Panel B

What are PISA and TIMSS? What do they tell us?

Lee Peng Yee (*moderator*)

Jan de Lange and William Schmidt (*panelists*)

Abstract. This is a panel discussion on PISA and TIMSS, two international comparative studies in educational achievement. The panelists are Jan de Lange of the Freudenthal Institute, the Netherlands, for PISA, and William Schmidt of Michigan State University, the United States, for TIMSS, with Lee Peng Yee of National Institute of Education, Singapore, as a moderator. They are to explain the nature, the aims, and the conclusions of the two studies, and to argue over their relative merits. This document contains three initial statements from the above-mentioned participants respectively.

Mathematics Subject Classification (2000). Primary 00A35; Secondary 00A05.

Keywords. PISA, TIMSS, comparative studies.

Introduction

by *Lee Peng Yee*

One area of interest in education is comparative studies in educational achievement, in particular, in mathematics, science and reading. There are two such international studies involving mathematics, namely, PISA and TIMSS. PISA stands for the Programme for International Student Assessment. It is better known in Europe. TIMSS stands for the Trends in International Mathematics and Science Study. TIMSS was previously known as the Third International Mathematics and Science Study. Each study of PISA or TIMSS involves approximately 50 countries and thousands of students in each participating country. The studies generated volumes of publication and numerous related research projects.

The fact that some Asian countries topped the achievement list in TIMSS amazed many people and drew the attention of the industrial countries. Consequently it induced the study on these high-performing Asian countries, namely, China, Korea, Japan, and Singapore. Further a country could do well in TIMSS but not in PISA. This phenomenon is now known as PISA shock. Hence in addition people are also interested in the comparison of these two international studies. The impact of PISA and TIMSS has gone way beyond the mathematics and science community. It even

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

influences the policy makers of a country. It is timely that we have a panel discussion on the topic.

TIMSS. The study was commissioned by IEA, the International Association for the Evaluation of Educational Achievement. The first round of TIMSS took place in 1995 and the second round in 1999. It was the third round that made TIMSS famous world wide It collects data on educational achievement from students at the fourth and eighth grades. It also collects extensive information from students, teachers and school principals about the teaching and learning of mathematics. The test items are matched against those in the standards or syllabus. Then the data are analyzed and the reports published. The next round will take place in 2007. For details, see [1].

PISA. The study was initiated by the OECD countries. OECD stands for Organisation for Economic Co-operation and Development whose member countries were originally countries from Western Europe but now they are all over the globe. PISA was conducted every three years in 2000, 2003 and the next one in 2006. The tests are administered to 15-years-old students. The tests are supposed to assess how well students are prepared for their full participation in society. Similarly, the data are analyzed and the reports published. As we can see, PISA differs from TIMSS in methodology and aims. For details, see [2].

Benchmarking. Both PISA and TIMSS have been used by many countries for benchmarking. Roughly speaking, TIMSS is grade-based, that is, testing students of Grade 4 and Grade 8, whereas PISA is age-based, that is, testing the 15-years-old students. The 15-years-old students are those who are near the end of their compulsory education. Test items in TIMSS are more content or standards orientated, whereas those in PISA are more literacy orientated. TIMSS assesses how much students have achieved in schools. PISA assesses how well students are prepared for the outside world. Of course, this is an over-simplified view of the differences between the two studies. It does give a general idea about the two studies.

Panelists. They are Jan de Lange of the Freudenthal Institute, the Netherlands, speaking for PISA, and William Schmidt of Michigan State University, the United States, speaking for TIMSS. Jan de Lange is Director of the Freudenthal Institute and a full professor at University of Utrecht, the Netherlands. He was a member of the National Advisory Board of the Third International Mathematics and Science Study, and is currently Chair of the Mathematical Functional Expert Group of the OECD-PISA. William Schmidt is a professor at the College of Education, Michigan State University, and the national research coordinator and executive director of the United States National Research Center which oversees the United States' participation in the Third International Mathematics and Science Study. At the panel discussion, they are to present what PISA and TIMSS are respectively, and what they are for. Then they will discuss and possibly answer questions from the audience.

Issues for discussion. The issues for discussion include at least some or all of the following questions. The questions are divided into three categories. First, what are PISA and TIMSS?

