Introduction

Performance Assessment

The Third International Mathematics and Science Study (TIMSS), conducted by the International Association for the Evaluation of Educational Achievement (IEA), is the largest international comparative study of student achievement to date.[1] The purpose of the study, like that of IEA studies generally, was to learn more about the nature and extent of student achievement and the context in which it occurs, in order to inform policy decisions about schooling and its organization in the participating countries. TIMSS tested students in mathematics and science at five grades and collected contextual data from students, their teachers, and the principals of their schools.

Although student achievement was measured in TIMSS primarily through written tests of mathematics and science, participating countries also had an opportunity to administer a performance assessment, which consisted of a set of practical tasks in mathematics and science.[2] The performance assessment was available for administration to a subsample of the fourth- and eighth-grade students that completed the written tests.[3] Table 1 presents the countries that participated in the TIMSS performance assessment. Table 2 shows, for each country, the name of the assessed grades, together with the number of years of formal schooling that students in that grade had been exposed to, and their average age at the time of the TIMSS assessment.

This report presents the initial findings from the TIMSS performance assessment. Some 1,500 schools and 15,000 students from 21 countries participated, making it the largest international performance assessment yet conducted. The study was an enormous undertaking that has yielded an unprecedented store of information on how students around the world perform on a selection of practical tasks in mathematics and science.

[1] See Appendix A for a description of TIMSS.

[2] The development of the TIMSS performance assessment was greatly facilitated by the work of the Performance Assessment Committee.

[3] More specifically, the written tests were to be given to the two adjacent grades with the largest proportion of 9-year-olds, the two adjacent grades with the largest proportion of 13-year-olds, and students in the final year of secondary schooling. The performance assessment was administered to subsamples of students at the upper grade tested for 9-year-olds and the upper grade tested for 13-year-olds. For most countries, these were the fourth and eighth grades.

Countries Included in the TIMSS International Performance Assessment Report[1] Table 1

Eighth Grade

- Australia
- Canada
- Colombia
- Cyprus
- Czech Republic
- England
- Hong Kong
- Iran, Islamic Republic
- Israel
- Netherlands
- New Zealand
- Norway
- Portugal
- Romania
- Scotland
- Singapore
- Slovenia
- Spain
- Sweden
- Switzerland
- United States

Fourth Grade

- Australia
- Canada
- Cyprus
- Hong Kong
- Iran, Islamic Republic
- Israel
- New Zealand
- Portugal
- Slovenia
- United States

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

[1] Please see Appendix A, Figure A.1, for countries participating in other components of the TIMSS testing. Because low school participation led to a small sample size, performance assessment results at the eighth grade for Hong Kong are presented in Appendix B. Results for Israel are presented in Appendix B because within-school sampling procedures were not documented at the fourth and eighth grades; in addition, Israel had a small sample size at the eighth grade.

Table 2 Information About the Grades Tested

Country	Eighth Grade			Fourth Grade		
	Country's Name for Grade	Years of Formal Schooling Including Grade Tested[1]	Average Age*	Country's Name for Grade	Years of Formal Schooling Including Grade Tested[1]	Average Age*
[2] Australia	8 or 9	8 or 9	14.3	4 or 5	4 or 5	10.2
Canada	8	8	14.1	4	4	10.0
Colombia	8	8	15.8	.	.	.
Cyprus	8	8	13.8	4	4	9.8
Czech Republic	8	8	14.4	.	.	.
England	Year 9	9	14.0	.	.	.
Hong Kong	Secondary 2	8	14.2 **	Primary 4	4	10.1
Iran, Islamic Rep.	8	8	14.6	4	4	10.4
Israel	8	8	14.1 **	4	4	10.0 **
[3] Netherlands	Secondary 2	8	14.3	.	.	.
[4] New Zealand	Form 3	8.5 - 9.5	14.0	Standard 3	4.5–5.5	10.0
Norway	7	7	13.9	.	.	.
Portugal	Grade 8	8	14.6	4	4	10.3
Romania	8	8	14.6	.	.	.
Scotland	Secondary 2	9	13.7	.	.	.
Singapore	Secondary 2	8	14.5	.	.	.
Slovenia	8	8	14.7	4	4	10.9
Spain	8 EGB	8	14.3	.	.	.
Sweden	7	7	13.9	.	.	.
Switzerland (German)	7	7	14.1	.	.	.
United States	8	8	14.2	4	4	10.1

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95. Information provided by TIMSS National Research Coordinators.

* Computed from TIMSS performance assessment sample.

** Due to performance assessment sampling issues, average age is computed based on the main assessment sample (see Appendix A).

[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 4 and 8; in the other four states/territories students were sampled from grades 5 and 9.

[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.

A dot (.) indicates country did not participate in performance assessment at the fourth grade.

