert Jin, Computer Programmer
Ce Shen, Computer Programmer
William J. Crowley, Fiscal Administrator
Thomas M. Hoffmann, Publications Coordinator
José Rafael Nieto, Senior Production Specialist

International Study Center (Continued)

Ann G.A. Tan, Conference Coordinator
Mary C. Howard, Office Supervisor
Diane Joyce, Secretary
Joanne E. McCourt, Secretary
Kelvin D. Gregory, Graduate Assistant
Kathleen A. Haley, Graduate Assistant (former)
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

Jens Brockmann, Research Assistant
Michael Bruneforth, Senior Researcher (former)
Jedidiah Harris, Research Assistant
Dirk Hastedt, Senior ResearcherSvenja Moeller, Research Assistant
Knut Schwippert, Senior Researcher
Heiko Sibberns, 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

IEA Secretariat

Tjeerd Plomp, Chairperson
Hans Wagemaker, Executive Director
Barbara Malak-Minkiewicz, Manager Membership Relations
Leendert Dijkhuizen, Financial Officer
Karin Baddane, Secretary

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 in initiating 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.

NATIONAL RESEARCH COORDINATORS

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
Universite 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

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

*Past National Research Coordinator.

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 (CIEP)
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
Wilfried Bos
Rainer Waterman
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 Kontogiannopoulou-Polydorides
Department of Education (Nipiagogon)
University of Athens
Navarinou 13A, Neochimio
Athens 10680, Greece

Joseph Solomon
Department of Education
University of Patras
Patras 26500, 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

*Past National Research Coordinator.

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

Ruth Zuzovsky
Tel Aviv University
School of Education
Ramat Aviv
PO Box 39040
Tel Aviv 69978, Israel

Italy

Anna Maria Caputo
Ministero 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
Klaas Bos
Anja Knuver
University of Twente
Faculty of Educational Science
and Technology
Department of Curriculum
P.O. Box 217
7500 AE Enschede, Netherlands

New Zealand

Megan Chamberlain
Steve May
Hans Wagemaker*
Ministry of Education
Research and International Section
P.O. Box 1666
45-47 Pipitea Street
Wellington, New Zealand

*Past National Research Coordinator.

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

Wong Cheow Cher
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
Barbara Japelj
Pedagoski Institut Pri Univerzi v Ljubljana
Gerbiceva 62, P.O. Box 76
61111 Ljubljana, Slovenia

South Africa

Sarah Howie
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

*Past National Research Coordinator.

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 Erziehungs-
direktion 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

*Past National Research Coordinator.

TIMSS ADVISORY COMMITTEES

The TIMSS International Study Center was supported in its work by several advisory committees. The TIMSS 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 mathematics 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
William Schmidt, United States
Benny Suprapto Brotosiswojo, Indonesia

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 F. Robitaille (Chair), Canada
Albert Beaton, International Study Director
Paul Black, England
Svein Lie, Norway
Rev. Ben Nebres, Philippines
Judith Torney-Purta, United States
Ken Travers, United States
Theo Wubbels, the Netherlands