- Does PISA or TIMSS really serve the purpose intended?
- Why do we need PISA when we already had TIMSS?
- What are the good points or bad points of PISA and TIMSS?

Secondly, what do they tell us?

- Is it really meaningful to use PISA or TIMSS for benchmarking?
- Some countries did well in TIMSS but not in PISA. Why?
- Both PISA and TIMSS have collected a vast amount of data. Are they useful for other researchers? What can they do with the data?

The last question above was previously raised at the International Round Table in Tokyo 2000 [3]. Thirdly, what is the future?

- The learning process of a student is a long-term affair. Perhaps the three-year cycle or four-year cycle is simply too short to measure the progressive achievement of a student. Do we need to measure so frequently?
- Will there be PISA or TIMSS 20 years from now?

This short statement serves as an introduction to the panel discussion to be held on 28 August 2006 in Madrid, Spain. Other statements from the panel speakers follow.

References

[1] TIMSS 2003, Trends in International Mathematics and Science Study. International Mathematics reports released 14 December 2004. Website: www.timss.com.

[2] PISA 2003 technical report, OECD Programme for International Student Assessment, Website: www.pisa.org.

[3] Lee, Peng Yee, International Round Table, Proceedings of the International Congress on Mathematics Education, Tokyo 2000.

TIMSS as a study of education: why should we care?

by *William H. Schmidt*

Comparative studies of education often seem to evoke a "so what?" or "who cares?" reaction. Studies of students' achievement in different countries may leave one wondering what practical importance such differences hold in the real world or work and commerce. Descriptions that highlight differences in common educational practices may appear intriguing and stimulate curiosity but may leave one wondering what the relevance is to what happens (or should happen) at the school down the street.

The value of such studies is almost assumed to be self-evident given, it seems, by the sort of attention the media frequently affords them. Reports of rankings along with comparisons of scores with countries x, y, and z reduce the entire endeavor to a sort of education Olympics or horserace. The value, obviously, lies in the comparisons! Who is on first? Who is doing it right?

As intriguing and entertaining as some comparisons may be – "Wow! Teachers in country x *never* assign homework!" or "Students in country z have to go to school on *Saturday*!" – these are practices that must be understood within their particular social, cultural, and educational contexts. Attempting to copy or transplant the practices of one country into another will not likely have the desired effect: alien grafts rarely take without extensive preparation and effort.

Value of international comparative studies. The real value in international studies lies not in the comparisons themselves, but in the insights we may gain into our own common practices. International comparisons hold up and frame what's familiar against a background of a considerable range of alternatives. This can lead to a thoughtful reconsideration of our rationale for doing things the way we do – or even initiate a thoughtful evaluation of something never before considered.

Many people, for example, are familiar with schools. They know what schools are and what happens in the classrooms inside the schools. Schools are schools; whether they are urban Paris or rural Montana. We began our involvement with international education research in the days leading up to TIMSS with a similar assumption about the nature of schools in various countries. We discovered that school has commonalities everywhere. What is common wherever schools are found are students, teachers, and textbooks. How these commonalities interact and work within a larger education system, however, can vary considerably. We discovered, for example, that in Norway primary teachers typically stay with the same group of students for the first five or six years of students' formal school experience. We also learned that in Switzerland, 'schools' only exist in large cities. The majority of students and teachers meet together in rooms located in buildings that are not necessarily dedicated to housing educational activities. Furthermore, school administrators and other support personnel are only found in such dedicated facilities which generally house the upper secondary grades or are located in the cities.

Clearly there are a number of ways to conduct school. The examples mentioned here were not selected to suggest that all countries change either the nature of their school buildings nor the length of time primary teachers typically work with the same group of students. Some consideration of these issues may be fruitful, but the point to be made here is that these schooling practices represent options – choices that have been made about how school gets done. The more we can see the way we do things as choices, the better position we are in to consider and construct profitable change or reform.

What we can learn from TIMSS. In the Third International Mathematics and Science Study (TIMSS) the focus was not on the structural aspects of school such as the previously mentioned examples, although these were a part of the study. The focus, rather, was on the substance of education, the school curriculum, the content that's at the center of what teachers and students do in schools.