The Nature of Performance Assessment

Performance assessment refers to the use of integrated, practical tasks, involving instruments and equipment, as a means of assessing students' content and procedural knowledge, as well as their ability to use that knowledge in reasoning and problem solving. The assessment task may be as simple as the routine use of a piece of equipment or as complex as an investigation combining manipulative and procedural skills and requiring higher-order thinking and communication. Performance assessment aims to provide students with a testing environment which is more "true to life" and "authentic" than the traditional paper-and-pencil written test, and, by providing them with equipment and materials to manipulate in a realistic problem-solving situation, attempts to elicit performances or behaviors which will be a more valid indication of the students' understanding of concepts and potential performance in real life situations.

Proponents of performance assessment argue that the practical nature of the tasks utilized in this mode of assessment permits a richer and deeper understanding of some aspects of student knowledge and understanding than is possible with written tests alone. These aspects include skills like weighing and measuring, the use of experimental or mathematical procedures, designing and implementing approaches to solve problems or investigate phenomena, and synthesizing knowledge, application, and personal experience into an interpretation of data. [4]

Performance assessment has captured the attention of teachers and policymakers for a variety of reasons. It reflects the current trend in many countries towards active, inquiry-oriented, hands-on teaching and learning. It is seen as a means of assessment that is educationally valid, psychologically and developmentally appropriate, and congruent with "constructivist" pedagogies. Performance assessment is particularly attractive to those science educators who conceive the subject not just as a body of knowledge to be assimilated, but also as a process of enquiry rooted in the subject matter of science, and heavily dependent on the effective use of tools and technology.

A well-designed performance task, with appropriate scoring rubrics, can elicit a rich variety of student performances, and offers the possibility of deeper understanding of cognitive processes and problem-solving strategies. For example, students asked to solve an interesting problem in a practical situation may draw on whatever content knowledge appears relevant, revealing both prior knowledge and misconceptions. The students may try several approaches, each demonstrating knowledge about different attributes of the phenomenon. The students have an opportunity to demonstrate their grasp of conceptual and procedural issues, and their reasoning ability. At the conceptual level they may do so by recognizing what data to collect, what variables to control, and how many data points they may need for an adequate picture of the phenomenon they are asked to investigate; and later, by developing explanations

[4] See for example:

Tamir, P. and Doran, R. (1992). Conclusions and Discussion of Findings Related to Practical Skills Testing in Science. *Studies in Educational Evaluation, 18* (3), pp.393-406.

Shavelson, R.J., Baxter, G.P., and Pine, J. (1991). Performance Assessment in Science. *Applied Measurement in Education, 4* (4), pp.347-362.

Haertel, E.H. and Linn, R.L. (1996). "Comparability" in G.W. Phillips (Ed.), *Technical Issues in Large-Scale Performance Assessment.* Washington, D.C.: National Center for Education Statistics.

for the trends they find in their data. Students may exhibit procedural knowledge through the use of appropriate equipment, through collecting and organizing data in tables, lists or graphs, by applying algorithms, or by reading data tables and comparing and computing differences between entries. Students may demonstrate reasoning ability by identifying trends and patterns, drawing conclusions, predicting and extrapolating to new data points, and relating findings to the original question.

Few would argue against the premise that the detailed study of student performance on practical tasks in life-like assessment situations offers greater potential for understanding student achievement than paper-and-pencil tests alone. However, in very large-scale assessments the benefits of performance assessment in terms of the extra information it may provide about student achievement must be balanced against the extra cost and complexity inherent in this mode of assessment. As the largest and most ambitious international study of student achievement in mathematics and science to date, TIMSS provided a unique environment in which to develop and implement the ideas of performance assessment within the constraints of a large-scale international comparative study.

Performance Assessment in TIMSS

The major challenge in developing a performance assessment for TIMSS was to identify a series of tasks in mathematics and science which could elicit a wide range of student performances, both from a subject matter perspective and from the perspective of the student behaviors necessary to complete the tasks ("performance expectations" in the terminology of TIMSS), yet which could be performed with inexpensive and readily available materials, and be adaptable to standardized administration procedures in many different cultures and languages. In addition, because the performance assessment was to be part of a much larger written assessment which made considerable demands on the time of students, teachers, and principals, it was essential that the performance assessment keep the student response burden to a minimum.

Following an extensive field-trial, a set of 13 tasks (12 for each grade level) were identified as suitable for the main assessment. These tasks could be assembled from widely-available materials, and translated readily into different languages. The issue of response burden was addressed by assigning a subset of the tasks to each student so that each student was asked to attempt only about one third of the tasks. The performance assessment was administered in a "circus" format in which a student completed three to five tasks by visiting three stations at which one or two tasks were assembled.[5] The assignment of students to stations was determined according to a predetermined scheme.

