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

2. DEVELOPMENT OF THE TIMSS ACHIEVEMENT TESTS 2-1
Robert A. Garden and Graham Orpwood

2.1 OVERVIEW.....2-1
2.2 ITEM TYPES.....2-1
2.3 DEVELOPING THE ITEM POOLS.....2-3
2.4 TEST BLUEPRINT FINALIZATION.....2-9
2.5 THE FIELD TRIAL.....2-13
2.6 PREPARATION FOR THE MAIN SURVEY.....2-15
2.7 CALCULATORS AND MEASURING INSTRUMENTS.....2-19

2. Development of the TIMSS Achievement Tests

Robert A. Garden
Graham Orpwood

2.1 OVERVIEW

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

2.2 ITEM TYPES

Large-scale surveys of student achievement have traditionally used, either exclusively or mainly, multiple-choice items. Well constructed tests composed of such items typically have high reliability and high validity. In addition, practical considerations make multiple-choice items popular in many applications: testing conditions can be easily standardized, the administration costs are low, and where machine scoring is appropriate, very large samples may be processed economically and efficiently.

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

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

1. Multiple-Choice Items — Multiple-choice items used in TIMSS consist of a stem and either four or five answer choices, of which only one is the best or the correct answer. Neither “I don’t know” nor “None of the above” is an option in any of the items. In the instructions at the front of the test booklets, students are encouraged to choose “the answer [they] think is best” when they are unsure. The instructions do not suggest or imply that students should guess where they do not know the answer.

2. Free-Response Items — For the free-response items–both short-answer and extended-response types–students write their responses, and these are coded using the two-digit coding system developed TIMSS. See Chapter 7 for a discussion of the coding system.

3. Performance Tasks — Some of the skills and abilities that mathematics and science programs are intended to transmit to students are not easily assessed by the kinds of items usually found in a written test. Only “hands-on” activities allow students to demonstrate their ability to make, record, and communicate

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

2.3 DEVELOPING THE ITEM POOLS

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

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

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

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

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

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

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

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

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

2.3.1 ITEM PILOT

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

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

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

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

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

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

2.3.2 AUGMENTATION OF THE ITEM POOLS

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

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

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

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

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

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

2.3.3 PREPARATION FOR THE FIELD TRIAL/ITEM PILOT

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

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

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

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

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

Table 2.1 Schedule of Test Development From August 1993

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

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

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

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

2.4 TEST BLUEPRINT FINALIZATION

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

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

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

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

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

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

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

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

In Population 2 science, 'Earth features' and 'Earth science: other content' were combined to form 'Earth science.' 'Life science' was composed of 'Human biology' and 'Life science: other content.' 'Energy types,' 'Light,' and 'Physics: other content' were combined to form ‘Physics,’ while the content category ‘Chemistry’ remained a separate reporting category. ‘Environment’ and ‘Other content’ were combined to form ‘Environmental issues and the nature of science.’

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

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

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

Table 2.2 Reporting Categories for Mathematics and Science Content Areas

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

Table 2.3 Reporting Categories for Performance Expectations

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

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

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

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

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

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

- *Coverage of subject-matter content.* At the time the blueprints were developed, preliminary data were available from about 20 countries from the modified topic trace mapping and document analyses data collected for the curriculum analyses. These data showed the proportion of each country's curriculum that was allocated to each content topic. Rough averages of these numbers provided a basis for determining the proportion of total test time to be allocated to each content topic. These were then adjusted to ensure that adequate test time was given to in-depth topics. The resulting grids were prepared for mathematics and science separately.

- *Coverage of performance expectations.* Once the total number of minutes had been allocated to a given content topic, it was distributed across performance categories using the best professional judgment of the group. It was intended that no more than 70% of the total testing time would be allocated to multiple-choice items. In the case of mathematics, the number of items by type was allocated to each cell of the grid. In the case of science, the total number of minutes per cell was allocated, leaving the specific numbers of each type of item in each cell to be determined later. This procedure gave science item selection more flexibility.

2.5 THE FIELD TRIAL

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

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

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

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

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

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

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

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

Table 2.5 Participation in the TIMSS Field Trial

Population 1

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

Population 2

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

Population 3

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

2.6 PREPARATION FOR THE MAIN SURVEY

2.6.1 ITEM SELECTION FOR THE MAIN SURVEY

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

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

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

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

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

2.6.2 Free-Response Item Coding and Test Development

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

The coding guides for the 1993 item pilot and for the 1994 field trial were designed to produce a single “correctness” score on a three- or four-point scale. There was, however, considerable interest in obtaining more informative "diagnostic" data from the free-response items. Accordingly, following the field trial, researchers in some of the Nordic countries collaborated to prepare and trial an alternative coding system of double-digit codes that provided not only "correctness" scores for each response but also qualitative distinctions among different responses having the same score. The TIMSS codes finally developed were based on that proposal.

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

2.6.3 Item Clustering and Test Booklet Preparation

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

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

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

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

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

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

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

2.6.4 Linking Items Across Populations

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

Table 2.6 Link Items

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

2.7 CALCULATORS AND MEASURING INSTRUMENTS

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

Population 1 — calculating devices NOT permitted

Population 2 — calculating devices NOT permitted

Population 3 — calculating devices permitted.

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

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

REFERENCES

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

Robitaille, D.F., Schmidt, W.H., Raizen, S.A., McKnight, C.C., Britton, E., and Nicol, C. (1993). *TIMSS Monograph No. 1: Curriculum Frameworks for Mathematics and Science.* Vancouver, Canada: Pacific Educational Press.