Previous international studies led us to suspect that the achieved curriculum, what students demonstrate that they know, varies from one country to another. TIMSS assessed this aspect of curriculum in the context of an extensive investigation of the intended curriculum, what systems intend their students to learn, along with the implemented curriculum, what is taught in the classroom. Measurements of these curriculum aspects led to one unmistakable conclusion: the mathematics taught and studied in the schools of one country can differ substantially from what exists in the schools of another. In short, there are many ways to do mathematics education.

More specifically, this curriculum measurement in TIMSS led to some thought provoking insights into the U.S. mathematics curriculum. For example, the U.S. intends teachers and students to study two to three times the number of topics in the first through eighth grade as is typical in other countries. Consistent with this breadth, U.S. textbooks are truly first in the world in their size, weight, and scope. Not too surprisingly, given these intentions and resources, the U.S. teachers tend to spend some time on every intended topic typically without emphasizing any small number of topics as is common in other countries. All of this contributes to the "mile wide, inch deep" nature of the U.S. curriculum.

These insights were possible because TIMSS was designed from the start to examine the relationship among the various aspects of the curriculum: the intended, the implemented, and the attained. These insights have also led to several efforts to thoughtfully revise the U.S. mathematics curriculum.

So, what is the value of international study? Certainly not to obtain bragging rights for the top spot on some list nor even to identify specific practices that we may want to copy. The real value stems from obtaining a fresh perspective on the array of choices embedded in our own approach to education. Thoughtful and principled insights stimulated by examples from other systems can lead to powerful revision in our quest to provide a challenging and equitable education for all students.

PISA: promises, problems and possibilities

by *Jan de Lange*

PISA versus TIMSS. According to the OECD:

> The OECD's Programme for International Student Assessment (PISA) is a collaborative effort among the member countries of the OECD to measure how well young adults, at age 15 and therefore approaching the end of compulsory schooling, are prepared to meet the challenges of today's knowledge societies. The assessment is forward looking, focusing on young people's ability to use their knowledge and skills to meet real-life challenges, rather than on the extent to which they have mastered a specific school curriculum. This orientation reflects a change in the goals and objectives of curricula themselves, which are increasingly concerned with what students can do with what they learn at school, and not merely whether they have learned it. The term 'literacy' is used to encapsulate this broader conception of knowledge and skills.

The first PISA survey was carried out in 2000 in 32 countries, including 28 OECD member countries. Another 13 countries completed PISA 2000 in 2002, and from PISA 2003 onwards more than 45 countries will participate 'representing more than one third of the world population'. PISA 2000 surveyed reading literacy, mathematical literacy, and scientific literacy, with the primary focus on reading. In 2003 the main focus was on mathematical literacy (published in 2004), and in 2006 scientific literacy will be highlighted.

It will be clear that TIMSS and PISA have a lot of similarities resulting in improper identification of the two series of studies in the media, which is undesirable and confusing. But the descriptions of the organizations that are responsible, show that they both claim similar relevance for the studies. Even for the expert it will be difficult to relate the following either to TIMSS or to PISA: 'Countries participating in this study will have information at regular intervals about how well their students read and what they know and can do in mathematics and science.' Both studies do this and do it, methodologically speaking, in a very similar way (based on Item Response Theory, IRT). Even the reporting tables in the respective reports look very similar.

If there is a problem that both studies share, it is the design of the measuring instrument in relation to the validity of the outcomes. Traditionally, validity concerns associated with tests have centered about test content, meaning how the subject domain has been sampled. Typically evidence is collected through expert appraisal of alignment between the content of the assessment tasks and the curriculum standards (in case of TIMSS) and 'subject matter' assessment framework (PISA). Nowadays, empirical data are often used before an item is included in a test.

Traditionally validation emphasized consistency with other measures, as well as the search for indirect indicators that can show this consistency statistically. More

recently is the recognition that these data should be supplemented with evidence of the cognitive or substantive aspect of validity. Or as *Knowing What Student Knows* (2001) summarized: 'The trustworthiness of the interpretation of test scores should rest in part on empirical evidence that the assessment tasks actually tap the intended cognitive process.'