Ideally, the performance assessment would have included observations of students as they worked through the tasks, as well as evaluation of written responses. However, such observations were prohibited by cost and time constraints. Instead, structured response sheets were created with questions (items) worded to elicit evidence of specific skills and thinking processes.[6] After completing the tasks at each station, students submitted their work booklets to the performance assessment administrator, together with any products. The work recorded in the booklets and any products created during the assessment were evaluated by coders specially trained to use the TIMSS scoring rubrics.[7] The coding system developed for TIMSS allowed for the identification of common approaches and types of errors in student responses.

The TIMSS performance assessment was conducted with a subsample of fourth- and eighth-grade students that had participated in the main assessment.[8] Of the 45 countries that took part in the written assessment at the eighth grade, 21 chose also to administer the performance assessment. At the fourth grade, 10 of the 26 countries that participated in the written assessment also took part in the performance assessment. For many of these countries, this was their first experience conducting a large-scale performance assessment, and was therefore a useful model with tasks, administration procedures, and coding schemes that could help them explore the feasibility of performance assessment in their own countries.

[5] For more information on the performance assessment design see Appendix A of this report. See also Harmon, M. and Kelly, D.L. (1996). "Performance Assessment" in M.O. Martin and D.L. Kelly (Eds.), *Third International Mathematics and Science Study Technical Report, Volume I.* Chestnut Hill, MA: Boston College.

[6] See Baxter, G.P., Shavelson, R.J., Goldman, S.R., and Pine, J. (1992). Evaluation of Procedure-based Scoring for Hands-on Science Assessment. *Journal of Educational Measurement, 29* (1), pp. 1-17, on the use of "notebooks" as a reasonable surrogate for process observation.

[7] See Appendix A for more details on the coding procedures and reliability.

[8] See Appendix A for a more complete description of the TIMSS performance assessment sample.

The TIMSS Performance Assessment Tasks

Of the 13 tasks, 11 were similar in some sense across both the fourth and eighth grades. One task was unique to fourth grade, and one task to eighth grade. Each set of 12 tasks included five science tasks, five mathematics tasks, and two combination tasks, integrating mathematics and science content and skills areas. Although more than half the tasks required both science and mathematics knowledge and skills, tasks were classified according to the primary content area addressed. The tasks classified as addressing primarily science content are: Pulse, Magnets, Batteries, Rubber Band, and Solutions (eighth grade only) or Containers (fourth grade only). The mathematics tasks are Dice, Calculator, Folding and Cutting, Around the Bend, and Packaging. The two combination tasks are Shadows and Plasticine. While some tasks are identical for the fourth and eighth graders, most differ either by providing more structure for the younger students or by including additional items for the older students.

In developing the performance assessment tasks, considerable effort was expended in ensuring that the tasks would elicit a wide range of performance expectations. The term "performance expectations" is used in TIMSS to describe the cognitive or manipulative skills that students are expected to use in working on the items in a task. Performance expectations include recalling and using simple or complex information; using equipment, routine procedures, and experimental processes; problem solving; designing and conducting an investigation; analyzing and interpreting findings; formulating and justifying conclusions; and communicating scientific or mathematical information (see Table A.1 in Appendix A). Items measuring these thinking and experimental skills were distributed across all the tasks.

Each TIMSS performance assessment science task began with a primary problem or investigation to be completed by the student, followed by a series of items that required, successively, a solution to the problem, and a description of problem-solving strategies; or for the more extensive investigations, an experimental plan, data display, and students' analyses and interpretations of their own data, sometimes with predictions based on their hypotheses. In mathematics, students began with applications of routine procedures and proceeded through more complex procedures requiring data organization and analysis to creating their own problem-solving strategies, with predictions and conjectures based on their solutions.

Structure of the Performance Assessment Report

This report describes the TIMSS performance assessment and provides a detailed summary of the performance of the students in each participating country on every item of every task. In the interests of making the results available in the shortest possible time, this report presents only descriptive summaries of student performance on the assessment tasks, and makes no attempt to relate student achievement on the performance assessment to achievement in the written assessment, or to any of the myriad background variables available in TIMSS.

Chapter 1 of this report presents a description of the tasks administered to the students in the TIMSS performance assessment, together with examples of student work and the criteria used to evaluate the work. For each task and each item within the task, results are presented for each country and for the international average. Chapter 2 displays the national differences in student achievement across all performance assessment tasks and separately for mathematics and science tasks at eighth and fourth grades. This chapter also displays results for boys and girls separately on each task for both grades. Chapter 3 displays national differences in student achievement by performance expectation at both the eighth and fourth grades. This chapter also compares the international performance of eighth-grade students on example items selected to illustrate the performance skills subcategories contained in the broader performance expectation categories.