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

6. Development and Design of the TIMSS Performance Assessment

Maryellen Harmon
Dana L. Kelly

6.1 OVERVIEW

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

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

6.2 CONSIDERATIONS FOR THE DESIGN

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

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

6.3 TASK DEVELOPMENT

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

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

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

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

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

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

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

6.3.1 PERFORMANCE ASSESSMENT FIELD TRIAL

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

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

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

---

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

6.3.2 Selection of Tasks for the Main Survey

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

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

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

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

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

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

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

6.3.3 Tasks Selected for the Main Survey

Science Tasks

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

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

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

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

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

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

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

Mathematics Tasks

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

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

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

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

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

Combined Science and Mathematics Tasks:

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

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

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

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

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

Table 6.1 Science Tasks[5]

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

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

Table 6.2 Mathematics Tasks[6]

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

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

Table 6.3 Combined Science and Mathematics Tasks[7]

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

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

6.4 PERFORMANCE ASSESSMENT DESIGN

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

Table 6.4 Assignment of Tasks to Stations

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

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

6.4.1 SAMPLING SCHOOLS[8]

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

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

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

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

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

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

6.4.2 SAMPLING STUDENTS

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

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

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

6.4.3 Assigning Students to Stations

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

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

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

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

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

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

Table 6.5 Assignment of Students to Stations

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

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

Figure 6.1 Example Performance Assessment Tracking Form

Performance Assessment Tracking Form

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

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

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

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

Table 6.6 Assignment of Students to Tasks

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

6.4.4 Summary of Sampling of Schools and Students

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

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

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

6.5 ADMINISTRATION PROCEDURES

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

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

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

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

6.6 CONCLUSION

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

REFERENCES

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

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

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

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