One method to do this is a protocol analysis in which students are asked to think aloud as they solve problems; another is an analysis of reasons in which students are asked to provide rationales for their responses; and a third method is an analysis of errors in which one draws inferences about processes from incorrect procedures, concepts, or representations of problems. Although some of these methods are applied only after the test is administered, there is a trend that large-scale assessments like TIMSS and PISA use these methods as well. The use of cognitive laboratories to gauge whether students respond to the items in ways the developers intended has become a new instrument in the developmental process. The use of double-digit coding is another sign of interest in the process of problem solving instead of just judging whether an answer is incorrect or correct. A 'correct' or 'partly correct' score given not only to each work of the student, but also to which strategy was used or where in the process the students 'lost track'.

Validity. The validity of the test instrument remains a complex issue. It goes without saying that there is an inherent tension between the traditional choice of item formats, usually with very restricted time (1–2 minutes per item), and the rather ambitious definitions of what the instrument is intended to measure. But not only the concern about 'errors' plays an important role in relying so much on multiple-choice, it is also an economic issue: Many countries participating in these large cooperative studies are unwilling or unable to fund much more expensive multiple marker studies, even if such studies have demonstrated their efficacy.

PISA 2003 also had a problem solving component. Many of the items would fit the mathematics Framework, and given the fact that the instrument for problem solving (PS) had much more open 'constructive' items, a study relating the math items and the PS items could be very helpful in advancing the discussion on item instruments and their restrictions in large-scale international studies. According to the PISA report on problem solving: 'The items for problem solving give a first glimpse of what students can do when asked to use their total accumulated knowledge and skills to solve problems in authentic situations that are not associated with a single part of the school curriculum.'

One can easily argue that this is always the case in a curriculum: For mathematical literacy, mathematics as taught at school will not suffice. Students need to read, need to interpret tables and graphs (seen by many as belonging to reading literacy), and, indeed, need problem-solving strategies. But seen from the perspective of promising developments on item formats and item quality, the problem-solving component of PISA is interesting, at least. And if TIMSS implements their intent to 'place more emphasis on questions and tasks that offer better insights into students' analytical,

problem-solving, and inquiry skills and capabilities,' innovation in large-scale assessments could materialize.

PISA versus TIMSS. The main differences between TIMSS and PISA seem to be the following:

- curricular emphasis for TIMSS versus functional aspect (literacy) for PISA;
- grade-specific structure of TIMSS versus age-specific structure of PISA.

TIMSS uses the curriculum as the major organizational aspect. The TIMSS curriculum model has three aspects: the intended curriculum, the implemented curriculum, and the achieved curriculum. These represent, respectively, the mathematics and science intended for students to learn, and how the education system should be organized to facilitate this learning: what is actually taught in the classrooms, who teaches it, and how is it taught; and finally, what it is that students have learned, and what they think about those subjects.

International curricular diversity was a serious point of concern to the TIMSS study. The goal was to develop an international test that would be equally fair to all participating countries. Therefore subject-matter specialists from all countries were consulted and asked to contribute to the process of test development. Most countries participating in TIMSS had an intended mathematics curriculum that matched with more than 90% of the items. The outliers were the United States and Hungary with 100% matching, and the Netherlands, with 71% matching.

Insiders have discussed the procedure and its validity of this equally unfair analysis. The question not satisfactorily answered is how the mathematics education communities in the different countries were involved, and how representative they were. But if these numbers are accepted, in this context it is worth looking at the minimal matching result of the Netherlands.

It was expected that students of other countries would outperform Dutch students. However, contrary to expectations, in 1995 Dutch grade 8 students performed well on the TIMSS test. Their score was significantly above the international average, just below the four Asian top-scoring countries. After some additional research it was concluded that somehow the Dutch students were knowledgeable about the 29% of test items that were remote from their intended curriculum. In the end it was concluded that the students had the abilities for transfer of their knowledge and skills to items that did not match with their intended curriculum. It can be very appropriate to test students on material they have not been taught, if the test is used to find out whether the schools are doing their job.

PISA takes this point even further: It is based on a dynamic model of lifelong learning in which new knowledge and skills necessary for successful adaptation to a changing world are continuously acquired throughout life. It focuses on young people's ability to use their knowledge and skills to meet real-life challenges, rather than on the extent to which they have mastered a specific school curriculum.

The two different approaches can both be critiqued: What does it mean that the Netherlands scored so high with the minimal relation with its curriculum? What does

it mean if PISA will not constrain itself to any national curricula? It is clearly not true that international studies of student achievement may be unintentionally measuring little more than the degree of alignment between the test instrument and the curriculum. What it does measure is still a question open to interpretation.

Another indication that shows how difficult it is to make statements that go beyond well-intended opinions can be found in the observation of Westbury in 1992, in relation to SIMS, when he observed that the lower achievement of the United States is the result of curricula that are not as well matched to the SIMS test as are the curricula of Japan. But in TIMSS the match was 100% (see earlier), and still the United States did not perform very well.

Impact. The Germans produced a national PISA 2000 report of 550 pages, the international OECD report was 330 pages, and the Dutch report a mere 65 pages. Most countries had something around 150 pages. It is not the statistics that are interesting here, but the message from the report and what has been selected to be included. Even a superficial analysis, which was carried out for this article with the reports mentioned and the one from the United States, makes significant differences visible. There is a common myth that numbers do not lie. It is now widely accepted that data can be gathered, processed, mathematized, and interpreted in a variety of ways. So a key issue is the question of who influences this process, for what reasons, and through what means. The studies just mentioned underscore this concern apart from the fact that even numbers can lie.

Back to the very *gründliches* German report. Not only did the German PISA Konsortium do an excellent and thoughtful job, it also made recommendations for immediate improvement, including ones that directly affect the content. The changes should include:

- more integration of inner- and outer-mathematical 'networks';
- fewer calculations;
- more thinking activities and student mental 'constructions';
- more reflection;
- more flexible use of schoolbooks.

These goals can be reached when the recommendations that were formulated after TIMSS are implemented:

- development of a different math-problems culture: more open-ended, more 'real-world';
- a new teaching-and-learning culture, with a more exiting cognitive school environment;
- more and different professionalization of teachers, emphasizing teamwork.

PISA adds to these recommendations a 'very different conceptualization' of mathematical concepts and emphasis of modeling and mathematization, situated in contexts. And, argued the report, the Germans have definitely not reached the optimum in using different representations as a tool to build better conceptual understanding.

Mathematics education is in a state of transition, in part because of the fact that both TIMSS and PISA were taken seriously. Surprisingly the shock and catastrophe that struck Germany as some kind of natural disaster, if one had only the popular media as a resource, has resulted in a government-supported nationwide action-plan with a very strong content part that will result in a different mathematics education culture at schools. Of course, the success of these changes will be measured by PISA 2003, 2006, 2009, and so on. At least in part.

The future of PISA. It is very hard to predict the future of PISA. Of course it is a very successful project if one looks at the number of countries participating: 58 in 2006 and growing. And there are many opportunities to make PISA more successful from the content point of view. If PISA is able to include longer and more complex items, as it did with its Problem Solving study in 2003, if technology gets a proper place (as is intended), if group-work can be included in some way PISA would make itself much more rewarding for policy makers and practitioners alike.

PISA will also start a study for the 9-year olds, in the near future. In short the OECD definitely has the intention to continue PISA for the next decade at least. And if the instrument keeps improving, it seems worth the effort – although OECD has to be more clear about the fact that PISA measures mathematical literacy, and not curricular mathematics – and how to deal with this principle in the future.

PISA will have to address the problem of the Horse-race – a very undesirable aspect that draws a lot of criticism – and rightfully so. Another format of the international report with portraying country by country would not only be more informative, but also would give a more valid picture: one number cannot represent the quality of an educational system.

Validity issues have to be addressed, even if PISA is using state-of-the-art methodology. Not only the methodology should be of the highest quality, but also the content – and improvement should be on the agenda continuously.

And of course: communication between all parties should improve: Math educators and research mathematicians feel as being watchers of a game they hardly feel any ownership for. This is undesirable: PISA should not address just policy makers if it really wants to make a difference: the data of PISA are in the public domain and any country can analyse these data for its own purpose. This opportunity should not be lost. The meaning of PISA can be co-defined by its users.

National Institute of Education of Singapore, Singapore 637616, Singapore
E-mail: pylee@nie.edu.sg

Freudenthal Institute, University of Utrecht, 3561 GE Utrecht, The Netherlands
E-mail: J.deLang@fi.uu.ni

Center for the Study of Curriculum, Michigan State University, East Lansing, MI 48824-1034, U.S.A.
E-mail: bschmidt@msu.edu